CITY, UNIVERSITY OF LONDON

Quantifying and Modelling Online Decentralised Systems: A Complex Systems Approach

Author:
Abeer ElBahrawy

Supervisors:
First supervisor:
Dr. Andrea Baronchelli
Second supervisor:
Dr. Mark Broom

Examiners:
Dr. Lucas Lacasa
and
Prof. Tobias Preis

A doctoral dissertation submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Department of Mathematics
School of Mathematics, Computer Science and Engineering

January 11, 2020

CITY, UNIVERSITY OF LONDON

Abstract

Quantifying and Modelling Online Decentralised Systems: A Complex Systems Approach

Cryptocurrencies are unique examples of decentralised socioeconomic systems. All the transactions, trading, and development are traceable and publicly available. Bitcoin, the first cryptocurrency, was introduced in 2009 launching a market of more than 2500 cryptocurrencies and has a value of more than 200 billion dollars. In comparison to the rising importance of cryptocurrencies in the financial world, the research on cryptocurrencies is still limited. In this thesis, we analyse three novel datasets namely, cryptocurrencies' market data, cryptocurrencies' Wikipedia page views and edits, and illicit transactions on Bitcoin. We study the cryptocurrencies ecosystem, including the market dynamics, the social attention and the transaction network. We find that the ecological neutral model can capture the market dynamics, hinting at the extent to which technological differences between cryptocurrencies are considered in investment decisions. We also investigate the relationship between information production and consumption and cryptocurrency market dynamics. We find that a small community of tightly connected editors is responsible for most of the production of information about cryptocurrencies in Wikipedia. Finally, we assess dark markets' Bitcoin transactions showing the ability of the markets to adapt to multiple closures, including law enforcement raids. We expect that our contribution will be of interest to researchers working on either cryptocurrencies or complex systems. We anticipate that adopting a complex systems approach, will spark more research that interweaves both the technological and socioeconomic aspect of cryptocurrencies.

Publications

This thesis is based on the following publications:

I. Abeer ElBahrawy, Laura Alessandretti, Anne Kandler, Romualdo Pastor-Satorras, and Andrea Baronchelli. Evolutionary dynamics of the cryptocurrency market. Royal Society Open Science, 4(11), 2017.

II. Abeer ElBahrawy, Laura Alessandretti and Andrea Baronchelli. Wikipedia and Cryptocurrencies: Interplay Between Collective Attention and Market Performance. Frontiers in Blockchain, 2019, doi: 10.3389/fbloc.2019.00012.

III. Abeer ElBahrawy, Laura Alessandretti, Leonid Rusnac, Daniel Goldsmith, Alexander Teytelboym, Andrea Baronchelli. Collective Dynamics of Dark Web Marketplaces. arXiv preprint arXiv:1911.09536, 2019. Presented at Harvard Big Data 2019.

IV. Sam Miller, Abeer ElBahrawy, Martin Dittus, Joss Wright, and Mark Graham. Predicting drug demand with Wikipedia views: Evidence from darknet markets. Accepted in The World Wide Web Conference.

V. Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, and Andrea Baronchelli. Anticipating cryptocurrency prices using machine learning. Complexity, 2018.

Other publications:

VI. Mike Seiferling, Abeer ElBahrawy, Tales Padilha, and Keith Chan. Cryptocurrencies and the future of money, 2019. To appear soon.

Acknowledgements

I have worked on this thesis, surrounded by a vibrant and warm community. I am very thankful for:

Andrea Baronchelli, who has guided me through my PhD with care and challenged me to work to my full potential.

The staff of the Mathematics Department at City, who have welcomed me in a friendly research environment. Among them, a special thank goes to Anne Kandler, Alessandro De Martino, and Mark Broom.

My office mates, who shared with me a windowless room and many hard moments. Thanks to Cecilia de Fazio, Lleonard Rubio, Valdo Tatitscheff, Johann Bauer, Patrick Serwene, Abrar Ali, Adam Varga, Hamish Forbes, Julia Cen and Roberta Amato. A special thanks to Laura Alessandretti, for working closely with me, I will remember you with every graph!

My colleagues at Chainalysis who welcomed me in a whole new environment for me. Thanks to Philip Gradwell, Leonid Rusnac, Daniel Goldsmith and Kim Grauer.

The Turing Institute for the support and allowing me to meet great colleagues and collaborators, Among them Sam Miller and Martin Dittus.

My collaborators Romualdo Pastor-Satorras, Alex Teytelboym, and Luca Maria Aiello for the insightful discussions and contribution.

My Master's thesis supervisor Mohammed El-Beltagy for always inspiring and encouraging me.

Tobias Preis and Lucas Lacasa, who have kindly accepted to act as examiners for this thesis and my defence.

My flatmates; Siobhan, James, and Li who filled our home with good cooking and Catan.

Nazla Eid for the non stop "oumy zakry". Sahar Khaled for never ending CD repeats. Malte Probst, for the roasting and all the laughters.

My family, especially my little nieces Aya, Menna and Kinda for filling my life with joy and silliness.

Declaration of Authorship

I, Abeer ElBahrawy, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signed:

Date:

Contents

1 Introduction

Cryptocurrencies are unique examples of decentralised socioeconomic systems. The currencies allow everyone - with an Internet connection - to issue, maintain and store the transactions' ledger. Moreover, all the transactions, the trading, and the development are transparent, traceable and publicly available. This innovative approach to economy introduced new concepts of money issuing and ledger keeping which delayed cryptocurrencies' regulation, allowing them to be used in illicit transactions and exhibit complicated and precarious dynamics. The decentralised nature, the transparent fully public system, the coupling of economic nature and innovative technology, and the growing value made cryptocurrencies particularly interesting systems to study.

Bitcoin, the first cryptocurrency, was introduced in 2009 as a borderless, decentralised and anonymous digital currency [1]. In Bitcoin, all transactions are recorded in a decentralised public ledger - the blockchain - while relying on cryptography to guarantee users anonymity; hence the name cryptocurrency. Within 8 years, Bitcoin price exceeded $20,000$ dollars and became accepted in more than 15000 venues. The anonymity of Bitcoin transactions also provided suitable cover for the online trade of illicit goods and services. Soon after its release, in 2011, an online illicit drug market called Silk Road started its operation relying on Bitcoin as its official currency [2]. It grossed more than 300 million dollars in two years. On the other hand, Bitcoin is a promising solution to countries suffering from sanctions, or a failure of the domestic currency, for example the case in Iran and Venezuela [3].

Right now, the cryptocurrency market includes more than 2500 cryptocurrencies [4]. Some of these cryptocurrencies are replicas of Bitcoin while others depend on novel technologies. Alongside the growth of the number of cryptocurrencies, the total market capitalisation grew to exceed 700 Billion dollars in December 2017 [4]. As a result of the market growth, cryptocurrencies audience became broader and more diverse. Instead of an audience mainly composed of technology enthusiasts, cryptocurrencies drew the attention of many, including investors and regulatory entities.

Even though cryptocurrencies were gaining more attention, research on Bitcoin and cryptocurrencies is scarce, primarily because the field is still in its infancy. The earliest "crypto enthusiasts" were attracted by the technological novelty, which consequently drew the attention of researchers [5]. Later on, as the economic importance of cryptocurrencies started to unravel, more research focused on the economic nature of Bitcoin [6, 7, 8] and its price prediction [9, 10, 11, 12]. Whether Bitcoin is a medium of exchange or an asset was investigated [7]. The connection between the cryptocurrency market and the stock market enabled researchers to deploy the same techniques used in stocks' price prediction, including online social traces [13] and machine learning [14]. However, the research was limited to Bitcoin and few cryptocurrencies (7 at most) [15].

In this thesis, we build upon the limited yet growing research on cryptocurrencies. Our contribution can be summarised in four main points. Firstly, at the time of our publication, we were the first to *expand the analysis to more than* 2500 *cryptocurrencies taking into consideration the whole market dynamics.* Secondly, *we adopt a complex system approach to the study of the cryptocurrencies ecosystem*, which facilitates a better understanding of the interplay between cryptocurrencies economic, technological and social aspects. We also rely on tools coming from probability theory, random walks, agent-based modelling, numerical simulations and complex networks. Thirdly, *we analyse several dimensions of the cryptocurrencies ecosystem*, namely, the market dynamics, social attention and the blockchain transactions. Finally, our analysis *spans different periods of significant changes in the market and social attention toward cryptocurrencies.* Our analysis captures four significant periods. Our analysis includes the period before the market surge in December 2017 as well as the period of the surge. It also includes the period of the sharp drop in prices later in January 2018. We further examine the period of relative stability from January 2019 to May 2019. Through all these fluctuations, we review our results, conclusions and even adapt to the changes in the data sources.

Our results and contributions come from the analysis of three primary novel datasets covering three dimensions of cryptocurrencies ecosystem, in Chapter 2, we present an overview of cryptocurrencies technology, ecosystem and the current state of research. In Chapter 3, we describe the datasets and the techniques used to collect and to pre-process them.

Our research addresses questions that have raised growing interest within the scientific community [6, 16, 17]. Below, we present how the thesis is structured around these three questions. Our main findings are presented in chapters 4 to 8. Robustness and sensitivity tests are reported at the end of the thesis; in appendices A to D.

Chapter 4: What are the competition dynamics of the cryptocurrency market?

Despite the cryptocurrency market increasing relevance in the financial world, a comprehensive analysis of the whole system was lacking at the time of our analysis, as most studies had focused exclusively on the behaviour of one (Bitcoin) or few cryptocurrencies. We contribute to the field by considering the history of the entire market and analyse the behaviour of more than 2000 cryptocurrencies introduced between April 2013 and May 2019. Our work sparked interest in the field and triggered more work on the cryptocurrency market properties [18, 19, 20, 21]. In Chapter 4, we reveal that, while new cryptocurrencies appear and disappear continuously and their markets capitalisation is increasing (super-)exponentially (up to December 2017), several statistical properties of the market have been stable for years. These include the number of active cryptocurrencies, the distribution of market share and the turnover of cryptocurrencies. Adopting an ecological perspective, we show that the so-called neutral model of evolution can reproduce several key empirical observations, despite its simplicity and the assumption of no selective advantage of one cryptocurrency over another. Our results shed light on the properties of the cryptocurrency market and establish a first formal link between ecological modelling and the study of this growing system. We also revisit the results after the sudden collapse of the market in January, 2018 showing that the neutral model continues to describe the observed patterns in the data well. This chapter is based on publication [I].

Chapter 5: What is the connection between online sources, their quality and the community driving them and the cryptocurrency market dynamics?

The production and consumption of information on Bitcoin and other cryptocurrencies have grown, along with their market capitalisation. Earlier research focused on predicting Bitcoin and a limited number of cryptocurrencies prices using data generated from several social media platforms and forums. Only one publication investigated the different communities in play that discuss Bitcoin price and their influence on price fluctuations. In Chapter 5, we tackle both issues by quantifying the interplay between the attention paid to cryptocurrencies in Wikipedia and their market performance. We consider the entire edit history of currency-related pages and their views history from July 2015. First, we quantify the evolution of cryptocurrency pages presence on Wikipedia by analysing the editorial activity and the network of co-edited pages. We found that a small community of tightly connected editors are responsible for most of the production of information about cryptocurrencies in Wikipedia. Then, we show that a simple trading strategy informed by Wikipedia views performs better than baseline strategies, in terms of returns on investment, for most of the covered period, although the ”buy and hold strategy” dominates during the periods of explosive market expansion. Work presented in this chapter is based on publication [II].

Chapter 6: How is the usage of Bitcoin in illicit activities -especially in dark markets-influenced by the multiple closures?

The earliest work examined Bitcoin regulations focused on dark markets [6]. Dark markets are commercial websites which are accessible via darknets (e.g., Tor which is a software enables anonymous communication) and often specialise in selling drugs, weapons, and other illicit goods. Bitcoin is the standard currency for trading on dark markets. Recently, dark markets have seen a dramatic increase in their customer base and transaction volume [22]. Multiple successful police raids and scams have shut down many of the largest dark markets [23]. Despite the use of Bitcoin, all the work conducted on dark markets did not investigate their Bitcoin transaction. In Chapter 6, we contribute to the work done both on Bitcoin blockchain transactions and dark markets by analysing the Bitcoin transactions network for 74 dark markets. We investigate the dynamics of 31 dark markets during several shutdowns. First, we show that users migrate quickly to other dark markets following shutdowns resulting in a quick recovery for the total dark markets sales. Second, we describe the characteristics of migrating users. Finally, we study how migrant users coordinate to move to a coexisting dark market following shutdowns. This chapter is based on publication [III].

We then extend the analysis to include drug sales on dark markets and cryptocurrencies' price prediction. Chapter 7 and Chapter 8 are based on these further contributions.

Chapter 7 focuses on the study and prediction of online drug sales using Wikipedia data (publication [IV]). Online drug markets are the primary revenue of dark markets, as roughly three out of four dark markets deal primarily in drugs. A study showed that in early 2016, dark markets drug sales were between 170 USD million and 300 USD million per year [24]. While general drug use statistics are hard to obtain in granular frequency, online drug sales can be inferred from dark markets transaction or websites. In Chapter 7, we introduce a novel method to predict drug markets sales using several drugs' Wikipedia page views. We measure our models' predictive capability by the out-of-sample - "nowcast" - errors. We show that overall Wikipedia pages views enhance prediction accuracy with a mean average error of 43%. We also show that this result is consistent if we split the prediction per country and per drug type. For this analysis, we relied on dataset scraped from 4 dark markets.

In Chapter 8, we analyse cryptocurrency market efficiency and price prediction using machine learning (publication [V]). Cryptocurrencies trading strategies were not only based on online social traces, but algorithmic trading and machine learning algorithms were also used to aid price prediction. The focus of the prediction attempts, however, had been Bitcoin's price. In Chapter 8, we present a machine learning approach to predict cryptocurrencies prices. Our work investigates more than 1600 cryptocurrencies, providing the most comprehensive price prediction study. We show that simple trading strategies assisted by state-of-the-art machine learning algorithms outperform standard benchmarks. Our results show that non-trivial, but ultimately simple, algorithmic mechanisms can help anticipate the short-term evolution of the cryptocurrency market.

In summary, this thesis advances our understanding of the cryptocurrencies and its ecosystem by answering these crucial questions. However, as this is a nascent field of research, our results beg further questions. Future work can move in several directions, for example investigate which factors decide cryptocurrencies survival in the market and what is the influence of developers on cryptocurrencies market dynamics. Another research direction can be examining the different characteristics of dark markets' Bitcoin transactions compared to exchanges and other services on Bitcoin. We discuss in detail these research directions in the last chapter; Chapter 9.

2 Background

Bitcoin was created in 2009 as a medium of exchange [1, 25]. Relying on a decentralised public ledger and cryptography - hence the name cryptocurrency-, Bitcoin offers a borderless, anonymous and transparent currency. The currency is transparent since the entire transaction ledger is publicly available, yet all the addresses are encrypted, and no registration is needed to use Bitcoin. Bitcoin anonymity encouraged the use of Bitcoin in illicit activities [26, 27, 28] , in particular, online drug markets (dark markets) which found in Bitcoin the missing piece of the puzzle to secure its payment process.

After Bitcoin's appearance, many other cryptocurrencies joined the market. More than 2500 cryptocurrencies are being traded at the time of writing these words. Cryptocurrencies are nowadays used both for payments and as speculative assets [29, 30]. Other uses include cross-borders money transfer and various non-monetary uses such as time stamping [25].

In December 2017, the price of Bitcoin reached $20,000$ dollars [4], drawing the attention of a broader audience. The growth of the cryptocurrency market to 750 billion dollars [4] made a clear case for the need to understand and regulate the market. However, cryptocurrencies' dual nature as a payment system and innovative technology challenged economic and regulatory institutions [31]. To date, there is no unifying agreement on the nature of cryptocurrencies or how the markets should be regulated. Nevertheless, the past year witnessed increasing effort from several institutions to regulate and rationalise the whole cryptocurrencies phenomenon. This lack of precise regulations, together with the decentralised virtual aspect of cryptocurrencies created a fertile decentralised environment of innovative, collaborative actors. These actors are often technology enthusiasts and farsighted opportunistic individuals. As a first mover, they benefited from the developing space and built different virtual components that enriched cryptocurrencies systems.

The aim of this chapter is to discuss the cryptocurrencies ecosystem. The first section will discuss how Bitcoin works, delineate its limitations, and outline one of its earliest and most critical applications in online drug markets. The next section will introduce an overview of the leading cryptocurrencies introduced after Bitcoin. The third section is dedicated to the cryptocurrencies' market, discussing the exchanges where cryptocurrencies are traded and the social platforms dedicated or utilised by cryptocurrencies community. The section will also discuss the attempts in research to rationalise Bitcoin and cryptocurrencies' price and market behaviour. Finally, the last section will describe the typical data sources available, which will be discussed in detail later in the next chapter; data collection and preparation.

2.1 Bitcoin, the first cryptocurrency

In 2008, a white paper authored by unknow author using the alias Satoshi Nakamoto titled "Bitcoin" circulated on the cryptography mailing list Cypherpunks [1]. Later that year, the same author launched as an open-source project a realisation of this idea. In the paper, Nakamoto proposed a borderless, decentralised and secured digital currency. Bitcoin technology is based on decades of innovation in databases, cryptography, and network protocols [32]. It intertwines three pieces of technology, namely, Blockchain [33], hash functions [34], and peer to peer networks [35].

Similar to banks, Bitcoin needs to record all transactions in a secure ledger. The Blockchain is Bitcoin's transactions ledger where the entire history of transactions is recorded. One block contains a group of transactions' details and a unique pointer which refers to the previous block in the chain. In contrast to centralised systems or banks, in the case of Bitcoin, the ledger is not stored in one safe place. Instead, everyone using Bitcoin (runs the core software) is connected through a peer to peer network and is storing a replica of the Bitcoin Blockchain. These many replicas are living on multiple machines guaranteeing its safety against system failures or attacks.

While everyone can store and view a version of the ledger, cryptography guarantees the anonymity of the users. No names or personal information are recorded anywhere in Bitcoin's ledger, nor is it necessary to provide private details to download the software and open a wallet. Alternatively, each user has a wallet that is identified by cryptographic address or addresses. A cryptographic address is a collection of numbers and letters. Only these addresses appear in the ledger, ensuring the users' identity to be anonymous.

When a user makes a Bitcoin transfer, a message circulates the network notifying users that a transaction has been made. This transaction is stored in a pool of unconfirmed transactions. For the transaction to be confirmed, it needs to be written in the blockchain.

When it comes to how to write a transaction in the ledger and thus render it valid, Nakamoto proposed to tie the validation of a block of transactions to a computational puzzle, its solver (the miner) would be rewarded with some fixed amount of Bitcoin. The computational puzzle requires finding a block whose SHA-256 hash (a cryptographic hash function) is less than a target value. A hash function digests a string of data of an arbitrary length and returns a fixed-length sequence of characters and numbers. In the case of the SHA-256 hash function, the output data size will be 256-bit (32-byte). The puzzle is solved by trial and error, using a machine to generate many proposals as quickly as possible until it finds a match. For a miner to solve the computational puzzle, s/he groups a selection of transactions from the unconfirmed pool together with a randomly chosen sequence to generate a proposed block. Then the SHA-256 hash function processes the proposed block and returns the 256-bit string. If the output string is less than the specified target value, the block will be confirmed and added to the blockchain. If the sequence does not match the target, the miner keeps trying until a correct hash is found.

Miners are incentivised through a reward determined by a fixed schedule pre-programmed in Bitcoin. Right now, the reward amounts to 12.5 BTC; however, every $210,000$ blocks Bitcoin halves this reward thus regulating the total supply of Bitcoin and controlling its inflation rate. Miners also can be rewarded by fees attached to the transactions they helped record in the blockchain. However, users do not necessarily pay fees. The process of achieving consensus through solving the computational puzzle (mining) is known as proof of work.

If two miners solved the puzzle roughly at the same time, both blocks are appended to the chain, and a fork occurs. Later blocks can join either branch in the chain, and the most extended branch is eventually (after six blocks) accepted while the shorter one's transactions are revoked. This mechanism ensures the integrity of the blockchain. For a dishonest miner to confirm a fraud transaction (double spending), they first must generate a block involving said transaction which solves the hash puzzle, and do so before any competitor in the network. Second, they have to ensure that their branch remains dominant by receiving the majority of subsequent blocks. Both conditions can be satisfied if this miner has more than 50% of the total ability of the network to generate blocks (known as the network hash rate). The speed at which a miner's machine operates to find a correct hash is known as the hash rate. With time, approximately every two weeks the difficulty of the puzzle is adjusted according to the network hash rate. An increase in the difficulty means a smaller number of possible solutions, which means more work for the machines to find the right one. The adjustment of difficulty is essential to keep the transaction validation process within a reasonable time. Right now, mining a new block requires significant processing power; thus, mining pools collectively mine Bitcoin. A mining pool is a combined processing power shared by several miners.

Finally, Bitcoin core is an open-source software developed and maintained by core developers. The software is the realisation of the Bitcoin consensus protocol. It also includes a set of rules which define a valid transaction. The developers of the Bitcoin core software are the gatekeepers of the code. At the beginning of Bitcoin, if a modification of the protocol was suggested or required, developers had to unilaterally accept changes in the protocol rules to be included in the code. In 2012, a mechanism to propose and accept changes was introduced; it is known as Bitcoin Improvement Proposal (BIP) [36]. The mechanism allows anyone to propose changes to the protocol and miners to be included in the decision. If 95% of the miners adopted the proposed change, it is confirmed.

In Bitcoin, everyone can download the software, transfer fee-free money, store the ledger and even maintain it, democratising the control over the system. While the original whitepaper paints an ideal currency, research and reality revealed many drawbacks in the proposal which will be discussed in the next section.

2.1.1 From the white paper to reality

Since Bitcoin appearance, there were concerns and questions of whether Bitcoin does fulfil the goals set out by Nakamoto in the white paper or not [37]. Bitcoin technological stability is crucial to its survival. Stability for Bitcoin is broadly defined as the system ability to continue functioning as a currency while adapting to the growth and malicious attacks [37]. Researchers investigated several aspects of Bitcoin stability, in particular, the stability of the transaction validation rules (part of the core software) [38] and the consensus protocol [39]. As Bitcoin relies on decentralisation to guarantee its integrity and stability, the question of how decentralised Bitcoin is, became closely related to Bitcoin stability. Research also challenged Bitcoin anonymity claims showing that some addresses' identity can be revealed, rendering Bitcoin pseudonymous at best [40]. Finally, Bitcoin limited transaction speed, excessive usage of electricity and its complex technology stand as barriers to its adoption.

In the case of the transaction validation rules, as we discussed in the previous section, they were hardcoded in the Bitcoin initial code by Nakamoto with no precise changing mechanism until 2012. Such a limitation motivated Gervais et al. [38] to argue against Bitcoin decentralised nature, citing the fact that only a few developers can change the transaction validation rules. While a mechanism [36] was introduced to suggest and accept proposals, the core developers remain the only entity capable of applying these changes and whether this mechanism guarantee decentralisation in Bitcoin or not is unclear.

The consensus protocol stability, on the other hand, relies heavily on the miners' behaviour. First, in case of a majority miner (miner with 51% or more of the computational power of the network), the miner can decide which transactions to accept and which to ignore. Briefly, in July 2014, the mining pool Ghash.io exceeded 50% of Bitcoin computational power. This is not the only option to break the honest mining assumption; another possibility is a collusion between miners. In fact, there are different possible attack strategies, incentives and conditions in which the stability of the consensus mechanism is under threat [41, 39, 42].

This ability of miners, along with influential players to alter and force specific rules on the Bitcoin network challenges the assumption of decentralisation. Recently, the Bitcoin community witnessed a panic when the CEO of Binance (the biggest cryptocurrency exchange) suggested to "reorganising the chain" after the exchange was hacked and lost $40 million dollars [43]. However, influential actors advised against the idea in fear of squandering trust in Bitcoin. This conclusion was also the main argument of Nakamoto against the majority miner attack (51% miners attack), where s/he argued that in the long term, miners would be better off playing by the rules [1].

Researchers also investigated the wealth distribution in Bitcoin as another aspect of Bitcoin decentralisation nature. By adopting a complex network perspective, several studies analysed the Bitcoin transaction network. The studies showed that the network exhibits a power-law distribution with few addresses own most of the Bitcoins and send/receive most of the transactions [44, 45]. This finding motivated crypto enthusiasts to monitor the transactions of particularly wealthy addresses - known as whales - and their impact on Bitcoin's price [46]. Another study on the Bitcoin transaction network uncovered that 59.7% of the addresses contain "old coins"; which are coins received by addresses 3 months before the time of the study (May 13th, 2012) [47] and were never used. This phenomenon of old coins, also known as dormant addresses, is complicating many measurements needed to understand Bitcoin's behaviour, for example, the total number of active addresses in Bitcoin and the total Bitcoin in circulation.

Concerning the anonymity, recent research uncovered multiple heuristics that can be used to cluster addresses [48, 49]. Clustering the addresses reveal which transactions were made by one user aiding the process of money tracing. Furthermore, the identity of these addresses can be known through direct interaction, as shown in [48].

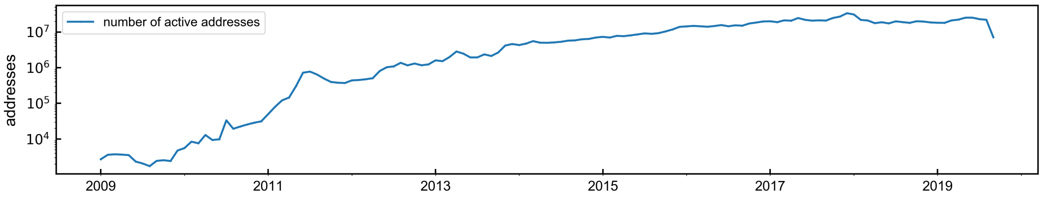


FIGURE 2.1: **Bitcoin addresses evolution.** The total number of active addresses (blue line). Values are aggregated using a time window of one month. Active addresses are addresses sent/received during this time period. Data was extracted from coinmetrics.io website [50].

Finally, there are adoption barriers facing Bitcoin as evidenced by the number of active addresses interacting on the blockchain, shown in Figure 2.1. It is important to note that a new address does not necessarily correspond to a new user since one user can own multiple addresses [1]. The actual number of Bitcoin active users and the cumulative value of their transactions are still an open research question [51]. However, the number of active addresses in 2018 reached $\sim$ 243 million [50] compared to 3.3 billion visa cards [52]. The low adoption rate of Bitcoin as a currency can be attributed to the limited speed of transactions. The transaction confirmation rate, which is currently only 4.5 transactions per second, approximately 380 times longer than the peak capacity of the Visa network [53]. Furthermore, transactions with no fees can suffer from even more delays compared to transactions with fees [54]. In addition to the transaction speed, the extensive use of electricity to mine Bitcoin has been criticised. The work in [55] reported that the Bitcoin mining process requires about as much electricity as the country of Ireland. These issues challenge Bitcoin's adoption and scalability.

Despite Bitcoin's adoption barriers, it was quickly adopted to be the primary currency for illicit online trading. According to a study [56], around 76 billion dollars of illicit activity per year involve Bitcoin. The next section will detail

Bitcoin usage in illicit payments; specifically in online drug markets known as dark markets.

2.1.2 Dark markets and illicit activities

Dark markets are online trading platforms for illicit goods, only accessible through the darknet; which is a restricted access part of the Internet dedicated mainly to illegal peer-to-peer sharing. The first dark market, Silk Road, was established in 2011 [2]. Due to the nature of the commodities on sale, Bitcoin - the anonymous currency - was the obvious currency choice. It continues to be the universally accepted currency in dark markets which are also known as cryptomarkets. The market grossed $\sim$ 313 million dollars of sales in 2 years till it was shut down by law enforcement in 2013. However, on February 2013, other markets opened providing different illicit goods, including drugs and weapons, as well as fake IDs and credit cards. Right now, there are approximately 18 active dark markets where Bitcoin is the primary accepted currency [57].

Dark markets operate mainly as an eBay for illicit goods where vendors advertise their products and consumers request the shipment through the website. Some consumers leave reviews which contribute to vendors reputation [58]. Markets vary in their specialisation, technology, and primary supported language. Silk Road's market terms instituted a ban on trading harmful or fraudulent goods, which limited its sales to drugs. Contrarily, other dark markets allow trading in weapons. Markets such as Unicc and Berlusconi Market are specialised in stolen credit cards and fake IDs.

The closure of Silk Road by law enforcement challenged the security and anonymity claims of dark markets. Later in the same year another market (November 2013), Sheep Marketplace also was closed. This time the market was closed by its administrator who escaped with 100 million US dollars [59]. This type of closure, commonly referred to as scam closure, raised more concerns regarding users' security and trust in the platforms. Because of these concerns, markets started to deploy technologies such as I2P [60], multisig [61], tumblers and rely more often on escrow services [58]. I2P is an anonymous network layer designed to overcome censorship and multisig enables users to authorize a transaction through multi signatures. Tumblers (also known as mixers) are services which obscure the trail back to Bitcoin payments. Escrow services guarantee that markets do not hold users money, instead a trusted third party holds the money until users confirm they have received the shipment. Despite the adoption of different services and technologies, markets' closure continued to occur, especially the scam ones [62, 63].

Finally, most dark markets have an English user interface except for a few markets that use Russian as their primary language. Regardless of all these differences, Bitcoin is the universally accepted currency.

One of the earliest studies relevant to Bitcoin was dedicated to dark markets [2]. However, research on dark markets did not rely on their Bitcoin transactions. Instead, the research analysed dark markets webpages. Using web scraping, researchers collected reviews from markets webpages and used the review time and details as a proxy for the transactions. Along with the reviews, data regarding vendors and their products were also collected. However, not all markets enforce its users to leave reviews. Also, the actual quantity bought is hard to be inferred from the review, which motivated researchers to consider one feedback to correspond to one unit [63]. Furthermore, dark market admins began to actively fight web scraping as a possible security threat, complicating this type of research and reducing the number of markets under study and the studying period [63, 64].

In early 2013 and based on the dark markets web scrapes, different methods were investigated to estimate Silk Road market volume [2, 65]. Similarly, in the case of Silk Road 2 - another dark market opened in November 2013 -, researchers provided two different estimations of the market volume [63, 66]. While using the same data collection approach, other research approached questions related to vendors and drug's supply chain [67] and the social interaction of opioid drug dealers" [68]. On the other hand, the work in [69, 70] relied on individual participants and customer surveys to address questions related to customers base and experience on the Silk Road market.

Similar to the traditional drug market, the question of the effectivity of police takedowns was raised and studied relying on the markets web scrapes.

Studies such as [71, 62] analysed markets which existed at a time of police enforcement shut down yet were not targeted. They found an increase in the number of these markets drug listing suggesting possible displacement from other markets. The most comprehensive study, analysing 12 dark markets, concluded that "the effect of law enforcement takedowns is mixed at best" [63]. In 2014, law enforcement agencies closed several dark markets in what later came to be known as operation Onymous. An investigation of the operation Onymous showed that the operation had an impact on the drugs' supply and demand but not the prices [64]. It also showed limited displacement of vendors. Whether law enforcement closures are effective or not, it is still an open question [28].

In terms of Bitcoin transactions, research efforts focused only on identifying which addresses are affiliated with a dark market [48]. Overall, a comprehensive study of dark markets evolution over their entire history and their responses to closures - whether law enforcement raid, scam or voluntary due technical issues is still lacking.

The usage of Bitcoin in illicit transactions was not exclusive to dark markets. The anonymity guaranteed to the users of Bitcoin and the lack of well-defined regulations motivated the use of Bitcoin in other illicit activities such as money laundering [72, 26, 73] and ransomware [74, 75]. For instance, over 2 years 16 million USD ransom payments made by 19,750 potential victims [74].

Although a widely used currency in dark markets and illicit online trading, Bitcoin is not the only cryptocurrency in the market. In the next section, we will discuss the cryptocurrencies that followed Bitcoin along with the technological improvements and changes in Bitcoin.

2.2 Bitcoin is not alone

The research on Bitcoin highlighted many vulnerabilities and concerns in Bitcoin protocol and led to changes in Bitcoin and the introduction of new cryptocurrencies. Innovation in cryptocurrencies typically proceeds in the following manners: first, changes of the Bitcoin core protocol or the introduction of services on Bitcoin blockchain to remedy some issues. A mixer is an example of a service which helps users cover their transactions in return of mixing fee. Second, a hard fork in Bitcoin, which means introducing an altered version of Bitcoin that no longer follows the rules of the old one. Third, the introduction of a new cryptocurrency, which can be either radically different from Bitcoin or simply a replica with a new name and minor protocol change.

The scalability problem discussed in Section 2.1.1 led to two recent deployments, Segwit (segregated witness) [76] and lighting network in Bitcoin [77]. Segwit is a protocol upgrade intended to solve a minor bug and reduce the transaction size. The reduction of the transaction size led to increasing the number of transactions a block can include which in return sped up the processing of the transactions. The upgrade to SegWit was activated on July 2017. Another proposal for the scalability problem is the lightning network, which provides another layer to Bitcoin's blockchain where users can transact through specific channels created by them. Since these channels are between the two users, transactions are going to be quicker. While these deployments have clear advantages, it also introduces security hazards [78].

Another approach to solving the scalability problem was increasing the block size from 1MB to 8MB [79]. This suggestion was debated for months and concluded with the introduction of the hard fork Bitcoin Cash. Hard forks occur whenever a new rule is introduced, which renders the new blocks unacceptable. Changes which are backwards compatible on the other hand are considered soft forks they, usually involve further limitations on which blocks are considered valid. Bitcoin forks occurred multiple times, starting from 2014 with the introduction of Bitcoin XT and ending with Bitcoin Gold in late 2017. Many hard forks did not survive, which raises the question of which changes were, in fact, essential [80].

Alternative cryptocurrencies to Bitcoin - altcoins - were created to address certain flaws in Bitcoin; however, some of them are just replicas of Bitcoin. According to coinmarketcap.com [4] - a leading cryptocurrencies data aggregator -, there are $\sim$ 3000 cryptocurrencies actively trading at the time of writing this thesis. In October 2011, Litecoin [81] was one of the first cryptocurrencies to join the market after Bitcoin. The main difference between Litecoin and Bitcoin at the time of the introduction was the block size limit, other than that both cryptocurrencies rely on the same proof of work protocol. Currently, Litecoin is the sixth largest cryptocurrency in terms of the total market capitalisation. Dogecoin [82] and Freicoin [83] are other examples of cryptocurrencies with minor differences from Bitcoin the main difference being their inflation rate.

Other altcoins, on the other hand, proposed more radical changes, in particular in the consensus protocol computational puzzle. The increasing energy consumption - discussed in Section 2.1.1 - encouraged what is being referred to as "useful puzzles" as an alternative. Primecoin [84] is a realisation of this idea, where the computation puzzle is to find a sequence of large prime numbers of mathematical interest. Another alternative to the proof of work is the popular proposal of a proof of stake (PoS), implemented in many successful cryptocurrencies such as Ethereum [85] and NEO [86] The basic idea behind the proof of stake is that, instead of weighting the allocation of block confirmation to miners according to their computational power, the algorithm weights them according to their wealth. There are many variations on how to determine the wealth of the miner, for example, the number of coins accumulated in an account (Ethereum) or coin's age (Peercoin) [87].

A significant departure from the main ideas outlined in Nakamoto's white paper was suggested in [88] where the usage of designated authorities was proposed to back the cryptocurrency. An instantiation of this idea is Ripple. The proposal of an institution backed cryptocurrency was also recently adopted by Facebook [89], and other companies and banks declared their interest as well [90, 91, 92].

Finally, what is known as "stable coins" were introduced as a solution for cryptocurrencies' price fluctuations. These coins use dollars or gold as an anchor for their price. The most popular stable coin -Tether- was introduced in 2012, and the first token was issued in 2015. Regardless of the technological difference, other cryptocurrencies were created to be used within specific geographical boundaries, other to support several applications in healthcare

(SOLVE cryptocurrency [93]), combat fraud in digital advertising (Basic Attention cryptocurrency [94]), content verification (Steem [95]) and even the peer review of academic papers [96].

The previously mentioned cryptocurrencies, together with more than 3000 cryptocurrencies, form a market with a total value exceeding 200 billion dollars. The next section will discuss the cryptocurrency market, its regulations, and the efforts to understand cryptocurrencies prices.

2.3 Cryptocurrency market

In May 2010, Bitcoinmarket.com was launched as the first cryptocurrency exchange. An exchange is a platform that allows users to trade cryptocurrencies for other cryptocurrencies or fiat currencies. Exchanges are essential for cryptocurrencies liquidity and price discovery [97]. Now, there are at least 302 active exchanges. These exchanges vary in their location, the number of cryptocurrencies and assets they trade, and security features. The lack of central authority and regulations governing these exchanges introduced variation in legal consideration and even Bitcoin price. As noted in [98], the differences in price among exchanges allowed for profitable arbitrage trading

There is no uniform definition or regulatory framework followed across the world [6]. While some countries, such as Switzerland and Japan, consider cryptocurrencies a legal tender, others do not. However, for instance, in the UK and the US, cryptocurrencies and exchanges remain legal, subject to official registration [99]. In the case of China, on the other hand, both cryptocurrencies and exchanges are illegal. Surprisingly, 81% of Bitcoin's network hash rate (power to validate transactions) is controlled by Chinese mining pools [100], and exchanges continue to trade in China [101]. Other countries such as Russia, did not provide any regulations yet. Due to these variations in regulations, till 2017, only 52% of the small (in terms of the trading volume) exchanges and 35% of the large exchanges held a formal government license [97].

Most exchanges support trading of a small number of cryptocurrencies. Until 2017, 73% of the small exchanges listed at most 2 cryptocurrencies, while

72% of large exchanges supported 2 or more cryptocurrencies [97]. Each exchange has its own distinct set of cryptocurrency listing rules, and they are often undeclared.

Exchanges can support trading of Bitcoin and other cryptocurrencies without any fiat currency involved. However, many exchanges support fiat currencies, and 63% of the exchanges support the US dollar [97]. Recently, a few exchanges introduced margin trading to Bitcoin and a limited number of cryptocurrencies. Margin trading is a practice of borrowing funds from a broker in order to trade financial assets one does not yet own.

Security features are an essential aspect of exchanges, especially after a spate of hacks in which at least 13 exchanges were targeted since 2011 [102]. Some exchanges provide a feature of cold and hot storage, where the hot wallet is available online, and the cold one is stored offline. This division between hot and cold wallets allows users to store their savings offline while keeping only the amount they are actively trading online. Other security features include two-factor authentication and peer to peer trading.

As the number of exchanges grew along with the number of cryptocurrencies and their respective prices, speculation became a more dominant aspect of the cryptocurrencies' economy. Although Bitcoin was created as a digital currency, its nature is a subject of ongoing debate [6, 7]. The work in [7] analysed the exchange trading volume of Bitcoin in comparison to its transaction volume and showed that Bitcoin's users are mostly using Bitcoin as a speculative asset and not as a means of payment. Comparisons between Bitcoin and fiat currencies [103, 104] and gold [105] have also been drawn. In terms of its role in the financial market, Bitcoin showed some similarities to gold and was concluded to be "something" in-between gold and the dollar [104]. The argument that Bitcoin is not being used for payment is countered by the increase of venues accepting Bitcoin and digital banks that offer Bitcoin as one method of payment.

The comparison between Bitcoin and traditional assets also motivated the usage of algorithmic trading, online social traces and machine learning to predict Bitcoin price, similar to the previous successful attempts in stock markets [106, 107, 108, 109]. This connection opened the door to investigating many factors that could influence Bitcoin's price, be they economical, such as the financial stress index, or social, such as word of mouth. Online social traces were particularly important due to the hype around the Bitcoin bubble after its price exceeded $20,000$ dollars in December 2017.

Google search trends and tweets were found to have a positive correlation with Bitcoin prices [9]. While the previous work focused on the number of tweets, the sentiment expressed in the tweets was investigated in [10]. The study reported a positive correlation between emotional tweets and Bitcoin's price and exchange volume. However, this correlation was not confirmed with Granger causality; on the other hand, emotional sentiment in the tweets was caused (Granger causality) by higher trading volume. In [13], twitter sentiment was used to predict the increase or decrease of Bitcoin prices with 76.23% accuracy. The relationship between the price of Bitcoin and Wikipedia page views was also studied in [11], a bidirectional relationship between price and both Wikipedia and Google Trends was found.

The relationship between these social factors and the price of Bitcoin motivated researchers to build a trading strategy guided by them. In [8], a trading strategy was proposed for Bitcoin based on several social factors such as word of mouth volume, polarisation and emotional valence, all measured using Twitter data. It also included other factors, namely, search volume on Google, Bitcoin transaction volume on the blockchain and number of downloads of the most important Bitcoin client. They were able to introduce a trading strategy that outperforms many baseline strategies, including momentum trading and the buy and hold strategy. Using similar social and economical factors, the work in [17] found a positive feedback loop between Bitcoin price bubbles and word of mouth. The study also reported another positive feedback loop between Bitcoin's price and the number of its new adopters. Deep learning was also used to build a successful Bitcoin trading algorithm [110]. The algorithm relied on data from Wikipedia page views, Google search trends, the Bitcoin forum [111] and cryptocurrencies news website access [112]. Other algorithmic trading strategies were also developed relying on Google search volume and Wikipedia page views and were shown to outperform other benchmark trading strategies [113]. Finally, machine learning algorithms and algorithmic trading were used to predict Bitcoin price independent from social or economical factors [14, 114, 9] showing the efficacy of the methodology and the inefficiency of the market. Bitcoin prices were predicted using random forests [114], Bayesian neural network [115], long short-term memory neural network [116] and other algorithms [117, 118]. These studies were able to anticipate, to different degrees, the price fluctuations of Bitcoin, and revealed that best results were achieved by neural network-based algorithms.

With the increasing relevance of other cryptocurrencies in the market [4], a few papers investigated the most notable cryptocurrencies besides Bitcoin. The work in [15] analysed the change in the prices of seven cryptocurrencies for the period from May 2013 to June 2014, investigating whether the upon introduction of a new cryptocurrency, the market follows the law of reinforcement or substitution. While the reinforcement effect describes a system in which competition does not endanger, and even boosts, the position of the dominating actor, the substitution effect, by contrast, denotes the opposite dynamics, in which new competitors gradually eat away at the market share of the established actors, until an equilibrium is reached. The work found different dominating effects along the period of study. At first, the market exhibited no signs of either effect. Later, around October 2013, the market exhibited a substitution effect, but quickly shifted to reinforcement effect, favouring Bitcoin against all other cryptocurrencies. The work was pioneering in its questions given the little research work on the economics of cryptocurrencies done up to that point and its coverage of different cryptocurrencies. Although the research covered one year only and only the early life of the fledgeling cryptocurrencies, the rapidly shifting competitive dynamics of the market could be observed.

Price prediction, as had previously been attempted, with some measure of success, for Bitcoin, quickly became a focus in the study of these secondary cryptocurrencies. Social platforms and online search data were also used for prediction. Comments and replies on Bitcoin, Ethereum and Ripple forums [111, 119, 120] were found to Granger cause their respective price changes [121]. However, Bitcoin price was more predictable using activity on these forums compared to Ethereum and Ripple. The social news aggregator Reddit has also been the focus of several studies. The platform has dedicated forums (”subreddits”) for various cryptocurrencies [122]. In [123], the Bitcoin and Ethereum subreddit were analysed using topic modelling [124] to identify the mention of which topics preceded price changes. The work showed that discussions on the BitcoinMarkets Reddit related to ”Risk/investment vs trading” precedes price drops. In [125], the authors applied a hidden Markov model which had previously been used to forecast influenza outbreaks to the prediction of cryptocurrency price bubbles based on data. The work relied on data from Reddit for 4 different cryptocurrencies, namely, Bitcoin, Litecoin, Ethereum, and Monero.

Another stream of research employed the framework of wavelet coherence analysis, which was developed to study short, middle, and long term correlations of two or more time series, and has been widely employed, and in finance especially [126, 127, 128, 129, 130, 131]. In [12, 132] wavelet coherence analysis was used to investigate the relationship between several online factors and cryptocurrencies prices’. The focus of both papers was 4 main currencies; Bitcoin, Ethereum, Litecoin, and Monero. However, different factors were examined. While [12] focused on more financial factors, such as gold price and financial stress index with only data from Google search trends to indicate social interest, the work in [132] had social factors focus. They examined data from Reddit [133], Wikipedia and Google search trends. Both papers concluded that the relationships between the social and financial factors and cryptocurrencies prices are stronger in the time of ”explosive prices ”. Machine learning algorithms were also used for cryptocurrencies price prediction [134, 135, 136, 137, 138, 139], however, the attempts came from outside the academic field.

The dual nature of Bitcoin as technological innovation and currency attracted a diverse audience [140]. Technology enthusiasts were the first to register their interest, utilising platforms such as GitHub for collaborative development, and, Reddit, Twitter and Telegram for discussions. They also created their forums, Bitcointalk was created by Nakamoto to discuss Bitcoin and blockchain-related issues [141]. Another forum dedicated especially to dark markets discussion called Dread was created on the darknet [141].

Later on, especially after the Bitcoin price spike at the end of 2017, Bitcoin attracted a more diverse audience [140]. Both consequence and driver of this growth is the attention it has progressively attracted from a broader and broader public. Cryptocurrencies, no longer a niche technology, were increasingly recognised as an investment opportunity. This shift in interest resulted in a shift in discussions and the materials covering cryptocurrencies. Exposés pitched at a less technically interested audience proliferated, a great number of articles about related subjects appeared on Wikipedia, and a dedicated blockchain and cryptocurrency Wiki was created. Financial newspaper such as The Economist and Bloomberg also started covering cryptocurrencies related news frequently [142]. Despite the increasing coverage, 35% of survey participants in the USA said that they did not own any cryptocurrency because they did not know how to buy [143].

While the work mentioned previously showed the importance of the online social traces to limited cryptocurrencies price prediction, an investigation of this diverse community is understudied. The only work that focuses on the community discussions nature and the contributors’ behaviour was introduced in [140]. Using data from the Bitcointalk forum, they found that discussion in innovative coin-related forums tend to be more technical. Their analysis showed that there are two distinct groups of contributors: those who are driven by the market hype and investors on one side and technology enthusiasts who are interested in the advancement of the cryptocurrency system on the other.

Finally, altcoins usage in the illicit economy is still unexplored, even though some of these altcoins were designed to address concerns over Bitcoin anonymity. For example, Monero was designed to be an untraceable shadow coin and is increasingly adopted for illicit transactions [56]. In dark markets, only a few accept other cryptocurrencies such as Litecoin, Dogecoin, Darkcoin and Monero. Among 87 dark markets, only four markets use a cryptocurrency other than Bitcoin exclusively, and ten dark markets allow usage of other cryptocurrencies along with Bitcoin [57].

The research conducted on the cryptocurrency market was made possible by the wealth of publicly available data. Whether it is indicators of online social interactions, numerical data on market developments, or the details of transactions as stored on the blockchain itself, the systems' transparency allowed for a vast amount of public data to be available for study. However, the lack of structure and regulation is a challenge. Many data sources are unconventional, and its quality is heterogeneous. In the next section, we discuss the typical data sources on cryptocurrencies ecosystem.

2.4 Data sources

The entire transaction ledger of Bitcoin (the blockchain) is available to anyone who downloads the "core" software. However, the raw data is not downloadable in a readily accessible format. The Blockchain explorer [144] is a website that provides an API where Bitcoin's blockchain can be retrieved in the form of transactions, which is easier to process and analyse.

The data on the blockchain is anonymous; addresses can not be directly mapped to specific entities. Online services such as Wallet Explorer [145] and specialised startups such as Chainalysis [146] and Elliptic [147] use several heuristics to identify influential entities interacting on the blockchain. Ethereum, Litecoin and Bitcoin Cash also have their own blockchain explorers. For other cryptocurrencies, less support is provided. General data on cryptocurrencies transactions such as fees, number of blocks, active addresses and number of transactions is readily available for a limited number of cryptocurrencies [50]. In the following chapter we discuss Bitcoin blockchain transactions retrieval and clustering.

In addition to the blockchain transactions, startups and crypto enthusiasts maintain an ecosystem of social platforms, blogs, and news venues dedicated to cryptocurrencies, for example Bitcointalk [111], a forum dedicated to cryptocurrencies; to announce new cryptocurrencies and websites that support cryptocurrencies ecosystem. Cryptocurrencies are also often discussed using traditional social media venues such as Reddit [133] and Twitter. Cryptocurrencies also have dedicated news websites such as Coindesk [112] and LongHash [148]. These websites provide more in-depth reporting on several aspects of the cryptocurrencies ecosystem, including the technical advances, market behaviour, major startups and regulations. Along with the news, the websites often provide datasets for the notable cryptocurrencies.

For market data, at the time of starting the thesis, October 2016, data was available on only a few cryptocurrencies and provided by websites also maintained by technology enthusiasts. Coinmarketcap had the most comprehensive public dataset both in terms of the number of cryptocurrencies covered and their details. The website aggregates data from exchanges to one website. In 2018, the website started to provide a limited free API for accessing the data. Now, many exchanges provide data through an API of their order book and price updates for a limited number of cryptocurrencies. Due to the lack of official data provided by the blockchain mechanism itself, websites and exchanges offer these services for a fee.

While cryptocurrencies' data is publicly available, the format is unconventional and unstructured. Through this chapter, we showed how the novelty of the cryptocurrencies technology challenged regulations and sparked interest among technology enthusiasts to build and maintain a vibrant decentralised ecosystem. Due to the novelty of the technology, the definition of key measures, such as the circulating supply, have not yet been universally agreed upon. Instead, data providers defined their measurements themselves and changed them according to new insights within the community. Navigating these data sources and measurements is a challenging task that previously obstructed a comprehensive analysis of the entire market and system dynamics. In the next section, we discuss our data collection process, that led to a unique, comprehensive dataset spanning cryptocurrency market dynamics, Wikipedia activity and on-chain transactions.

3 Data collection and preparation

Our research is based on three novel datasets; cryptocurrency market data, activity on cryptocurrencies' Wikipedia pages and Bitcoin transactions. The cryptocurrency market data was collected by web scraping the data aggregator platform "coinmarketcap". It includes $\sim$ 3074 cryptocurrencies' prices, market capitalisations and exchange volumes and covers a time period from April 2013 until May 2019. The second dataset is Wikipedia pages' views and edits. The data was collected using the Wikipedia API and included 38 cryptocurrencies pages. For each page, views from July 2015 until January 2019 were collected. The data also covers all the page edits since its creation. The third dataset is dark markets' Bitcoin transactions. The data covers 74 dark markets transactions and covers a period from April 2011 until July 2019 .

The decentralised, virtual, and growing nature of cryptocurrencies was reflected in the data sources available. There are multiple sources of data mostly provided by "crypto enthusiasts". For example, coinmarketcap.com, a leading cryptocurrency markets data provider, started in 2013. The website began only in 2018 to publish a blog discussing how they collect and validate the data. The website also released a paid API in 2018, introduced a mobile app, and recently began to provided data on three cryptocurrencies' blockchain transactions. Throughout our data collection we took into consideration this changing and developing nature of the data sources.

This chapter aim is to introduce in detail the datasets upon which our analysis was conducted. It will also introduce the data collection and preprocessing methods. The first section will be dedicated to discussing the cryptocurrency market data. It will introduce the possible data sources for market data, the reasoning behind our choice, the details of the information provided by the source and finally how the data was collected using web scraping. The following section will discuss the Wikipedia data collection and preprocessing details. The third section will introduce the dark markets transactional data. The section will first explain how blockchain data is typically stored and what preprocessing heuristics we follow in order to retrieve dark markets data from the blockchain, then will discuss the details of our dataset. Finally, we will describe the dark market websites data which we relied upon for the drug sales prediction in Chapter 8.

3.1 Market data

At the time of starting this thesis (October, 2016), information on the cryptocurrency market was limited. There was no institute that provided, validated and released the market data systematically. Data providers are typically cryptocurrency exchanges or websites created by "crypto enthusiasts". By now, some of the early data providers have documentation and criteria of cryptocurrencies listing. The lack of a comprehensive data source, APIs and uniformly formatted data had its influence on the research produced. For more than 60 finance papers, a recent study showed a discrepancy in the Bitcoin returns analysis results obtained due to data choice [149].

Finding the most reliable source in a collection of websites was the first step. The initial choice was between extracting data from an exchange or a market data aggregator platform. In case of exchanges, until now, they have no unified listing criteria which results in a limited number of cryptocurrencies and some variation of the cryptocurrencies listed between exchanges. A cryptocurrency's price can vary from an exchange to another [98]. Due to these reasons, we preferred using market data aggregator platforms. In order to determine which platform to use, four criteria were considered. Firstly, the source should include as large a number of cryptocurrencies and as long a period as possible. Secondly, it should cover many exchanges, and finally, information on how the website collected the data should be available.

Table 3.1 shows the most notable available providers of cryptocurrency data. Some of these websites initially were not available or as reliable as they are now, for example Coinmetrics, which only started in 2017. Coinmarketcap.com provided coverage for more than 2000 cryptocurrencies extracted from 273 exchanges and with documentation of the data quality. In 2016, the website did not have the blog section along with data quality section; however, the creator was available for questions.

TABLE 3.1: **Details on online sources for market data.** The table shows information on the publicly available data providers. For each provider the table lists the website used to access the data, the number of cryptocurrencies covered by the provider, the number of exchanges the data aggregated from, whether or not the website provides documentation on the data collection process and the launch date of the data provider. The data shown in the table was collected on the 28th of October, 2019

Website	Cryptocurrencies	Exchanges	Documentation	Start date
coinmarketcap.com	3,047	302	Yes	2013
coincap.io	1,516	71	Yes	2014
coingecko.com	5968	392	No	2014
cryptocompare.com	2,000	196	No	2014
coinlore.com	2,818	331	No	2016
coincodex.com	6,275	252	Yes	2017
coinmetrics.io	75	coinmarketcap.com data	Yes	2017
coinratecap.com	2,788	90	No	2017
coinlib.io	5,970	124	No	2017
coincheckup.com	2,403	120	Yes	2017
onchainfx.com	1,131	12	Yes	2017
livecoinwatch.com	2,165	154	only few sentences	2017
coinorderbook.com	654	12	No	2019

Coinmarketcap.com provides different market measures, namely price, market capitalisation, (24h) volume and circulating supply. The definition of some of these measures evolved across time.

A cryptocurrency's circulating supply at time x was measured by the total number of tokens issued by its protocol up till time x. Cryptocurrencies with proof of work protocol issue tokens as a reward to miners for each block generated, as discussed in Section 2.1. Given this mechanism, the circulating supply will increase at a stable, protocol controlled rate, regardless of whether the coins are used or not. According to [150], one particular Bitcoin address holds over 600 million dollars and made a single transfer. Such addresses often called "dormant" addresses since they have no recent sending activity. Proof of stake cryptocurrencies, for example Ethereum, issue the entire token reserve at the system initiation. For these cryptocurrencies, the calculated circulating supply remains fixed at any time. Due to these variations, the definition has been adapted. Coinmarketcap.com manually contact cryptocurrencies support teams to investigate the fraction of locked addresses and private allocation to be excluded from the circulating supply. Cryptocurrencies which overestimate their supply are removed from the website.

Another measurement provided by coinmarketcap is the total exchange volume in 24h for a cryptocurrency, which is defined as the sum of all trading volume of this cryptocurrency in all exchanges over the last 24 hours. A cryptocurrency's price is measured by coinmarketcap as the weighted average price of a cryptocurrency across all exchanges, where exchanges are weighted according to their total exchange volume. Finally, a cryptocurrency's market capitalisation is given by the product of price and circulating supply.

For an exchange to be considered for listing on Coinmarketcap, it has to meet several prerequisites. Among others, it needs to have a functional website, have been operating for no less than 60 days, and traders must have the option of placing buy and sell orders on an order book. The exchange has to provide a representative for further clarifications as well, especially to respond to claims of exaggerated exchange volumes [151, 149]. In contrast to stock markets, there is no unified regulation for cryptocurrencies to be listed on an exchange or market data aggregator. Cryptocurrencies to be listed on coinmarketcap need to meet certain prerequisites. For example, a cryptocurrency has to be traded publicly on at least two exchanges and has to have a functional website.

Throughout the thesis, we used two datasets scraped from Coinmarketcap.com. One is based on a weekly data readout (used in Chapter 4) while for the more granular analysis in Chapter 5 and Chapter 8 we increased the resolution to daily. The website provides weekly historical snapshots which is a collection of snapshots of the week taken on Sunday nights which are readily available on the website. Using Python, we developed a web crawler which accesses the different dates and scrapes the tables of data. A web crawler is a computer program that accesses web URLs and navigates the webpage to extract data embedded in the HTML. Appendix A.1 shows a schematic illustration of how the crawler works. Given that these are historical snapshots, cryptocurrencies which disappeared from the market were recorded in the data, which is essential for the consideration of the birth and death process of cryptocurrencies (analysed in Section 4.4).

Scraping the daily data was less straightforward. The landing page of the website has the list of all active cryptocurrencies. Each active cryptocurrency has a dedicated page with a section of all the historical data. Through this section, all daily historical data is available for the given cryptocurrency. The web crawler goes through each cryptocurrency page and extracts its data, see Appendix A.2 for an illustration. Following this process, the crawler only retrieves historical data for active cryptocurrencies at the time of collection, which was suitable for building trading strategies and the study of selected cryptocurrencies (Chapter 5 and Chapter 8).

The total number of cryptocurrencies, exchanges and time period covered will be detailed in each chapter since they all cover different time period and have a different objective.

3.2 Wikipedia data

3.2.1 Cryptocurrencies page

The second dataset this thesis relies on is cryptocurrencies' Wikipedia pages data. It was collected through the Wikipedia API [152] and includes the daily number of page views and the page edit history of the 38 cryptocurrencies with a page on Wikipedia. Table 3.2 shows the cryptocurrencies with Wikipedia pages along with some market characteristics. Using Wikipedia

API, we can choose which page views to be considered. For example, the parameter *all-access* filter by access method such as desktop, mobile-app or mobile-web. On the other, the parameter *agent* filter by agent type, for example user or bot. For the page views we use the API call: `https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/en.wikipedia.org/all-access/user/wiki_page/daily/start_date/end_date`, where *wiki_page* is the page name and *start_date* and *end_date* are the requested dates.

Page-view data range from July 1st 2015 until January 23rd 2019, since earlier data are not accessible through the API. On the other hand, the full editing history is accessible through the API, and includes the content of each edit, the editor, the time of creation and possible comments meant to highlight the nature and the reason of the edit. To retrieve the edit history, we use the API call: `https://en.wikipedia.org/w/api.php?action=query&format=json&prop=revisions&rvprop=timestamp%7Cuser%7Ccomment%7Ccontent&&titles=wiki_page`. Since edits retrieved from Wikipedia contain XML and HTML tags, we cleaned each edit by removing all those tags and keeping only the text of the edit.

Automated tools known as "bots" often carry out repetitive tasks to maintain pages. Wikipedia requires bots to have separate accounts and names which include the word "BOT", in order to make their edits identifiable. We excluded all edits from bots from our analysis.

We classified edits into two categories, namely edits with new content and maintenance edits. Maintenance edits aim to keep consensual page content by restoring a more accurate old version (reverts) and fighting malicious edits (vandalism). We identified reverts by selecting edits comments containing the word "rv" or "revert" [153], and by employing an MD5 hashing scheme. MD5 (message digest) is a hashing scheme which digests a variable-length sequence and outputs a fixed-length (128 bits) sequence [154]. MD5 is commonly used to identify identical files. We created an MD5 hash for all edits, and we identified edits sharing the same hash with a previous edit as reverts. Reverts which were made specifically to fight vandalism were identified by selecting edits labelled in their associated comment as fighting "vandalism" [153]. We considered edits which are not classified as vandalism nor reverts as new content.

We also collected data on the activity of the most active editors in other Wikipedia pages. To retrieve this data, we used Xtool [155], a web tool providing general statistics on the editors and their most edited pages. The tool does not provide an API, but through calling the URL of each editor, we accessed the page of the editor and scraped the data.

TABLE 3.2: **Cryptocurrencies with a page in Wikipedia.** The table is generated using data collected on January 23rd, 2019. The table shows, for the cryptocurrencies considered, their name, wikipedia page identifier, the date of their first appearance on the market, page creation date, rank (based on market capitalisation), and whether they can be marginally traded or not (see Appendix C.1 for information on exchanges that support margin trading). Currencies are ordered alphabetically

Name	Wikipedia page identifier	Trading start date	Wikipedia page creation date	Rank	Margin trading
Auroracoin	Auroracoin	2014 – 02 – 27	2014 – 03 – 16	578	No
Bitcoin	Bitcoin	2013 – 04 – 28	2009 – 03 – 08	1	Yes
Bitcoin Cash	Bitcoin_Cash	2017 – 07 – 23	2017 – 07 – 28	6	Yes
Bitcoin Private	Bitcoin_Private	2018 – 03 – 10	2018 – 01 – 18	121	No
Bitconnect	Bitconnect	2017 – 01 – 20	2017 – 06 – 28	Delisted	No
Bitcoin Gold	Bitcoin_Gold	2017 – 10 – 23	2017 – 10 – 15	28	Yes
Cardano	Cardano_(platform)	2017 – 10 – 01	2018 – 01 – 10	12	Yes
Dash	Dash_(cryptocurrency)	2014 – 02 – 14	2014 – 06 – 01	15	Yes
Decred	Decred	2016 – 02 – 10	2017 – 10 – 22	32	No
Dogecoin	Dogecoin	2013 – 12 – 15	2013 – 12 – 14	24	Yes
EOS	EOS.IO	2017 – 07 – 01	2017 – 11 – 30	5	Yes
Ethereum	Ethereum	2015 – 08 – 07	2014 – 01 – 27	2	Yes

Table 3.2 – *Continued from previous page*

Name	Wikipedia page link	Trading start date	Wikipedia page creation date	Rank	Margin trading
Ethereum Classic	Ethereum_Classic	2016 − 07 − 24	2016 − 07 − 25	18	Yes
Filecoin	Filecoin	2017 − 12 − 13	2017 − 08 − 11	1744	No
Gridcoin	Gridcoin	2015 − 02 − 28	2016 − 08 − 30	1179	No
Litecoin	Litecoin	2013 − 04 − 28	2012 − 10 − 20	4	Yes
MazaCoin	MazaCoin	2014 − 02 − 27	2014 − 02 − 28	Delisted	No
Monero	Monero_(cryptocurrency)	2014 − 05 − 21	2015 − 03 − 19	13	Yes
Namecoin	Namecoin	2013 − 04 − 28	2012 − 06 − 27	242	No
NEM	NEM_(cryptocurrency)	2015 − 04 − 01	2014 − 12 − 11	19	No
NEO	NEO_(cryptocurrency)	2016 − 09 − 09	2017 − 12 − 27	16	Yes
NuBits	NuBits	2014 − 09 − 24	2015 − 11 − 03	900	No
Nxt	Nxt	2013 − 12 − 04	2014 − 03 − 09	124	No
OmiseGO	OmiseGO	2017 − 07 − 14	2017 − 09 − 14	30	Yes
Peercoin	Peercoin	2013 − 04 − 28	2013 − 04 − 10	188	No
Petro	Petro_(cryptocurrency)	2014 − 04 − 15	2017 − 12 − 03	1208	No
PotCoin	PotCoin	2014 − 02 − 10	2014 − 08 − 06	413	No
Primecoin	Primecoin	2013 − 07 − 11	2013 − 07 − 29	438	No
Ripple	Ripple_(payment_protocol)	2013 − 08 − 04	2005 − 08 − 06	3	Yes
Stellar	Stellar_(payment_network)	2014 − 08 − 05	2014 − 09 − 04	9	Yes
Tether	Tether_(cryptocurrency)	2015 − 02 − 25	2017 − 12 − 05	7	Yes
Tezos	Tezos	2017 − 10 − 02	2018 − 06 − 28	23	No
Titcoin	Titcoin	2014 − 08 − 26	2014 − 09 − 13	1602	No
Vechain	Ven_(currency)	2018 − 08 − 03	2012 − 12 − 04	34	No

Table 3.2 – *Continued from previous page*

Name	Wikipedia page link	Trading start date	Wikipedia page creation date	Rank	Margin trading
Verge	Verge_(cryptocurrency)	2014 − 10 − 25	2018 − 01 − 24	49	No
Vertcoin	Vertcoin	2014 − 01 − 20	2014 − 03 − 12	175	No
Waves platform	Waves_platform	2016 − 06 − 02	2017 − 04 − 27	21	No
Zcash	Zcash	2016 − 10 − 29	2016 − 10 − 03	20	Yes

We also extracted data of drugs' Wikipedia pages for the prediction of drug sales on dark markets in Chapter 7. The next section will detail the process of data extraction.

3.2.2 Drug pages

We collect Wikipedia page views data through the Wikipedia API (similar to cryptocurrencies Wikipedia pages), which runs from July 2015 onward [156]. The raw data is available at daily frequency, but we aggregate to monthly frequency to match the sales data (discussed in Section 3.3.3).

We further split the data by language, and use that as a proxy for the country of the viewer. This is likely a reasonable assumption for some languages. For example, viewers of the Polish language page are probably located in Poland. However there may be a measurement error, particularly for the English language page which we assume to cover several countries. We limited our analysis to the languages of the top 9 countries trading on the dark market. In order to retrieve data using API for several languages we use the following call `https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/lang.wikipedia.org/all-access/user/wiki_page/daily/start_date/end_date`, where *lang* is the language desired to be retrieved, *wiki_page* is the drug page name and *start_date* and *end_date* are the requested dates. Table 3.3 shows the list of drugs Wikipedia pages retrieved.

TABLE 3.3: **Drug Wikipedia pages** The table shows the list of drugs Wikipedia pages included in our analysis. For each drug the table shows the street name and the Wikipedia page identifier.

Drug street name	Wikipedia page
MDMA	MDMA
Cannabis	Cannabis_(drug)
Amphetamine	Amphetamine
Diazepam	Diazepam
LSD	Lysergic_acid_diethylamide
2CB	2C-B
Ketamine	Ketamine
Methamphetamine	Methamphetamine
Alprazolam	Alprazolam
DMT	N,N-Dimethyltryptamine
Cocaine	Cocaine
Heroin	Heroin
Fentanyl	Fentanyl

Figure 3.1 describes the Wikipedia data over the sample period. Figure 3.1A shows that total monthly views across the period of study are relatively stable over time. Figure 3.1B shows the distribution of views across languages, of which the English pages are unsurprisingly by far the most popular. Figure 3.1C shows the distribution of views across drugs, which is more evenly spread.

3.3 Dark markets data

For the analysis of dark markets behaviour, we used data extracted from the Bitcoin blockchain as provided by [146]. The Bitcoin blockchain (ledger) is

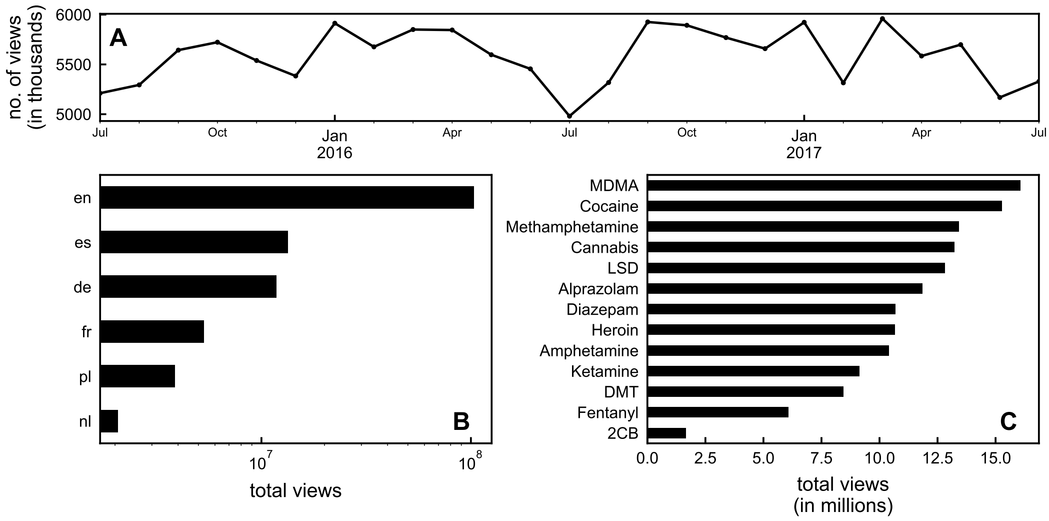


FIGURE 3.1: **Wikipedia data description.** (A) The total number of drugs Wikipedia pages views for all languages and all pages, aggregated to a monthly frequency. (B) Total Wikipedia views over the entire period of study for each language understudy across all drugs Wikipedia pages. (C) Total Wikipedia views over the entire period of study across all languages for each drug Wikipedia page.

publicly available. Data can be retrieved through two methods, downloading the Bitcoin core software or using a third party API.

Bitcoin core software downloads automatically the ledger which amounts to 250 Gigabytes, as of October 2019. The ledger, however, is encrypted and needs processing to be in a readable transactional data format. The ledger is a set of blocks where each block has a pointer that refers to the previous block. A block has the metadata and transaction section. A transaction is listed in a block only once confirmed; there is no record of the actual time a transaction was made. In 2015, a study showed that transactions with zero fees wait for a median of $\sim$ 22 minutes for confirmation [157]. Transaction fees have been predicted to become increasingly important as the mining reward decreases [158]. Blocks are not timestamped either; two methods are used to infer the time a block was registered in the ledger. Mining pools self report their timestamping when they mine a block. Another method is through block height which indicates the block position in the chain, and since a block confirmation takes $\sim$ 10 minutes, block confirmation time is inferred using this assumption.

In the transactions section of each block, transactions are represented as inputs and outputs where inputs are a reference to a previous transaction output. The out section also contains, values in Satoshi, one Bitcoin is divided into 10^8 satoshi.

Third-party APIs process and decrypt the blocks and provide the ledger in a conventional transactional database format. Blockchain.com [159] is the first website to offer an API to retrieve the blockchain in a processed form. For each block, transactions are listed in the form: from address, to address, the value in Bitcoin (BTC) and timestamp. Addresses are only a sequence of numbers and letters (public keys) with no connection to their actual identity. The API also provides which mining pool mined the block.

Identifying the actual identity of the addresses is an open research question. The next section will be dedicated to discussing the techniques first to clustering several addresses to one shared address and second to map addresses to an identity.

3.3.1 Clustering and identification techniques

In Bitcoin, multiple addresses can belong to one user; grouping these addresses will reduce the complexity of the ledger and Bitcoin anonymity [47]. Clustering techniques rely on how Bitcoin's protocol works, user behaviour on the blockchain, Bitcoin's transaction graph structure and finally, machine learning.

Methods which are relying on Bitcoin's protocol specifically exploited what is known as change addresses. Bitcoins available in an address have to be spent as a whole. Figure 3.2 shows an example of a created change address. User A's wallet has two addresses, one contains 1BTC and another has 2BTC. User A would like to transfer 0.25BTC to user B as shown in Figure 3.2A. After transferring the 0.25BTC to B, the change (0.75BTC) will not stay in the same address. Bitcoin protocol will create another address, also assigned to A, where the 0.75BTC change will be stored. By observing this pattern, a heuristic technique proposed in [49] suggested that these addresses can be grouped as they belong to one user.

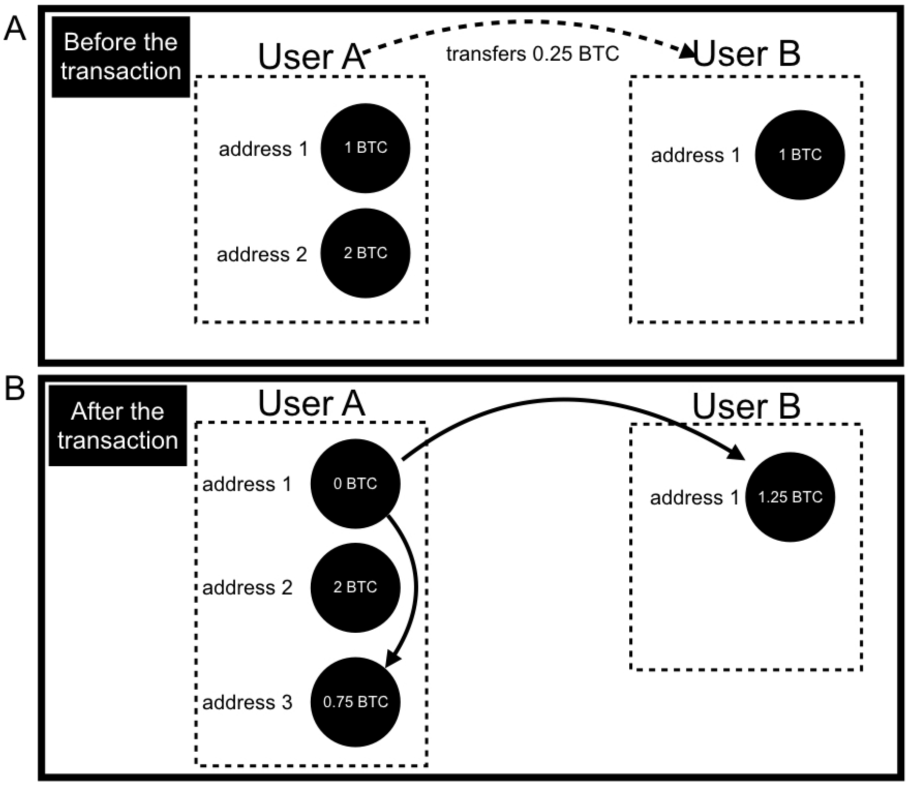


FIGURE 3.2: **How Bitcoin's protocol handls transactions with change.** (**A**) A desired transaction between users *A* and user *B* where *A* would like to transfer 0.25 Bitcoins to *B*. User A has two addresses, one has 1 Bitcoin and the other has 2 Bitcoins. User B has one address and contains 1 Bitcoin. (**B**) How a transaction will be conducted using Bitcoin protocol. User *A* address 1 will transfer 0.25 Bitcoin to user *B* address 1. The change of 0.75 Bitcoin will not stay in User A's address 1 instead will appear as another transaction to a new address from user A address 1. The dotted boundaries in both figures represent a grouping of these addresses as they belong to one user. A solid arrow represents an already executed Bitcoin transaction while the dotted arrow represents a desired transaction.

Since users can have multiple addresses, they can use multiple of these addresses to transfer money in a single transaction. For example, Figure 3.3A shows a case where user A controls 3 different addresses. Each address has a different amount of Bitcoins, 1, 4 and 2.5 respectively. User *A* would like to transfer 5 Bitcoins to user *B*, and two addresses will be used to complete the transaction as shown in Figure 3.3B. This observation allowed the grouping of these two addresses as a single user [49].

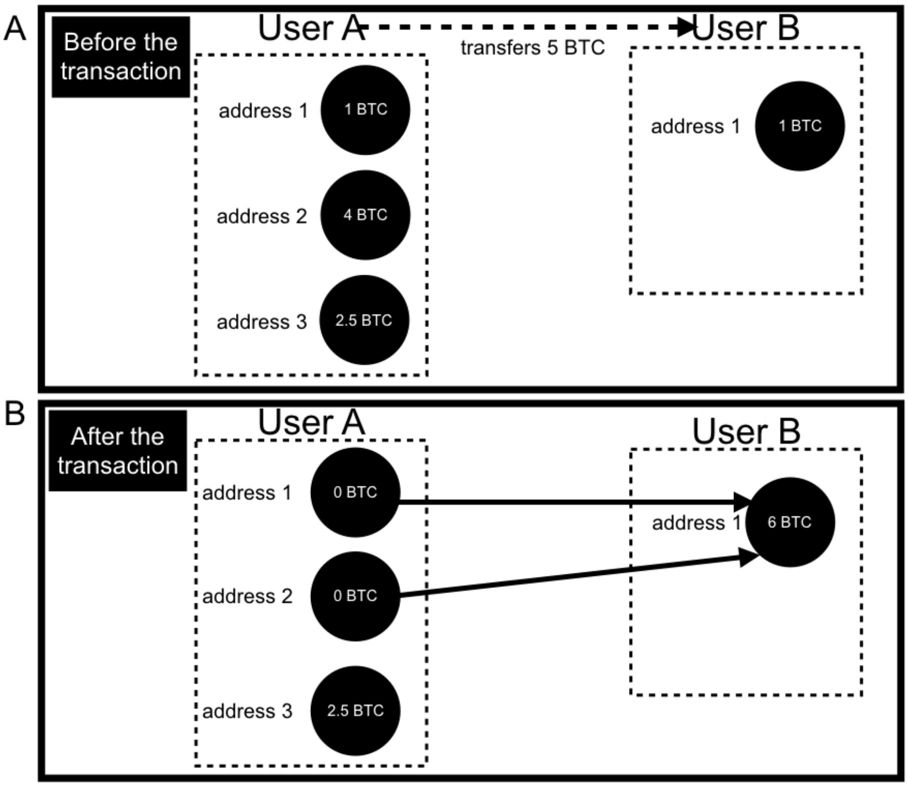


FIGURE 3.3: **Sending from multiple inputs in Bitcoin** (**A**) A desired transaction between user *A* and user *B* where *A* would like to send 5 Bitcoins to user *B*. User *A* has 3 different addresses with 1, 4 and 2.5 Bitcoins respectively. User *B* has one address containing 1 Bitcoin. (**B**) How the transaction will be conducted according to the Bitcoin protocol. User *A* will use two addresses to complete the transaction. Both addresses will send to one address belonging to user *B*. The doted boundaries in both figures represent a grouping of these addresses as they belong to one user. The solid arrows represent an already executed Bitcoin transaction while the dotted arrow represents a desired transaction.

The work in [48] challenged these heuristics showing the possibility of having false positives and not taking into consideration changes in the protocol. The work suggests instead a manual process where the behaviour of each entity is investigated. Page rank (network centrality measure [160]) was also used to identify important addresses [161]; however, the addresses were already grouped using the heuristics introduced by [49]. Machine learning as well was shown to identify addresses which should be grouped as one with 77% accuracy.

Mapping addresses to an actual identity is more challenging. Some entities already publish their public key for donation and payment, such as Wikimedia Foundation [162]. The only research which introduced a method for mapping a collection of addresses to a real-world identity is [48] through direct interaction with the address. In this work, researchers directly engaged in 344 transactions with different services including mining pools, exchanges, dark markets and gambling websites.

The introduction of these heuristics did not only challenge Bitcoin's anonymity but also eased the regulation of Bitcoin. Companies specialising in blockchain analytics started to capitalise on these heuristics and provide tools for exchanges and law enforcement entities to facilitate regulatory efforts. For our analysis of dark markets, our data was provided by Chainalysis [146], which is a blockchain analytics company. Chainalysis aided several investigations led by different law enforcement entities, including the United States Internal Revenue Service (IRS) [163].

3.3.2 Our dataset

Our dataset sampling approach (from the entire Bitcoin transactions) deploys a complex network perspective. Transactions on the blockchain can be modelled as a directed weighted graph where a node represents a user, and a directed edge between A and B represents a transaction from user A to user B. Depending on the clustering algorithm, a node can represent one address or multiple addresses. A node can also be labelled as a specific entity or unlabelled (unnamed). Figure 3.4 shows a sketch of the network and the different possible meanings of a node. For example in Figure 3.4, the black unnamed node on the right side is a representation of two different addresses clustered together, however, they were not attributed to an entity thus remained unnamed.

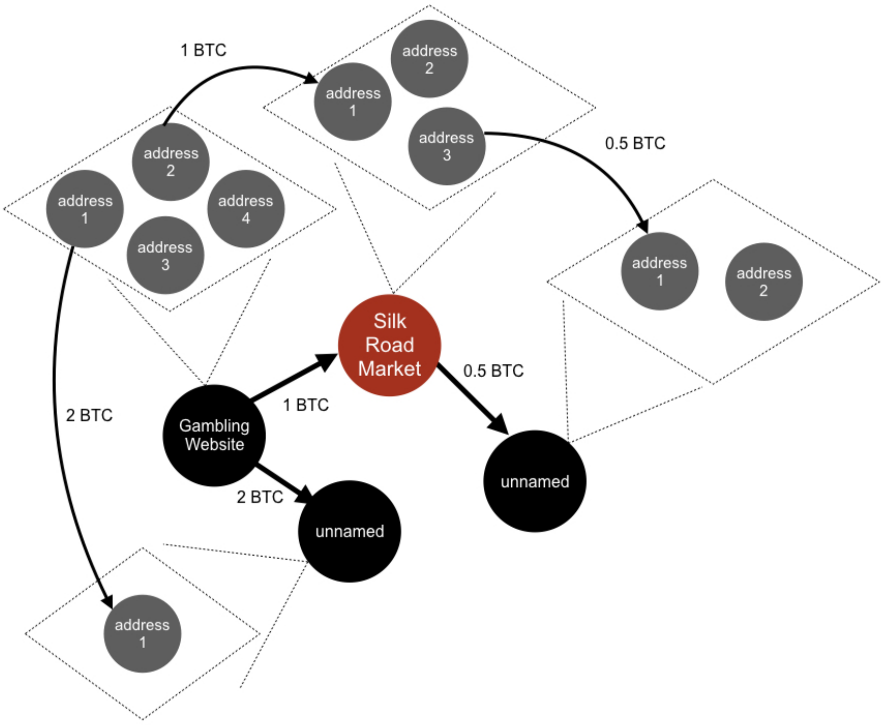


FIGURE 3.4: **A dark market's Bitcoin transaction network.** A schematic representation of our dataset as a complex network. Nodes represent users, and a directed edge between two nodes represents a transaction in the direction of the edge. Nodes can represent different abstractions as shown by the dotted rhombus. Starting from the right side, the unnamed black node represents a cluster of two different addresses which however, was not attributed to a specific entity. The dark market (in dark red) node (Silk Road Market), is a representation of 3 addresses and attributed by the algorithm to the market. The black named node on the left side of Silk Road Market node is a representation of 4 addresses and named to belong to a specific entity. Finally, the black unnamed node at the bottom left side of the figure, represents one address.

For the purpose of our study, we needed to collect only those transactions related to dark markets. Figure 3.5 shows a schematic sketch of the data sampling. We sampled the data by ranking nodes according to their proximity to a dark market. First, list all nodes which are labelled as a dark market. Second, we list all nodes which directly interacted with dark markets; we consider these nodes the nearest neighbours of dark markets. We exclude the exchanges form our nearest neighbours nodes list since the study focuses on markets users' adaptation to closures. For each node in the nearest neighbour list, we compute the first time a node interacted with a dark market. Finally, we extract the entire history of transactions of both dark markets nodes and nearest neighbours. A nearest neighbour node only appears in the dataset after interacting with a dark market. Using this process, we have all the transactions made by dark markets nodes and all the nearest neighbours transactions since their first interaction with a dark market. Other nodes will only appear through interaction with a nearest neighbour; however, their full transaction history is not considered in the data. Other nodes' transactions are only recorded in our dataset if they involve a nearest neighbour of a dark market. Overall our data can be thought of as a collection of sub-networks of different dark markets. Each market sub-network can be represented as an egocentric network [164] of radius 2.

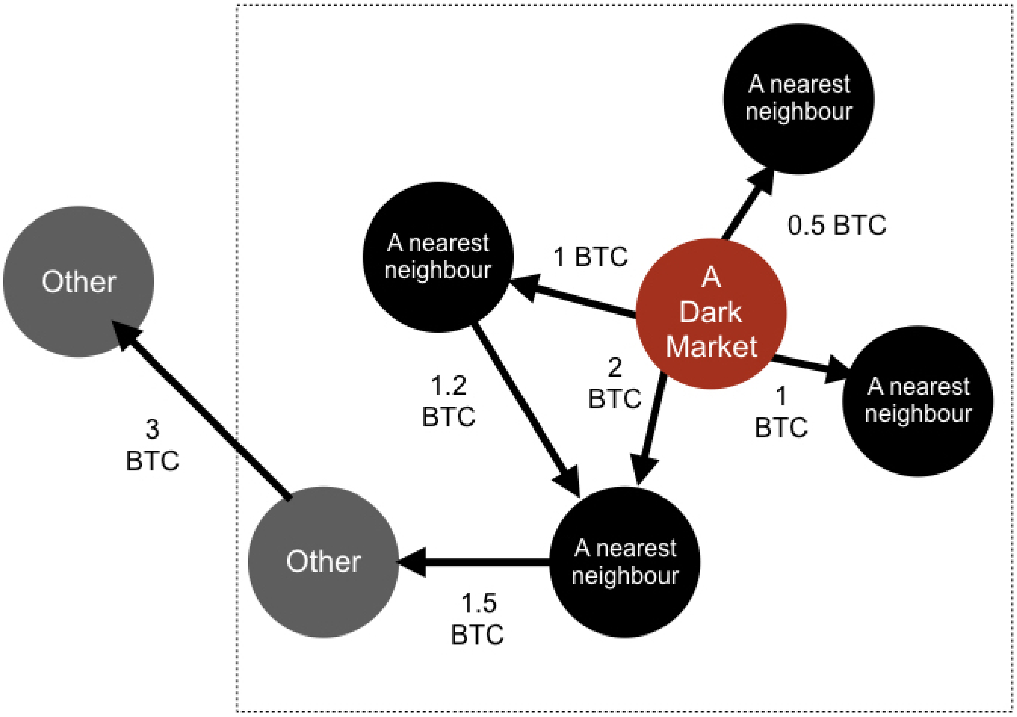


FIGURE 3.5: **Data sampling process.** The included transactions in our dataset are those within the dotted square. The transaction between A dark market node (in red) and the nearest neighbours nodes (in black) are included. The transaction between the nearest neighbours nodes (in black) and Bitcoins to other nodes (in grey) is included. The transaction between two "other" nodes is excluded from our dataset.

Through the aforementioned process, we collected data on 74 dark markets. The dataset covers the entire transaction history of these dark markets from 18th of June 2011 to 24th of July 2019. The data contains around 143 million transactions among over 42 million users. The markets covered in our dataset include the major markets on the dark net as discussed by law enforcement agencies reports [23, 165] and th World Health Organization [28].

We also scraped data from [57] to collect metadata on the dark markets, Appendix A.3, Table A.1 shows general information on the markets included in our dataset. From the website, we gathered data on closed dark markets closure reason, closure date, and the start of their activity.

Among the dark markets, 21 markets are specialised in stolen and fake credit cards and IDs. The rest of the markets are either specialised only in drugs (31 markets) or sell drugs along with other illicit goods (22 markets). Considering only transactions sent and received by dark market addresses, the total transaction volume amounts to 4.5 billion dollars. Markets specialised in selling cards are responsible for 13% of the volume. Out of the 74 dark markets, at least 18 are active, 12 markets were shut down as part of a scam, 12 were raided by law enforcement, and 3 were closed voluntarily by the market administrator.

In our analysis, we selected 31 markets to focus on due to limitation of data availability on markets starting and closure dates (See Table 3.4). Several reasons drove the choice of these markets. First, the existence of publicly available information on their lifetime, closure reason and time. Second, the markets represent most dark markets sales in terms of transactions volume. Finally, given that markets differ in their specialisation, language and closure reason, our chosen markets represent this diversity. Figure 3.6 shows the chosen 31 markets, their lifetime and the reason behind the closure if a market was closed.

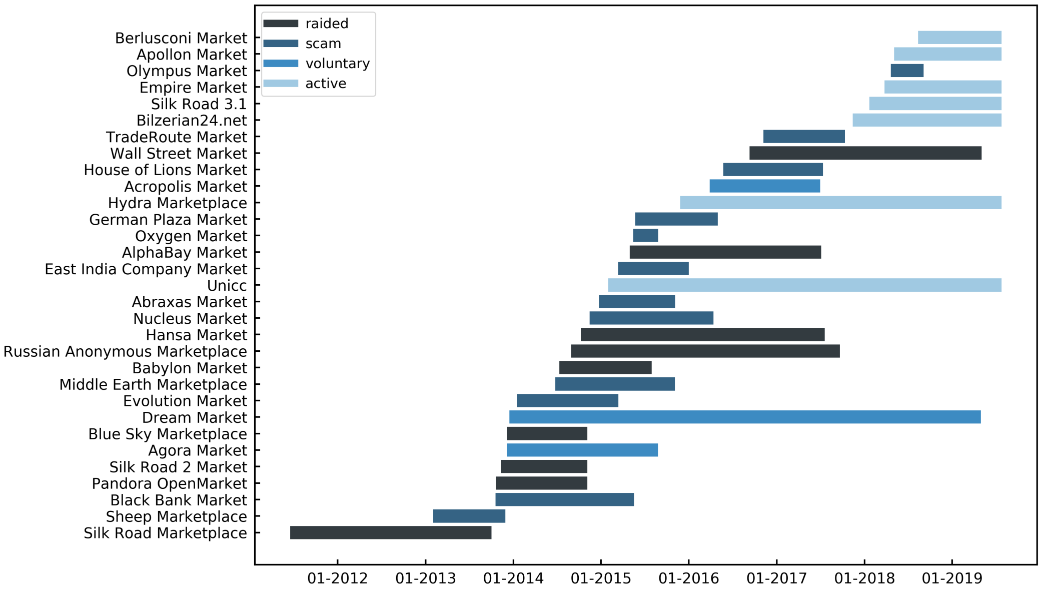


FIGURE 3.6: **Dark markets lifetime** The selected markets beginning and end of the activity as observed in our dataset. Bars are colored according to the closure reason. Silk Road, Hansa and Alphabay are an examples of markets which were raided by authorities (darkest blue), Agora market is an example of a market closed voluntarily (dark blue), Evolution market is an example of a scam closure (light blue) and Empire market is as an example of active market (lightest blue).

The 31 selected markets cover 92% of the total transaction volume of dark markets. Figure 3.7 shows percentage of volume included in our selected markets across time, on average the selected markets have 93% of the total volume.

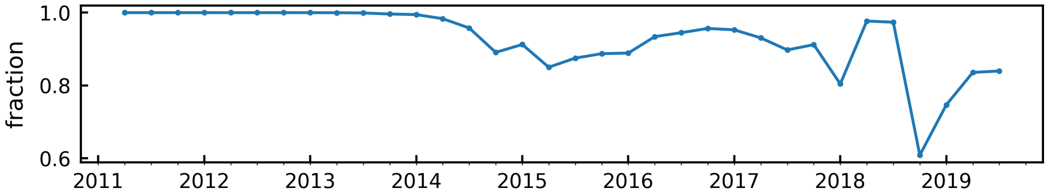


FIGURE 3.7: **Selected markets total volume share.** The fraction of total dark markets volume covered by the 31 markets we are focusing on in our analysis. Values are calculated using a time window of 3 months.

12 of the selected markets were closed in a scam, while 9 were raided, 3 were voluntarily closed by their administrators and 7 are still active. The total number of addresses which directly interacted with all dark market is $\sim$ 10 million users, 84% of these addresses are nearest neighbours of the dark markets we focused on. Among our selected markets 2 markets official language is Russian and 3 markets are specialised in stolen cards. The 31 markets' dataset contains $\sim$ 133 million transactions among over 38 million users. The total number of addresses which directly interacted with dark market is $\sim$ 8.3 million. The volume of transactions sent and received by dark markets addresses amount to $\sim$ 4.2 billion dollars.

TABLE 3.4: **Dark markets information.** Information on the 31 selected dark markets included in our dataset. For each dark market, the table states the name of the market, the start and end dates of its operation, the closure reason if applicable and the type of products sold by the market. "drugs" indicates that the primary products sold on the market are drugs while "credits" indicates the market specialty is fake IDs and credit cards and "mixed" indicates the market sells both types of products.

Name	Start date	End date	Closure reason	Sales
Abraxas Market	2014 − 12 − 13	2015 − 11 − 05	scam	drugs
Acropolis Market	2016 − 03 − 27	2017 − 07 − 01	voluntary	mixed
Agora Market	2013 − 12 − 03	2015 − 08 − 26	voluntary	mixed
AlphaBay Market	2014 − 12 − 22	2017 − 07 − 05	raided	mixed
Apollon Market	2018 − 05 − 03	active	active	drugs
Babylon Market	2014 − 07 − 11	2015 − 07 − 31	raided	drugs
Berlusconi Market	2018 − 08 − 12	active	active	mixed
Bilzerian24.net	2017 − 11 − 13	active	active	credits

Table 3.4 – *Continued from previous page*

Name	Start date	End date	Closure reason	Sales
Black Bank Market	$2014-02-05$	$2015-05-18$	scam	mixed
Blue Sky Marketplace	$2013-12-03$	$2014-11-05$	raided	drugs
Dream Market	$2016-03-19$	$2019-04-30$	voluntary	mixed
East India Company Market	$2015-04-28$	$2016-01-01$	scam	drugs
Empire Market	$2018-02-01$	active	active	mixed
Evolution Market	$2014-01-14$	$2015-03-14$	scam	drugs
German Plaza Market	$2015-05-22$	$2016-05-01$	scam	mixed
Hansa Market	$2014-03-09$	$2017-07-20$	raided	drugs
House of Lions Market	$2016-05-23$	$2017-07-12$	raided	drugs
Hydra Marketplace	$2015-11-25$	active	active	mixed
Middle Earth Marketplace	$2014-06-22$	$2015-11-04$	scam	mixed
Nucleus Market	$2014-10-24$	$2016-04-13$	scam	mixed
Olympus Market	$2018-04-20$	$2018-09-04$	scam	mixed
Oxygen Market	$2015-04-16$	$2015-08-27$	scam	drugs

Table 3.4 – *Continued from previous page*

Name	Start date	End date	Closure reason	Sales
Pandora OpenMarket	2013 – 10 – 20	2014 – 11 – 05	raided	drugs
Russian Anonymous Marketplace	2014 – 08 – 29	2017 – 09 – 21	raided	mixed
Sheep Marketplace	2013 – 02 – 28	2013 – 11 – 29	scam	drugs
Silk Road Marketplace	2011 – 01 – 31	2013 – 10 – 02	raided	mixed
Silk Road 2 Market	2013 – 11 – 06	2014 – 11 – 05	raided	mixed
Silk Road 3.1	2018 – 01 – 21	active	active	drugs
TradeRoute Market	2016 – 11 – 06	2017 – 10 – 12	scam	mixed
Unicc	2015 – 01 – 30	active	active	credits
Wall Street Market	2016 – 09 – 09	2019 – 05 – 02	raided	mixed

3.3.3 Drug sales on dark markets

We also analysed dark markets using data scraped from the websites. We relied on data collected for a previous study [67] which contains a snapshot of product listings and buyer reviews of several dark markets, namely Alphabay, Dream Market, Hansa, Traderoute, and Valhalla in the summer of 2017.

For each market, a single snapshot of the full catalogue of their website was scraped in late June to late July 2017. Trading volumes were estimated from buyers reviews for the listed products, similar to previous work in the literature [2, 16]. Every review is taken to correspond to one purchase, even if multiple items were mentioned in the review. Buyer reviews are not mandatory on all markets, thus the resulting estimates represent a lower bound of the number of trades. Through this approach, the data provided the reviews of Alphabay and Traderoute, the last 6 months of Hansa's reviews, and the last 3 months of Valhalla reviews. In total, the collected data contains almost 1.5M trades.

Vendors list their location which later on was standardised by mapping them to ISO 3166 country codes. Based on the product's title and category, products were recategorised to wider category. Products labelled under terms such as "bud", "weed","hash", "cannabis", "cannabis concentrates", or similar were labelled as Cannabis. Finally products that were labelled or title contained "heroin", "morphine", or "opium" were titled Opiates.

In Chapter 6 we investigate the ability to predict drug sales using Wikipedia page views. For this analysis, it is important to investigate the data stationarity. Drug sales on the darknet have risen over time. This means the sales data is nonstationary, which is problematic for assessing time series model performance [166]. Figure 3.8 shows global darknet sales for MDMA where the sales over time are growing rapidly, so they are not stationary. We formally test for stationarity in Section 7.3.1.

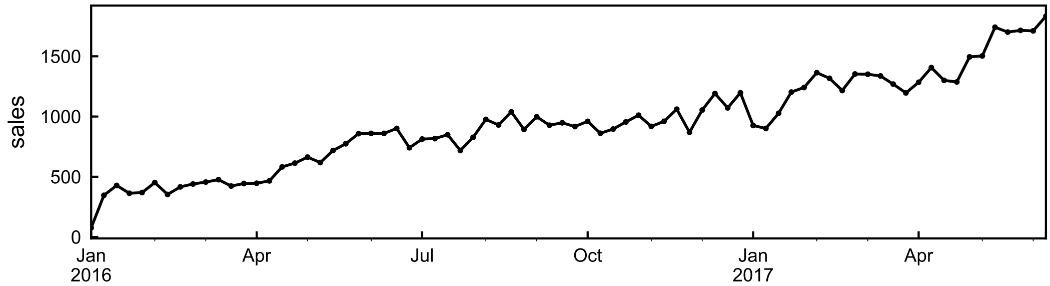


FIGURE 3.8: **Darknet MDMA sales over time. The weekly MDMA sales across the** 5 **dark markets.**

The sales timestamps are continuous, so we could conduct our analysis at different levels of time aggregation. The higher frequency the aggregation, the more granular the measure of drug demand would be. However, higher frequencies make the sales data sparser with more zero observations (see Figure 3.9).

To manage this trade-off, we aggregate the sales data to monthly frequency, which is still much more frequent than the annual official drug surveys (for more details see Section 7.1).

A potential limitation of the scraped review data is that it only captures drug listings that were still available in June-July 2017. If a vendor were to create a listing and remove it before that point, we would not observe any of the sales in the scrape. We could reduce the impact on our analysis by limiting our data to be as close to the scraping period as possible. For example, if we only consider sales from May - July 2017 then there would be far fewer removed listings. However this would also reduce our sample size. Instead, we use all available data for our analysis and assess the impact of restricting the sample period.

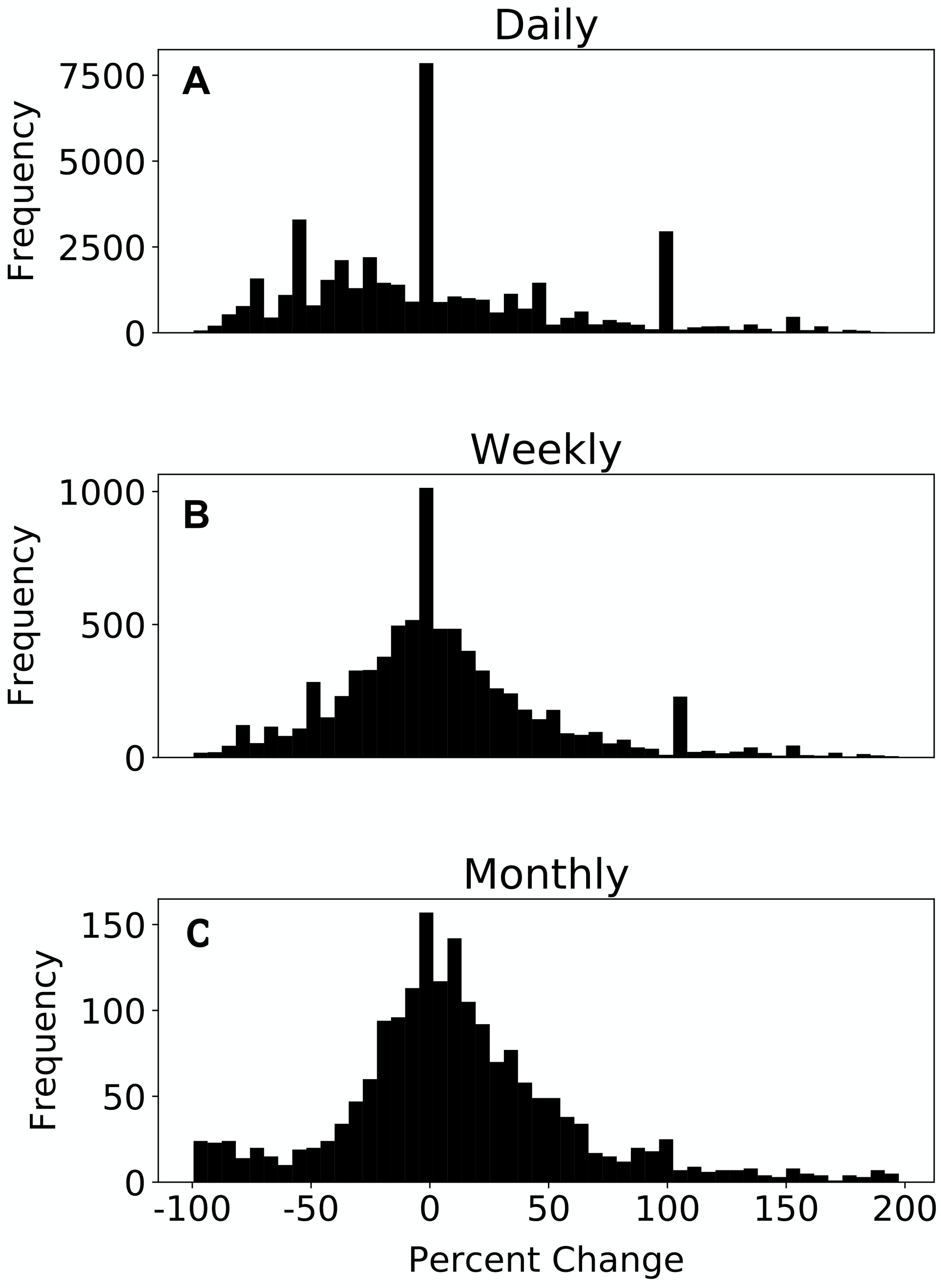


FIGURE 3.9: **Further analysis of sparsity in the drug sales data.** The distributions of the percentage of changes in drug demand at daily (**A**), weekly (**B**) and monthly f(**C**) requencies. The higher frequencies are problematic because the data is more sparse. For example, if we aggregate to daily frequency then 18% of percentage changes are zero. For weekly frequency, this falls to 8% and for monthly to 3% are zeroe. The analysis in Chapter 7 is therefore conducted at monthly frequency.

4 Evolutionary dynamics of the cryptocurrency market

Cryptocurrency market at October 2019 included more than 3000 cryptcourrencies and had a value of $\sim$ 240 billion dollars. All cryptocurrencies share the underlying blockchain technology and reward mechanism, but they typically live on isolated transaction networks. Many of them are basically clones of Bitcoin, although with different parameters such as different supplies, transaction validation times, etc. Others have emerged from more significant innovations of the underlying blockchain technology [97] (see Chapter 2).Bitcoin currently dominates the market but its leading position is challenged both by technical concerns [167, 168, 169, 38, 79] and by the technological improvements of other cryptocurrencies [170], see more details on Bitcoin challenges in Section 2.1.1.

Despite the theoretical and economic interest of the cryptocurrency market [25, 171, 172, 173], however, a comprehensive analysis of its dynamics was lacking. Existing studies have focused either on Bitcoin or on a restricted group of cryptocurrencies (typically 5 or 7) of particular interest (see Chapter 2). But even in this case there is disagreement as to whether Bitcoin dominant position may be in peril [97] or its future dominance as leading cryptocurrency is out of discussion [15].

Here we present a first complete analysis of the cryptocurrency market, considering its evolution between April 2013 and May 2019. We first analyse the period from April 2013 until May 2017. We focus on the market shares of the different cryptocurrencies (see Section 4.1) and find that Bitcoin has been steadily losing ground to the advantage of the immediate runners-up. We then show that several statistical properties of the system have been stable for the past few years, including the number of active cryptocurrencies, the market share distribution, the stability of the ranking, and the birth and death rate of new cryptocurrencies. We adopt an "ecological" perspective on the system of cryptocurrencies and notice that several observed distributions are well described by the so-called "neutral model" of evolution [174, 175], which also captures the decrease of Bitcoin market share. We believe that our findings represent a first step towards a better understanding and modelling of the cryptocurrency market.

Finally, in Section 4.6 we extend the results to the period from May 2017 untill May 2019, reflecting on our results after the work publication in December 2017. We show that despite the prices fluctuation, the market is still well described by the neutral model. On the other hand, ranking dynamics are becoming more stable. The work presented in this chapter is based on publication [I].

4.1 Materials and methods

Cryptocurrency data was extracted from the website Coin Market Cap [4]. The website has changed the definition of some of the measurement (see Section 3.1 for more details on the changes) ; however, these changes did not impact our results. The dataset which covers the period from April 28, 2013 up to May 13, 2017 was extracted before the changes. Results shown in Section 4.6 rely on dataset following the new measurements.

For the first dataset, the website collected data from 157 exchange markets. Now the website provide data relying on 285 exchanges. For all active cryptocurrencies, the website provides the market capitalisation, the price in U.S. dollars and the volume of trading in the preceding 24hours. Data on trading volume was collected starting from December 29, 2013.

The website lists cryptocurrencies traded on public exchange markets that are older than 30 days and for which an API as well as a public URL showing the total mined supply are available. Information on the market capitalisation of cryptocurrencies that are not traded in the 6 hours preceding the weekly release of data is not included on the website. Cryptocurrencies inactive for

7 days are not included in the list released. These measures imply that some cryptocurrencies can disappear from the list to reappear later on.

The *circulating supply* is the number of coins available to users. In the second dataset the website update the calculation of the market capitalisation to consider the dormant coins, see Section 3.1 for more details. The *price* is the exchange rate, determined by supply and demand dynamics. The *market capitalisation* is the product of the circulating supply and the price. The *market share* is the market capitalisation of a currency normalized by the total market capitalisation.

4.2 Market description

Our analysis focuses on the market share of the different cryptocurrencies and is based on the whole history of the cryptocurrency market between April 28, 2013 and May 13, 2017. For this period, our dataset includes 1,469 cryptocurrencies, of which around 600 were active by that time.

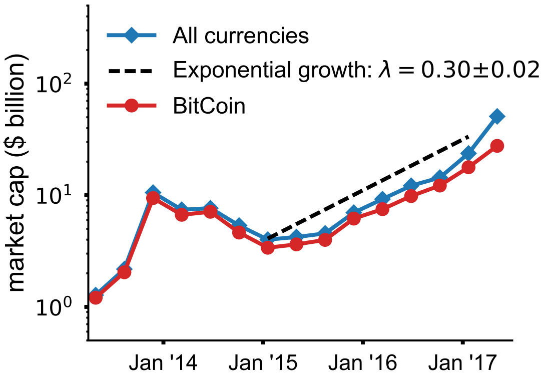


FIGURE 4.1: **Evolution of the market capitalisation.** Evolution of the market capitalisation over time (starting from April 2013), for all cryptocurrencies (blue line,diamonds) and for Bitcoin (red line, dots). The dashed line is an exponential curve $f(t) \sim e^{\lambda t}$, with $\lambda = 0.3$, shown as a guide for the eye. Data is averaged over a 15-week window.

The total market capitalisation of cryptocurrencies (C) has been increasing since late 2015 after a period of relative tranquillity (Figure 4.1). As of May 2017, the market capitalisation is more than 4 times its value compared to

May 2016 and it exhibits an exponential growth $C \sim \exp(\lambda t)$ with coefficient $\lambda = 0.30 \pm 0.02$, where t is measured in units of 15 weeks.

4.3 Decreasing Bitcoin market share

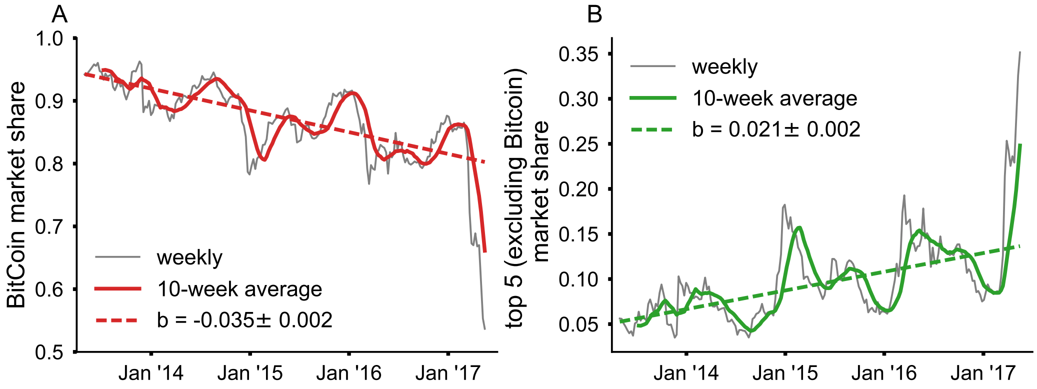


FIGURE 4.2: **Evolution of the market share of top-ranking cryptocurrencies. (A)** The market share of Bitcoin across time sampled weekly (gray line) and averaged over a rolling window of 10 weeks (red line). The dashed line is a linear fit with slope $b = -0.035 \pm 0.002$ (the rate of change in 1 year) and coefficient of determination $R^2 = 0.63$. **(B)** Total market share of the top 5 cryptocurrencies excluding Bitcoin sampled weekly (gray line) and averaged over a rolling window of 10 weeks (green line). The dashed line is a linear fit with slope $b = 0.021 \pm 0.002$ (the rate of change in 1 year) and coefficient of determination $R^2 = 0.45$.

Bitcoin was introduced in 2009 and followed by a second cryptocurrency (Namecoin, see Appendix B.1) only in April 18, 2011. This first-mover advantage makes Bitcoin the most famous and dominant cryptocurrency to date. However, recent studies analysing the market shares of Bitcoin and other cryptocurrencies reached contrasting conclusions on its current state. While Gandal and Halaburdain in their 2016 study concluded that "Bitcoin seems to have emerged - at least in this stage - as the clear winner" [176], the 2017 report by Hileman and Rauchs noted that "Bitcoin has ceded significant market cap share to other cryptocurrencies" [97].

To clarify the situation, we consider the whole evolution of the Bitcoin market share over the past 4 years. Figure 4.2A shows that Bitcoin market share has been steadily decreasing for the past years, beyond oscillations that might mask this trend to short-term investigations. The decrease is well described by a linear fit $f(t) = a + bt$ with slope $b = -0.035 \pm 0.002$ representing the change in market share over $t = 1$ year. Neglecting the impact of non-linear effects and potential changes in the competition environment, the model indicates that Bitcoin market share can fluctuate around 50% by 2025. Conversely, Figure 4.2B shows that the top 5 runners-up (see Appendix B.1) have gained significant market shares and now account for more than 20% of the market.

4.4 Stability of the cryptocurrency market

Figure 4.3A shows the evolution of the number of active cryptocurrencies across time, averaged over a 15-week window. The number of actively traded cryptocurrencies is stable due to similar birth and death rates since the end of 2014 (Figure 4.3B). The average monthly birth and death rates since 2014 are 1.16% and 1.04%, respectively, corresponding to approximately 7 cryptocurrencies appearing every week while the same number is abandoned.

In order to characterize the cryptocurrencies dynamics better, we now focus on the statistical properties of the market. We find that while the relative evolution of Bitcoin and rival cryptocurrencies is tumultuous, many statistical properties of the market are stable.

Interestingly, the market share distribution remains stable across time. Figure 4.4A shows that curves obtained by considering different periods of time are indistinguishable. This is remarkable because the reported curves are obtained by considering data from different years as well as data aggregated on different time spans - from one week to the entire $\sim$ 4 years of data. The obtained distribution exhibits a broad tail well described by a power law $P(x) \sim x^{-\alpha}$ with exponent $\alpha = 1.58 \pm 0.12$ (Figure 4.4A), where the fit coefficient is computed using the method detailed in [177]. The expected relationship between the probability distribution and the frequency rank

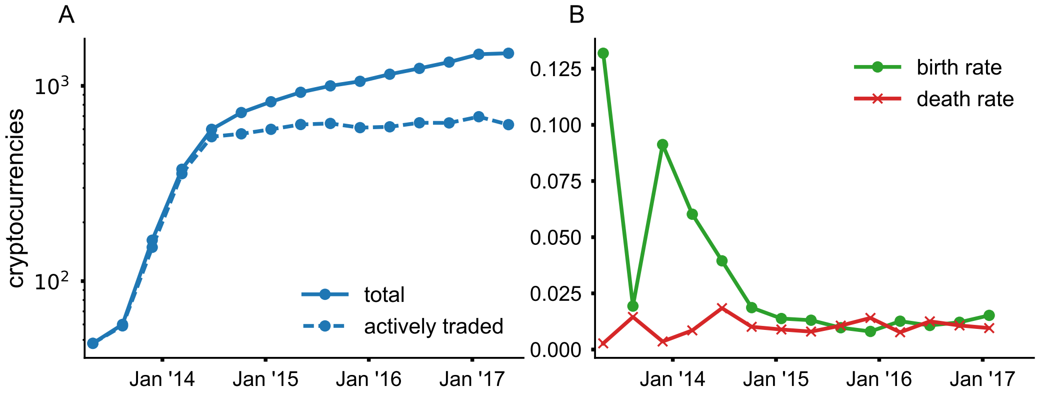


FIGURE 4.3: **Evolution of the number of cryptocurrencies.** **(A)** The number of cryptocurrencies that ever entered the market (filled line) since April 2013, and the number of actively traded cryptocurrencies (dashed line). **(B)** The birth and death rate computed across time. The birth (resp. death) rate is measured as the fraction of cryptocurrencies entering (resp. leaving) the market on a given week over the number of living/active cryptocurrencies at that point. Data is averaged over a 15 weeks window.

distribution predicts the latter is a power-law function $P(r) \sim r^{-\beta}$ with exponent $\beta = 1/(\alpha - 1)$ [178], yielding in our case $\beta = 1.72$ Figure 4.4B. The empirical fit coefficient $\beta = 1.93 \pm 0.23$ is consistent with this prediction. (Figure 4.4B). This was also verified for each year individually (see Appendix B.4).

We further investigate the stability of the market by measuring the average occupation time (l_r) of rank r (Figure 4.4C), defined as the amount of time a cryptocurrency typically spends in a given rank before changing it. This can be calculated using Equation 4.1.

$$l_r = \frac{t_r}{\sum_{i=1}^{N} e_i^r}, \tag{4.1}$$

where N is the total number of cryptocurrencies and e_i^r is equal to one if cryptocurrency i occupied rank r otherwise zero. We find that the time spent in a top-rank position decays fast with the rank, while for low-rank positions such time approaches 1 week. Again, this behaviour is stable across years

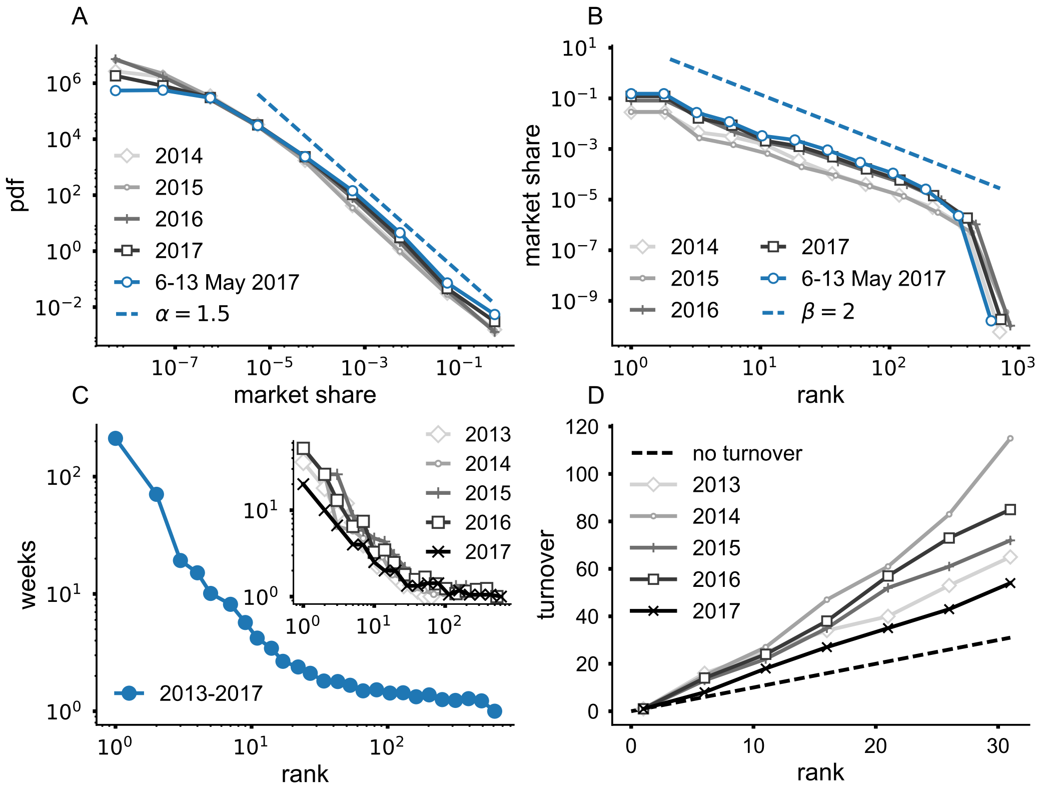


FIGURE 4.4: **Stable properties of the cryptocurrency market.** **(A)** Distribution of market share computed aggregating across a given year (gray filled lines), and over the week 6-13 May 2017 (blue thick line). The dashed line is a power law $P(x) \sim x^{-\alpha}$ curve with exponent $\alpha = 1.5$, shown as a guide for the eye. **(B)** Frequency-rank distribution of cryptocurrencies, computed aggregating across a given year (gray filled lines), and over the week 6-13 May 2017 (blue thick line). The dashed line is a power law curve $P(r) \sim r^{-\beta}$ with exponent $\beta = 2$, shown as a guide for the eye. **(C)** Average amount of time (in weeks) a cryptocurrency occupies a given rank computed averaging across all years (blue line), and across given years (gray lines, inset). **(D)** Turnover of the ranking distribution, defined as the total number of cryptocurrencies ever occupying rank higher than a given rank. The measure is computed averaging across given years (gray filled lines). The 2013 and 2017 curves must be taken purely as an indication as they are computed on less than 12 months (approximately 8 and 4 months, respectively). The dashed line has slope 1, and corresponds to the case in which the ranking of cryptocurrencies is fixed (i.e., the variable turnover captures only the initial size of the toplist).

(Figure 4.4C - inset). We also consider the turnover profile defined as the total number of cryptocurrencies ever occupying rank higher than a given rank in period t (see [179] for a similar definition). To measure the turnover we first compile a list C_r, which include all cryptocurrencies that occupied rank r over a period t. The turnover T_s where s is the number of ranks considered will be equal to Equation 4.2.

$$T_s = |C_1 \cup C_2 \cup .... \cup C_s| \,. \tag{4.2}$$

Figure 4.4D shows that also this quantity is substantially stable across time.

The first rank has been always occupied and continues to be occupied by Bitcoin, while the subsequent 5 ranks (i.e., ranks 2 to 6) have been populated by a total of 33 cryptocurrencies with an average life time of 12.6 weeks. These values change rapidly when we consider the next set of ranks from 7 to 12 to reach 70 cryptocurrencies and an average life time of 3.6 weeks. At higher ranks, the mobility increases and cryptocurrencies continuously change position.

4.5 A simple model for the cryptocurrency ecology

The Wright-Fisher model of neutral evolution describes a fixed size population of N individuals where each individual belongs to one of m species. At each generation, the N individuals are replaced by N new individuals. Each new individual belongs to a species copied at random from the previous generation, with probability $1 - \mu$, or to a species not previously seen, with probability μ, where μ is a mutation parameter that does not change over time [180]. Despite its simplicity, the neutral model is able to reproduce the static patterns of the competition dynamics of many systems including ecological [181] and genetics [182] systems, cultural change [183], English words usage [184] and technology patents citations [185].

In order to account for the empirical properties of the dynamics of cryptocurrencies we have discussed above, we adopt the view of a "cryptocurrency ecology" and consider the neutral model of evolution, a prototypical model in population-genetics and ecology [174, 175]. Figure 4.5 shows a schematic explanation of how the model work.

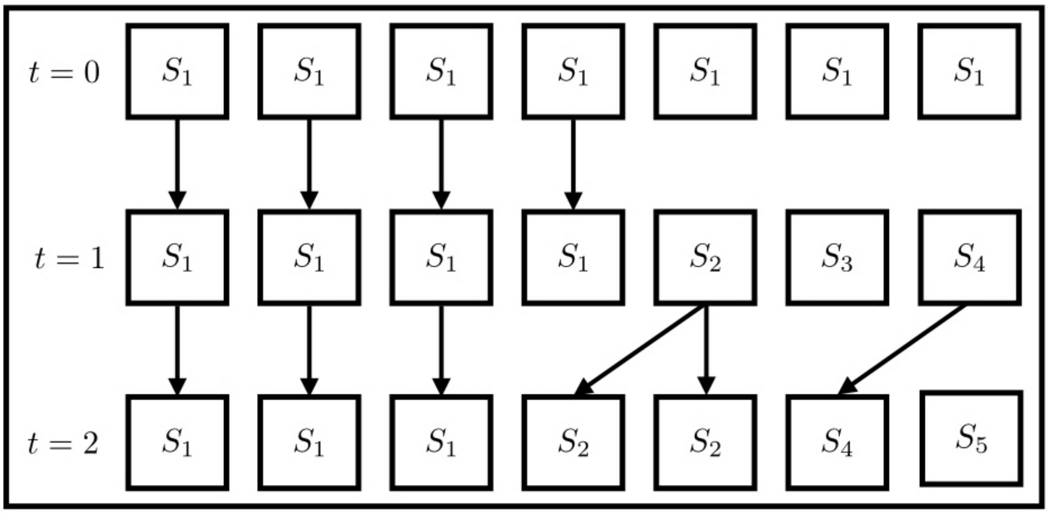


FIGURE 4.5: **Neutral model schematic description.** Dynamics of the neutral model across three generations. The initial condition is one species S_1 at $t = 0$. At $t = 1$, three different species namely S_2, S_3 and S_4 through mutation μ. Specie S_1 number of individuals is declining to 4. At $t = 2$, S_1 still dominating the population due to the higher probability to be selected. Specie S_2 number of individuals increase with 1, while specie S_3 disappeared completely. Specie $S4$ stayed in the third generation with the same number of individual similar to the second generation. The population size is fixed across all generations, J.

In our mapping of the ecological model to the cryptocurrency market, each individual corresponds to a certain amount of dollars, while species correspond to different cryptocurrencies (see Appendix B.2). The copying mechanism represents trading, with μ denoting the probability that a new cryptocurrency is introduced. μ is represents the innovation rate. Our choice of μ is informed by the data to yield a number of new cryptocurrencies per unit time corresponding to the empirical observation. We thus fix $\mu = \frac{7}{N}$, where

N is the population size in the model. Thus, one model generation corresponds to 1 week of observations, the choice of μ guaranteeing an average of 7 new cryptocurrencies entering the system every week, as empirically observed. Finally, in contrast to most neutral models, we assume that a new species does not enter the system with a single individual but with a size proportional to the empirical average market share of new cryptocurrency (see Appendix B.2).

The neutral model translates in the simplest way three main assumptions [186]: (i) interactions between cryptocurrencies are equivalent on an individual per capita basis (i.e., per US dollar); (ii) the process is stochastic; and (iii) it is a sampling theory, where the new generation is the basis to build the following one. In other words, the neutral model assumes that all species/cryptocurrencies are equivalent and that all individuals/US dollars are equivalent.

Testing the consistency between observed patterns of the cryptocurrency market and theoretical expectations of neutral theory revealed that neutrality captures well at least four features of the cryptocurrency ecology, namely:

1. The exponent of the market share distribution (Figure 4.6A);
2. The linear behavior of the turnover profile of the dominant cryptocurrencies (Figure 4.6B);
3. The average occupancy time of any given rank (Figure 4.6C);
4. The linear decrease of the dominant cryptocurrency (Figure 4.6D).

The neutral model generates in fact an aggregated species distribution (i.e., obtained when all generations up to the i^{th} are combined together and analysed as a single population of size $N * i$ [188, 185]) that, at equilibrium, can be described by a power law distribution $P(x) \sim x^{-\alpha}$ with $\alpha = 1.5$ [187] (see Appendix B.5 for more details on the analytical derivation of the exponent), in agreement with the empirical value $\alpha = 1.58 \pm 0.12$ obtained by the fitting procedure described in [177]. Figure 4.6A shows the agreement between simulations and data (same behaviour of the long tail), where simulations results

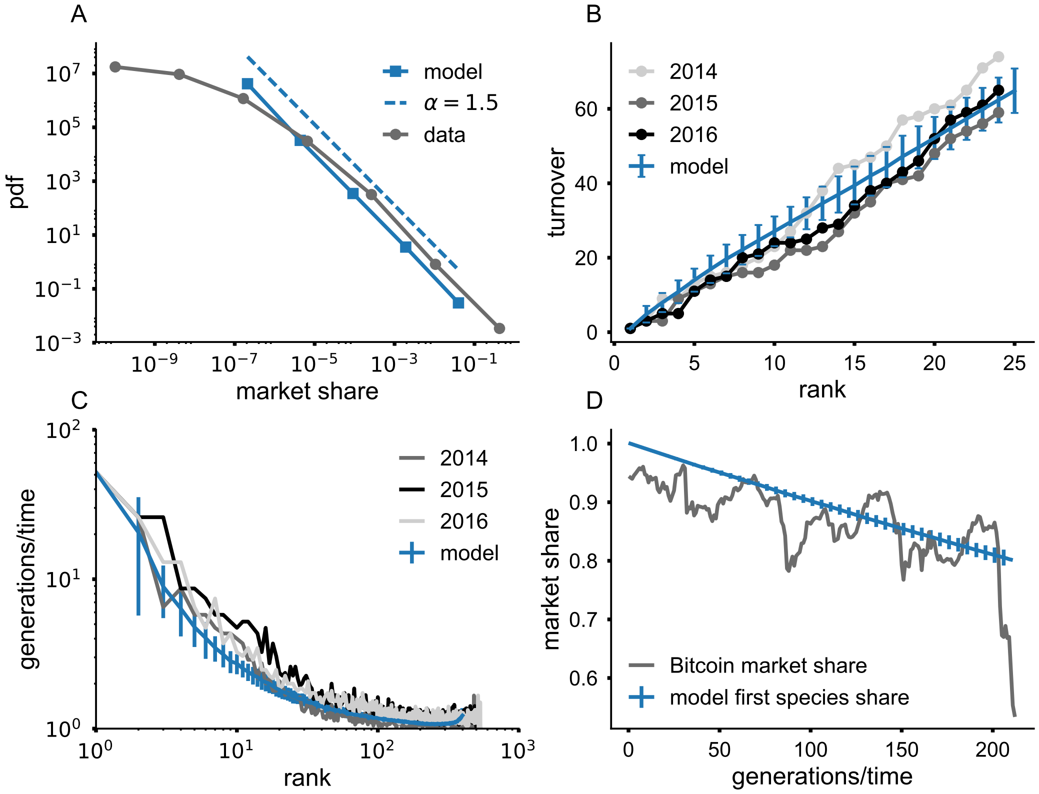


FIGURE 4.6: **Neutral model for evolution and empirical observations. (A)** Distribution of cryptocurrencies market shares aggregated over all years (gray line, dots) and the equilibrium distribution resulting from numerical simulations (blue line, squares) aggregated over 210 generations. The dashed line is the power law curve $P(x) \sim x^{-\alpha}$ predicted analytically with exponent $\alpha = 1.5$ [187]. **(B)** Turnover of the ranking distribution computed considering 52 generations for the cryptocurrencies data (gray lines, dots) and for numerical simulations (blue line), **(C)** Average number of generations a cryptocurrency (gray lines) and a species in the neutral model (blue line) occupies a given rank. Averages are computed across 52 generations. **(D)** Evolution of the market share of Bitcoin (gray line) and the expected market share of the first species in numerical simulations (blue line). All simulations are run for $N = 10^5$ and $\mu = 7/N$ starting from 1 species in the initial state. The size of entering species m, whose average $m = 15$ is informed by the data, is taken at random in the interval $m = [10, 20]$. Error bars are standard deviations, computed across 100 simulations. For panels (B) and (C) measures start at generation $g_1 = 105$ (see Appendix B.2 for variations of this parameter).

are aggregated over $i = 210$ generations, corresponding to 4 years of empirical observations under our choice of μ. The existence of a power law phase with exponent 1.5 in the model is independent of μ (see Appendix B.2) [187].

Furthermore, when we account for the fact that Bitcoin was originally the only cryptocurrency by setting 1 species in the initial state, the model captures also the remaining properties. In Figure 4.6B and 4.6C, we compare the turnover profile and the ranking occupation times with the corresponding simulation results. We compute these quantities over a period of 52 generations, corresponding to one year of observations. The curves reported in Figures 4.6B and 4.6C correspond to measures performed between generation $g_1 = 105$ and $g_2 = 156$, corresponding to year 3 (2015) in the data. Crucially, however, both measures are stable in time, i.e. they do not depend on the choice of g_1 (but for an initial period of high rank variability for the very first generations, see Appendix B.2). It is worth noting that the linearity of the turnover profile in Figure 4.6B corresponds to a similar behaviour observed in [179] when the measure is performed between two consecutive generations. Figure 4.6D shows the observed linear decrease of the leading cryptocurrency market share (Figure 4.6C), indicating that newborn cryptocurrencies mostly damage the dominating one.

4.6 After the publication

The results discussed in the previous section were published in 2017 (publication [I]). After the publication, in December, Bitcoin price exceeded $20,000$ dollars and fell to below $12,000$ dollars in less than a month. These fluctuations led to a question of whether our results are still relevant. In this section, we comment on the changes in the markets and our results in the light of these changes.

The total market capitalisation continued to grow exhibiting an exponential growth $C \sim \exp(\lambda t)$ with coefficient $\lambda = 1.00 \pm 0.06$, where t is measured in units of 15 weeks. The market capitalisation then dropped and continued to decrease until January 2019, where it has been still steadily growing.

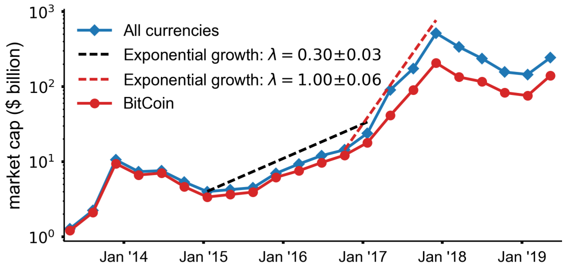


FIGURE 4.7: **Evolution of the market capitalisation.** Evolution of the market capitalisation over time (starting from April 2013), for all cryptocurrencies (blue line,diamonds) and for Bitcoin (red line, dots). The dashed black line is an exponential curve $f(t) \sim e^{\lambda t}$, with $\lambda = 0.3$, shown as a guide to the eye. The dashed red line is an exponential curve $f(t) \sim e^{\lambda t}$, with $\lambda = 1$, for the period after publication until Jan 2018 and shown as a guide to the eye. Data is averaged over a 15-week window.

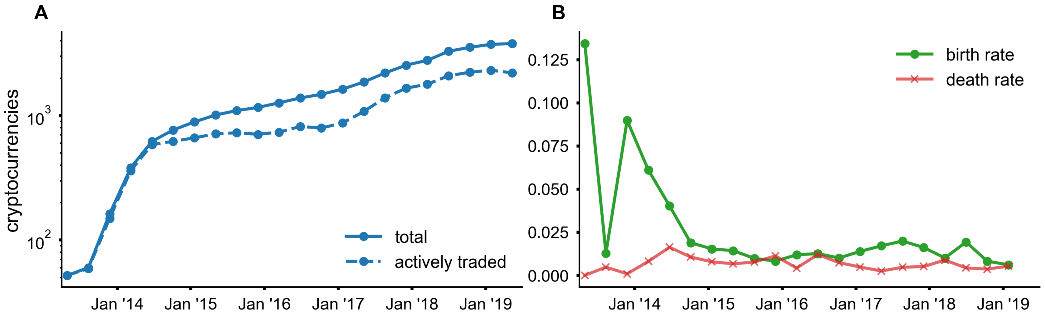


FIGURE 4.8: **Evolution of the number of cryptocurrencies.** **(A)** The number of cryptocurrencies that ever entered the market (filled line) since April 2013, and the number of actively traded cryptocurrencies (dashed line). **(B)** The birth and death rate computed across time. The birth (resp. death) rate is measured as the fraction of cryptocurrencies entering (resp. leaving) the market on a given week over the number of living/active cryptocurrencies at that point. Data is averaged over a 15 weeks window.

In terms of the number of cryptocurrencies, the number of cryptocurrencies has been increasing since the publication, Figure 4.8A. This is mainly due to an increase in the birth rate (Figure 4.8B). From approximately 7 cryptocurrencies entering each week the number of new cryptocurrencies jumped to 16. Death rate on the on the other hand did not significantly change.

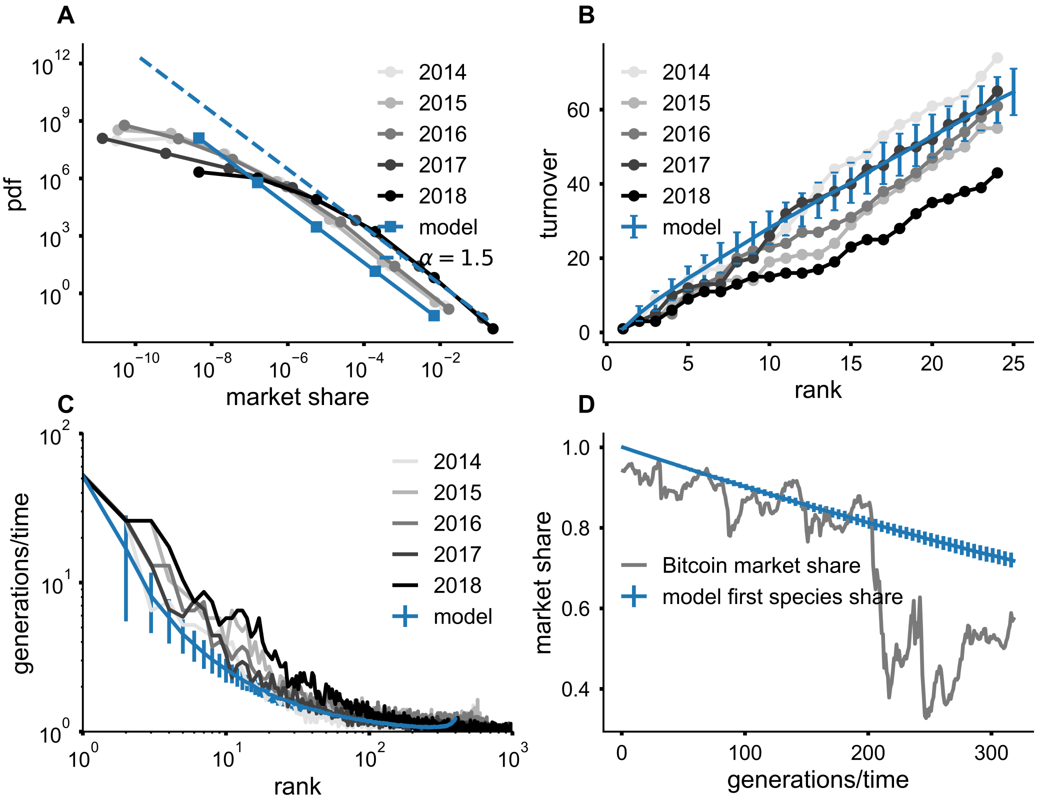


FIGURE 4.9: **Neutral model for evolution and empirical observations. (A)** Distribution of cryptocurrencies market shares over different years from 2014 until 2018 (gray lines, filled dots) and the equilibrium distribution resulting from numerical simulations (blue line, filled squares) aggregated over 318 generations. The dashed line is the power law curve $P(x) \sim x^{-\alpha}$ predicted analytically with exponent $\alpha = 1.5$ [187]. **(B)** Turnover of the ranking distribution computed considering 52 generations for the cryptocurrencies data (gray lines, dots) and for numerical simulations (blue line), **(C)** Average number of generations a cryptocurrency (gray lines) and a species in the neutral model (blue line) occupies a given rank. Averages are computed across 52 generations. **(D)** Evolution of the market share of Bitcoin (gray line) and the expected market share of the first species in numerical simulations (blue line). All simulations are run for $N = 10^5$ and $\mu = 7/N$ starting from 1 species in the initial state. The size of entering species m, whose average $m = 15$ is informed by the data, is taken at random in the interval $m = [10, 20]$. Error bars are standard deviations, computed across 100 simulations. For panels (B) and (C) measures start at generation $g_1 = 105$ (see Appendix B.2

Finally, the neutral model is still capable of replicating the market dynamics, Figure 4.9. In terms of the market shares distribution, the distribution can be described by a power-law distribution $P(x) \sim x^{-\alpha}$ with exponent $\alpha = 1.5$, Figure 4.9A. Figure 4.9B and C show the ranks dynamics similar to Figure 4.6B and C. As shown by the figures, the market is becoming more stable, cryptocurrencies tend to stay on their rank for a longer number of weeks, and the turnover process is slower than the market in 2017. The first rank is still occupied by Bitcoin; however, ranks from 2 to 6 have been occupied by 11 cryptocurrencies with an average life time of 17.4 weeks. According to the neutral model prediction, Bitcoin market share was to drop under 50%, which happened immediately after the publication. Bitcoin stayed for the entire year of 2018 has a share of less than 50%; however it is market share started to increase at the beginning of 2019 as shown in Figure 4.9D.

4.7 Conclusion and discussion

In this chapter we have investigated the whole cryptocurrency market between April 2013 and May 2019. We have shown that the total market capitalisation has entered a phase of exponential growth from January 2015 to January 2017 and continued to grow until January 2019. The market share of Bitcoin has been steadily decreasing until June 2017 to recover in 2018 however since then Bitcoin's market share has been fluctuating. We have identified several observables that have been stable since the beginning of our time series, including the number of active cryptocurrencies, the market-share distribution and the rank turnover. By adopting an ecological perspective, we have pointed out that the neutral model of evolution captures several of the observed properties of the market.

The model is simple and does not capture the full complexity of the cryptocurrency ecology. However, the good match with at least part of the picture emerging from the data does suggest that some of the long-term properties of the cryptocurrency market can be accounted for based on simple hypotheses. In particular, since the model assumes no selective advantage of one cryptocurrency over the other, the fit with the data shows that there is no detectable population-level consensus on what is the "best" currency or that different currencies are advantageous for different uses. Furthermore, the matching between the neutral model and the data implies that the observed patterns of the cryptocurrency market are compatible with a scenario where technological advancements have not been key so far (see Appendix B.3) and where users and/or investors allocate each packet of money independently. Future work will need to consider the role of an expanding overall market capitalisation and, more importantly, try to include the information about single transactions, where available, in the modelling picture.

Another possible direction for future work is to focus on competition in terms of price changes instead of overall market share distribution similar to [15]. Cryptocurrencies market analysis can benefit from adopting a time series analysis approach; firstly, by identifying competition regimes of different maturity level across time [189, 190]; secondly, investigate markets efficiency through price predictability or other statistical properties of the time series and time irreversibility [191]. Another direction could be adopting a complex network approach similar to the work in [192].

In the immediate and mid-term future, legislative, technical and social advancements will most likely impact the cryptocurrency market seriously and our approach, together with recent results in computational social science dealing with the quantification of financial trading and bubble formation [193, 194, 195, 196], could help make sense of the market evolution. In April 2017, for example, Japan started treating Bitcoin as a legal form of payment driving a sudden increase in the Bitcoin price in US dollars [197] while in February 2017 a change of regulation in China resulted to a $100 price drop [198]. Similarly, the exponential increase in the market capitalisation (Figure 4.1) will likely attract further speculative attention towards this market while at the same time increasing the usability of cryptocurrencies as a payment method. While the use of cryptocurrencies as speculative assets should promote diversification [15], their adoption as payment method (i.e., the conventional use of a shared medium of payment) should promote a winner-take-all regime [199, 200]. How the self-organized use of cryptocurrencies will deal with this tension is an interesting question do be addressed in future studies.

5 Wikipedia and cryptocurrencies: interplay between collective attention and market performance

As we have shown in the previous chapter, the cryptocurrency market grew super-exponentially for more than two years until January 2018, before suffering significant losses in the subsequent months. Consequence and driver of this growth is the attention it has progressively attracted from a larger and larger public. In this chapter, we quantify the evolution of the production and consumption of information concerning the cryptocurrency market as well as its interplay with the market behavior. Capitalizing on recent results showing that Wikipedia can be used as a proxy for the overall attention on the web [201], our analysis relies on data from the popular online encyclopedia.

Social media platforms nowadays provide researchers with vast amount of data that can signal public opinions or interests. Since stock markets are highly influenced by the rationale of the investors and their interests, several studies investigated the link between online social signals and stock market prices. Pioneering studies showed how signals from Google Trends and Wikipedia [106, 107] or Twitter sentiment [108, 109] can help anticipate stock prices.

This approach has been recently extended to investigate the relationship between social digital traces and the price of Bitcoin [132, 113, 13, 17, 11, 125, 121, 202, 110], or few top cryptocurrencies [132]. While these studies showed the importance of relying on different digital sources, a systematic investigation of multiple cryptocurrencies has been lacking so far. Furthermore, only in few cases [8, 13, 113], mostly centred on Bitcoin, the analysis incorporated social media signals into an investment strategy in the spirit of the work in [106]. Finally, an analysis of the community driving the discussions and the information on cryptocurrencies was limited to few cryptocurrencies and to discussion platforms such as Bitcointalk forum [140] and Reddit [203].

Here, we investigate the interplay between the consumption and production of information in Wikipedia and market indicators. Our analysis focuses on all cryptocurrencies with a page on Wikipedia, from July 2015 until January 2019. The chapter is organized as follows: In Section 5.1 we describe the data collection and preparation briefly. In the following sections, we present the results of our analysis. Namely, we study the interplay between cryptocurrencies' Wikipedia pages and market properties in Section 5.2; we study in details the evolution of cryptocurrencies pages in Section 5.3; we investigate the role of cryptocurrency pages edits in Section 5.4, and, finally, we explore and investment strategy based on Wikipedia in Section 5.5. The work presented in this chapter is based on publication [II].

5.1 Materials and methods

Wikipedia data was collected through the Wikipedia API [152] and include the daily number of views and the page edit history of the 38 cryptocurrencies with a page on Wikipedia (see Table 3.2 in Section 3.2.1).

Page-view data range from July 1st, 2015 until January 23rd, 2019, since earlier data are not accessible through the API. On the other hand, full editing history is accessible through the API, and includes the content of each edit, the editors, the time of creation and the comments to the edits. For more details on the API calls and data cleaning see Section 3.2.1.

We classified edits into two categories, namely edits with new content and maintenance edits. Maintenance edits aim to keep consensual page content by restoring more accurate old version (reverts) and fighting malicious edits (vandalism). In Section 3.2.1 we describe how we identified reverts and vandalism edits. We considered as new content all edits that were not classified as vandalism nor reverts.

We also collected data on the activity of the most active editors in other Wikipedia pages. To retrieve this data, we used Xtool [155], a web tool providing general statistics on the editors and their most edited pages.

Market data include daily price, exchange volume and market capitalisation of cryptocurrencies, and was collected from the "Coinmarketcap" website [4]. The price of a cryptocurrency represents its exchange rate (with USD or Bitcoin, typically) which is determined by the market supply and demand dynamics. The exchange volume is the total trading volume across exchange markets. The market capitalisation is calculated as a product of a cryptocurrency circulating supply (the number of coins available to users, see Section 3.1) and its price. The market share is the market capitalisation of a cryptocurrency normalized by the total market capitalisation of the market.

The Wikipedia-based investment strategy we implement in this paper can be applied only to "margin traded" cryptocurrencies. We compiled a list of 17 such cryptocurrencies from active exchange platforms including Poloniex and Bitfinex (see Appendix C.2). Note that these are also the most widely traded currencies [4]. In our analysis, we consider that cryptocurrencies can be traded once their trading volume exceeds 100, 000USD. We excluded days where the reported volume did not lie within 2 standard deviations from the average trading volume, which are likely due to how market exchanges report their exchange volumes [204].

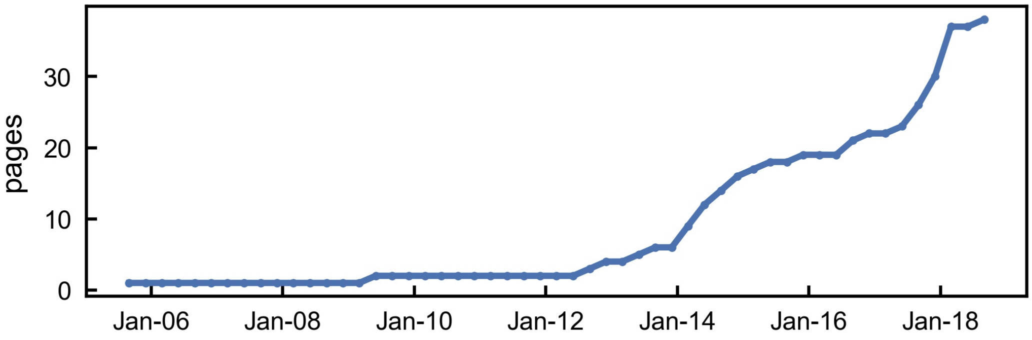


FIGURE 5.1: **Cryptocurrencies on Wikipedia.** Evolution in time of the cumulative number of cryptocurrencies with a Wikipedia page.

5.2 Wikipedia pages and market properties

In this section, we investigate the connection between a cryptocurrency performance in the market and the attention it attracts on Wikipedia. Wikipedia is the 5^{th} most visited website on the Internet [205], attractive to a non-expert audience seeking compact and non-technical information. Previous work has shown that Wikipedia traffic can help to predict stock market prices [106].

The number of cryptocurrency pages on Wikipedia has grown together with their overall market capitalisation. In August 2005, Ripple became the first cryptocurrency with a page. At that point, it was not identified as a cryptocurrency, but as the idea of a monetary system relying on trust. Bitcoin appeared only in March 2009, followed by other 36 currencies (see Figure 5.1).

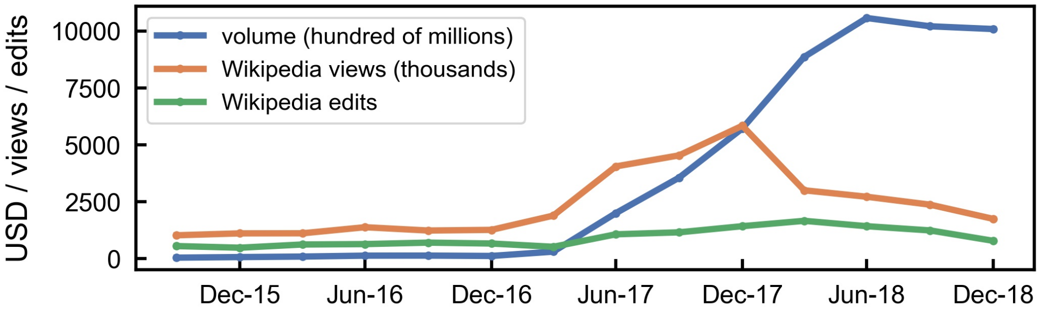


FIGURE 5.2: **Market volume and attention to cryptocurrency pages.** The market volume (USD) for all cryptocurrencies with a page in Wikipedia (solid blue line), the total number of views to cryptocurrency pages (solid orange line) and the total number of edits to cryptocurrency pages (solid green line). Values are aggregated using a time window of 3 months.

The number of views received daily by a Wikipedia page is a good proxy for the overall attention on the web [201]. We find that the number of views to cryptocurrency pages has overall increased from 2015 until January 2018 (see Figure 5.2). In 2016, the 23 cryptocurrency pages were viewed $\sim 4 \cdot 10^6$ times. While in 2017, 34 cryptocurrecies pages received $\sim 16 \cdot 10^6$ views. In 2018, the sudden drop in cryptocurrency prices impacted the number of views. The total number of views received by 38 cryptocurrency pages in 2018 was $\sim 9 \cdot 10^6$. A second aspect characterizing the evolution in time of

Wikipedia pages is their edit history. We find that, on average, pages are more edited than in the past. Cryptocurrency pages (38 pages) were edited $\sim 5 \cdot 10^3$ times in 2018. In 2016, the 23 cryptocurrency Wikipedia pages were edited in total $\sim 2 \cdot 10^3$ times (see Figure 5.2). Bitcoin, in 2016 was the most viewed cryptocurrency page, with views and edits share of $\sim 74\%$ and $\sim 37\%$ overall other cryptocurrency pages, respectively. However, these numbers dropped to $\sim 46\%$ and $\sim 16\%$ in 2018. The fraction of editors active on Bitcoin's page over all other cryptocurrency pages have also dropped from $\sim 34\%$ in 2016 to 10% in 2018. On the other hand, the fraction of views to the 5 most visited pages compared to all other cryptocurrencies has grown from $\sim 20\%$ in 2016 to $\sim 27\%$ in 2018.

Interestingly, Bitcoin's share of the total market capitalisation declined during the same period [206] suggesting a possible connection between the properties of the market and the evolution of attention for cryptocurrencies (see Figure 5.3A). We test this connection considering all cryptocurrencies (see Figure 5.3B) and focusing on other market properties. We find that there is a positive correlation between the average share of views and (i) the average price (Spearman correlation $\rho = 0.37$, $p = 0.02$), (ii) the average share of volume (Spearman correlation $\rho = 0.71$, $p < 10^{-7}$), and (iii) the average market share (Spearman correlation $\rho = 0.71$, $p < 10^{-6}$) of a cryptocurrency. Moreover, these correlations are robust in time (see Appendix C.2).

We also find that the average share of edits of a currency is connected to the overall cryptocurrency performance in the market (see Figure 5.3C). We observe a positive correlation between the average fraction of edits and (i) the average price of a given currency (Spearman correlation $\rho = 0.38$, $p = 0.017$), (ii) the average share of exchange volume for a given currency (Spearman correlation $\rho = 0.67$, $p < 10^{-6}$) and (iii) its market share (Spearman correlation $\rho = 0.68$, $p < 10^{-5}$). These correlations are robust in time (see Appendix C.2).

Note that the observed correlations suggest only a connection between the relative attention to a given currency and its market properties relative to other currencies. Granger casuality test (see Appendix C.8, Table C.6), do not allow to conclude that changes (differences) in Wikipedia views explain

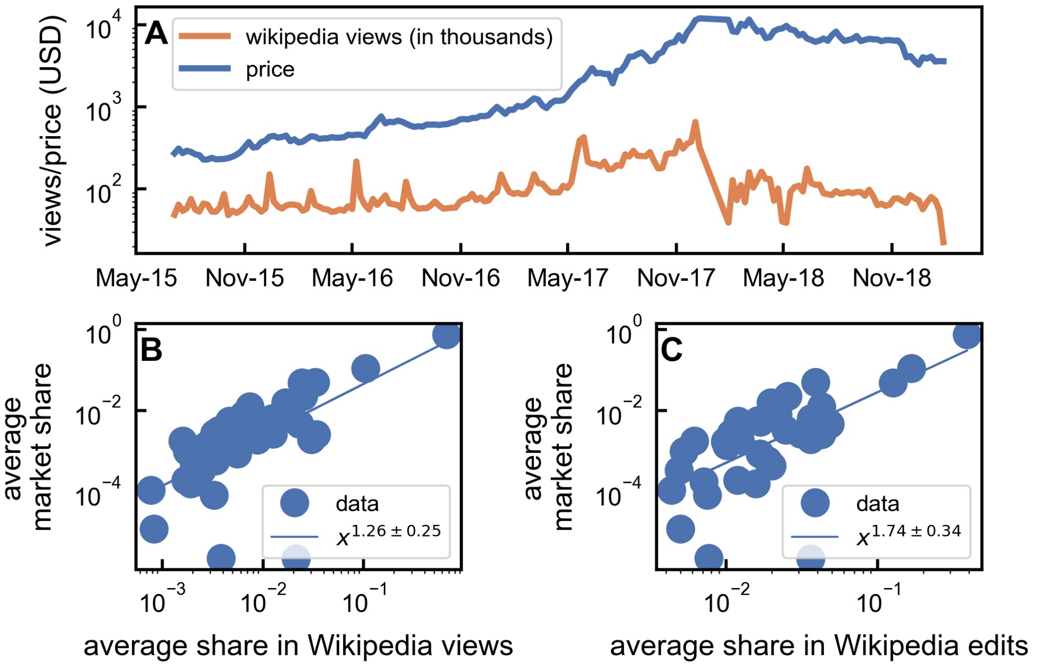


FIGURE 5.3: **Overall correlation between attention on Wikipedia and market performance. (A)** The temporal evolution of price (blue line) and number of Wikipedia views (orange line) for Bitcoin. Averages are computed using a time window of 1 week. **(B)** Values market share in *USD* vs the average Wikipedia views share. Each dot is a different cryptocurrency. (Spearman correlation $\rho = 0.71$, $p < 10^{-6}$). The solid line represents a power law fit of the data with exponent $\beta = 1.26 \pm 0.25$. **(C)** Average market share vs the average Wikipedia edits share. (Spearman correlation $\rho = 0.68$, $p < 10^{-5}$). The solid line represents a power law fit of the data with exponent $\beta = 1.74 \pm 0.34$

changes in prices for individual currencies (the test is passed at $p < 0.05$ by 5 currencies out of 17, considering the difference time series).

5.3 Evolution of cryptocurrency pages

The demonstrated connection between cryptocurrency's success in the market and the overall consumption of information on its Wikipedia page sheds light on the important role played by the latter. In the following sections, we focus on the production of information contained in Wikipedia pages,

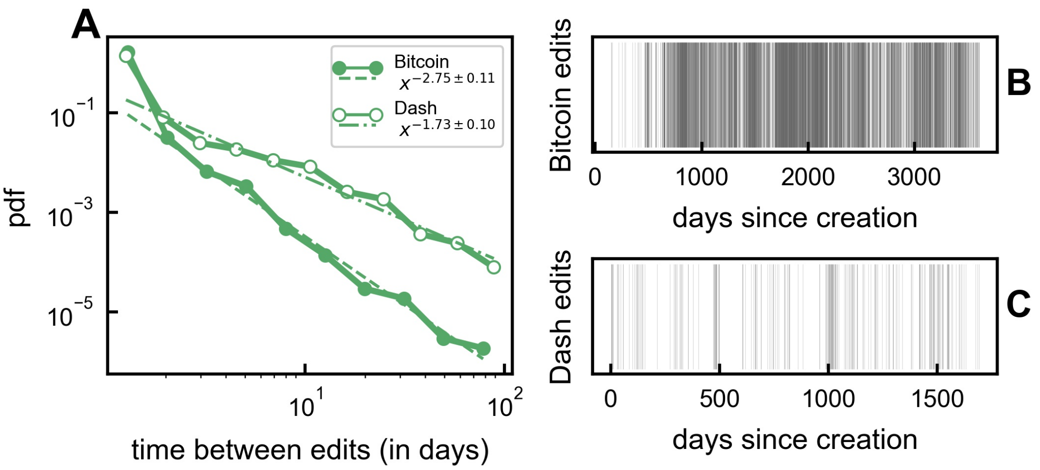


FIGURE 5.4: **Example of edit histories.** **(A)** Distribution of the inter-event time between two consecutive edits for Bitcoin (line with filled circles) and Dash (line with white circles). The dashed line is a power-law ($P(x) \sim x^{-\beta}$) with exponents $\beta =$ 2.75 and $\beta = 1.73$ for Bitcoin and Dash, shown as a guide to the eye. Edits are shown as vertical black line as a function of time for Bitcoin **(B)** and Dash **(C)**

by analyzing the evolution of cryptocurrency pages and the role played by Wikipedia editors.

Frequency of edits and editors diversity are considered reliable indicators of the quality of information included in a Wikipedia page [207]. Cryptocurrency pages differ with respect to their edit history (see Figure 5.4). Some pages, including those of Bitcoin and Ethereum, experience continuous edits throughout their history, while for other pages, including Dash and Cardano, contributions are intermittent in time, with periods of higher activity followed by calmer ones. For example, the change of the Dash logo in April 2018 triggered a spike in the number of edits.

The nature of edits changes over a Wikipedia page life. While at the beginning, editors focus largely on new content, as the page ages more efforts are dedicated to fighting vandalism and misinformation (maintenance work) [208, 153]. We quantify maintenance work by looking at "reverts", edits that restore a previous version of the page, and at the number of edits reporting vandalism (see Chapter 3, Section 3.2.1 for more details on

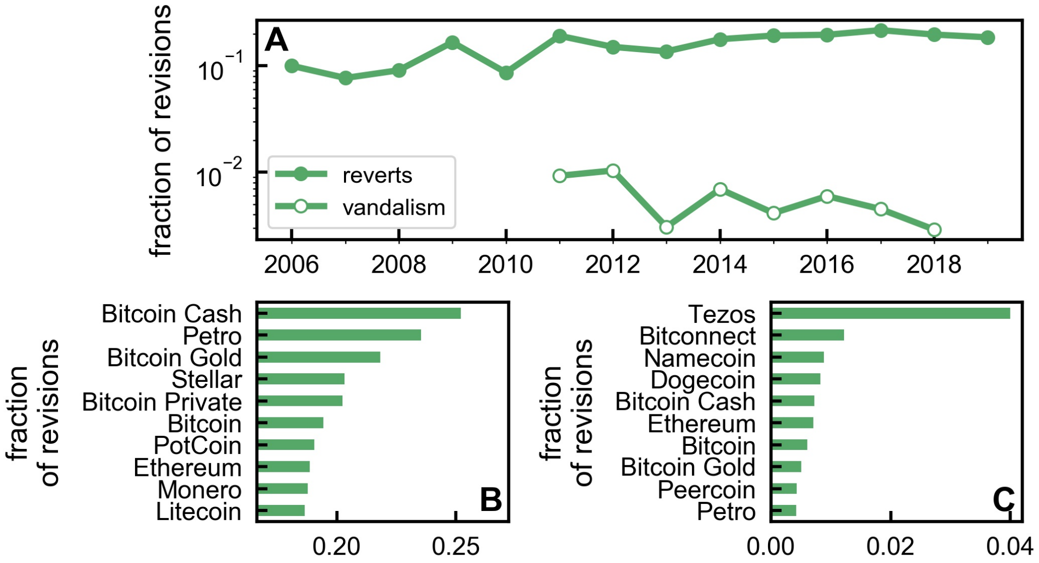


FIGURE 5.5: **Reverts and vandalism revisions. (A)** The fraction of "revert" edits (line with filled circles) and edits reported as vandalism (line with white circles) over time. Values are aggregated using a time-window of one year. B-C) The fraction of reverts **(B)** and vandalism **(C)** edits for the top 10 cryptocurrencies sorted by number of reverts and vandalism edits, respectively.

reverts detection). We find that reverts constitute the 18.2% of all edits, and that, on average, they constitute the 15.3% ± 4.5 of contributions to a cryptocurrency page. The fraction of reverts is stable in time (see Figure 5.5A). Cryptocurrency pages experience higher rates of reverts than an average page in Wikipedia (8% of the edits at the end of 2016 [209]), suggesting there is more debate around their content. Only 0.5% of edits were reported as acts of vandalism and their occurrence is constant in time since mid 2011 (see Figure 5.5A). Well established cryptocurrency pages are less subject to maintenance edits than other pages (see Figure 5.5B and C). Pages of cryptocurrencies forked from Bitcoin such as Bitcoin Cash, Bitcoin Private and Bitcoin Gold were the source of many debates [210] resulting in a high number of maintenance edits (see Figure 5.5B).

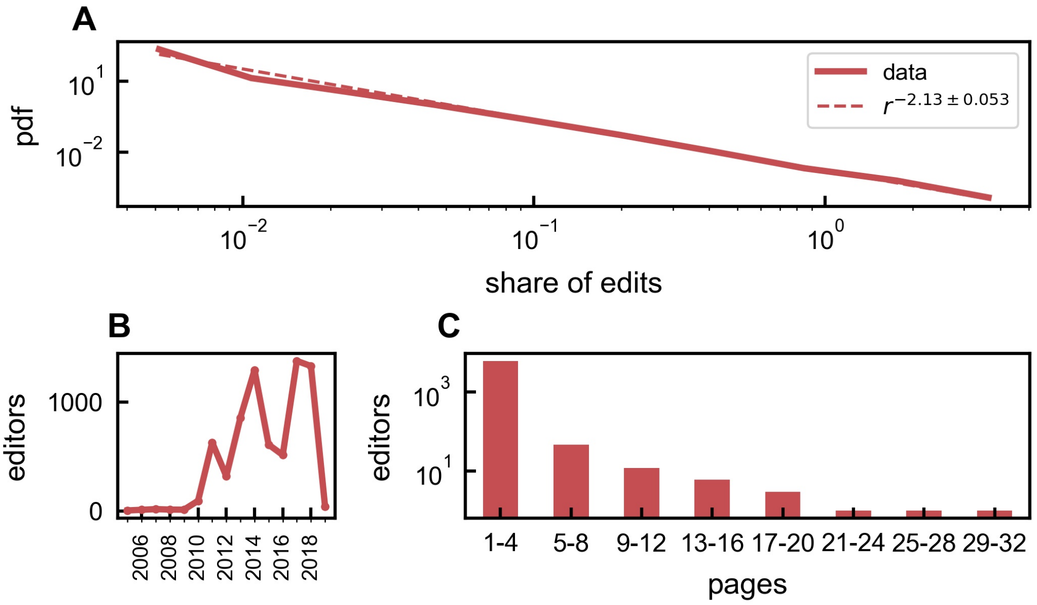


FIGURE 5.6: **Uneven distribution of contributions of Wikipedia editors.** **(A)** Distribution of share of edits between 2005 and 2018 (red solid line). The dashed line is a power-law fit $(P(r) \sim r^{-\beta})$ with exponent $\beta = 2.135 \pm 0.053$, shown as a guide to the eye. **(B)** The number of editors contributing to cryptocurrency pages. Values are aggregated using one year time window. **(C)** Histogram of editors based on the number of Wikipedia pages they have contributed.

5.4 Role of editors

Our dataset includes ~ 6170 editors who contributed $\sim 29,000$ total edits. Although the number of new editors/year fluctuates (see Figure 5.6B, and Appendix C.5), the number of editors has overall increased from 2006. Only in 2017, when 10 new cryptocurrency pages were created, $\sim$ 1200 new editors joined. Interestingly, this growth does not characterize all pages on Wikipedia. For example, in [211], the authors show that the number of editors in medical related article has been decreasing (see Appendix C.3 for a comparison between cryptocurrencies Wikipedia page and pages reported in previous research).

The editing activity is heterogeneously distributed, as we find by ranking the editors according to the number of edits (see Figure 5.6A). This result

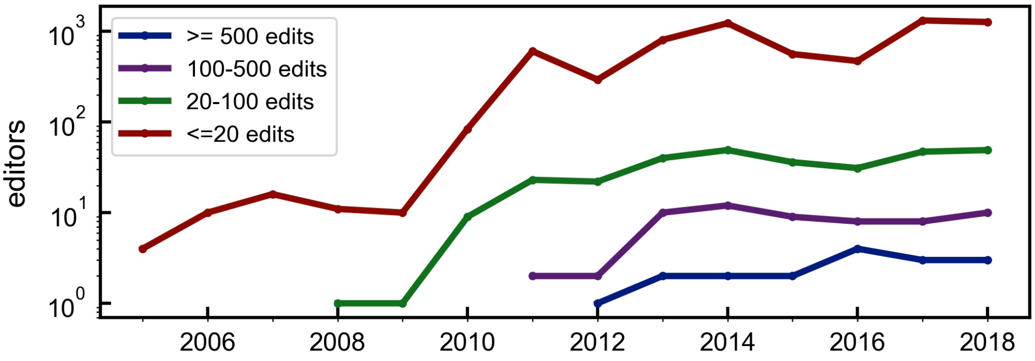


FIGURE 5.7: **Active editors per group.** The number of active editors per group from 2005 until 2018. Results are computed using a temporal window of one year. Editors are divided into four groups based on their total number of edits: More than 500 edits (blue line), 100 to 500 edits (purple line), 20 to 100 edits (green line), less than 20 edits (red line). Editors were classified according to their total contributions at January 23rd 2019, then traced back.

is in line with what generally observed in Wikipedia [212], and consistent across time (see Appendix C.4). In particular, the most active editor alone is responsible for $\sim$ 10% of the edits (see Appendix C.5 for more details on the most active editor) and only $\sim$ 9.6% of the editors (596) have edited at least 2 pages (Figure 5.6C). This group is responsible for 50% of the total number of edits for all cryptocurrency Wikipedia pages.

Then, we study the evolution of editors' activity in time. We classify editors into four groups based on their total number of edits at the end of the study, in January 2019 (see Figure 5.7): Contributors who made more than or equal to 500 edits (6 editors, responsible for 23% of edits), contributors who made 100 to 500 edits (23 editors, responsible for 15% of edits), contributors who made 20 to 100 edits (142 editors, responsible for 19% of the edits), editors who made less than 20 edits (97% of editors, responsible for 43% of the edits). We find that the higher the cumulative activity of a group, the most recently they started editing the pages (see Figure 5.7), in contrast to what is generally observed on Wikipedia [213, 214]. Note that the group of most active contributors started editing in August 2012, 3 years after the creation of Bitcoin's page. Furthermore, Figure 5.7 shows that editors with the largest

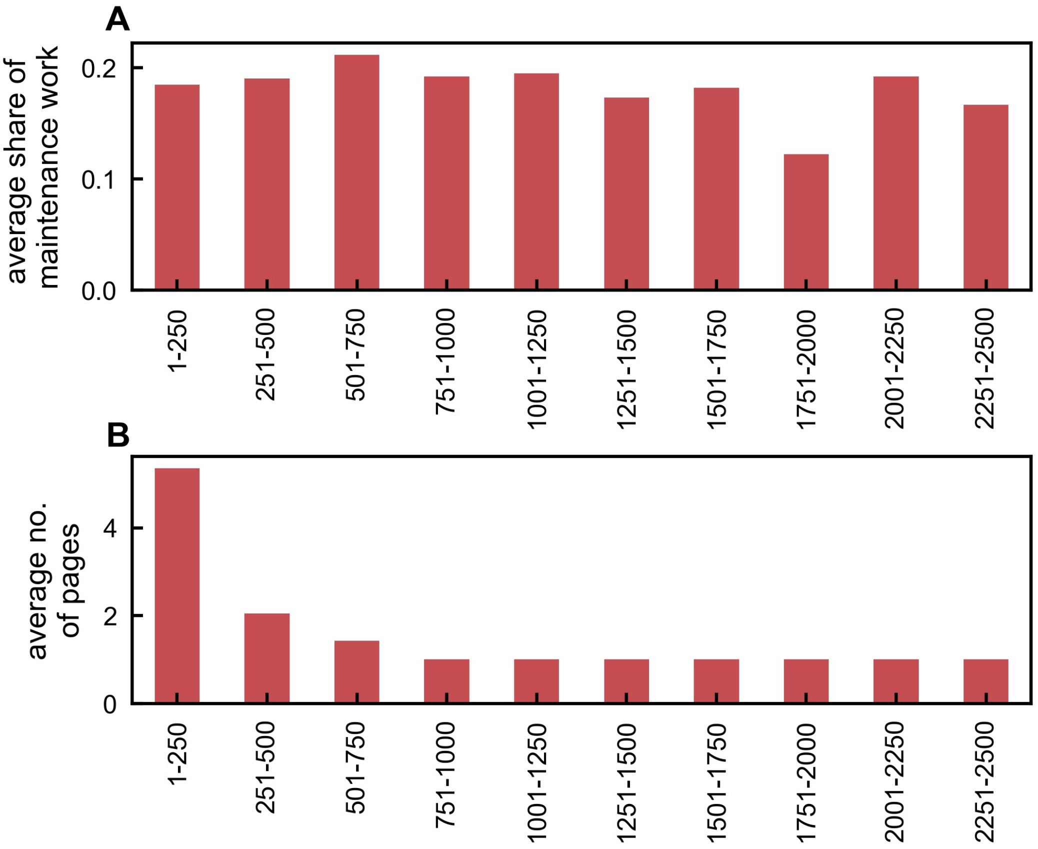


FIGURE 5.8: **The focus of editors.** Editors are ranked based on the total number of edits in descending order and grouped based on their rank. **(A)** The fraction of maintenance edits for each rank group. **(B)** The average number of contributed pages for each rank group. Only editors with more than one edit are considered.

number of edits are responsible for the most extensive contributions in terms of number of edited words. Some of their edits, however, may be for maintenance. By ranking editors in descending order according to their total number of edits across the entire period of study, we find that, for the top 10 contributors, maintenance edits amount to 20% of their edits. On average, $\sim$ 18% of the edits written by top 250 editors are maintenance work (see Figure 5.8A). This value is consistent among different rank groups. Finally, top ranked editors tend to contribute in more than one page (see Figure 5.8B), on average $\sim$ 4 pages.

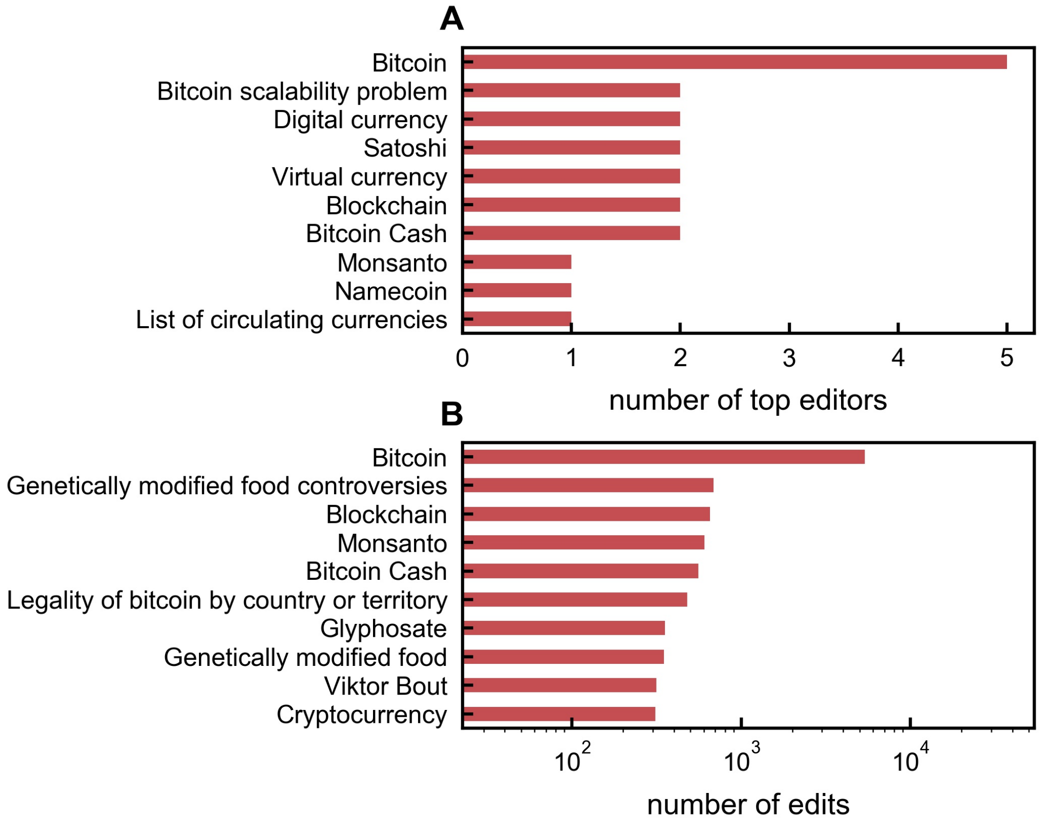
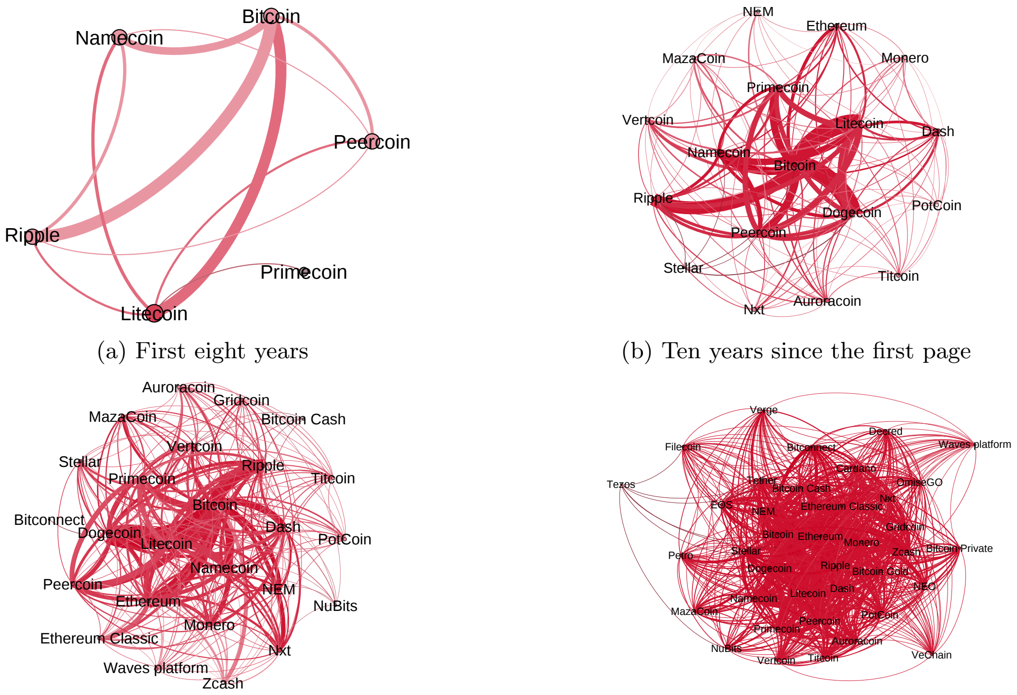


FIGURE 5.9: **The activity of the top** 6 **cryptocurrency pages editors. (A)** The top 10 pages by the number of editors. The x-axis shows the number of top editors who had this page in their top edited pages. Note that here we consider only the top 10 pages per editor. **(B)** The top 10 pages by the number of edits. The x-axis shows the total number of edits per page. Results are obtained for the subset of 6 most active editors.

To understand the general interests and the specialization of the top editors of cryptocurrency Wikipedia pages, we focus on a subset of 6 editors that have contributed at least 500 edits each. We studied in details their interests by considering their contribution over the entire Wikipedia. Our results show that the main interests of these editors are cryptocurrencies and blockchain (see Figure 5.9). Results are consistent when we extend the analysis to the top 29 editors, who are responsible for 37% of the edits (see Appendix C.4, Figure C.3). Top editors also contribute in other non-cryptocurrency related pages, however, these pages are less homogeneous and include several different interests such as; genetically modified food, musicians and motor company.



(a) First eight years

(b) Ten years since the first page

(c) Twelve years since the first page

(d) The entire period of study

FIGURE 5.10: **Evolution of the network of cryptocurrency pages.** Nodes represent Wikipedia pages and edge exist between two nodes if they have at least one common editor. The radius of a node is proportional to the sum of weights of incoming links and the edge thickness is proportional to the edge weight, measured as the number of common editors. The network is aggregated over a different period of times: **(A)** from July 2005 until July 2013, **(B)** from July 2005 until July 2015, **(C)** from 2005 until July 2017, **(D)** for the entire period of study.

We further study the network of co-edited Wikipedia pages. We construct an undirected weighted graph, where nodes are Wikipedia pages, an edge exists between two nodes if they have at least one common editor, and link weights correspond to the number of common editors. By the end of July 2014, the network had 13 nodes (see Figure 5.10B) and the average node weighted degree was $\langle s \rangle = 78.3$ with a total of 2691 editors. The weighted degree was heterogeneously distributed: Bitcoin had the largest strength, $s_{BTC} = 207$, while recently introduced nodes (Dash, Auroracoin and Nxt) had the lowest weighted degree. These properties have persisted in time (see Figure 5.10C and 5.10D) and a cryptocurrency page age is positively correlated with its network weighted degree (Pearson correlation $\rho = 0.40$, $p = 0.015$, see Appendix C.6). Bitcoin has the highest degree centrality throughout the entire period considered (see Appendix C.6).

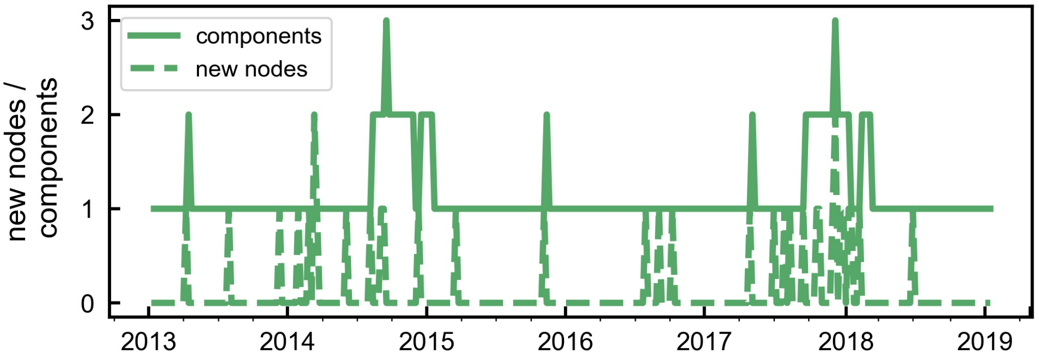


FIGURE 5.11: **Short-term dynamics of the Wikipedia network evolution.** The Cumulative number of new nodes (dashed line) and the total number of network components (solid line). Values are aggregated using a 1 week time window.

A giant component (see Figure 5.10) emerges in the network, implying each node is connected to all other nodes when we analyse its evolution under large time-windows ($\sim$ years). Instead, if weekly time windows are considered, we find that the network is disconnected (see Figure 5.11). Typically, new pages are created by new editors. On average, new pages connect to the giant component within 5.2 weeks from creation (see Figure 5.11), in most cases thanks to experienced editors who contribute to the newly created page.

5.5 An investment strategy based on Wikipedia attention

The demonstrated connection between how successful a cryptocurrency is and the attention it draws on Wikipedia suggests the latter could help in informing a successful investment strategy. We investigate this possibility by testing a Wikipedia-based strategy similar to the one proposed in [106, 107] for stock markets investments.

For a given page and a given day t, the Wikipedia investment strategy relies on the difference $\Delta n(t) = v(t) - v(t-1)$ between the number of page views $v(t)$ at day t and the number of views $v(t-1)$ at $t-1$. According to the strategy, if $\Delta n(t) > 0$, the investor sells the asset (at price $p(t+1)$) at time $t+1$ and then she buys at time $t+2$ (at price $p(t+2)$). This trading position is formally known as short position. On the other hand, if $\Delta n(t) \leq 0$ the investor buys at time $t+1$ (at price $p(t+1)$) and sells at time $t+2$ (at price $p(t+2)$), which is known as long position. We consider the price and the total number of views calculated over the entire day. The intuition behind the strategy is that if attention and information gathering has been rising, prices will drop, and vice-versa [106, 215]. We consider Wikipedia views rather than edits, since the latter do not vary on a daily basis (the average time between edits is 10.12 days). Considering a longer period would overlook the cryptocurrencies' price volatility [216]. Here, we make the assumption that investors influence is negligible, e.g. they will be "price-takers" [217].

We also consider three baseline strategies. The first is based on the price difference $\Delta p(t) = p(t) - p(t-1)$ rather than the page views difference $\Delta n(t)$ [218]. In all other aspects, it is identical to the Wikipedia-based strategy. This will allow us to test which indicator (price or Wikipedia page views) has better predictive capabilities under the same conditions. The rationale behind the first baseline strategy is that if the price has been rising, a drop will follow, and vice-versa. As a second baseline, we choose a random strategy, where, at every time t, one chooses either to buy or to sell an asset with 50% probability [106]. Finally, we test a "buy and hold" strategy (see also [107]), implemented by buying all currencies in the beginning of a period (or when they are born) and selling them at the end of the period under study.

The performance of the different strategies is assessed by computing the cumulative return R, defined as the summation of log-returns obtained under the proposed strategies. When $\Delta n(t) > 0$ the log-return is computed as $log(p(t+1)) - log(p(t+2))$, while, in the opposite case, the log-return is $log(p(t+2)) - log(p(t+1))$. The use of the log returns is motivated by the ease of calculation of the short and long positions and since we are considering multi-period returns [219].

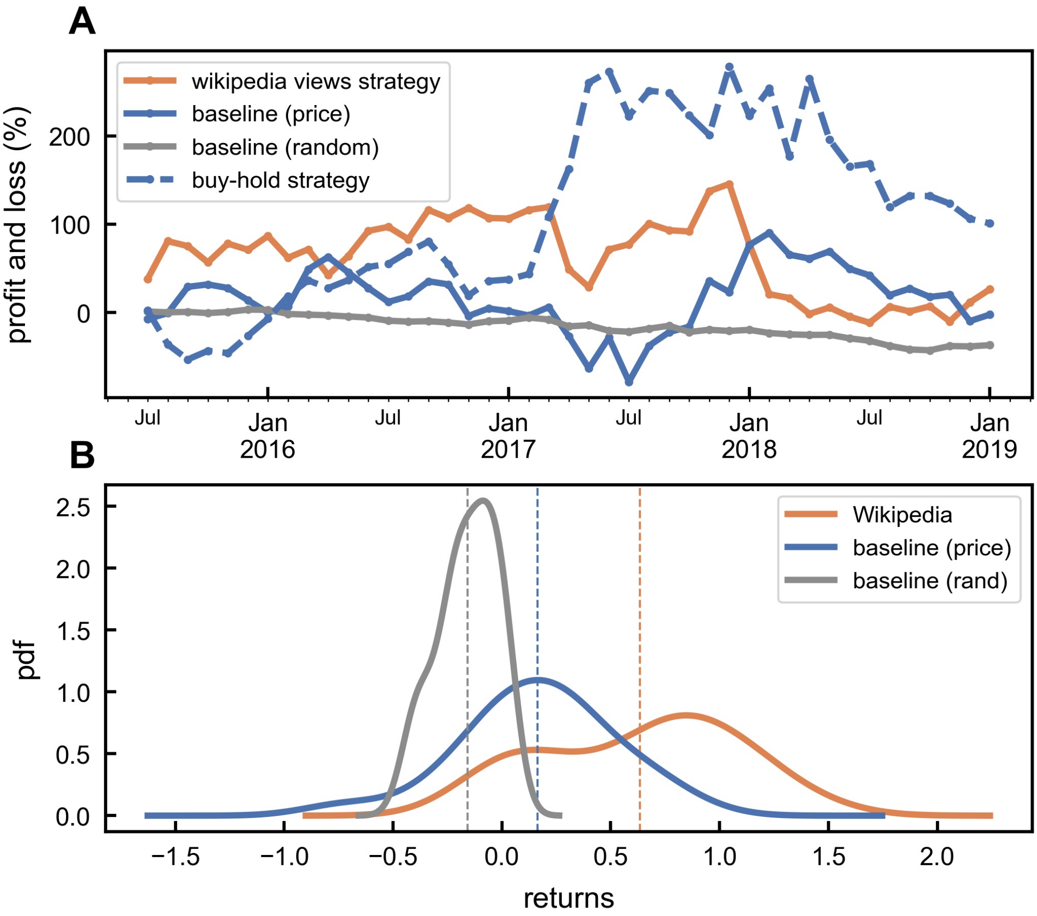


FIGURE 5.12: **The Wikipedia based investment strategy outperforms the baseline. (A)** The cumulative return obtained using four investment strategies: the Wikipedia-based strategy (orange line) the baseline strategy based on prices (blue solid line), the "buy and hold" strategy (blue dashed line) and the random strategy (grey line). **(B)** The distributions of the daily returns obtained using the Wikipedia-based strategy (orange line), the baseline strategy based on prices (blue line) and the random strategy (grey line). The average returns are $\langle r_w \rangle = 0.62 \pm 0.42$ (dashed orange line), $\langle r_p \rangle = 0.16 \pm 0.36$ (dashed blue line), $\langle r_r \rangle = -0.15 \pm 0.13$ (dashed grey line) for the Wikipedia-based strategy, the price based baseline, and the random strategy, respectively. Data is displayed using a kernel density estimate, with a Gaussian kernel and bandwidth calculated using Silverman's rule of thumb. Data for the random strategy is obtained from 1000 independent realizations. All results are shown for investments between July 2015 and January 2019 for all cryptocurrencies which can be margin traded combined.

We test the Wikipedia-based strategy against the baselines for the 17 cryptocurrencies that have a Wikipedia page and can be margin traded (see list of exchanges with margin trading support in Appendix C.1 and list of cryptocurrencies in Table C.1, Chapter 3). Margin trading is a practice of borrowing fund from a broker to trade financial assets, that rely on selling assets one does not yet own. We test the strategies considering a period from July 1st, 2015 until January 23rd, 2019.

We find that the Wikipedia based strategy outperforms the price based and the random baseline strategies, when one considers the period between July 2015 and January 2018 (see Figure 5.12A). However, it outperforms the "buy and hold" strategy only up to January 2017, when the explosive growth of the market made holding extremely profitable. On average, the return obtained following the Wikipedia based strategy is $\langle r_w \rangle = 0.62 \pm 0.42$, while the average return obtained under the random strategy is $\langle r_r \rangle = -0.15 \pm 0.13$ (see Figure 5.12B). The distributions of returns obtained under the two strategies are significantly different under Kolomogorov-Smirnov test, with $p \ll 0.05$. The price baseline strategy produces lower mean returns compared to the Wikipedia strategy ($\langle r_p \rangle = 0.16 \pm 0.36$). To evaluate the risk factor in the three strategies we calculate the Sharpe ratio. The Sharpe ratio is defined as

$$S = \frac{\bar{R}}{S_R},$$

where $\bar{R}$ represents the average annual return and S_R the standard deviation of the annual returns. We find that the Wikipedia based strategy yields a Sharpe ratio $S_w = 0.066$, higher than the ones obtained under the baseline strategies: $S_p = -0.022$ and $S_r = -0.799$ for the price and random strategy respectively. However, the Sharpe ratio of the Wikipedia strategy is not consistently outperforming the baseline strategies along the entire period of study (see Appendix C.8, Figure. C.8).

A closer inspection shows that there are consistent differences between cryptocurrencies, with respect to the cumulative returns (see Figure 5.13), with some even yielding overall negative returns. The Wikipedia-based strategy yields a positive cumulative returns of $\sim 300\%$ for Ethereum Classic, but for other currencies, including Ripple and Ethereum, investing based on

Wikipedia leads to negative returns.

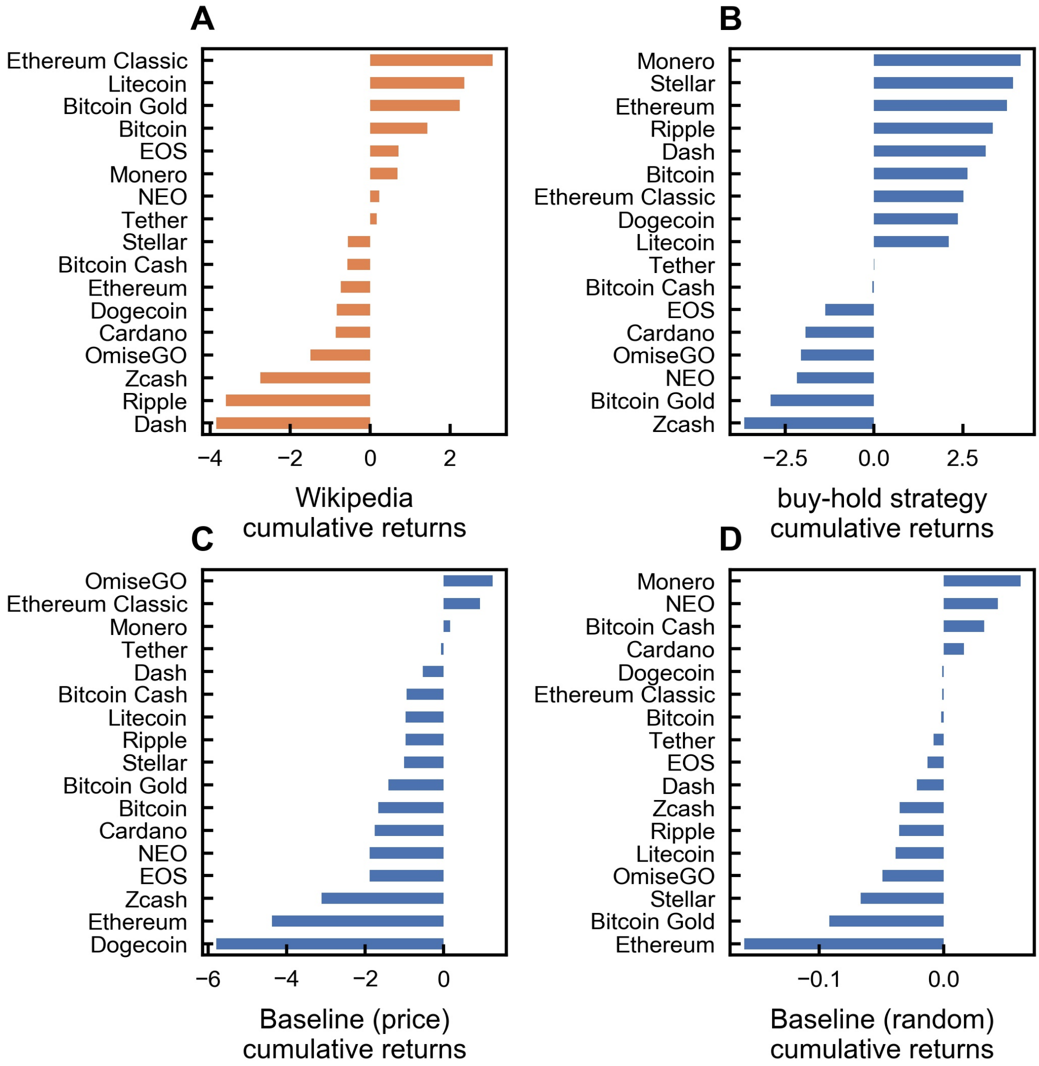


FIGURE 5.13: **Performance of the strategies for different cryptocurrencies.** The cumulative returns along the whole period of investment, following the Wikipedia based strategy **(A)** the buy hold strategy **(B)**, the price-based baseline strategy **(C)** and the random strategy **(D)** for the 17 cryptocurrencies considered.

The observed differences could be potentially explained by the correlation or causality between changes in daily price and in Wikipedia views (see more details on the correlation and Granger causality for each cryptocurrency in Appendix C.8). Instead, we observe that, neither the correlation nor the Granger causality explains the results observed, suggesting other mechanisms could be in play [220]. For example, our proposed strategy does not simply map to buying a cryptocurrency when its Wikipedia page views increases. In order to gain positive returns using our proposed strategy, an increase in the number of views at time t, should be followed by an increase in price in the next day $t + 1$ and a decrease of the price in the day after $t + 2$. Positive returns will also occur in case of a decrease in the number of views at time t if it was followed by a decrease in the price at time $t + 1$ and an increase in price at time $t + 2$.

Finally, we investigate the role of the start and end times of the investment period (see Figure 5.14). We find that, for most of the choices, the Wikipedia-based strategy has a higher cumulative returns than the random and the price baseline strategy. It outperforms both baseline strategies for the majority of the periods ending before January 2018, when the market entered a period of dramatic losses. Instead, the "Buy and hold" strategy yields higher returns for start dates before March 2017, especially for long hold periods. The Wikipedia strategy outperforms the "Buy and hold" strategy when trading starts after November 2017.

5.6 Conclusion and discussion

In this chapter, we have investigated the interplay between the production and consumption of information about digital currencies in Wikipedia and their market performance. We have shown that there is a positive correlation between a cryptocurrency's overall success in the market, as measured by its price, volume, and market share and the overall attention gained by its Wikipedia page, measured by the number of page views and the number of page edits. This result suggests that the production and consumption of information in Wikipedia is relevant for investment purposes.

We have analysed the edit history of cryptocurrency pages in Wikipedia. We have shown that contributions to cryptocurrency pages are bursty in time, with periods of high activity followed by calmer ones. We have found that cryptocurrency pages have experienced a higher number of revert edits (18%) compared to other pages, suggesting they have been subject to vivid

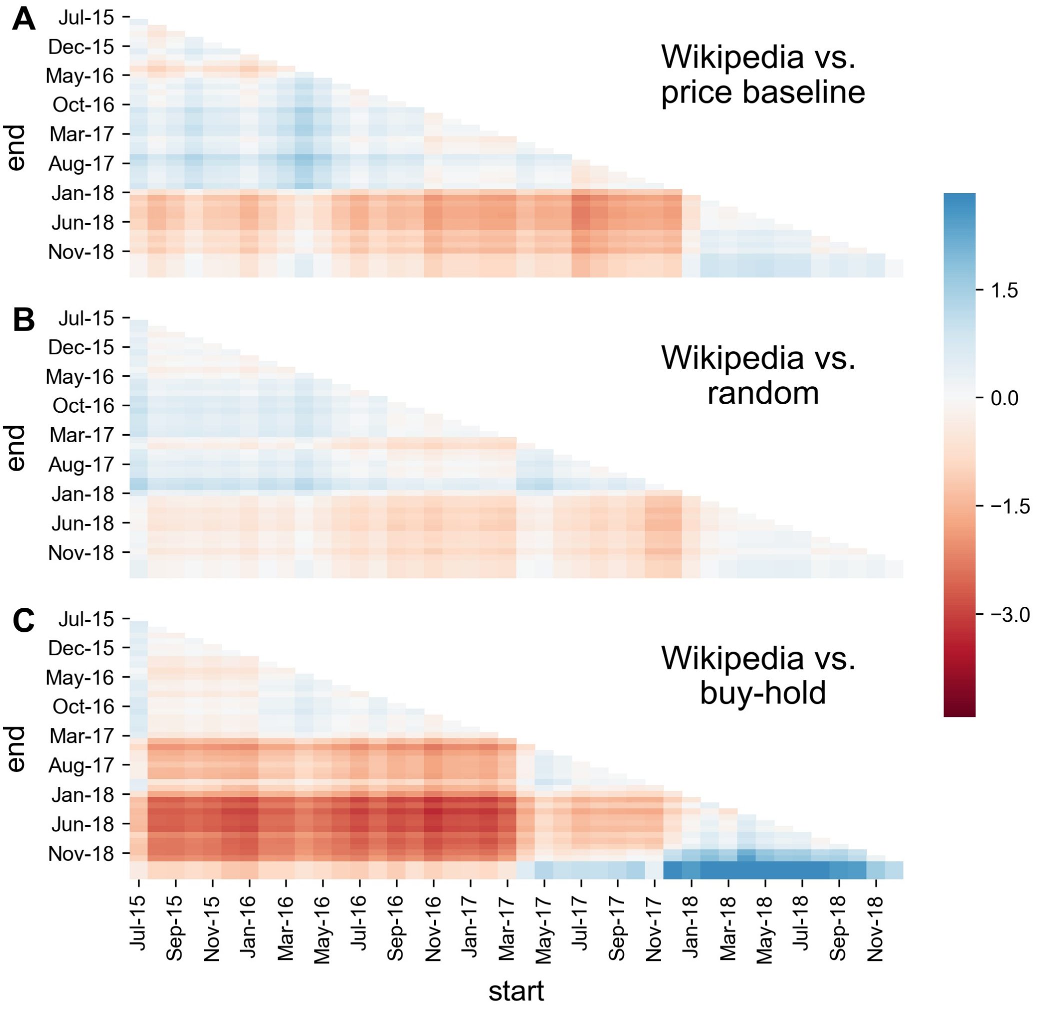


FIGURE 5.14: **Comparison between strategies across different periods of time.** Difference between the cumulative log returns of the Wikipedia based strategy and the price based baseline **(A)** or the random baseline **(B)** or "buy and hold" strategy (**C**) given a different start and end dates.

debates around their contents. Also, we have found that the number of cryptocurrency pages editors has increased in the period considered, and this is not the case for editors of other topics in Wikipedia. However, very few editors are responsible for most of the edits, consistently with the rest of Wikipedia. Interestingly, this subset of editors has started contributing relatively recently (after 2012), also in contrast with the rest of Wikipedia. We have shown that the information in Wikipedia is, to a large extent, provided by cryptocurrency and technology enthusiasts. In fact, we have found that editors who are very active on cryptocurrency pages focus their editing activity almost exclusively on cryptocurrencies and blockchain. We have found that the community of cryptocurrency editors is tight: On average, each page is connected to 37 other pages through an average of 7 editors and active contributors tend to edit many pages. New cryptocurrency pages are typically created by new editors, but then also edited by more experienced ones. For this reason, we find that older pages have a higher degree in the co-editing network. Further investigation of the nature of edits which arises as a response to price changes could uncover another interesting dimension of the relationship between Wikipedia editors and the market. In our analysis of page edits we considered all edits that had ever been made in a page. Considering edits with specific number of words or significance might impact the the distribution. However, choosing a threshold can be a non trivial task, especially given the difference between pages maturity.

Finally, we have proposed a trading strategy relying on Wikipedia page views, similar to the Wikipedia based strategy proposed for the stock market [106] and found it yields significant returns compared to baseline strategies. However, the strategy is less profitable that the simple ”buy and hold” approach after the explosive growth of the market that started in January 2017 and becomes generally unsuccessful after January 2018, when the cryptocurrency market started suffering major losses. To further enrich the picture, we have discussed the relative performance between different strategies also by considering the effect of the hypothetical starting and ending period of trading, showing that the Wikipedia strategy is a valid option to be considered. In order to delimit the scope of our findings, it is important to note that, although our strategy yields overall positive returns, when considering currencies individually, returns are positive only for 8/17 of them. It is important to note that our trading strategy is mostly meant to demonstrate Wikipedia pages relevance to investment decisions and market performance. Hence there are certain caveats worth mentioning. First, the price reported in our dataset is not tradable price since it’s an average price across different exchanges. Second, Bitcoin margin trading started only in 2017 and more exchanges supported margin trading for cryptocurrencies later on. However in our analysis we considered the period from July 2015 which could explain the positive returns in this period. Furthermore, that our strategy neglects the role played by fees, which could significantly decrease profits in real scenarios. Finally, for the sake of simplicity and it is customary for a study like ours, we have assumed that investors influence is too small to perturb the market; relaxing this assumption could be an interesting aspect to include in future works.

Characterizing the production and consumption of information around cryptocurrencies is key to understand the market dynamics and inform investment decisions [221]. Although our study was limited to the analysis of Wikipedia data, other sources of information including traditional news outlets , Twitter, Reddit or bitcointalk could reveal important information about the cryptocurrency market dynamics.

6 Coordinating in the dark: the rise and fall of Bitcoin's marketplaces

Dark markets are commercial websites, accessible only via the darknet and specialised at trading illicit goods. Dark markets are often referred to as cryptomarkets since cryptocurrencies are the universally acceptable payment method. In Section 2.1.2 we discussed the markets' sales, technology and the research investigated them. The markets drew the research attention as well as law enforcement entities due to their growing trading volume and user base [23, 2].

Here, we present the first complete analysis of dark markets evolution and adaptation dynamics to closures from June 2011 to July 2019 relying on novel transaction data. In Section 6.2, we show that dark markets across the time are resilient to closures and capable of a quick recovery. We then investigate the adaptation dynamics of the market to closures; specifically, users migration to other coexisting active markets after closure (in Section 6.3). We focus and 31 dark markets and show that for each closure a flux of users migrates to coexisting markets. Most importantly, coexisting markets gain, on average 7.3 times their total USD volume at the time of closing within 12 days of market closure. We then provide an analysis of the migrating user's behaviour (in Section 6.4) showing that overall migrating users are more active. Finally, in Section 6.5, we study how migrant users coordinate to migrate to a coexisting dark market following shutdowns. The work presented in this chapter is based on publication [III].

6.1 Materials and methods

Our analysis relies on a novel dataset of dark markets transactions on the Bitcoin's blockchain. The ledger of Bitcoin transactions (the blockchain) is publicly available and can be retrieved through Bitcoin core [222] or a third-party API such as Blockchain.com [223]. It consists of the entire list of transaction records, including time, amount transfered in USD, origin and destination addresses.

Addresses are identifiers of $26-35$ alphanumeric characters that can be generated at no cost by any user of Bitcoin, such that a single Bitcoin wallet can be associated to multiple addresses. In fact, to ensure privacy and security, most Bitcoin software and websites help users generate a new address for each transaction. Thus, blockchain data has to be pre-processed to map groups of addresses to individual users. In Section 3.3.1 we discuss some of these clustering techniques. We used data pre-processed by Chainalysis [146] following the same approach discussed.

We considered the entire transaction data of 31 dark markets (see Section 3.3.2) between June 18th, 2011 and July 24th, 2019. We also considered data for users who interacted with one of these markets, where, in this case, the data includes all transactions starting from their first interaction with a dark market. Thus, each market ecosystem can be represented as an egocentric network [164] of radius 2, where the market is the central node, its nearest neighbours represent market users, and direct edges represent transaction occurring either between the market and one of its neighbours, or between two neighbours. See Section 3.3.2 for more details on the data sampling and dark markets considered in our analysis.

6.2 Markets resilience

The capacity of dark markets to recover following closures can be studied quantifying the evolution of the total volume traded by dark markets in time.

Despite recurrent closures, we find that the number of markets has been relatively stable from 2014 (see Figure 6.1A). In addition, the total weekly volume sent/received by dark market addresses has grown from 2014 until the end of 2019 (see Figure 6.1B), suggesting darknet markets were resilient to closures. Starting from the end of 2018, however, we observe a decrease in the total volume traded. Note that, here, we considered the total volume (in American dollars) sent/received across the entire dark market egocentric network as a proxy for the entire volume.

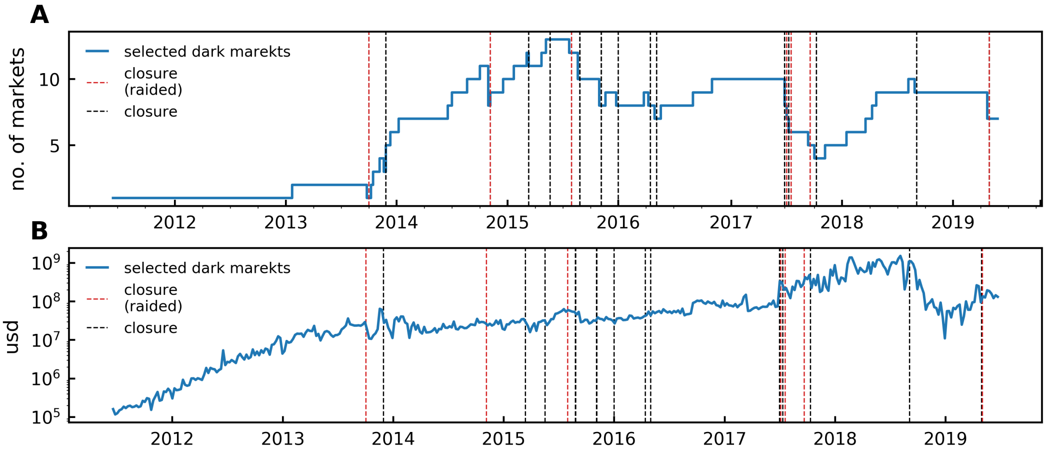


FIGURE 6.1: **Dark markets resilience.** (**A**) The total number of dark markets active across time (blue line). (**B**) The total USD in circulation among all 31 dark markets and their nearest neighbours (blue line). Values are calculated using a time window of one week. Red dashed lines represent market closure due to law enforcement raid, and the black dashed line represents market closure due to any other reason.

6.3 Users migration

The observation that dark markets are resilient to closure suggests that their users may move to other markets. We refer to this phenomenon as *migration*.

It has been suggested that users could migrate to other markets following market closure [64, 68]. In fact, migration was observed [224] after the closure of the AlphaBay market, when other markets, Hansa Market and Dream market, experienced an abnormal spike in activity. In this section, we provide the first systematic investigation of dark market users migration, by studying the effects of multiple closures.

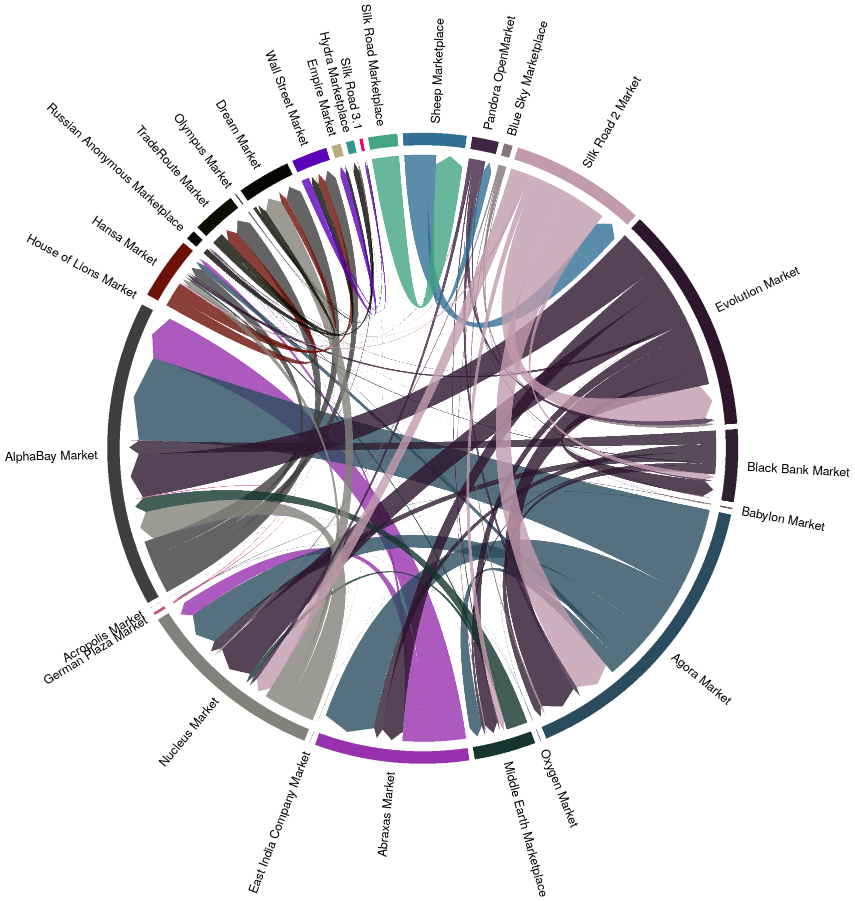


FIGURE 6.2: **Migration between coexisting markets.** The total number of nodes migrating from a dark market after closure to another coexisting market. The arrowhead points to the direction of migration from market x to market y and the width of the arrow represent the number of nodes. Markets are ordered clockwise according to the closing date in ascending order starting from Silk Road Marketplace.

We identify migrant users in the following way. For each market that was shut down, we identify users who started trading with another coexisting market after the closure. Note that users who were trading simultaneously on multiple markets before closure are not considered migrants.

In Figure 6.2, we show the number of migrant users, there is a consistent behaviour of migration after each market closure.

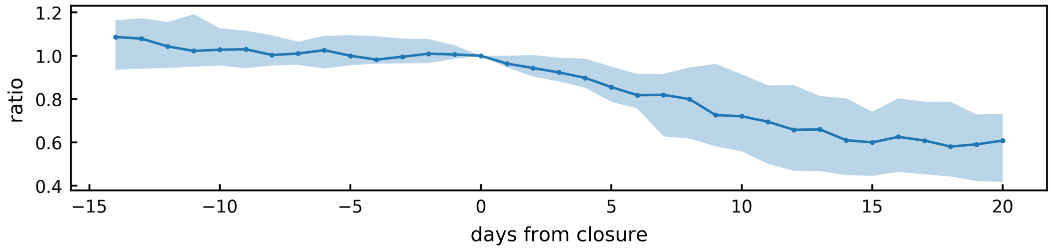


FIGURE 6.3: **Closure influence on Activity.** The median ratio of number of returning users to number of returning users at time of closures across all dark markets. The shaded area is the 50% quantiles. Values calculated 14 days before closure and 20 days after.

An important question is which fraction of users involved in illicit trading continue to exchange with dark markets following a closure.

The prediction of whether a user will return to activity on Bitcoin is still an open question [47]. The work in [47] found that most of the minted Bitcoins were accumulated in addresses which never sent, see Section 2.1.1 for more details on the dormant addresses problem.

We approach this problem by computing the fraction of "returning users" over time, meaning the fraction of all users active in a given week that are active also in the following week.

After compute the fraction of returning users over time, we normalize it by the fraction of returning users at the time of closure. Then, we consider the median across market closures. We find that the ratio of returning users drops down to 85%, 5 days after the closure. After 20 days, the fraction of returning users drops to 60% (see Figure 6.3).

6.4 Who is migrating?

The observations that some users stop trading following a dark market closure, but the total volume traded in dark markets does not decrease could indicate that migrant users are on average more active than others.

We test this hypothesis by computing the activity of migrant users before and after closure. We refer to the first dark market a user was interacting with as its *home market*. For all users (migrant and non-migrant), we measure the total volume exchanged with any other user. We find that the median volume exchanged by migrant users is $\sim$ 10 times larger than the volume exchanged by non-migrant users (see Figure 6.4A), with the median volume exchanged summing to 3882.9USD for migrant users and to 387.2USD for non-migrant users. Similar conclusions can be drawn by considering the volume exchanged with the home market only, which has median value of 263USD and for non-migrant users and 74.3USD for migrant users (see Figure 6.4B).

The activity distribution of migrants is significantly different from the non-migrant users' distribution (using Kolmogorov Smirnov test, $p < 0.01$, see Table 6.1).

TABLE 6.1: **Kolmogorov Smirnov test results** d statistic and p value for the Kolmogorov Smirnov test between the distributions of migrant and non migrant users transactions with dark markets and all their transactions. The test was adjusted for multiple comparisons using Bonferroni correction method [225].

Dark market	d-statistic for all transactions	d-statistic for dark market transactions	p-value for dark market transactions	p-value for all transactions
Abraxas Market	0.32288	0.1907	$< 10^{-82}$	$< 10^{-240}$
Agora Market	0.29796	0.23973	0.0	0.0
AlphaBay Market	0.522167	0.326691	0.0	0.0
Babylon Market	0.628725	0.29961376	0.0075899	$< 10^{-6}$
Black Bank Market	0.368434766	0.189521445	$< 10^{-40}$	$< 10^{-157}$
Blue Sky Marketplace	0.56419	0.3263753986	$< 10^{-20}$	$< 10^{-65}$
Dream Market	0.5959877	0.31426	$< 10^{-16}$	$< 10^{-64}$
Evolution Market	0.3949466	0.30032	0.0	0.0
German Plaza Market	0.60932	0.381227638	$< 10^{-15}$	$< 10^{-42}$
Hansa Market	0.5943	0.369	$< 10^{-157}$	0.0
Middle Earth Marketplace	0.317158	0.1531	$< 10^{-18}$	$< 10^{-81}$
Nucleus Market	0.352989	0.28461	$< 10^{-172}$	$< 10^{-265}$
Olympus Market	0.543	0.20472	0.03	$< 10^{-20}$
Pandora OpenMarket	0.50238	0.2950	$< 10^{-62}$	$< 10^{-185}$
Russian Anonymous Marketplace	0.31653	0.4152695	$< 10^{-81}$	$< 10^{-46}$
Sheep Marketplace	0.4120	0.286	$< 10^{-110}$	$< 10^{-229}$
Silk Road 2 Market	0.408889	0.302276	0.0	0.0
Silk Road Marketplace	0.50883668	0.38742247	0.0	0.0
TradeRoute Market	0.524726	0.323715	$< 10^{-61}$	$< 10^{-164}$
Wall Street Market	0.462557	0.3100879	$< 10^{-53}$	$< 10^{-121}$

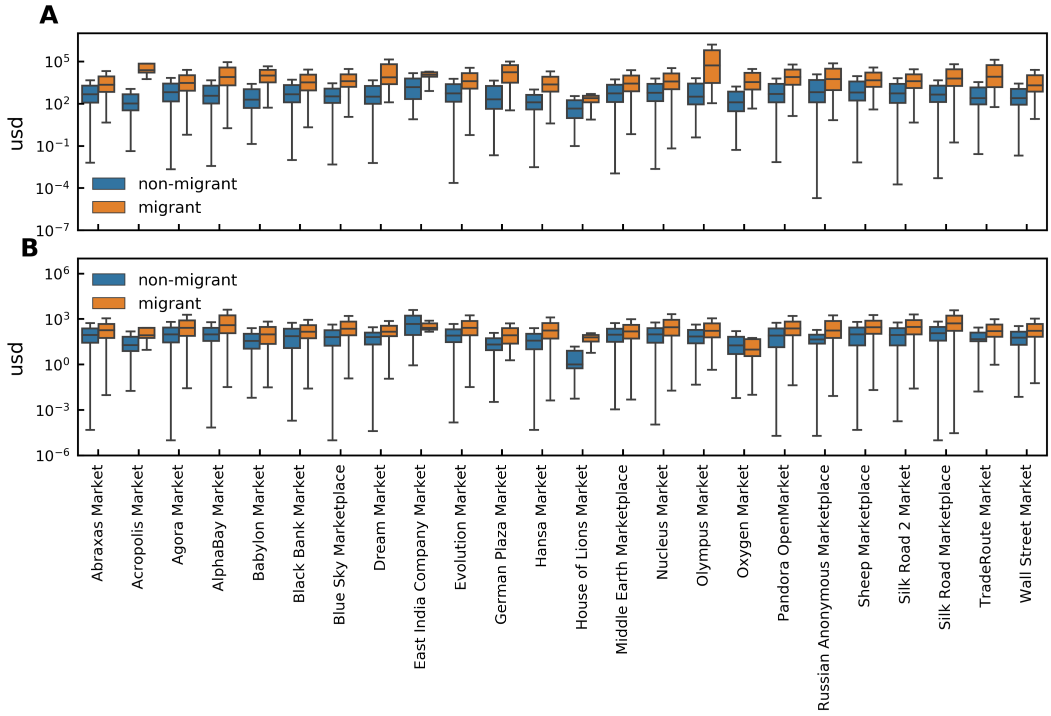


FIGURE 6.4: **Active migrants (A)** Box-and-whisker plot for migrant users (orange boxes) and non-migrant users (blue boxes) total US dollar sent and received across their time of interaction with the closed dark market. **(B)** Focusing on interactions with the closed dark market only, a Box-and-whisker plot for migrant users (orange boxes) and non-migrant users (blue boxes) total US dollar sent and received to/from the dark market closed and across their time of interaction with the closed dark market. The horizontal line in each box represents the median. The lower box boundary shows the first quartile, and the upper one shows the third quartiles. The whiskers show the minimum and maximum values within the 1.5 lower and upper interquartile rate.

6.5 Coordination in the dark

A natural question that follows our analysis is how users choose where a new market following a dark market closure. In all cases but one, at least two markets coexist following closure. In this section, we investigate how migrant users decide where to migrate whenever there are multiple options.

We find that, on average, one market absorbs 71% ± 20.6 of all migrant users (see Figure 6.5). Only 4% of the users migrate to multiple markets instead on one market.

In Figure 6.5, we show the evolution of the trading volume shares of the shut down market and the top two destination markets in the periods preceding and following a closure. We find that the top two destination markets experience an increase in share starting 2 days after the closure, and saturating after about 6 days.

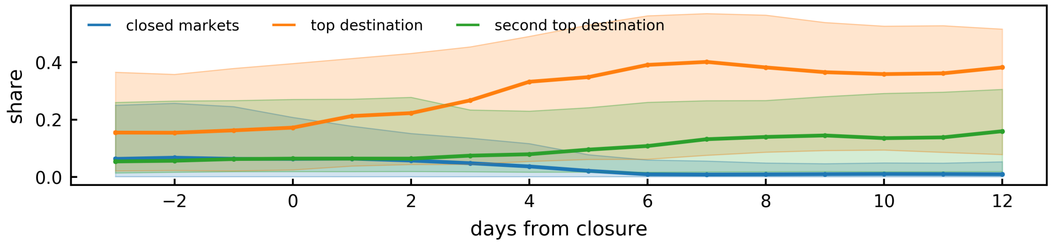


FIGURE 6.5: **Migration boosting coexisting markets value.** The median share of the total dark markets value of all closed markets (blue market), the market which was the top destination for the migrant users (orange line) and the second top destination for migrant users (green line). The shaded area represents a 50% quantile. The values are calculated 3 days after and 12 days after. For both figures, values are calculated using a rolling window of one week.

We investigate the characteristics of the first destination market for migrant users, by ranking coexisting markets according to (1) the total trading volume in USD at the time of closure and (2) total number of common users between the shut down and the coexisting market before closure. We find that, regardless of the reason behind closure, users chose to move to the market with the highest number of common users, which, in some cases, it is also the biggest market in terms of the USD value. Figure 6.6A shows that 33% of the times, users chose to move to the market with the largest number of common nodes (rank number one). Users are equally likely to move to the second and the third rank and, only 10% of the times, they move to the fourth rank. Users do not choose to migrate to markets with rank lower than fourth.

Figure 6.6B shows that almost 40% of the users migrate to the second-largest market. A closer look at the data reveales that the Russian market is always high in terms of value but it tends not to be selected given the language and geographical barriers. Once we correct for this effect by excluding the Russian market from the ranking, we find that 40% of users migrate to the largest market (see Figure 6.6C).

We compare the users' decisions with a null random model, where at each closure a user moves with equal probability to any of the existent markets. The random probability P of rank i to be chosen for migration after m closures is equal to

$$P_i = \frac{\sum_{j=1}^{m} 1/c_j}{m},$$

, where c_j is the number of coexisting markets at the time of closure j. We find that the results of the random procedure are very different from actual data (see Figure 6.6)

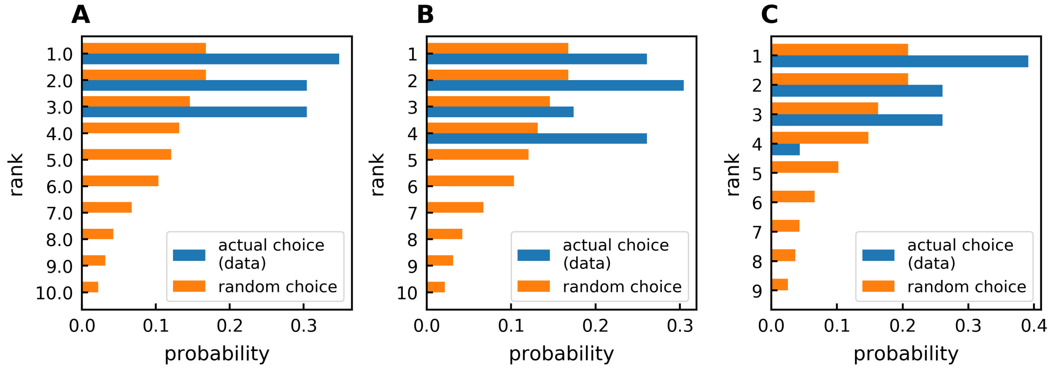


FIGURE 6.6: **Migration decision** Probability of a market to be chosen for migration given its rank at the time of coexisting market closure in comparison to the random model. Markets are ranked in descending order according to the number of overlapping users they have with the closed market excluding Russian markets (**A**), according to the total value in circulation in USD **(B)** and according to the total value in circulation in USD excluding Russian markets from the ranking (**C**). The random model in all figures represent a model where users can move to any existent market with equal probability.

6.6 Conclusion and discussion

In this chapter, we analyzed a novel dataset of Bitcoin transactions on 31 large dark marketplaces and investigated how the darknet market ecosystem was affected by the unexpected market closures between 2013 and 2019. The markets we considered were heterogeneous in many ways and 24 of them were closed abruptly due to police raids and scams. We found that the total volume traded on these dark markets dropped only temporarily following closures, revealing a remarkable resilience of the marketplace ecosystem. We identified the origin of this resilience, by focusing on individual users, and unveiled a swift and ubiquitous phenomenon of migration between recently closed markets and other coexisting ones. We found that migrating users were more active in terms of total transaction volume compared to users who did not migrate. Finally, we found that migrating users tended to migrate predictably to co-existing marketplaces which had the largest overall volume and the most numbers of users in common with the closed marketplace. Our findings shed new light on the consequences of sudden closure and/or police raids on dark market, which had been previously raised in the literature and among law enforcement entities [23, 24, 64]. Interesting future research directions include the role of market closure on the emergence of new markets, refining the analysis to investigate whether scam closures and police raids may have so-far neglected effects on user migration, delving deeper into the types of user behavior that can predict migration, and broadening the research to include the effect of online forums on the performance of existing markets as well as on the migration choices after a closure [141]. In our analysis we focused on the migration behaviour right after market closure (a week in particular). Extending the analysis to a longer period and user migration to other markets different from their first choice afterwards can indicate imitation behaviour.

7 Predicting dark markets' drug demand using Wikipedia views

Rapid changes in illicit drug use are a major public health concern. In the USA, $30,000$ people died from Fentanyl overdoses in 2018 alone [226]. It has become harder for authorities to monitor illicit drug markets. This is partly caused by shifts in production and distribution channels [227].

Changes in drug demand are also hard to observe. Traditionally, authorities have relied on annual surveys, such as the United Nations Office on Drugs and Crime (UNODC) World Drug Report [22]. The low frequency of these statistics means that authorities may miss opportunities to intervene early in drug crises, such as the US Fentanyl epidemic [228]. New drug categories may not appear in such surveys at all [229, 230]. For example, there were at least 36 Novel Psychoactive Substances (NPS) discovered between January and August 2019 alone [231]. A more frequent measure of drug use would enable public health authorities to intervene earlier, thereby using their limited resources more effectively.

To address these concerns, we present a novel method to predict drug demand based on high-frequency sales data from darknet markets. These are online markets that rely on encryption and digital currencies to enable anonymous trade of goods and services [232]. We measure predictive value by the out-of-sample "nowcast" errors of our predictive models. Nowcasting means estimating the current value of statistics that are usually released with a lag [233]. This method was first used for economic variables, such as GDP and inflation [233, 234, 235, 236], but has also been applied in epidemiology to predict Flu and Dengue outbreaks [237, 238, 239, 240]. Accurate nowcasts of drug demand would also be useful, given the long lag between current annual surveys.

We find that nowcasting models based on historic sales alone cannot accurately predict drug demand. It is also difficult to scrape the markets and there are frequent outages [241], so a measure based on darknet data alone could be unreliable. However, consumers may search for information on drugs before making a purchase [242, 243]. We therefore also collect data on Wikipedia page views for each drug, because these are reliably available in real-time [244]. We find that adding data on Wikipedia page views for these drugs dramatically improves the models' predictive accuracy. Therefore, we can construct a more frequent measure of drug demand using Wikipedia data. In turn, this could reduce public response times to future drug epidemics. The work presented in this chapter is based on publication [IV]. I contributed to this research through data collection and preparation, initial methodology testing (in sample regression model), methodology and study design and results interpretation.

7.1 Relevant literature

Predicting drug use with online behaviour is an emergent field with little directly comparable literature. One recent paper found a correlation between Google searches for NPS and their annual change in sales, as measured by the UNODC surveys [245]. However, they could not assess whether this relationship holds out-of-sample. Out-of-sample means predicting data that was not used to fit the model. due to the low time frequency of their sales data. Another study found a correlation between the volume of online comments about opioids, on the large forum Reddit, and the level of opioid abuse across US states [246]. They too were limited to in-sample analysis by the low time-frequency of their data, and were geographically restricted to the USA. Our global drug demand data is available at higher time frequency, which allows us to evaluate model performance out-of-sample. Moreover, the darknet data is actual sales rather than drug user surveys, which are particularly vulnerable to response bias [247]. This is therefore the first paper to predict darknet drug sales, and also the first to assess whether internet search can predict drug use out-of-sample.

Another growing strand of literature is using Wikipedia page views to predict economic variables. Previous papers have found relevance for traditional variables such as the stock market [248] and box office sales [249]. Some papers have also found Wikipedia data can predict cryptocurrency prices and Bitcoin trades [250, 251], which is particularly relevant given our use of darknet data. Finally, our results are consistent with prior findings that Wikipedia page views can predict other epidemics such as the Flu [252].

7.2 Materials and methods

Our analysis in this chapter relies on two datasets. The first dataset, is dark markets drug sales which comes from scraping the reviews from the four largest markets (Alphabay, Hansa, Traderoute and Valhalla) during June and July 2017. These covered 80% of global trade at the time [253]. For more details on the data collection and limitations see Section 3.3.3.

The second dataset is drugs Wikipedia pages daily views. We collect Wikipedia page views data through the Wikipedia API, which runs from July 2015 onward [156]. We further split the data by language, and use that as a proxy for the country of the viewer. This is likely a reasonable assumption for some languages. For more details on the data collection and limitations see Section 3.2.2.

We considered using Google Trends as an alternative indicator of interest expressed online. However Google Trends may be problematic due to language ambiguity for drugs. For example, a search for Magic Mushrooms could feasibly be expressed as ”mushrooms”, ”shrooms”, ”magic shrooms”, or ”truffles”. Wikipedia is much simpler in this regard, as there is a set page for each drug [254]. Furthermore, previous research has found a substantial correlation between Wikipedia and Google searches, so the added value in using both may be limited [255].

7.3 Results

7.3.1 Pooled model

For valid time-series inference, we require the distribution of our data to be stationary across time [166]. To formally test stationarity, we conduct Augmented Dickey Fuller (ADF) tests on the sales data for each drug. We find monthly sales to be non stationary, with an average ADF p-value of 0.42 across drugs. In contrast, the average ADF p-value after converting the sales data to percentage changes is approximately 0.00. We therefore conduct our analysis on the monthly percentage change in sales over time.

Our data is longitudinal with 3 dimensions: drug, country and time. Let $y_{i,j,t}$ denote the percentage change in sales of drug i in country j and time period t. Our baseline model is:

$$y_{i,j,t} = \beta_0 y_{i,j,t-1} + \alpha_i + \delta_j + \gamma_t \tag{7.1}$$

Where $y_{i,j,t-1}$ is an autoregressive term, in case of serial correlation. We also engineer binary variables (“dummies”) from the longitudinal data structure, which add complexity[1] to the baseline model:

- α_i are dummies for each drug.
- δ_j are dummies for each country.
- γ_t are dummies for each month, in case of seasonality.

In this specification (the ”pooled” model), we model all drugs jointly. The advantage is that we have more data to fit each of the pooled parameters, which makes overfitting less likely. For example, if we have N drugs and J countries then we have $N * J$ observations to fit each time dummy γ_t. Having fewer drug-specific parameters may also allow prediction of drugs that are not in our sample. However, the disadvantage is that we restrict complexity relative to modelling each drug separately, which we analyse in Section 7.3.2.

[1] Complexity means how much the model’s predictions can vary, rather than computational or time complexity.

To estimate the performance improvement from Wikipedia data, we add it to the baseline model. Letting $X_{i,j,t}$ be the percent change in Wikipedia views for drug i in country j and time period t, the "Wikipedia model" is:

$$y_{i,j,t} = \beta_0 y_{i,j,t-1} + \beta_1 X_{i,j,t} + \alpha_i + \delta_j + \gamma_t \tag{7.2}$$

Table 7.1 presents in-sample results comparing the pooled models. All models are unpenalised regression. Scores are adjusted R^2, which includes a penalty term for models with more features. The baseline score is the model accuracy without including Wikipedia views. The Wikipedia Model includes data on Wikipedia views. The models in the first column use only the autoregressive terms as predictors and Wikipedia views as predictors. The models in the second column add complexity with country, drug and month dummies.

The Wikipedia model outperforms the baseline by between 49 and 64 percentage points (pp), depending on the model choice. Therefore Wikipedia data is a strong in-sample indicator for drug demand. This effect is also much larger than the boost from adding the dummies, which we estimate at 7-20pp. However, in-sample performance may not reflect true predictive accuracy because of possible overfitting.

TABLE 7.1: **Pooled model - in-sample accuracy**

	Simple Model	All Dummies
Baseline R^2	0.003	0.22
Wikipedia Model R^2	0.64	0.71
Sample Size	1918	1918
Number of features	2	35

We cannot evaluate out-of-sample performance with a random train test split, as time series data is not independent and identically distributed (i.i.d). A random split would put some data in the training set that occurs chronologically after some of the testing set. We would therefore be using data from the future to fit a model predicting the past. This is clearly not possible when performing an actual prediction.

We instead use a one-step ahead nowcasting procedure to measure out-of-sample performance [238]. We first set a training window, w, that determines the size of the training set. Then for each period $t \in [w, T]$ in the data, the training set is data from periods $\in [t - w - 1, t - 1]$. To prevent overfitting, we penalise the model's coefficients using LASSO and 5-fold cross validation in the training set. The penalised model then predicts the test set from period t, which is completely held out from training. This procedure only predicts the present with data from the past, so it is truly out-of-sample.

We record the errors in period t and use the mean absolute error (MAE) to measure that period's accuracy. Each time we increase t, we slide the training window to update the data and re-fit the model. The model therefore "adapts" over time to new data, which helps maintain accuracy if the underlying relationship changes over time. We set a training window of 12 months, which allows the model to see each month in the training set and fit the seasonality dummies. The first period in our test set is therefore October 2016.

Figure 7.1 compares out-of-sample results from the pooled models. We include month, drug and country dummies in both models.

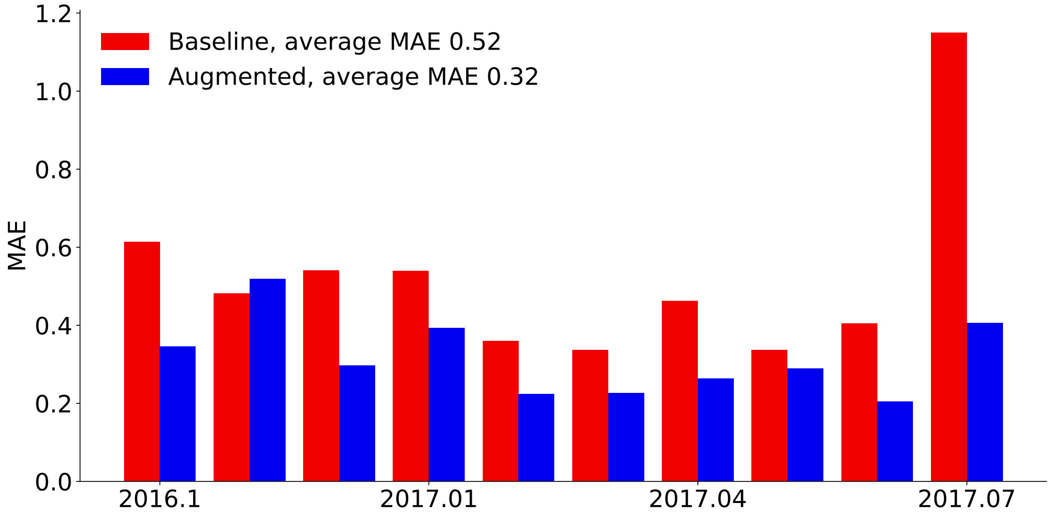


FIGURE 7.1: **Out-of-sample adaptive nowcasting results - pooled model.**

Adding Wikipedia data to the model reduces nowcast mean absolute error (MAE) in almost every time period. The average reduction in error across the sample is 43%. These results are robust to a range of training windows as shown in Table 7.2. Therefore, Wikipedia data is also a strong out-of-sample predictor for drug demand.

TABLE 7.2: Out-of-sample results with different training windows. The main text results use a 12 month window.

Training Window	Baseline MAE	Augmented MAE
10 months	0.51	0.30
11 months	0.52	0.30
13 months	0.52	0.31
14 months	0.53	0.29

We also address the potential data limitations of data sparsity and deleted review (discussed in Section 3.3.3) by considering different aggregation frequencies and different starting time. Table 3.9 shows the usage of different aggregation frequencies effect on the MAE, in all cases the augmented model is outperforming the baseline model. Similarly, in case of using different start dates, Table 7.3 shows that the augmented model is outperforming the baseline model.

TABLE 7.3: Out-of-sample results at different aggregation frequencies. The main text results aggregate data to 1 month frequency.

Aggregation Frequency	Baseline MAE	Augmented MAE
2 weeks	0.45	0.30
4 weeks	0.50	0.30
6 weeks	0.55	0.32
8 weeks	0.61	0.29

7.3.2 Modelling each drug separately

We have both country and time dimensions in the data, which increases the sample size for each individual drug. This allows us to fit separate models for each drug i:

TABLE 7.4: Out-of-sample results using different start dates for the data. The later the start date, the less the darknet scrape's coverage is affected by removed listings. The main text results use a start date of July 2015, which keeps all possible data.

Start Date	Baseline MAE	Augmented MAE
April 2016	0.47	0.29
July 2016	0.45	0.27
October 2016	0.50	0.27
January 2017	0.57	0.35

$$y_{i,j,t} = \beta_0^i y_{i,j,t-1} + \beta_1^i X_{i,j,t} + \delta_j^i + \gamma_t^i \tag{7.3}$$

The parameters β_0^i, β_1^i, α_j^i and γ_t^i now vary by drug. This allows for much more complexity in the models. The Wikipedia model from Equation 7.3 now has 299 features, compared to 35 features for the model from Equation 7.2. The models may be more accurate, but also more prone to overfitting. This is particularly true for the dummies, as there will be far fewer data points to fit each of them. We therefore focus on the out-of-sample model performance, but show in-sample performance in Table 7.5 where Wikipedia based model still outperform the baseline model.

We again assess out-of-sample performance using adaptive nowcasting, but we perform this separately for each drug. Figure 7.2 compares nowcast results between the baseline and Wikipedia models. The MAE is the mean across the entire nowcasting procedure for a given drug - we do not display results over time as in Figure 7.1. We do not include any dummies in these models: this actually reduces accuracy as there likely is not enough data to fit the dummies robustly when modelling each drug separately.

Adding Wikipedia data reduces nowcast errors for every drug relative to the baseline. The average MAE reduction across drugs is 42%. Therefore, Wikipedia data remains strongly predictive of demand when modelling each drug separately.

We then focus particularly on the Fentanyl due to the recent epidemic in the USA [228]. The US Fentanyl epidemic demonstrates how a more frequent

TABLE 7.5: In-sample results when modelling each drug separately. We only report out-of-sample results in the main text due to concerns about overfitting.

Drug	Baseline Model R^2	Wikipedia Model R^2
Cannabis	0.57	0.67
MDMA	0.76	0.89
DMT	0.81	0.92
Xanax	0.70	0.87
Cocaine	0.84	0.94
Diazepam	0.55	0.94
Heroin	0.78	0.81
Ketamine	0.81	0.94
Fentanyl	0.88	0.95
LSD	0.60	0.92
2CB	0.80	0.84
Speed	0.66	0.96
Meth	0.84	0.94

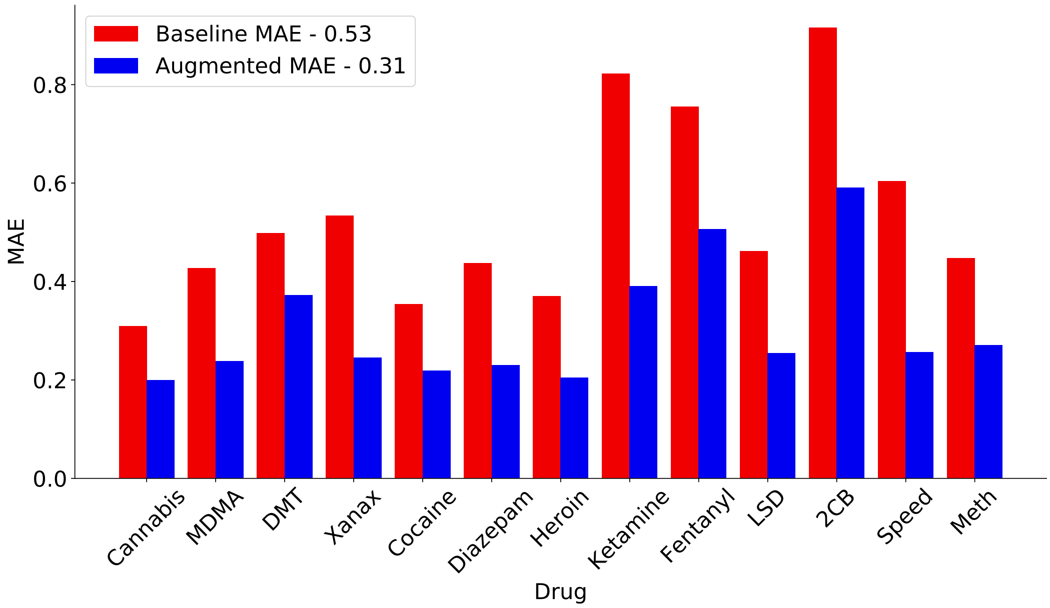


FIGURE 7.2: **Out-of-sample adaptive nowcast results, modelling each drug separately.**

measure of drug use could be highly valuable. The federal government only declared a national emergency in January 2017, which was arguably too late [228]. Figure 7.3 compares the Wikipedia model's predictions against the baseline for US demand for Fentanyl. The Wikipedia model makes more variable predictions and is therefore more able to detect shifts in demand. For example, the Wikipedia model correctly predicts the big demand spikes in June 2016 and January 2017, whereas the baseline model does not. Therefore, the Wikipedia model may have been able to provide early warning of the US Fentanyl epidemic.

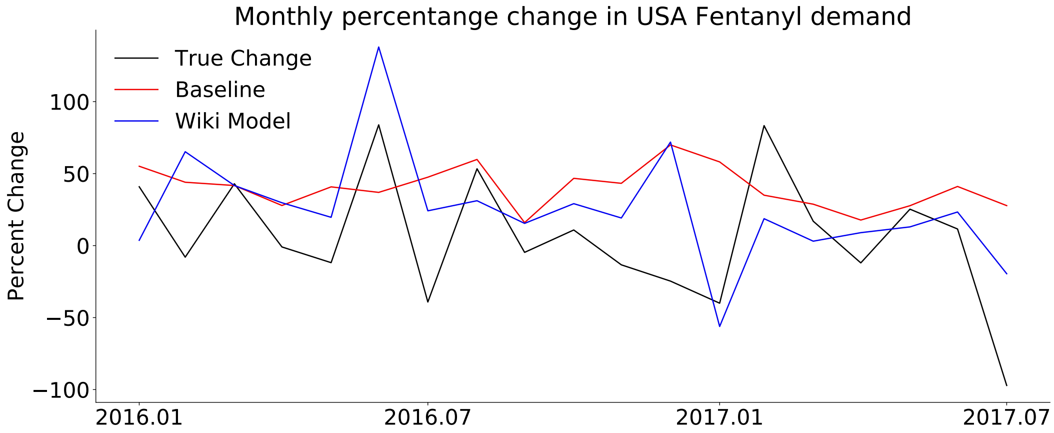


FIGURE 7.3: Out-of-sample predictions of US demand for Fentanyl in the USA. The black series is the true change in demand, red is the baseline model and blue is the Wikipedia model.

7.3.3 Modelling each country separately

Similarly to Section 7.3.2, we can also fit a separate model for each country j:

$$y_{i,j,t} = \beta_0^j y_{i,j,t-1} + \beta_1^j X_{i,j,t} + \alpha_i^j + \gamma_t^j \tag{7.4}$$

The parameters β_0^j, β_1^j, α_j^j and γ_t^j now vary by country, which again allows for greater model complexity.

Figure 7.4 presents out-of-sample adaptive nowcast results when modelling each country separately. We fit a separate model for each country, so each country has its own feature weights on the autoregressive term and Wikipedia page views. MAE is the mean error across the entire nowcasting procedure, for a given country.

Adding Wikipedia data to the baseline improves accuracy in every country relative to the baseline, which is consistent with previous models. The average MAE reduction across countries is 40%. Therefore, Wikipedia data remains a strong out-of-sample predictor when modelling each country separately.

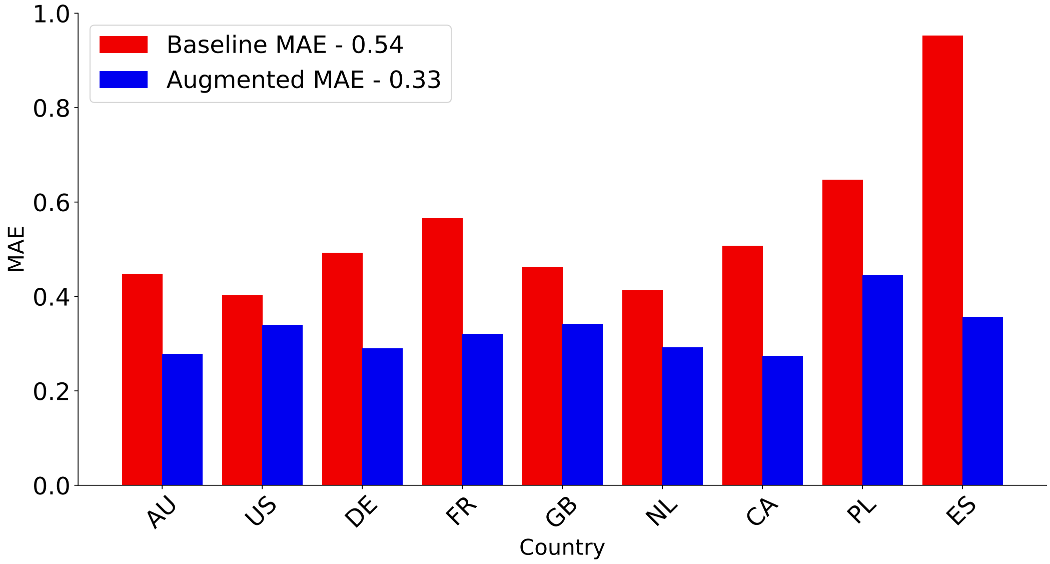


FIGURE 7.4: **Out-of-sample adaptive nowcast results, modelling each country separately.**

7.4 Conclusion and discussion

In this chapter, we have analysed whether darknet and internet search data can help construct a high frequency measure of drug demand. We first show that a model based on darknet data alone cannot accurately nowcast drug demand. We then show that adding data on Wikipedia views for each drug greatly improves predictive accuracy. These results hold out-of-sample across all drugs and are robust to a range of modelling choices.

Table 7.1 shows that past drug sales data alone cannot accurately predict current drug sales. However our results consistently show that adding Wikipedia data, which is reliably available in real-time, greatly boosts nowcast accuracy. We present results using two broad approaches: a pooled approach, where all drugs are modelled jointly, and a second approach where we model each drug and country separately. In all specifications, Wikipedia page views reduce nowcast errors by at least 40% relative to the baseline model.

The average nowcast errors in the pooled model, shown in Figure 7.1, are comparable to the errors when modelling each drug separately, shown in Figure 7.2. This suggests we should use the pooled model, as it may allow prediction of demand for drugs with little available data. This would be particularly useful when new drugs are entering the market, such as the NPS, which traditional surveys struggle to capture [229]. However, formal analysis of whether this would be accurate in practice is beyond the scope of this paper.

Given the daily frequency of our data, we can vary the time frequency of our predictive models. Higher time frequencies may be more useful for policymakers as they would get a faster estimate of drug use. However the sales data is sparser at higher frequency, as shown in Figure 3.9. Our results are qualitatively robust to all aggregation frequencies, but stronger at lower frequencies (see Table 7.3, Section 7.3.1). This suggests there may be a trade-off between model speed and accuracy. Nevertheless, the monthly frequency of our main results would still be much faster than the current annual survey data.

There may be an issue with the geographic link between Wikipedia page views and darknet sales. The Wikipedia data is split by language of the page (e.g. French), whereas the darknet sales are split by country of sale (e.g. France). The link between them is likely to be strong when the language is not widely spoken outside its origin country, such as Polish. However there are languages where the country of origin is less clear, such as English. We analyse this issue in Figure 7.4, which presents results from modelling each country separately. There is some evidence that the Wikipedia model performs worse for countries with a shared language, such as the US and Australia. Nevertheless, the difference is small and the Wikipedia model outperforms the baseline across all countries. Future research could use internet search data where the user's location is known, such as Google Trends, rather than inferred from language.

We acknowledge potential limits to the external validity of extrapolating our darknet results to predicting overall drug use. The lack of granular real data on drug usage renders our analysis both important and untested against actual usage. Another important limitation is that our analysis was considering historical data from markets which are no longer operating now, however other dark markets continue to trade on the dark web. There are also known demographic biases with internet usage [256], so darknet drug users may not be representative of drug users overall. If so, this may diminish the predictive power of Wikipedia data for overall drug use. However, previous research found that darknet demand geographically represents overall drug demand well for cannabis, cocaine and heroin [253]. Moreover, the Wikipedia model performs well across a range of drugs, as shown in Figure 7.2. If demographic bias were affecting our results, we may expect the Wikipedia model to perform better among drugs whose consumers use the internet more, such as DMT, LSD and 2C-B [257]. We cannot find strong evidence of this, with the Wikipedia model performing well for harder ”street” drugs such as heroin and cocaine, whose demographics are less represented among internet users. Furthermore, predicting darknet demand alone may be useful given its rapid growth over the last decade [16].

We acknowledge there are limits on extrapolating results from darknet data to wider drug consumption, due to demographic biases among internet users. Nevertheless, we believe there is strong evidence overall that internet search data may greatly improve the speed of official drug statistics, which are currently annual frequency. In turn, this may help policymakers respond more quickly to the next drug epidemic.

8 Anticipating cryptocurrencies prices using machine learning

The popularity of cryptocurrencies has skyrocketed in 2017 due to several consecutive months of super-exponential growth of their market capitalisation as we have shown in Chapter 4. Between 2.9 and 5.8 millions of private as well as institutional investors are in the different transaction networks, according to a recent survey [97], and access to the market has become easier over time. Major cryptocurrencies can be bought using fiat currency in a number of online exchanges (e.g., Binance [258], Upbit [259], Kraken [260], etc) and then be used in their turn to buy less popular cryptocurrencies. The volume of daily exchanges is currently superior to $15 billions. Since 2017, over 170 hedge funds specialised in cryptocurrencies have emerged and bitcoin futures have been launched to address institutional demand for trading and hedging Bitcoin [261].

The market is diverse and provides investors with many different products. Just to mention a few, Bitcoin was expressly designed as a medium of exchange [1, 25]. In Chapter 4 we have shown that the long-term properties of the cryptocurrency marked have remained stable between 2013 and 2017 and are compatible with a scenario in which investors simply sample the market and allocate their money according to the cryptocurrency's market shares [206]. While this is true on average, various studies have focused on the analysis and forecasting of price fluctuations, using mostly traditional approaches for financial markets analysis and prediction [262, 118, 263, 15, 264].

The success of machine learning techniques for stock markets prediction [265, 266, 267, 268, 269, 270, 271], suggests that these methods could be effective also in predicting cryptocurrencies prices. However, research was limited to Bitcoin or few cryptocurrencies.

Here, we test the performance of three models in predicting daily cryptocurrency price for 1,681 currencies. Two of the models are based on gradient boosting decision trees [272] and one is based on long short-term memory (LSTM) recurrent neural networks [273]. In all cases, we build investment portfolios based on the predictions and we compare their performance in terms of return on investment. We find that all of the three models perform better than a baseline 'simple moving average' model [274, 275, 276, 277] where a currency's price is predicted as the average price across the preceding days, and that the method based on long short-term memory recurrent neural networks systematically yields the best return on investment.

The chapter is structured as follows: In Section 8.1 we describe the data (see Section 8.1.1), the metrics characterizing cryptocurrencies that are used along the paper (see Section 8.1.2), the forecasting algorithms (see Section 8.1.3), and the evaluation metrics (see Section 8.1.4). In Section 8.2, we present and compare the results obtained with the three forecasting algorithms and the baseline method. In Section 8.3, we conclude and discuss results. The work presented in this chapter is based on publication [V]. I contributed to this research through data collection and preparation, initial methodology testing, methodology and study design and results interpretation.

8.1 Materials and methods

8.1.1 Data description and pre-processing

Cryptocurrency data was extracted from the website Coin Market Cap [4], collecting daily data from 300 exchange markets platforms starting in the period between November 11, 2015 and April 24, 2018. The dataset contains the daily price in U.S. dollars, the market capitalisation and the trading volume of $1,681$ cryptocurrencies, where the market capitalization is the product between price and circulating supply, and the volume is the number of coins exchanged in a day. The daily price is computed as the volume weighted average of all prices reported at each market. Figure 8.1 shows the number of currencies with trading volume larger than V_{min} over time, for different values of V_{min}. In the following sections, we consider that only currencies with daily trading volume higher than 10^5 USD can be traded at any given day.

The website lists cryptocurrencies traded on public exchange markets that have existed for more than 30 days and for which an API as well as a public URL showing the total mined supply are available. Information on the market capitalization of cryptocurrencies that are not traded in the 6 hours preceding the weekly release of data is not included on the website. Cryptocurrencies inactive for 7 days are not included in the list released. These measures imply that some cryptocurrencies can disappear from the list to reappear later on. In this case, we consider the price to be the same as before disappearing. However, this choice does not affect results since only in 28 cases the currency has volume higher than 10^5 USD right before disappearing (note that there are $124,328$ entries in the dataset with volume larger than 10^5 USD).

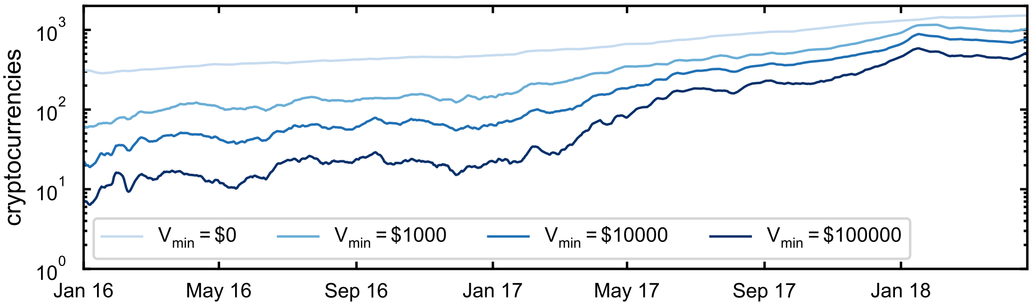


FIGURE 8.1: **Number of cryptocurrencies.** The cryptocurrencies with volume higher than V_{min} as a function of time, for different values of V_{min}. For visualization purposes, curves are averaged over a rolling window of 10 days.

8.1.2 Metrics

Cryptocurrencies are characterised over time by several metrics, namely

- Price, The exchange rate, determined by supply and demand dynamics.

- Market capitalization, The product of the circulating supply and the price.
- Market share, The market capitalization of a currency normalized by the total market capitalization.
- Rank, The *rank* of currency based on its market capitalization.
- Volume, Coins traded in the last 24 hours.
- Age, Lifetime of the currency in days.

The profitability of a currency c over time can be quantified through the *return on investment* (ROI), measuring the return of an investment made at day t_i relative to the cost [278]. The index i rolls across days and it is included between 0 and 844, with $t_0 =$ January 1, 2016, and $t_{844} =$ April 24, 2018. Since we are interested in the short-term performance, we consider the return on investment after 1 day defined as

$$ROI(c, t_i) = \frac{p_{(c,t_i)} - p_{(c,t_i-1)}}{p_{(c,t_i-1)}}. \tag{8.1}$$

In Figure 8.2, we show the evolution of the ROI over time for Bitcoin (orange line) and on average for currencies whose volume is larger than $V_{min} = 10^5$ USD at $t_i - 1$ (blue line). In both case, the average return on investment over the period considered is larger than 0, reflecting the overall growth of the market.

8.1.3 Forecasting algorithms

We test and compare three supervised methods for short-term price forecasting. The first two methods rely on XGboost [279], an open-source scalable machine learning system for tree boosting used in a number of winning Kaggle solutions (17/29 in 2015) [280]. The third method is based on the long short-term memory (LSTM) algorithm for recurrent neural networks [273] that have demonstrated to achieve state-of-the-art results in time-series forecasting [281].

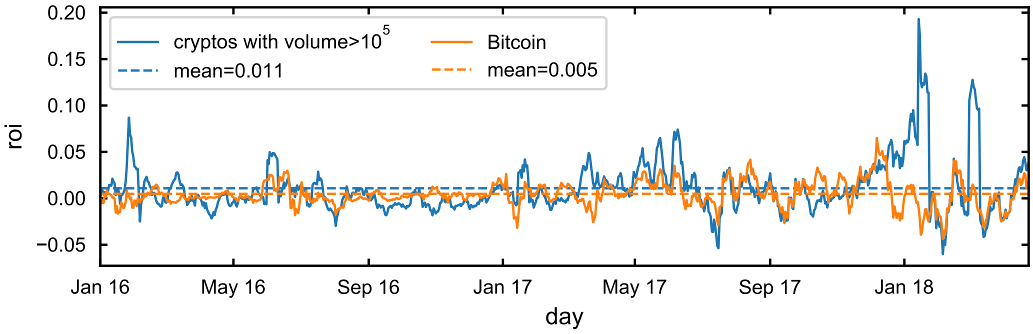


FIGURE 8.2: **Return on investment over time.** The daily return on investment for Bitcoin (orange line) and the average for currencies with volume larger than $V_{min} = 10^5$ USD (blue line). Their average value across time (dashed lines) is larger than 0. For visualization purposes, curves are averaged over a rolling window of 10 days.

Method 1: The first method considers one single regression model to describe the change in price of all currencies (see Figure 8.3). The model is an ensemble of regression trees built by the XGboost algorithm. The features of the model are characteristics of a currency between time $t_j - w$ and $t_j - 1$ and the target is the ROI of the currency at time t_j, where w is a parameter to be determined. The characteristics considered for each currency are: price, market capitalization, market share, rank, volume and ROI (see Equation 8.1). The features for the regression are built across the window between $t_j - w$ and $t_j - 1$ included (see Figure 8.3). Specifically, we consider the average, the standard deviation, the median, the last value and the trend (e.g. the difference between last and first value) of the properties listed above. In the training phase, we include all currencies with volume larger than 10^5 USD, and t_j between $t_i - W_{training}$ and t_i. In general, larger training windows do not necessarily lead to better results (see results section), because the market evolves across time. In the prediction phase, we test on the set of existing currencies at day t_i. This procedure is repeated for values of t_i included between January 1, 2016 and April 24, 2018.

Method 2: Also the second method relies on XGboost, but now the algorithm is used to build a different regression model for each currency c_i (see Figure 8.4). The features of the model for currency c_i are the characteristics

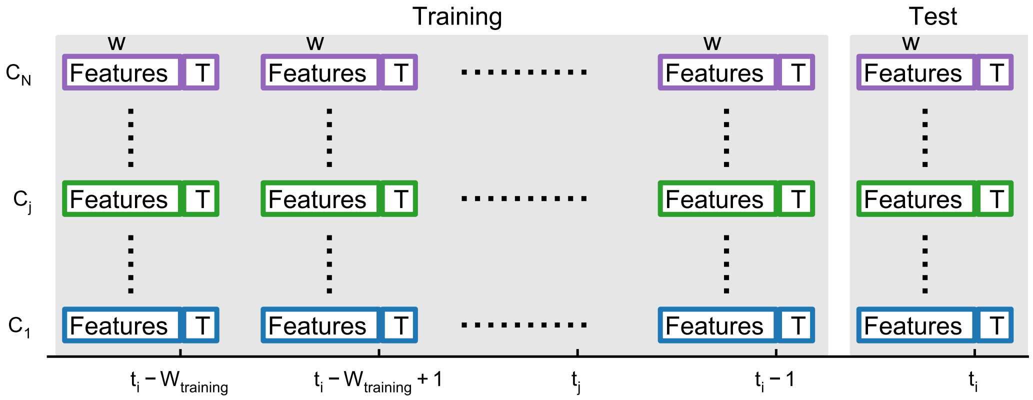


FIGURE 8.3: **Schematic description of Method 1.** The training set is composed of features and target (T) pairs, where features are various characteristics of a currency c_i, computed across the w days preceding time t_j and the target T is the price of c_i at t_j. The features-target pairs are computed for all currencies c_i and all values of t_j included between $t_i - W_{training}$ and $t_i - 1$. The test set includes features-target pairs for all currencies with trading volume larger than 10^5 USD at t_i, where the target is the price at time t_i and features are computed in the w days preceding t_i.

of all the currencies in the dataset between $t_j - w$ and $t_j - 1$ included and the target is the ROI of c_i at day t_j(i.e., now the algorithm learns to predict the price of the currency i based on the features of all the currencies in the system between $t_j - w$ and $t_j - 1$). The features of the model are the same used in Method 1 (e.g. the average, standard, deviation, median, last value, difference between last and first value of the following quantities: price, market capitalisation, market share, rank, volume and ROI) across a window of length w. The model for currency c_i is trained with pairs features target between times $t_i - W_{training}$ and $t_i - 1$. The prediction set include only one pair: the features (computed between $t_i - w$ and $t_i - 1$) and the target (computed at t_i) of currency c_i.

Method 3: The third method is based on Long Short Term Memory networks, a special kind of Recurrent Neural Networks, capable of learning long-term dependencies. As for Method 2, we build a different model for each currency. Each model predicts the ROI of a given currency at day t_i based on the values of the ROI of the same currency between days $t_i - w$ and $t_i - 1$ included.

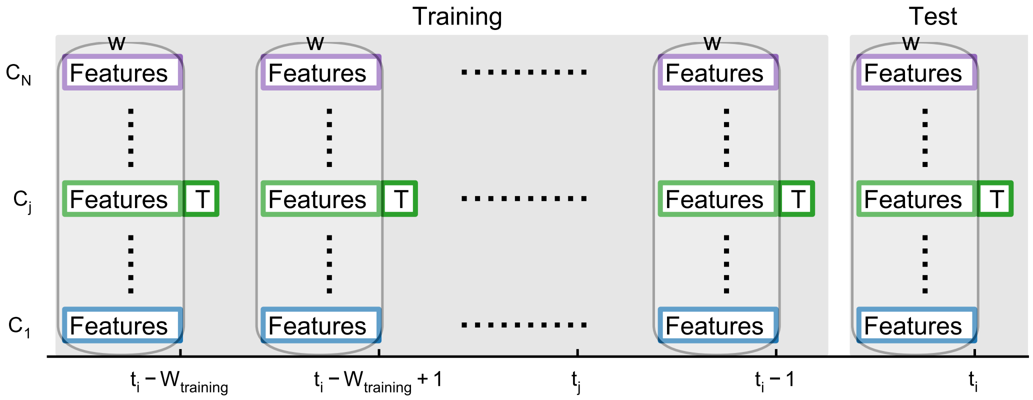


FIGURE 8.4: **Schematic description of Method 2.** The training set is composed of features and target (T) pairs, where features are various characteristics of all currencies, computed across the w days preceding time t_j and the target T is the price of c_i at t_j. The features-target pairs include a single currency c_i, for all values of t_j included between $t_i - W_{training}$ and $t_i - 1$. The test set contains a single features-target pair: the characteristics of all currencies, computed across the w days preceding time t_i and the price of c_i at t_i.

Baseline method: As baseline method, we adopt the simple moving average strategy (SMA) widely tested and used as a null model in stock market prediction [274, 275, 276, 277]. It estimates the price of a currency at day t_i as the average price of the same currency between $t_i - w$ and $t_i - 1$ included.

8.1.4 Evaluation

We compare the performance of various investment portfolios built based on the algorithms predictions. The investment portfolio is built at time $t_i - 1$ by equally splitting an initial capital among the top n currencies predicted with positive return. Hence, the total return at time t_i is:

$$R(t_i) = \frac{1}{n} \sum_{c=1}^{n} ROI(c, t_i). \tag{8.2}$$

The portfolios performance is evaluated by computing the Sharpe ratio and the geometric mean return. The Sharpe ratio is defined as:

$$S(t_i) = \frac{\overline{R}}{s_R}, \tag{8.3}$$

where $\overline{R}$ is the average return on investment obtained between times 0 and t_i, and s_R, the corresponding standard deviation.

The geometric mean return is defined as:

$$G(t_i) = \sqrt[t_i]{\prod_{t_j=1}^{t_i} 1 + R(t_j)}, \tag{8.4}$$

where t_i corresponds to the total number of days considered. The cumulative return obtained at t_i after investing and selling on the following day for the whole period is defined as $G(t_i)^2$.

The number of currencies n to include in a portfolio is chosen at t_i by optimising either the geometric mean $G(t_i - 1)$ (geometric mean optimisation) or the Sharpe ratio $S(t_i - 1)$ (Sharpe ratio optimisation) over the possible choices of n. The same approach is used to choose the parameters of Method 1 (w and $W_{training}$), Method 2 (w and $W_{training}$), and the baseline method (w).

8.2 Results

We predict the price of the currencies at day t_i, for all t_i included between Jan, 1st 2016 and Apr 24th, 2018. The analysis considers all currencies whose age is larger than 50 days since their first appearance and whose volume is larger than \$100000. To discount for the effect of the overall market movement (i.e., market growth, for most of the considered period), we consider cryptocurrencies prices expressed in Bitcoin. This implies that Bitcoin is excluded from our analysis.

8.2.1 Parameter setting

First, we choose the parameters for each method. Parameters include the number of currencies n to include the portfolio as well as the parameters specific to each method. In most cases, at each day t_i we choose the parameters that maximise either the geometric mean $G(t_i - 1)$ (geometric mean optimisation) or the Sharpe ratio $S(t_i - 1)$ (Sharpe ratio optimisation) computed between times 0 and t_i.

Baseline strategy: We test the performance of the baseline strategy for choices of window $w \geq 2$ (the minimal requirement for the ROI to be different from 0) and $w < 30$. We find that the value of w mazimising the geometric mean return (see Appendix D.1) and the Sharpe Ratio (see Appendix D.1) fluctuates especially before November 2016 and has median value 4 in both cases. The number of currencies included in the portfolio oscillates between 1 and 11 with median at 3, both for the Sharpe Ratio (see Appendix D.1) and the geometric mean return (see Appendix D.1) optimisation.

Method 1: We explore values of the window w in $\{3, 5, 7, 10\}$ days and the training period $W_{training}$ in $\{5, 10, 20\}$ days (see Appendix D.2). We find that the median value of the selected window w across time is 7 for both the Sharpe ratio and the geometric mean optimisation. The median value of $W_{training}$ is 5 under geometric mean optimisation and 10 under Sharpe ratio optimisation. The number of currencies included in the portfolio oscillates between 1 and 43 with median at 15 for the Sharpe Ratio (see Appendix D.2) and 9 for the geometric mean return (see Appendix D.2) optimisations.

Method 2: We explore values of the window w in $\{3, 5, 7, 10\}$ days and the training period $W_{training}$ in $\{5, 10, 20\}$ days (see Appendix D.3). The median value of the selected window w across time is 3 for both the Sharpe ratio and the geometric mean optimisation. The median value of $W_{training}$ is 10 under geometric mean and Sharpe ratio optimisation. The number of currencies included has median at 17 for the Sharpe Ratio and 7 for the geometric mean optimisation (see Appendix D.3).

Method 3: The LSTM has three parameters: The number of epochs, or complete passes through the dataset during the training phase; the number of neurons in the neural network, and the length of the window w. These parameters are chosen by optimising the price prediction of three currencies (Bitcoin, Ripple, and Ethereum) that have on average the largest market share across time (excluding Bitcoin Cash that is a fork of Bitcoin). Results (see Appendix D.4) reveal that, in the range of parameters explored, the best results are achieved for $w = 50$. Results are not particularly affected by the choice of the number of neurones nor the number of epochs. We choose 1 neuron and 1000 epochs since the larger these two parameters, the larger the computational time. The number of currencies to include in the portfolio is optimised over time by maximising the geometric mean return (see Appendix D.5) and the Sharpe ratio (see Appendix D.5). In both cases the median number of currencies included is 1.

8.2.2 Cumulative return

In Figure 8.5, we show the cumulative return obtained using the 4 methods. The cumulative returns achieved on April,24 under the Sharpe Ratio optimisation are $\sim$ 65 BTC (Baseline), $\sim 1.1 \cdot 10^3$ BTC (Method 1), $\sim$ 95 BTC (Method 2), $\sim 1.2 \cdot 10^9$ BTC (Method 3). Under geometric mean optimisation we obtain $\sim$ 25 BTC (Baseline), $\sim 19 \cdot 10^3$ BTC (Method 1), $\sim$ 1.25 BTC (Method 2), $\sim 3.6 \cdot 10^8$ BTC (Method 3). The cumulative returns obtained in USD are higher (see Appendix D.4, Figure D.9). This is expected, since the Bitcoin price has increased during the period considered. While some of these figures appear exaggerated, it is worth noticing that (i) we run a theoretical exercise assuming that the availability of Bitcoin is not limited and (ii) under this assumption the upper bound to our strategy, corresponding to investing every day in the most performing currency results in a total cumulative return of $6 \cdot 10^{123}$ BTC (see Appendix D.6). We consider also the more realistic scenario of investors paying a transaction fee when selling and buying currencies (see Appendix D.3). In most exchange markets, the fee is typically included between 0.1% and 0.5% of the traded amount [282]. For fees up to 0.2%, all the investment methods presented above lead, on average, to positive returns over the entire period (see Appendix D.3, Table D.1). The best performing method, Method 3, achieves positive gains also when fees up to 1% are considered (see Appendix D.3, Table D.1).

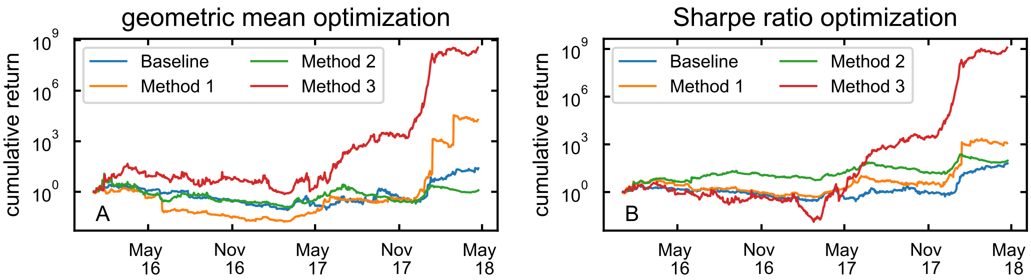


FIGURE 8.5: **Cumulative returns.** The cumulative returns obtained under the Sharpe Ratio optimisation (A) and the geometric mean optimisation (B) for the baseline (blue line), Method 1 (orange line), Method 2 (green line) and Method 3 (red line). Analyses are performed considering prices in BTC.

The cumulative return in Figure 8.5 is obtained by investing between January 1st, 2016 and April 24th, 2018. We investigate the overall performance of the various methods by looking at the geometric mean return obtained in different periods (see Figure 8.6). Results presented in Figure 8.6 are obtained under Sharpe ratio optimisation for the baseline (Figure 8.6A), Method 1 (Figure 8.6B), Method 2 (Figure 8.6C), and Method 3 (Figure 8.6D). Note that, while in this case the investment can start after January 1st, 2016, we optimised the parameters by using data from that date on in all cases. Results are considerably better than those achieved using geometric mean return optimisation (see Appendix D.10). Finally, we observe that better performance is achieved when the algorithms consider prices in Bitcoin rather than USD (see Appendix D.4 , Table D.2).

8.2.3 Feature importance

In this section we investigate features importance. Since XGBoost relies on decision trees for prediction, feature importance is calculated as the average importance of a feature across all decision trees in the model. For a single tree, feature importance is measured by the information gained (increase in the performance measure) by adding the new feature in the tree. In Figure 8.7, we illustrate the relative importance of the various features in Method 1 and Method 2. For Method 1, we show the average feature importance;

For Method 2, we show the average feature importances for two sample currencies: Ethereum and Ripple.

8.2.4 Portfolio composition

The 10 most selected currencies under Sharpe Ratio optimisation are the following:

Baseline: Factom (91 days), E-Dinar Coin (89 days), Ripple (76 days), Ethereum (71 days), Steem (70 days), Lisk (70 days), MaidSafeCoin (69 days), Monero (58 days), BitShares (55 days), EDRCoin (52 days).

Method 1: Ethereum (154 days), Dash (128 days), Monero (111 days), Factom (104 days), Ripple (94 days), Litecoin (93 days), Dogecoin (92 days), Maid Safe Coin (86 days), BitShares (73 days), Tether (59 days)

Method 2: Ethereum (63 days), Monero (61 days), Factom (51 days), Ripple (42 days), Dash (40 days), Maid Safe Coin (40 days), Siacoin (30 days), NEM (26 days), NXT (26 days), Steem (23 days).

Method 3: Factom (48 days), Monero (46 days), Ethereum (39 days), Lisk (36 days), Maid Safe Coin (32 days), E-Dinar Coin (32 days), BitShares (26 days), B3 Coin (26 days), Dash (25 days), Cryptonite (22 days).

8.3 Conclusion and discussion

We tested the performance of three forecasting models on daily cryptocurrency prices for $1,681$ currencies. Two of them (Method 1 and Method 2) were based on gradient boosting decision trees and one is based on long short-term memory recurrent neural networks (Method 3). In Method 1, the same model was used to predict the return on investment of all currencies; in Method 2, we built a different model for each currency, that uses information on the behaviour of the whole market to make a prediction on that single currency; in Method 3, we used a different model for each currency, where the prediction is based on previous prices of the currency.

We built investment portfolios based on the predictions of the different method and compared their performance with that of a baseline represented by the well known simple moving average strategy. The parameters of each model were optimised for all but Method 3 on a daily basis, based on the outcome of each parameters choice in previous times. We used two evaluation metrics used for parameter optimisation: The geometric mean return and the Sharpe ratio. To discount the effect of the overall market growth, cryptocurrencies prices were expressed in Bitcoin. All strategies, produced profit (expressed in Bitcoin) over the entire considered period and for a large set of shorter trading periods (different combinations of start and end dates for the trading activity), also when transaction fees up to 0.2% are considered.

The three methods performed better than the baseline strategy when the investment strategy was ran over the whole period considered. The optimisation of parameters based on the Sharpe ratio achieved larger returns. Methods based on gradient boosting decision trees (Method 1 and 2) worked best when predictions were based on short-term windows of 5/10 days, suggesting they exploit well mostly short-term dependencies. Instead, LSTM recurrent neural networks worked best when predictions were based on $\sim$ 50 days of data, since they are able to capture also long-term dependencies and are very stable against price volatility. They allowed to make profit also if transaction fees up to 1% are considered. Methods based on gradient boosting decision trees allow to better interpret results. We found that the prices and the returns of a currency in the last few days preceding the prediction were leading factors to anticipate its behaviour. Among the two methods based on random forests, the one considering a different model for each currency performed best (Method 2). Finally, it is worth noting that the three methods proposed perform better when predictions are based on prices in Bitcoin rather than prices in USD. This suggests that forecasting simultaneously the overall cryptocurrency market trend and the developments of individual currencies is more challenging than forecasting the latter alone.

It is important to stress that our study has limitations. First, we did not attempt to exploit the existence of different prices on different exchanges, the consideration of which could open the way to significantly higher returns on investment. Second, we ignored intra-day price fluctuations and considered an average daily price. Finally, and crucially, we run a theoretical test in which the available supply of Bitcoin is unlimited and none of our trades influence the market. Notwithstanding these simplifying assumptions, the methods we presented were systematically and consistently able to identify outperforming currencies. Extending the current analysis by considering these and other elements of the market is a direction for future work.

A different yet promising approach to the study cryptocurrencies consists in quantifying the impact of public opinion, as measured through social media traces, on the market behaviour, in the same spirit in which this was done for the stock market [106]. While it was shown that social media traces can be also effective predictors of Bitcoin [283, 12, 220, 11, 284, 17, 170] and other currencies [285] price fluctuations, our knowledge of their effects on the whole cryptocurrency market remain limited and is an interesting direction for future work.

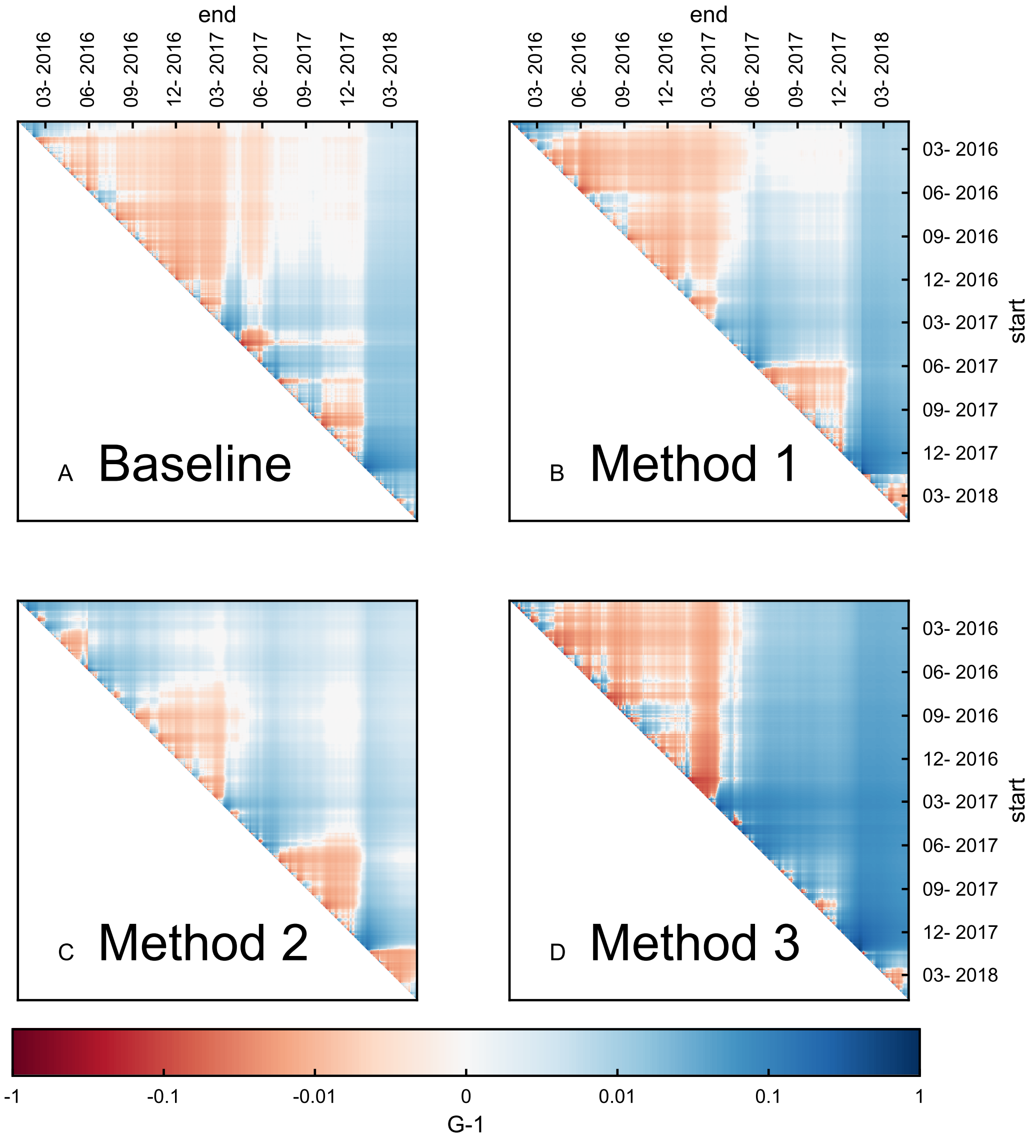


FIGURE 8.6: **Geometric mean return obtained within different periods of time.** The geometric mean return computed between time ”start” and ”end” using the Sharpe ratio optimisation for the baseline (A), Method 1 (B), Method 2 (C) and Method 3 (D). Note that, for visualization purposes, the figure shows the translated geometric mean return G-1. Shades of red refers to negative returns and shades of blue to positive ones (see colour bar).

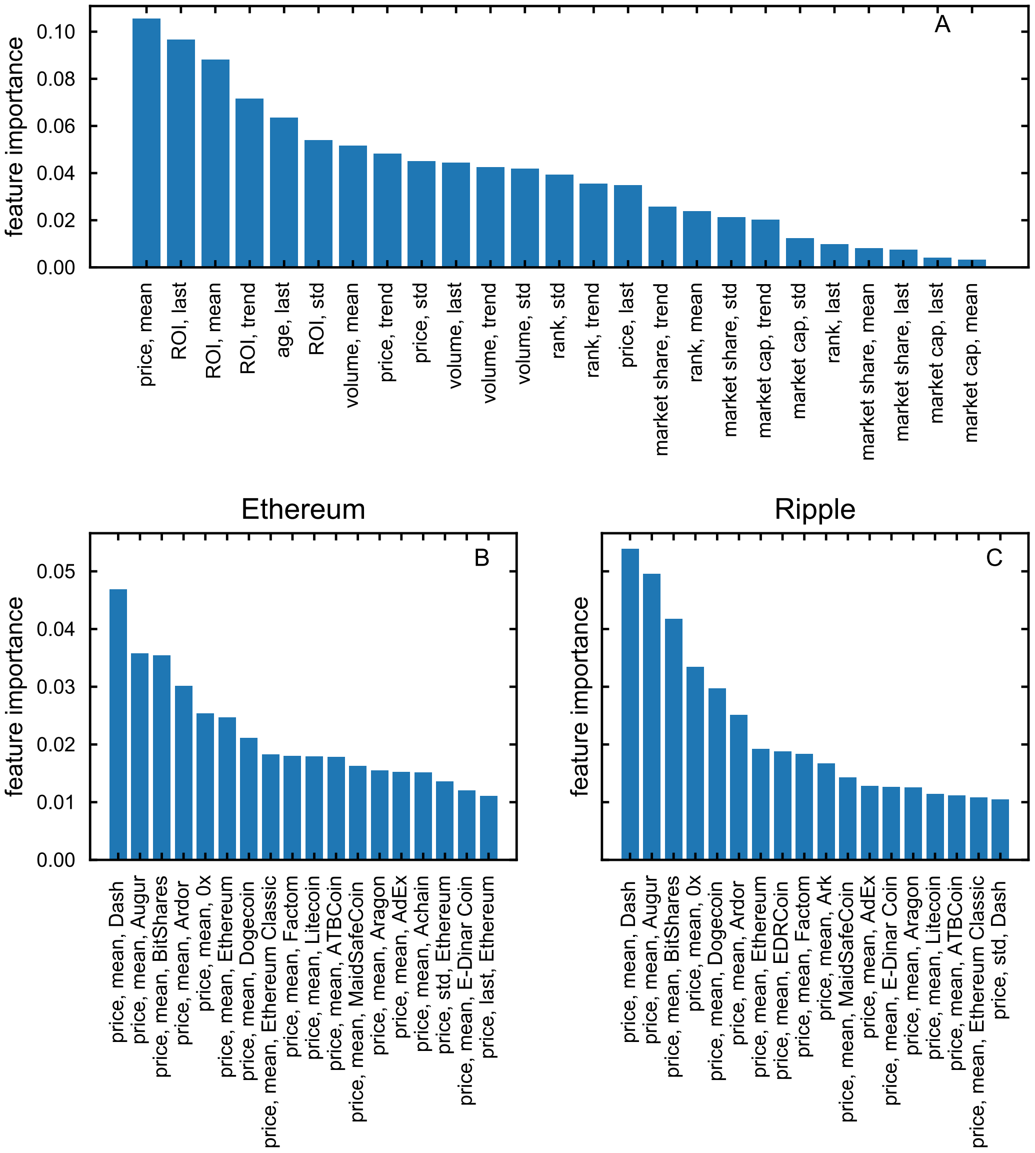


FIGURE 8.7: **Feature importance for Methods 1 and 2.** (A) The average importance of each feature for the XGBoost regression model of Method 1. Results are shown for $w = 7$ and $W_{training} = 10$. (B,C) Examples of average feature importance for the XGBoost regression model developed in Method 2. Results are shown for $w = 3$, $W_{training} = 10$, for Ethereum (B) and Ripple (C). For visualization purposes, we show only the top features.

9 Conclusion

This thesis was stimulated by the recent developments in cryptocurrencies. Compared to previous approaches, our contribution can be summarised in four main points. Firstly, we expanded the analysis of the cryptocurrency market from a few limited cryptocurrencies to more than 2000 cryptocurrencies taking into consideration the dynamics of the entire market. Secondly, we adopted a complex systems approach to studying the cryptocurrencies ecosystem, which aids a better understanding of the interplay between cryptocurrencies economic, technological and social aspects. Thirdly, we analysed several dimensions of the cryptocurrencies ecosystem, namely, the market dynamics, social attention and the blockchain transactions. Finally, our analysis spans different periods of significant changes in the market and social attention toward cryptocurrencies.

Our contribution took into consideration a vibrant decentralised ecosystem of exchanges, data aggregators and dark markets which we discussed in detail in Chapter 2. Our results were based on an analysis of three novel datasets that reflect the different layers of the cryptocurrencies ecosystem, and whose collection and processing we describe in Chapter 3.

Our research addressed three questions that had attracted growing interest in the scientific community, the following are the main findings of our analysis.

Cryptocurrency market exhibit stable statistical properties despite tumultuous growth and fluctuations and the simple neutral model of evolution can capture the market dynamics.

In Chapter 4, we investigated cryptocurrencies competition dynamics and showed that, while new cryptocurrencies appear and disappear continuously and their market capitalisation exhibits fluctuations, several statistical properties of the market have been stable for years. These include the number of active cryptocurrencies, market share distribution and the turnover of cryptocurrencies. Adopting an ecological perspective, we show that the neutral model of evolution can reproduce several key empirical observations, despite its simplicity and the assumption of no selective advantage of one cryptocurrency over another. These results hints at the limited effect of technological differences better cryptocurrencies in the investment decision. Our results relied on data set considering the history of the entire market and analyse the behaviour of 2000 cryptocurrencies introduced between April 2013 and May 2019.

Activity on cryptocurrencies' Wikipedia page is correlated to their overall market performance while a small tightly connected community is responsible for the information on these pages.

We tackled the question of how is social attention influence cryptocurrency market in Chapter 5. We considered the entire edit history of currency-related pages and their views history from July 2015. First, we quantified the evolution of cryptocurrency presence in Wikipedia by analysing the editorial activity and the network of co-edited pages. We found that a small community of tightly connected editors are responsible for most of the production of information about cryptocurrencies in Wikipedia. Then, we showed that a simple trading strategy informed by Wikipedia views, performs better than baseline strategies, in terms of returns on investment, for most of the covered period, although the "buy and hold strategy" dominates during the periods of explosive market expansion.

Dark markets are resilient to multiple closures through users migration to other coexisting markets.

In Chapter 6, we showed an analysis of the dark markets' Bitcoin transactions. We investigated the dynamics of 31 dark markets before and after dark market shutdowns. First, we showed that users migrate quickly to other dark markets following shutdowns. Second, we described the characteristics of migrant users. Finally, we studied how migrant users coordinate to move to a coexisting dark market following shutdowns.

In Chapter 7, We also analysed a selected number of dark markets drug sales and investigated the ability to predict the sales using Wikipedia data. Relying on 4 dark markets web scrapes, we showed that Wikipedia data enable better prediction of drug sales with mean error average of 43%. The results are consistent, considering different drug types and different countries.

Finally, in Chapter 8, we presented a machine learning approach to predict cryptocurrencies prices. Our work investigates more than 1600 cryptocurrencies, providing the most comprehensive price prediction study of cryptocurrencies. We showed that simple trading strategies assisted by state-of-the-art machine learning algorithms outperform standard benchmarks.

Present and future work will move in several directions. We list here briefly the topics we are currently addressing, or we plan to investigate in the near future.

Which factors decide cryptocurrencies survival in the market?

In Chapter 4, we showed that cryptocurrency chances to be invested in is proportional to its market share. However, before a cryptocurrency appears on an exchange and thus appear in the market data an initial coin offering (ICO) occurs. In an ICO, cryptocurrency's creators announce the currency details and aim to acquire initial fund that is typically collected through crowdfunding. A possible direction of research is to investigate the characteristics of successful ICO and whether this success continues to the market.

How other sources of information and social interactions shape cryptocurrencies prices?

Our analysis in Chapter 5 focused on Wikipedia as a source of information accessible by a general audience and maintained by a community. Our analysis showed that core crypto enthusiasts users edit cryptocurrencies pages. Other sources, such as source targeting experts on cryptocurrencies might exhibit different characteristics and different group dynamics. Already the work in [140] showed that there are two different groups of users and discussion on Bitcointalk forum. A systematic analysis and modelling of the groups' dynamics, interests and discussion across cryptocurrencies could reveal the influence of different groups on cryptocurrency market. It can also lead us to an understanding of how closed is the cryptocurrencies community.

What is the influence of developers on the financial side of cryptocurrencies?

While Bitcoin was promised to be fully decentralised, research in [38] showed that Bitcoin developers (11 of them) have full control on the cryptocurrency protocol rules. Although anyone can submit a suggestion or recommendation to alter the system, developers and miners must vote and accept these proposals. Similar mechanisms are adopted in other cryptocurrencies; however, the area is understudied. Understanding the impact of the developers' additions, alters and votes on the cryptocurrencies survival is crucial since they are the gatekeepers of the system.

How do other cryptocurrencies' transaction networks compare to the Bitcoin transaction network?

Early in 2013, a complex network approach was adopted to characterise Bitcoin blockchain. More work followed revealing the ability to identify influential users using page rank centrality, understand wealth distribution and even predict Bitcoin's price. A comparison between Bitcoin and Bitcoin cash network showed that for both networks fitness-based model describes the global structure. Extending the comparisons to other networks can aid characterising the actual usage of these currencies and add a differentiating dimension behind their technical differences.

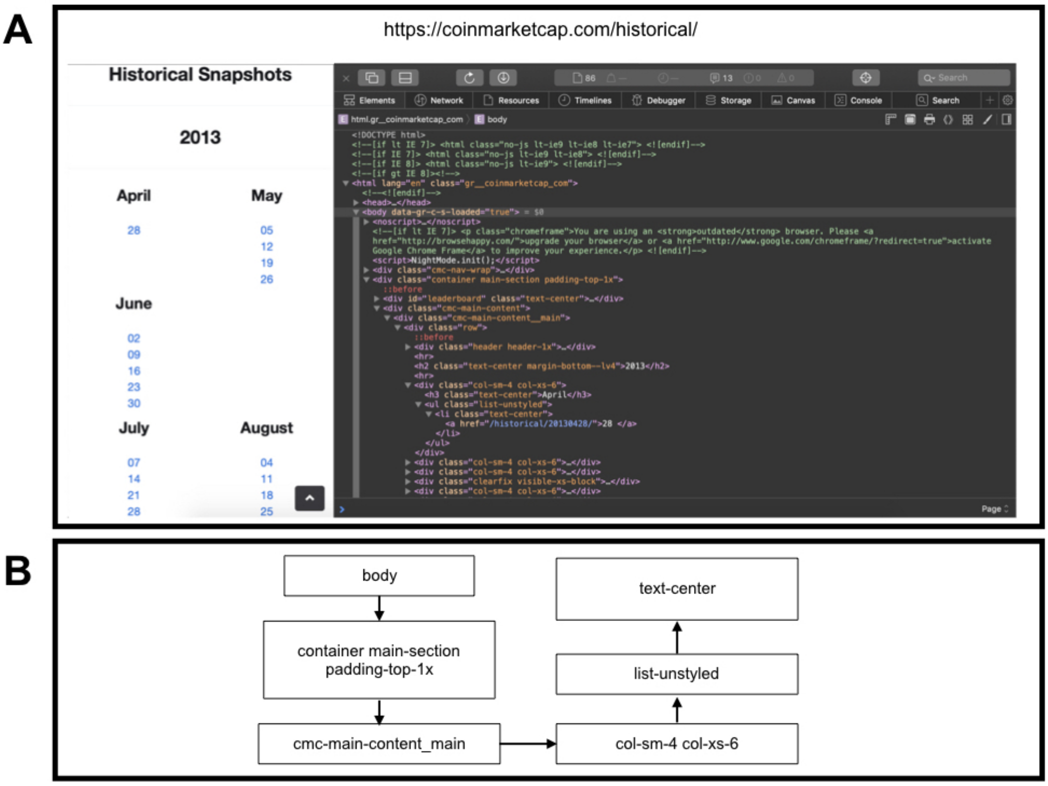
How is the activity in dark markets network different from other illicit transactions or services?

Our analysis in Chapter 6, focused on dark markets Bitcoin's transactions evolution. Our work showed that migrant users are active compared to non-migrants, and in general, the activity distribution is homogeneous. A natural question arising from our analysis is how different addresses in these network from other addresses engaged in other illicit transactions such as money laundering or non-illicit on such as trading. Finding such patterns can provide an additional method for identifying illicit transactions.

A Appendix to chapter 3

A.1 Weekly data collection

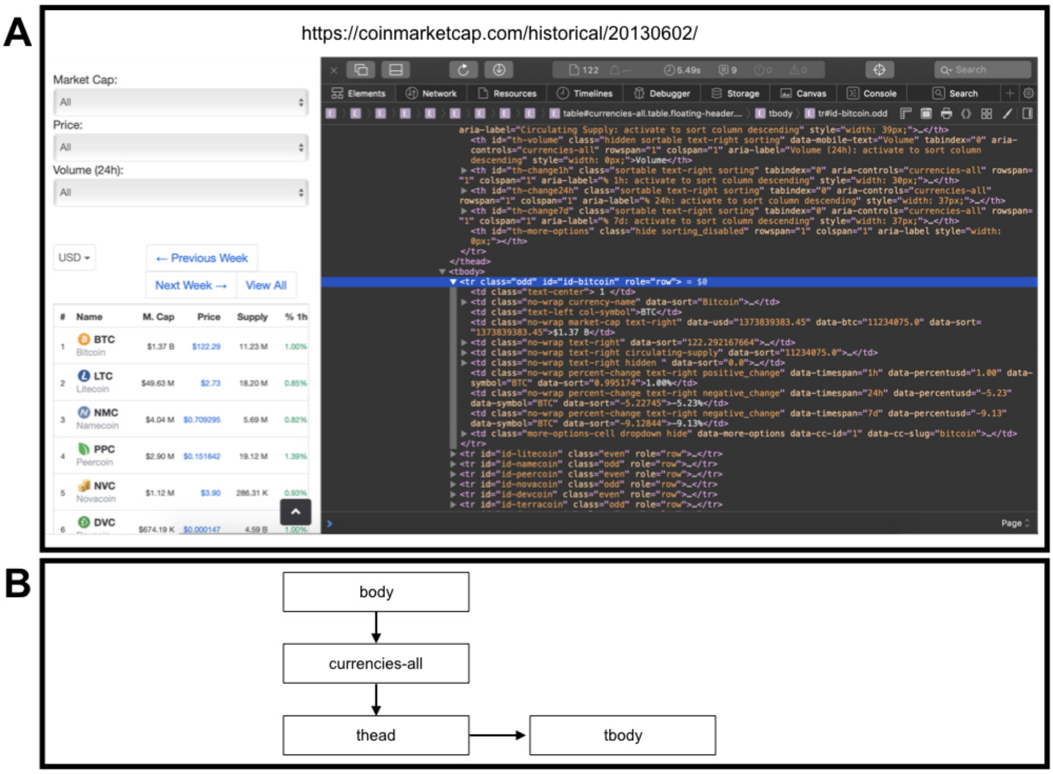
The first market data set we rely on for our analysis was the weekly data set scraped from “coinmarketcap.com” [4]. The scraping process follows two steps. Firstly we access the historical snapshots page as shown in Figure. A.1A. Using web inspection (a browser tool) we can see the HTML hierarchy of the page shown in Figure. A.1A on the right. From this structure the web crawler accesses the hierarchy shown in Figure. A.1B and extracts the list of snapshots recorded by the website. Through this step we gather a list of the available historical snapshots provided by the website and their web address to access.



```
<!DOCTYPE html>
<!--[if lt IE 7]> <html class="no-js lt-ie9 lt-ie8 lt-ie7"> <![endif]-->
<!--[if IE 7]> <html class="no-js lt-ie9 lt-ie8"> <![endif]-->
<!--[if IE 8]> <html class="no-js lt-ie9"> <![endif]-->
<!--[if gt IE 8]><!-->
▼ <html lang="en" class="gr__coinmarketcap_com">
    <!--<![endif]-->
  ▶ <head>…</head>
  ▼ <body data-gr-c-s-loaded="true"> = $0
    ▶ <noscript>…</noscript>
      <!--[if lt IE 7]> <p class="chromeframe">You are using an <strong>outdated</strong> browser. Please <a
      href="http://browsehappy.com/">upgrade your browser</a> or <a href="http://www.google.com/chromeframe/?redirect=true">activate
      Google Chrome Frame</a> to improve your experience.</p> <![endif]-->
      <script>NightMode.init();</script>
    ▶ <div class="cmc-nav-wrap">…</div>
    ▼ <div class="container main-section padding-top-1x">
        ::before
      ▶ <div id="leaderboard" class="text-center">…</div>
      ▼ <div class="cmc-main-content">
        ▼ <div class="cmc-main-content__main">
          ▼ <div class="row">
              ::before
            ▶ <div class="header header-1x">…</div>
              <hr>
              <h2 class="text-center margin-bottom--lv4">2013</h2>
              <hr>
            ▼ <div class="col-sm-4 col-xs-6">
                <h3 class="text-center">April</h3>
              ▼ <ul class="list-unstyled">
                ▼ <li class="text-center">
                    <a href="/historical/20130428/">28 </a>
                  </li>
                </ul>
              </div>
            ▶ <div class="col-sm-4 col-xs-6">…</div>
            ▶ <div class="col-sm-4 col-xs-6">…</div>
            ▶ <div class="clearfix visible-xs-block">…</div>
            ▶ <div class="col-sm-4 col-xs-6">…</div>
```

FIGURE A.1: **coinmarketcap weekly historical snapshot page.**(**A**) Left side shows a screen shot of the historical snapshot page on coinmarktecap.com, while right side shows the web inspection tool output showing the HTML hierarchy. (**B**) The HTML hierarchy of the page which we developed the web crawler to navigate. The data is embedded in the HTML structure of the page.

The second step is accessing the web address of each snapshot in our list. Figure. A.2A shows the first week of the historical snapshots and its web inspection. The web crawler later traverses the HTML hierarchy shown in Figure. A.2B and extracts the data table.



https://coinmarketcap.com/historical/20130602/

Market Cap: All
Price: All
Volume (24h): All

USD
← Previous Week
Next Week →
View All

#	Name	M. Cap	Price	Supply	% 1h
1	BTC Bitcoin	$1.37 B	$122.29	11.23 M	1.00%
2	LTC Litecoin	$49.63 M	$2.73	18.20 M	0.85%
3	NMC Namecoin	$4.04 M	$0.709295	5.69 M	0.82%
4	PPC Peercoin	$2.90 M	$0.151642	19.12 M	1.39%
5	NVC Novacoin	$1.12 M	$3.90	286.31 K	0.93%
6	DVC	$674.19 K	$0.000147	4.59 B	1.00%

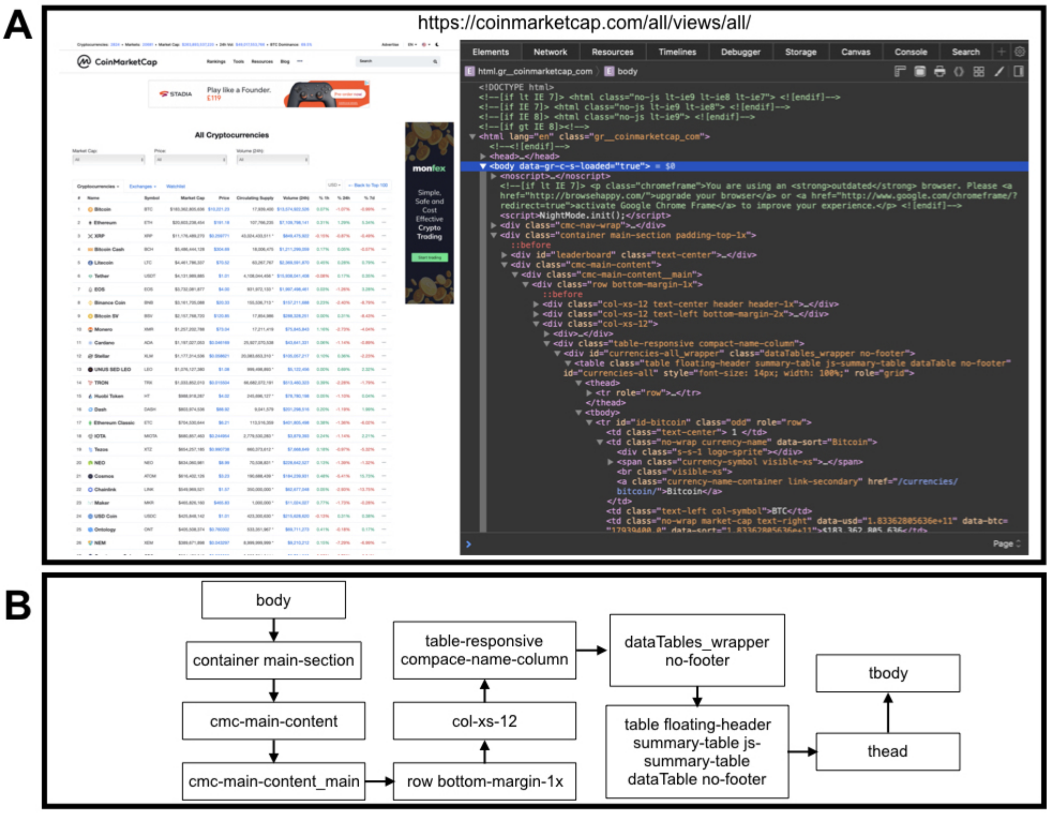
table#currencies-all.table.floating-header... 〉 tbody 〉 tr#id-bitcoin.odd

```
                aria-label="Circulating Supply: activate to sort column descending" style="width: 39px;">…</th>
                  <th id="th-volume" class="hidden sortable text-right sorting" data-mobile-text="Volume" tabindex="0" aria-
                  controls="currencies-all" rowspan="1" colspan="1" aria-label="Volume (24h): activate to sort column
                  descending" style="width: 0px;">Volume</th>
                ▶ <th id="th-change1h" class="sortable text-right sorting" tabindex="0" aria-controls="currencies-all" rowspan=
                "1" colspan="1" aria-label="% 1h: activate to sort column descending" style="width: 30px;">…</th>
                ▶ <th id="th-change24h" class="sortable text-right sorting" tabindex="0" aria-controls="currencies-all"
                rowspan="1" colspan="1" aria-label="% 24h: activate to sort column descending" style="width: 37px;">…</th>
                ▶ <th id="th-change7d" class="sortable text-right sorting" tabindex="0" aria-controls="currencies-all" rowspan=
                "1" colspan="1" aria-label="% 7d: activate to sort column descending" style="width: 37px;">…</th>
                  <th id="th-more-options" class="hide sorting_disabled" rowspan="1" colspan="1" aria-label style="width:
                  0px;"></th>
                </tr>
              </thead>
            ▼ <tbody>
              ▼ <tr class="odd" id="id-bitcoin" role="row"> = $0
                  <td class="text-center"> 1 </td>
                ▶ <td class="no-wrap currency-name" data-sort="Bitcoin">…</td>
                  <td class="text-left col-symbol">BTC</td>
                  <td class="no-wrap market-cap text-right" data-usd="1373839383.45" data-btc="11234075.0" data-sort=
                  "1373839383.45">$1.37 B</td>
                ▶ <td class="no-wrap text-right" data-sort="122.292167664">…</td>
                ▶ <td class="no-wrap text-right circulating-supply" data-sort="11234075.0">…</td>
                ▶ <td class="no-wrap text-right hidden " data-sort="0.0">…</td>
                  <td class="no-wrap percent-change text-right positive_change" data-timespan="1h" data-percentusd="1.00" data-
                  symbol="BTC" data-sort="0.995174">1.00%</td>
                  <td class="no-wrap percent-change text-right negative_change" data-timespan="24h" data-percentusd="-5.23"
                  data-symbol="BTC" data-sort="-5.22745">-5.23%</td>
                  <td class="no-wrap percent-change text-right negative_change" data-timespan="7d" data-percentusd="-9.13"
                  data-symbol="BTC" data-sort="-9.12844">-9.13%</td>
                ▶ <td class="more-options-cell dropdown hide" data-more-options data-cc-id="1" data-cc-slug="bitcoin">…</td>
                </tr>
              ▶ <tr id="id-litecoin" class="even" role="row">…</tr>
              ▶ <tr id="id-namecoin" class="odd" role="row">…</tr>
              ▶ <tr id="id-peercoin" class="even" role="row">…</tr>
              ▶ <tr id="id-novacoin" class="odd" role="row">…</tr>
              ▶ <tr id="id-devcoin" class="even" role="row">…</tr>
              ▶ <tr id="id-terracoin" class="odd" role="row">…</tr>
```

FIGURE A.2: **coinmarketcap first week.** (**A**) Left side shows a screen shot of the first week historical data provided by conimarketcap.com while right side shows the web inspection of the page. (**B**) The HTML hierarchy of the page which we developed the web crawler to traverse. The data is embedded in the HTML structure of the page.

A.2 Daily data collection

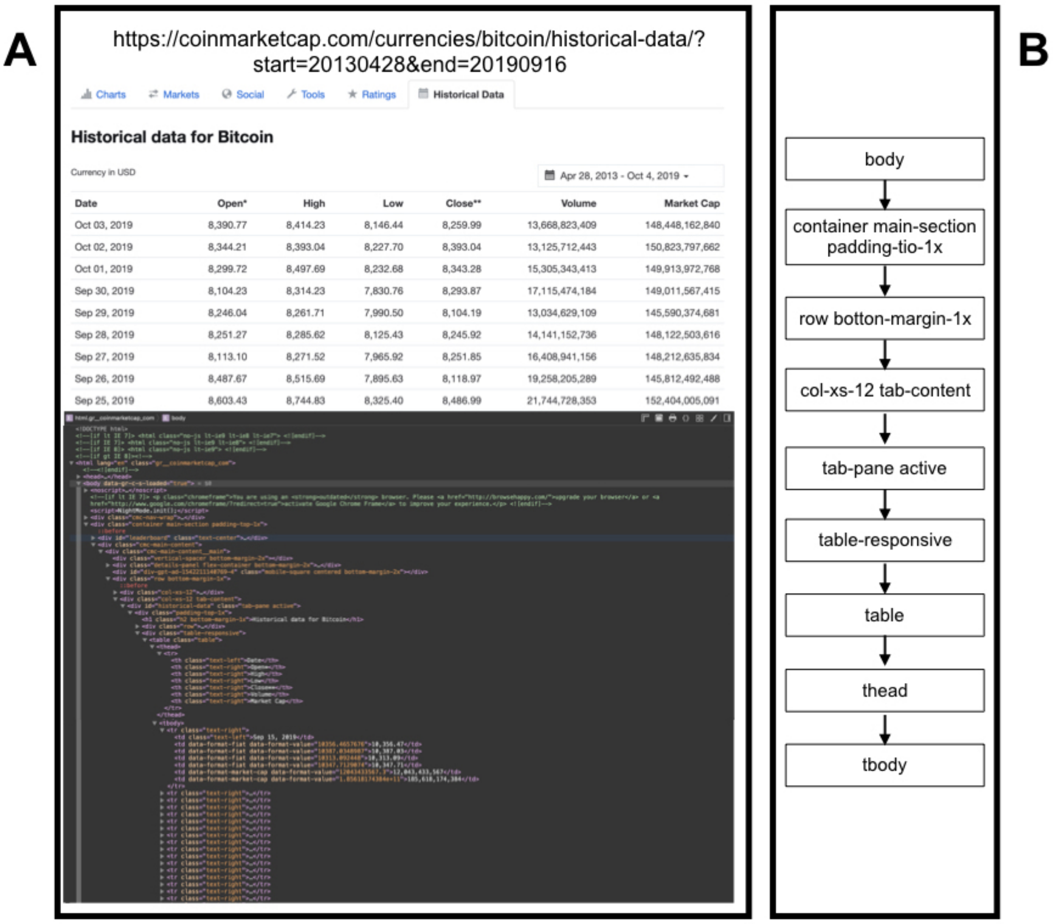
The second data set we rely on throughout the thesis is the daily dataset scraped from coinmarketcap.com. To scrape this data set we scrape all cryptocurrencies listed on the coinmarketcap landing page (Figure. A.3A) and traverse the HTML hierarchy shown in Figure. A.3B. Through this step we construct a list of all active cryptocurrencies listed and their web address on the website.



```
<!DOCTYPE html>
<!--[if lt IE 7]> <html class="no-js lt-ie9 lt-ie8 lt-ie7"> <![endif]-->
<!--[if IE 7]> <html class="no-js lt-ie9 lt-ie8"> <![endif]-->
<!--[if IE 8]> <html class="no-js lt-ie9"> <![endif]-->
<!--[if gt IE 8]><!-->
▼ <html lang="en" class="gr__coinmarketcap_com">
    <!--<![endif]-->
  ▶ <head>…</head>
  ▼ <body data-gr-c-s-loaded="true"> = $0
    ▶ <noscript>…</noscript>
      <!--[if lt IE 7]> <p class="chromeframe">You are using an <strong>outdated</strong> browser. Please <a
      href="http://browsehappy.com/">upgrade your browser</a> or <a href="http://www.google.com/chromeframe/?
      redirect=true">activate Google Chrome Frame</a> to improve your experience.</p> <![endif]-->
      <script>NightMode.init();</script>
    ▶ <div class="cmc-nav-wrap">…</div>
    ▼ <div class="container main-section padding-top-1x">
        ::before
      ▶ <div id="leaderboard" class="text-center">…</div>
      ▼ <div class="cmc-main-content">
        ▼ <div class="cmc-main-content__main">
          ▼ <div class="row bottom-margin-1x">
              ::before
            ▶ <div class="col-xs-12 text-center header header-1x">…</div>
            ▶ <div class="col-xs-12 text-left bottom-margin-2x">…</div>
            ▼ <div class="col-xs-12">
              ▶ <div>…</div>
              ▼ <div class="table-responsive compact-name-column">
                ▼ <div id="currencies-all_wrapper" class="dataTables_wrapper no-footer">
                  ▼ <table class="table floating-header summary-table js-summary-table dataTable no-footer"
                    id="currencies-all" style="font-size: 14px; width: 100%;" role="grid">
                    ▼ <thead>
                      ▶ <tr role="row">…</tr>
                      </thead>
                    ▼ <tbody>
                      ▼ <tr id="id-bitcoin" class="odd" role="row">
                          <td class="text-center"> 1 </td>
                        ▼ <td class="no-wrap currency-name" data-sort="Bitcoin">
                            <div class="s-s-1 logo-sprite"></div>
                          ▶ <span class="currency-symbol visible-xs">…</span>
                            <br class="visible-xs">
                            <a class="currency-name-container link-secondary" href="/currencies/
                            bitcoin/">Bitcoin</a>
                          </td>
                          <td class="text-left col-symbol">BTC</td>
                          <td class="no-wrap market-cap text-right" data-usd="1.83362805636e+11" data-btc=
                          "17939400.0" data-sort="1.83362805636e+11">$183,362,805,636</td>
```

FIGURE A.3: **coinmarketcap landing page.** (**A**) Left side shows a screen shot of the landing page of the conimarketcap.com while right side shows the web inspection of the page. (**B**) The html hierarchy of the page which we developed the web crawler to traverse. The data is embedded in the HTML structure of the page.

Later on, we go through one cryptocurrency at a time and access the web address of each cryptocurrency, specifically the historical data section. Figure. A.4A shows an example of a cryptocurrency historical data section, specifically for Bitcoin. Figure. A.4B shows the hierarchy of the HTML page which the web crawler was designed to scrape.


A
https://coinmarketcap.com/currencies/bitcoin/historical-data/?start=20130428&end=20190916
Charts
Markets
Social
Tools
Ratings
Historical Data
Apr 28, 2013 - Oct 4, 2019
html.gr__coinmarketcap_com
body
B
body
container main-section padding-tio-1x
row botton-margin-1x
col-xs-12 tab-content
tab-pane active
table-responsive
table
thead
tbody

Historical data for Bitcoin

Currency in USD

Date	Open*	High	Low	Close**	Volume	Market Cap
Oct 03, 2019	8,390.77	8,414.23	8,146.44	8,259.99	13,668,823,409	148,448,162,840
Oct 02, 2019	8,344.21	8,393.04	8,227.70	8,393.04	13,125,712,443	150,823,797,662
Oct 01, 2019	8,299.72	8,497.69	8,232.68	8,343.28	15,305,343,413	149,913,972,768
Sep 30, 2019	8,104.23	8,314.23	7,830.76	8,293.87	17,115,474,184	149,011,567,415
Sep 29, 2019	8,246.04	8,261.71	7,990.50	8,104.19	13,034,629,109	145,590,374,681
Sep 28, 2019	8,251.27	8,285.62	8,125.43	8,245.92	14,141,152,736	148,122,503,616
Sep 27, 2019	8,113.10	8,271.52	7,965.92	8,251.85	16,408,941,156	148,212,635,834
Sep 26, 2019	8,487.67	8,515.69	7,895.63	8,118.97	19,258,205,289	145,812,492,488
Sep 25, 2019	8,603.43	8,744.83	8,325.40	8,486.99	21,744,728,353	152,404,005,091

```
<!DOCTYPE html>
<!--[if lt IE 7]> <html class="no-js lt-ie9 lt-ie8 lt-ie7"> <![endif]-->
<!--[if IE 7]> <html class="no-js lt-ie9 lt-ie8"> <![endif]-->
<!--[if IE 8]> <html class="no-js lt-ie9"> <![endif]-->
<!--[if gt IE 8]><!-->
▼<html lang="en" class="gr__coinmarketcap_com">
   <!--<![endif]-->
 ▶<head>…</head>
 ▼<body data-gr-c-s-loaded="true"> == $0
   ▶<noscript>…</noscript>
    <!--[if lt IE 7]> <p class="chromeframe">You are using an <strong>outdated</strong> browser. Please <a href="http://browsehappy.com/">upgrade your browser</a> or <a
    href="http://www.google.com/chromeframe/?redirect=true">activate Google Chrome Frame</a> to improve your experience.</p> <![endif]-->
    <script>NightMode.init();</script>
   ▶<div class="cmc-nav-wrap">…</div>
   ▼<div class="container main-section padding-top-1x">
      ::before
     ▶<div id="leaderboard" class="text-center">…</div>
     ▼<div class="cmc-main-content">
       ▼<div class="cmc-main-content__main">
          <div class="vertical-spacer bottom-margin-2x"></div>
         ▶<div class="details-panel flex-container bottom-margin-2x">…</div>
          <div id="div-gpt-ad-1542211140769-4" class="mobile-square centered bottom-margin-2x"></div>
         ▼<div class="row bottom-margin-1x">
            ::before
           ▶<div class="col-xs-12">…</div>
           ▼<div class="col-xs-12 tab-content">
             ▼<div id="historical-data" class="tab-pane active">
               ▼<div class="padding-top-1x">
                  <h1 class="h2 bottom-margin-1x">Historical data for Bitcoin</h1>
                 ▶<div class="row">…</div>
                 ▼<div class="table-responsive">
                   ▼<table class="table">
                     ▼<thead>
                       ▼<tr>
                          <th class="text-left">Date</th>
                          <th class="text-right">Open*</th>
                          <th class="text-right">High</th>
                          <th class="text-right">Low</th>
                          <th class="text-right">Close**</th>
                          <th class="text-right">Volume</th>
                          <th class="text-right">Market Cap</th>
                        </tr>
                      </thead>
                     ▼<tbody>
                       ▼<tr class="text-right">
                          <td class="text-left">Sep 15, 2019</td>
                          <td data-format-fiat data-format-value="10356.4657676">10,356.47</td>
                          <td data-format-fiat data-format-value="10387.0340987">10,387.03</td>
                          <td data-format-fiat data-format-value="10313.092448">10,313.09</td>
                          <td data-format-fiat data-format-value="10347.7129074">10,347.71</td>
                          <td data-format-market-cap data-format-value="12043433567.3">12,043,433,567</td>
                          <td data-format-market-cap data-format-value="1.85618174384e+11">185,618,174,384</td>
                        </tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
                       ▶<tr class="text-right">…</tr>
```

FIGURE A.4: **coinmarketcap Bitcoin historical data page.** (**A**) Top side shows a screen shot of Bitcoin historical data page section on conimarketcap.com while bottom side shows the web inspection of the page. (**B**) The HTML hierarchy of the page which we developed the web crawler to traverse. The data is embedded in the HTML structure of the page.

A.3 Dark markets data

Table. A.1 shows a list of the dark markets covered in our dataset. The table show the market name, the dates of operation, closure reason if applicable, and the type of products typically sold on the market.

TABLE A.1: **Dark markets information.** Information on the 74 dark markets in our data set. For each dark market, the table states the name of the market, the start and end dates of its operation, the closure reason if applicable and the type of products sold by the market. "drugs" indicates that the primary products sold on the market are drugs while "credits" indicates the market specialty is fake IDs and credit cards and "mixed" indicates the market sells both types of products. "NA" indicates that the information in not available.

Name	Start date	End date	Closure reason	Sales
Abraxas Market	2014 − 12 − 13	2015 − 11 − 05	scam	drugs
Acropolis Market	2016 − 03 − 27	2017 − 07 − 01	voluntary	mixed
Agora Market	2013 − 12 − 03	2015 − 08 − 26	voluntary	mixed
AlphaBay Market	2014 − 12 − 22	2017 − 07 − 05	raided	mixed
Apollon Market	2018 − 05 − 03	active	active	drugs
Apoteksboden.com	2013 − 02 − 28	NA	active	drugs
Aviato Market	NA	NA	unclear	drugs
Babylon Market	2014 − 07 − 11	2015 − 07 − 31	raided	drugs
Berlusconi Market	2018 − 08 − 12	active	active	mixed
Bilzerian24.net	2017 − 11 − 13	active	active	credits
Black Bank Market	2014 − 02 − 05	2015 − 05 − 18	scam	mixed
BlackMart	NA	NA	active	mixed
BlackPass.name	NA	NA	active	credits
Blue Sky Marketplace	2013 − 12 − 03	2014 − 11 − 05	raided	drugs

Table A.1 – *Continued from previous page*

Name	Start date	End date	Closure reason	Sales
Bo-Bulk.cc	2017 − 07 − 07	NA	active	credits
Carder′s Paradize	NA	NA	active	credits
Cocaine Market	NA	NA	active	drugs
Core Market	Late 2018	NA	unclear	mixed
Darknet Heroes League Market	Late 2015	NA	active drugs	
DeDope Market	NA	NA	active	drugs
Doctor D Market	NA	NA	unclear	drugs
Dream Market	2016 − 03 − 19	2019 − 04 − 30	voluntary	mixed
Drug Market	NA	NA	active	drugs
DutchDrugz Market	2014	NA	active	drugs
East India Company Market	2015 − 04 − 28	2016 − 01 − 01	scam	drugs
Empire Market	2018 − 02 − 01	active	active	mixed
Entershop	NA	NA	active	drugs
Eviano Luxury Weed Market	NA	NA	active	drugs
Evolution Market	2014 − 01 − 14	2015 − 03 − 14	scam	drugs

Table A.1 – *Continued from previous page*

Name	Start date	End date	Closure reason	Sales
Flugsvamp Market	NA	November 2014	raided	drugs
French Connection Market	NA	NA	active	drugs
German Plaza Market	2015 − 05 − 22	2016 − 05 − 01	scam	mixed
GetSome.pw	2012 − 05 − 12	active	active	credits
Green Road Market	NA	active	active	drugs
Hansa Market	2014 − 03 − 09	2017 − 07 − 20	raided	drugs
House of Lions Market	2016 − 05 − 23	2017 − 07 − 12	raided	drugs
Hydra Marketplace	2015 − 11 − 25	active	active	mixed
Isellz.cc	Late 2011/Early 2012	active	active	credits
JokerStash Market	NA	active	active	credits
JustBuy.su	NA	NA	unclear	credits
LuxSocks.ru	2017 − 01 − 12	active	active	credits
McDuck.top	2016 − 10 − 22	NA	unclear	credits
Megapack Market	NA	NA	unclear	drugs
Middle Earth Marketplace	2014 − 06 − 22	2015 − 11 − 04	scam	mixed
Midland City Market	2016	active	active	mixed
MrGreen.ws	early 2014	active	active	credits

Table A.1 – *Continued from previous page*

Name	Start date	End date	Closure reason	Sales
MyFakeID.biz	2011 − 05 − 28	active	active	credits
N1Shop.cc	2014 − 10 − 19	active	active	credits
Nucleus Market	2014 − 10 − 24	2016 − 04 − 13	scam	mixed
Oasis Market	2015 − 12 − 20	2016 − 10 − 01	scam	mixed (mostly drugs)
Olympus Market Olympus Market	2018 − 04 − 20	2018 − 09 − 04	scam	mixed
Oxygen Market	2015 − 04 − 16	2015 − 08 − 27	scam	drugs
PP24.ws	2014 − 11 − 10	active	active	credits
Pandora OpenMarket	2013 − 10 − 20	2014 − 11 − 05	raided	drugs
Point Market (former Tochka)	early 2015	active	active	mixed
Rescator.at	NA	active	active	credits
Russian Anonymous Marketplace	2014 − 08 − 29	2017 − 09 − 21	raided	mixed
Russian Silk Road Market	NA	NA	unclear	mixed
San-Wells Market	NA	active	active	credits
Sheep Marketplace	2013 − 02 − 28	2013 − 11 − 29	scam	drugs

Table A.1 – *Continued from previous page*

Name	Start date	End date	Closure reason	Sales
Silk Road 2 Market	2013 − 11 − 06	2014 − 11 − 05	raided	mixed
Silk Road 3.1	2018 − 01 − 21	active	active	drugs
Silk Road Marketplace	2011 − 01 − 31	2013 − 10 − 02	raided	mixed
SlilPP Market	NA	active	active	credits
To You Team Market	2012	active	active	drugs
TradeRoute Market	2016 − 11 − 06	2017 − 10 − 12	scam	mixed
UAS-Shop.ru	2017 − 02 − 28	active	active	credits
Unicc	NA	active	active	credits
Valhalla Market (Silkkitie)	2013 − 10 − 20	2019 − 05 − 03	raided	drugs
ValidPins Market	NA	NA	NA	credits
Vendetta.cc	2019 − 03 − 17	active	active	credits
Wall Street Market	2016 − 09 − 09	2019 − 05 − 02	raided	mixed
Wholecelium.com	2009 − 12 − 10	active	active	drugs

B Appendix to chapter 4

B.1 some relevant cryptocurrencies

Table B.1 provides information on some relevant cryptocurrencies, either occupying high-rank positions or early introduced in the market. Data was collected in May 2017, see below for details on the Technology column.

TABLE B.1: **Details on the top runner cryptocurrencies in the market.** The table is generated using data collected on May 28, 2013

Name	Year	Technology	Market Cap ($)	Rank	Additional Info
Bitcoin	2009	Proof-of-work	35B	1	
Ethereum	2015	Proof-of-work	15B	2	Smart contracts
Ripple	2013	Distributed open source consensus ledger	8B	3	Widely adopted by companies and banks.
NEM	2015	Proof-of-importance	1B	4	
Ethereum Classic	2015	Proof-of-work	1B	5	DAO Hard-fork
Litecoin	2011	Proof-of-work	1B	6	
Dash	2014	Proof-of-work	809M	7	Gained market since early 2017. Privacy focused.
Monero	2014	Proof-of-work	535M	8	Gained momentum in late 2016. Privacy focused
NameCoin	2015	Proof-of-work	21M	58	

B.2 Simulations

Our choice of the mutation parameter μ is informed by the data to yield a number of new cryptocurrencies per unit time corresponding to the empirical observation. By choosing $\mu = \frac{7}{N}$, where N is the population size in the model it holds that 1 model generation corresponds to 1 week of observation (since on average 7 new cryptocurrencies enter the system every week, see Sec. 4.4). In Fig. B.1 we show that the distribution of species sizes (see Fig. 4.6A) has a very similar shape for a broad range of choices of μ [187].

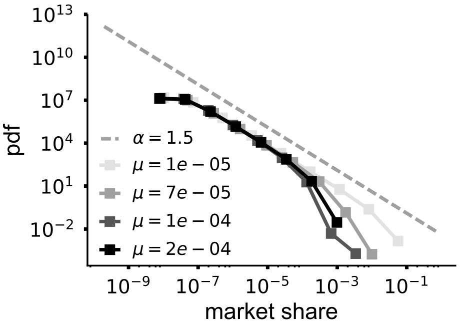


FIGURE B.1: **Distribution of species sizes for different values of μ.** Distribution of the species sizes resulted form numerical simulations given different values of μ.

All simulations are run starting with one species in order to capture the initial dominance of Bitcoin in the cryptocurrency market. This reflects the initial state of the cryptocurrencies market, when Bitcoin was the only existing cryptocurrency. Simulations are run using $N = 10^5$, implying that an individual in the model maps to $\sim$ \$100,000 (We verified that results do not depend on the choice of N, as long as N is large enough).

While in the neutral model a new species enters the system as a new individual, we further inform the model with the average size of a new cryptocurrency ($\sim$ \$1.5 million), corresponding to $m = 15$ individuals in the model when $N = 10^5$ as in our case. To consider the fact that new cryptocurrencies do not enter the market with exactly the same size, in our simulations, when a mutation occurs, the new species enters with a number m of individuals randomly extracted from the interval $[10, 20]$.

The exponent $\alpha = 1.5$ exhibited by the data and the simulations(see Fig.4.6A) are equilibrium properties of the neutral model, and hence obtained under a broad range of conditions (e.g., initial condition, time of start of measure and aggregation window) and robust to changes in the value of μ [187], Fig. B.1). Fig.4.6B and C are obtained starting from generation 104 and aggregating over 52 generations (i.e. performing the analysis over the single population obtained by aggregating the $N * 52$ individuals [188, 185]). Fig. B.2 shows the turnover profile (A) and average life time of a rank (B) when the measure is performed over 52 generations starting from different generations g_1 corresponding to the first year (measures start at generation $g_1 = 1$), second year (measures start at generation $g_1 = 53$), etc. It is clear that, with the exception of a high rank mobility characterizing the very first generations, the choice of g_1 has little effect on the curves produced by the model. Fig.4.6D is measured from generation 1 up to generation 210, corresponding to 4 years. Each point of the simulation curve corresponds to the instantaneous market share of the dominating cryptocurrency at that generation.

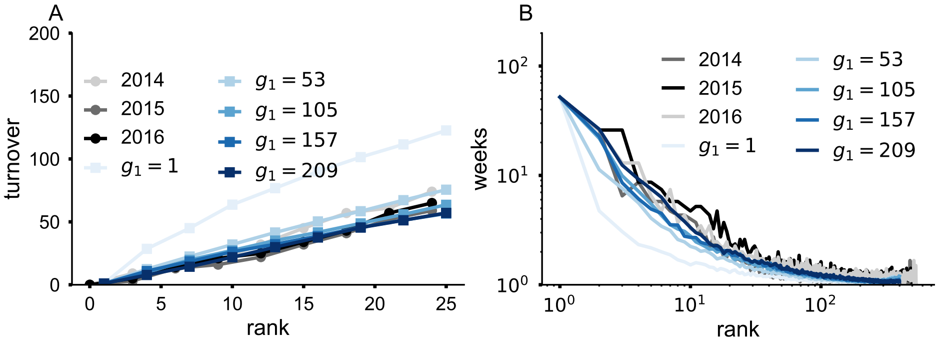


FIGURE B.2: **Neutral model ranks dynamics. (A)** Turnover profile computed considering 52 for the cryptocurrencies data (gray lines, dots) and for numerical simulations (blue lines). **(B)** The Average life time a cryptocurrency/species stays in a given rank computed considering 52 generations for the cryptocurrencies data (gray lines, dots) and for numerical simulations (blue lines). Simulation parameters are $\mu = 7/N$, $N = 10^5$ and 1 species in the initial state.

B.3 technologies, same distribution

In order to check whether technical differences leave any detectable fingerprint at the level of statistical distributions, we look at cryptocurrencies adopting one of the two main blockchain algorithms for reaching consensus on what block represents recent transactions across the network: Proof-of-work (PoW) or the Proof-of-stake (PoS) consensus algorithms.

The PoW scheme was introduced as part of Bitcoin in 2009 [1]. To generate new blocks, participating users work with computational and electrical resources in order to complete "proof-of-works", pieces of data that are difficult to produce but easy to verify. Block generation (also called "mining") is rewarded with coins. To limit the rate at which new blocks are generated, every 2016 blocks the difficulty of the computational tasks changes [286].

While the PoW mechanism is relatively simple, there are concerns regarding its security and sustainability. First, severe implications could arise from the dominance of mining pools controlling more than 50% of the computational resources and who could in principle manipulate the blockchain transactions. This scenario is far from being unrealistic: in 2014, one mining pool (Ghash.io) [287] controlled 42% of the Bitcoin mining power. Also, the energy consumption of PoW based blockchain technologies has raised environmental concerns: it is estimated that Bitcoin consumes about 12.76 TWh per year [288].

The PoS scheme was introduced as an alternative to PoW. In this system, mining power is not attributed based on computational resources but on the proportion of coins held. Hence, the richer users are more likely to generate the next block. Miners are rewarded with the transactions fees. While proof-of-work relies heavily on energy, proof-of-stake doesn't suffer from this issue. However, consensus is not guaranteed since miners sole interest is to increase their profit. Through the years both protocols have been altered to fix certain issues and continue to be improved.

Figure B.3 shows that the market shares of the two groups of cryptocurrencies follow the same behavior. The figure is generated using data collected from [289] and [4].

B.4 share and frequency-rank distributions for individual years

The power-law fit for the distribution of market share (Table B.2) and the frequency-rank distribution (Table B.3) are consistent with the theoretical predictions of the neutral model[178] also for individual years. Fits coefficient for the distribution of market share are computed using the methodology described in [177] (errors are obtained by bootstrapping 1000 times). Fit

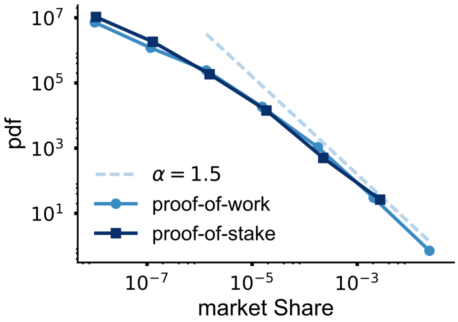


FIGURE B.3: **Distribution of market share.** Distribution of the market share for proof-of-work cryptocurrencies (light blue filled line) and distribution of market share of (proof-of-stake or hybrid) cryptocurrencies (dark blue filled line). The dashed line is power law curve with exponent $\alpha = 1.5$.

TABLE B.2: **Power-law fit coefficients of the market share distributions.**

Year	α
2013	1.37 ± 0.04
2014	1.54 ± 0.09
2015	1.62 ± 0.12
2016	1.59 ± 0.13
2017	1.60 ± 0.21
all years	1.58 ± 0.12

coefficients with errors for frequency-rank distributions are computed with the least-square method.

B.5 The progeny distribution exponent

Mathematically, the neutral model can be described by a process where each species take an independent, random walk based on a linear process. The following linear master equation describe that the probability $P(n|t)$ of a species has abundance n conditioned on its age.

TABLE B.3: **Power-law fit coefficients of the frequency-rank distributions.**

Year	β
2013	-1.98 ± 0.20
2014	-2.00 ± 0.13
2015	-1.83 ± 0.08
2016	-1.88 ± 0.08
2017	-1.86 ± 0.16
all years	-1.93 ± 0.23

$$\frac{dP}{dt} = b(n-1)P(n-1|t) - bnP(n|t) - dnP(n|t) + d(n+1)P(n+1|t)\,, \tag{B.1}$$

where d is the death rate and b is the birth rate. At the point of speciation process, the linear master equation will have a time-dependent solution as following,

$$P(n|t) = e^{(b-d)t}\frac{(b/(d-b))(1-e^{(b-d)t})^{n-1}}{(1+b/(d-b))(1-e^{(b-d)t})^{n+1}}\,. \tag{B.2}$$

By considering the assumption that in the steady state the rate of new species appearance will be equal to γJ. Where J is the fixed size population and γ is the per capita speciation rate which is equal to $\gamma = d - b$. It can be shown that the number of species with abundance k can be represented by a log series distribution

$$\langle S(k)\rangle \simeq \gamma J \int_0^\infty P(k|t)dt \simeq \frac{\theta}{k}(1-\frac{\theta}{J})^k, \tag{B.3}$$

where θ stand for "fundamental biodiversity number" equal to $\theta = (1 - b/d)J$.

Using the non zero sum approximation (NZS), and assuming that sufficient time has passed (T) that the progeny distribution reached stationarity, the progeny distribution can be described by,

$$q(k) = (-1)^{(k-1)} \binom{\frac{1}{2}}{k} \frac{2d}{b+d} \left(\frac{4bd}{(b+d)^2} \right)^{k-1}, \tag{B.4}$$

and the term $\binom{\frac{1}{2}}{k}$ can be defined by

$$\binom{\frac{1}{2}}{k} = \binom{2k}{k} \frac{-1^{k+1}}{2^{2k}(2k-1)}. \tag{B.5}$$

The distribution in equation B.4 is divided into two parts; the firs is a power law and the second is an exponential decay. For large enough value of k, the first term in equation B.4 can be approximated by

$$(-1)^{k-1} \binom{\frac{1}{2}}{k} = \frac{1}{\sqrt[2]{\pi} k^{3/2}}. \tag{B.6}$$

In summary, these results shows that the neutral progeny distribution tends towards a power law with an exponent of $-3/2$ not dependent on the neutral model parameters. However at $k = (b/v)^2$ the distribution exhibits an exponential cut-off at approximately. Figure. 4.5B shows the overall dynamics of the model.

C Appendix to chapter 5

C.1 Exchanges with margin trading support

Here, we provide data on the list of exchanges supporting margin trading. Margin trading is essential for our proposed investment strategy, since an investor can sell a cryptocurrencies which he does not own yet.

TABLE C.1: **List of exchanges supporting margin trading** The table is generated using data collected on January 23rd, 2019. It shows the names and webpage urls of the exchanges considered.

Name	Link
Bitmax	`https://www.bitmex.com`
Huobi	`https://www.huobi.co`
poloniex	`https://poloniex.com`
kraken	`https://www.kraken.com`
Bitfinex	`https://www.bitfinex.com`

C.2 Correlations between a cryptoccurrency success in the market and its Wikipedia attention

We show in the paper that the overall success of a cryptocurrency in the market is correlated to the attention it draws on Wikipedia (cryptocurrencies which have high price in the market, have high share in Wikipedia views and edits). In particular we show that the Spearman correlation between a cryptocurrency average share of page views and the market performance measured by its average market share (ρ_{vm}), average trading volume share

(ρ_{vv}) and average price (ρ_{vp}) across time is positive and consistent (see Figure C.1A). We show that the positive correlation between this quantities is consistent with time, with $0.65 \leq \rho_{vm} \leq 0.79$, $0.61 \leq \rho_{vv} \leq 0.83$, and $0.32 \leq \rho_{vp} \leq 0.51$.

In Figure C.1-B, we show the Spearman correlation between a cryptocurrecny average share of Wikipedia page edits and its market performance measured in average market share (ρ_{em}), average trading volume share (ρ_{ev}) and average price (ρ_{ep}) across time. We show that the positive correlation between this quantities is consistent with time, with $0.25 \leq \rho_{em} \leq 0.78$, $0.21 \leq \rho_{ev} \leq 0.79$, and $0.32 \leq \rho_{ep} \leq 0.66$. However the value of the correlation varies across the years which can be attributed to the variation in the number of pages created per year.

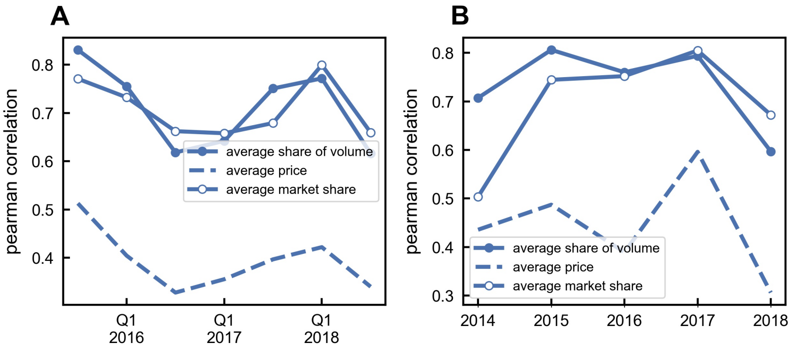


FIGURE C.1: **Persistency of the correlation between market properties and attention on Wikipedia.** (A) The Spearman correlation between average share of views and price (blue dashed line), volume (blue line with filled circles) and market share (blue line with white circles) across time. The correlation is computed over a window of 6 months. (B) The Spearman correlation between a cryptocurrency page average share of edits and price (blue dashed line), volume (blue line with filled circles) and market share (blue line with white circles) across time. The correlation is computed over a window of one year.

C.3 Literature review

Several studies have focused on Wikipedia pages and editors' activity. In Table C.2, we present a summary of their findings and a comparison with our results around cryptocurrencies Wikipedia pages.

TABLE C.2: **Comparison among our results and previous findings around Wikipedia pages and editors.** The table reports for each research paper: (1) Reference. (2) Focus of the article. (3) Key measurements. (4) Key findings relevant to our study. (5) Our findings around cryptocurrency pages in comparison to the previous findings.

Paper reference	Focus	Key measurements	Findings	Our findings
[153]	editors	Fraction of maintenance edits.	General increase in maintenance work, especially reverts.	Higher proportion of reverts. . No increasing trend in both reverts and vandalism (see main text, S2)
[214]	editors	Editors activity levels in relation to their life time	Highly active editors (Wikipedians) are active from two days after joining Wikpedia.	Similar findings for cryptocurrency pages (see main text, S3)
[213]	editors	Evolution of the contributions of editors given their activity levels.	Growth in the number of infrequent contributors and increase in their number of edits.	Infrequent editors have existed since . the beginning and their number of edits also increases (see main text, S3)
[211]	Medical related Wikipedia pages	Descriptive analysis of the general trends.	Decreasing number of editors	Increasing number of editors (see main text, S3

C.4 Robustness of the findings

The uneven distribution of edits across editors was show to be heterogeneous (see main text, S3). Here, we show that this result is consistent in time (see Figure C.2-A). We also test our results against saving mistakes by editors [214]. This often occurs when an editor mistakenly save an incomplete edit, producing multiple edits within a very short time. We solve this issue by excluding from the analysis edits that from the same editor on the same page, occurring within less than an hour from the prevopus one, as in [214]. In Figure C.2-B, we show that, our results are robust to this change.

We also study top editors contributions in all Wikipedia pages. For each editor with at least 100 edits in cryptocurrency pages, we collect data about the top 10 Wikipedia pages they contributed. This include pages outside the 38 cryptocurrency pages. For this task, we use a web tool [155], which provides the number of edits contributed by each editor to a given page. Figure C.3 shows that editors are mostly interested in cryptocurrencies and technology related pages. Compared to the set editors with more than 500 edits (see main text, Figure 10), the set of pages edited is more diverse.

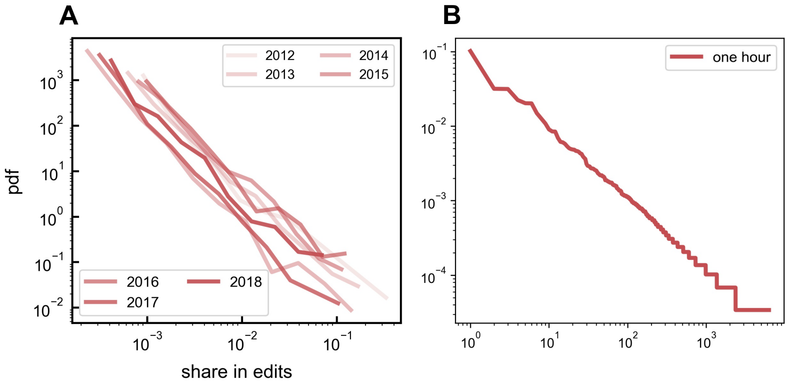


FIGURE C.2: **Users share of edits in different years** (A) The fraction of edits vs the rank *r* of an editor, computed over a year. Every line represents different year. (B) The fraction of edits vs the rank *r* of an editor, computed over the period between 2005 and 2018, after removing edits from the same editor on the same page, occurring within one hour from the previous.

C.5 The most active editor

Here, we provide information on the editor with the highest number of edits in cryptocurrency pages (10% of the edits). Table C.3 shows the editor general editing patterns in the entire English Wikipedia. Table C.4 shows the top pages edited by the top editor.

TABLE C.3: **Overall activity of cryptocurrency Wikipedia pages top editor.** The table reports for the editor with highest contribution in cryptocurrencies Wikipedia page ($\sim$ 10% of the edits): (1) number of pages edited, (2) total number of edits (English Wikipedia), (3) percentage of edits in cryptocurrency pages, (4) average number of edits per page, (5) date of the first edit. These data cover the editor activity across all pages in Wikipedia and it was collected on February 5th, 2019.

number of page	total number of edits	percentage of edits in cryptocurrency pages	average edits per page	date of first edit
442	9430	32%	2	2005 – 11 – 20

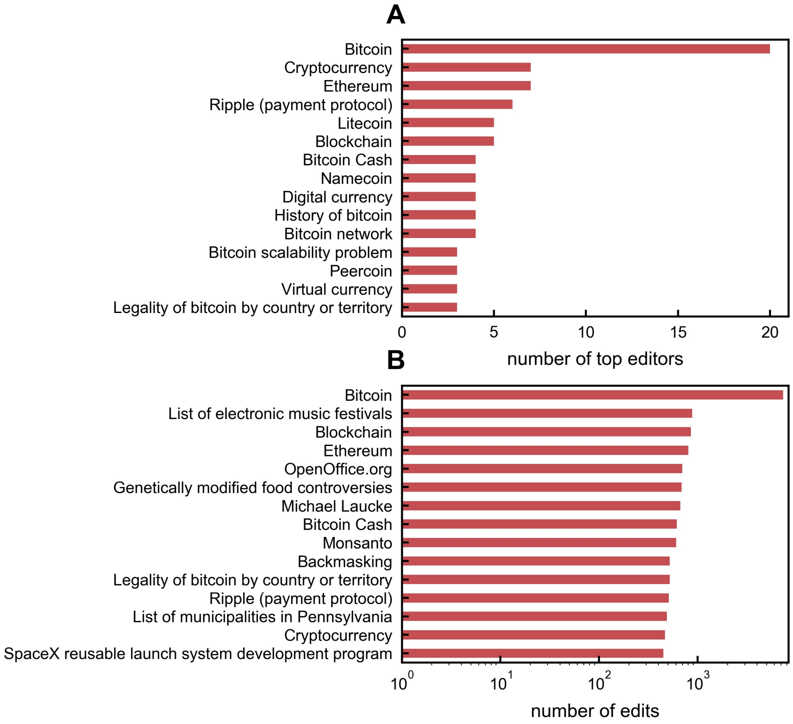


FIGURE C.3: **Activity of the top cryptocurrency pages editors.**(A) The top 15 pages by number of editors. The x-axis shows the number of top editors who had this page in their top edited pages. Note that here we consider only the top 10 pages per editor. (B) The top 15 pages by number of edits. The x-axis shows the total number of edits per page. Results are obtained for the subset of 23 most active editors.

TABLE C.4: **Top pages edited by the top editor** The top pages edited by the most active editor in cryptocurrency pages. The table shows the page name and number of edits. Data was collected on February 5th, 2019.

page name	number of edits
Bitcoin	2706
Blockchain	593
Legality of bitcoin by country or territory	467
Bitcoin Cash	349
Cryptocurrency	308
Rebol	209
Bitcoin scalability problem	187
History of bitcoin	177
Satoshi Nakamoto	109

C.6 New pages

Figure C.4 shows, for each of the years considered, the fraction of edits made to new pages and the fraction of editors contributing to new pages. On average, the $\sim$ 18% of editors contribute to the newly created pages within a given year, while only $\sim$ 10% of the edits are made to new pages.

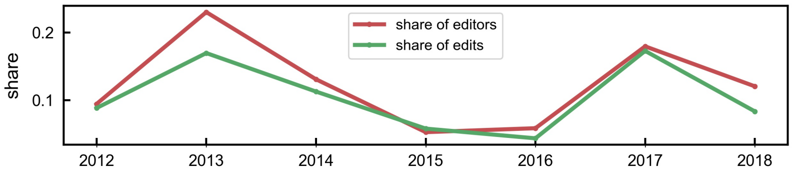


FIGURE C.4: **Editing activity on new pages.** Fraction of edits made to new pages (green solid line), and fraction of editors contributing to new pages (red solid line). Results are aggregated using a time window of one year.

C.7 Editing network

To characterize the co-editing activity in cryptocurrency Wikipedia pages, we constructed a weighted undirected network. A node represents a Wikipedia page and an edge exists between two nodes if they have at least one editor in common. Weights on edges represent the number of editors in common. We look at the evolution of the network across time and identify the most central pages according to the degree centrality. Figure C.5 shows the number of weeks each cryptocurrencies appeared in the top 5 ranks when cryptocurrencies are ranked according to their degree centrality in descending order. Figure C.6 shows the correlation between the age of a cryptocurrency page and its weighted degree ($\rho = 0.40, p = 0.015$).

C.8 Understanding the Wikipedia trading strategy behaviour

In this section we attempt to understand the behaviour of the trading strategy, specifically why it yields positive returns for some cryptocurrencies

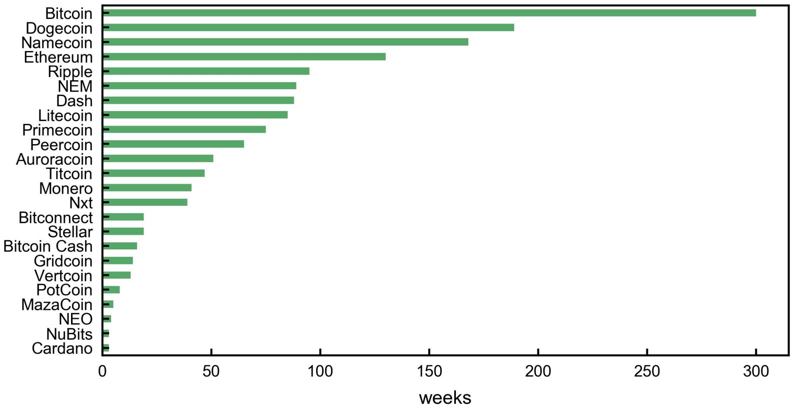


FIGURE C.5: **Ranking in degree centrality.** Number of weeks a cryptocurrency occupied one of top 5 ranks based on degree centrality in the co-editing Wikipedia pages network.

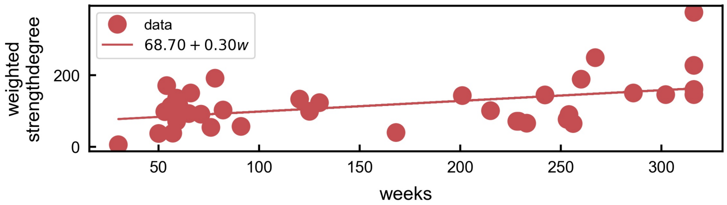


FIGURE C.6: **Correlation between page age and network strength.** Page age in weeks vs its weighted degree in the editing network. Each point represents a node (page). Pearson correlation $\rho = 0.40$, $p = 0.015$. The solid line represents a fit $a + bw$ where $b = 0.28 \pm 0.10$.

and negative for others. Firstly, we evaluate for each cryptocurrency the correlation between its change in Wikipedia views and change in price. Instead of looking at the magnitude value we transform the data to binary signal representing the up or down ticks, $+1$ for an increase and -1 for the decrease. The motivation behind this choice is to match the behaviour of the strategy where any increase or decrease in the Wikipedia views trigger a buy or sell action regardless its magnitude. We also limit our analysis to the 17 cryptocurrencies which are marginally traded and used in the Wikipedia strategy. Figure C.7A shows the Pearson correlation between changes in Wikipedia views changes and price changes values for cryptocurrencies where the $p < 0.05$. We extend the analysis to include change in volume too (Figure C.7B), however in both cases the correlation did not explain the behaviour of the trading strategy. Table C.5 also show the details of the Pearson correlation test for all 17 cryptocurrencies.

We further investigate the strategy behaviour through Granger causality test. We show a Granger causality analysis for the 17 cryptocurrencies under the study assuming a one day lag. Table C.6 shows the results of the Granger causality test for each cryptocurrency. The results show that changes in Wikipedia views Granger causes changes in price for 5 cryptocurrencies, namely Bitcoin, Ethereum Classic, Monero, Stellar, Tether. Changes in a cryptocurrency price on the other hand cause changes in Wikipedia views in case of Bitcoin, Dash, EOS, Ethereum, Ethereum Classic, Litecoin, Monero, Ripple, Stellar, Tether. For WIkipedia edits, a change in Wikipedia edits shown to Granger cause changes in price for only NEO cryptocurrency. Finally, change in price Granger cause changes in Wikipedia edits for Bitcoin and Monero. The Granger causality does not justify still the performance of all the cryptocurrencies (only 4 cryptocurrencies), which imply that another dynamics in play as we detailed in the main text.

Finally we show the evolution of the Sharpe ratio with time as shown in figure C.8. Although the Wikipedia based strategy has overall higher Sharpe ratio compared to the price and random baseline strategy (see main text), the performance fluctuates over time.

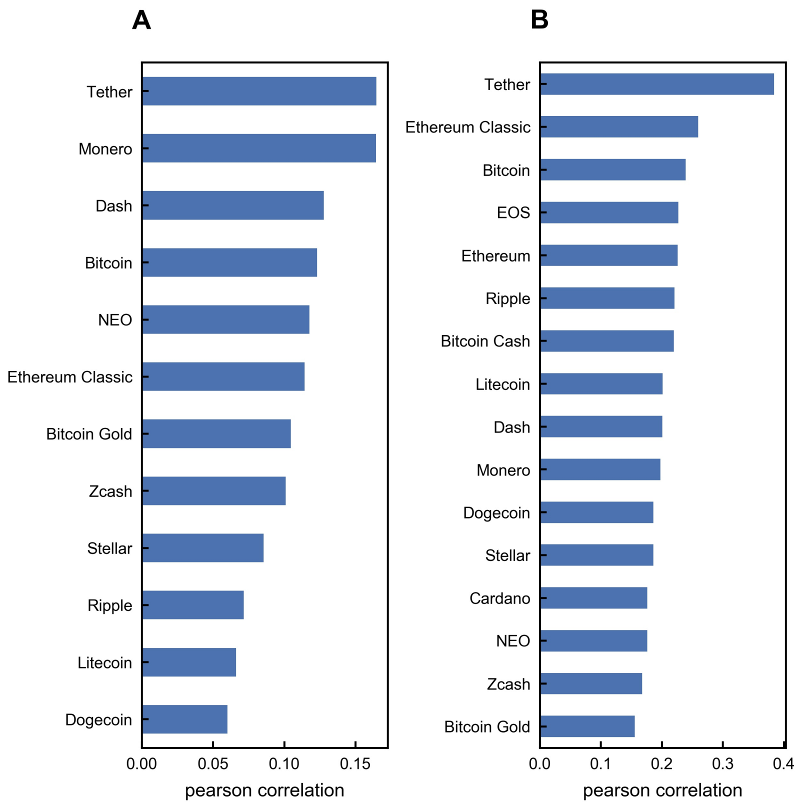


FIGURE C.7: **Correlation between crypotocurrency's market properties and Wikipedia page views.** The Pearson correlation between changes in Wikipedia page views and changes in trading volume (A) or price (B), measured across a cryptocurrency entire existence. Only significant correlations ($p < 0.05$) are shown.

TABLE C.5: **Pearson correlation between Wikipedia views and price and volume for each cryptocurrency.** The table reports the values for the Pearson correlation tests. r_p^2 and P_p are the r^2 and P values related to the Pearson correlation between changes in Wikipedia views and changes in price. r_v^2 and P_v are the r^2 and P values related to the Pearson correlation between changes in Wikipedia views and changes in price. Values are sorted in ascending order according to the value of P_p

Cryptocurrency	r_p^2	P_p	r_v^2	p_v
Monero	0.1645	$< 10^{-9}$	0.1976	$< 10^{-12}$
Dash	0.1278	$< 10^{-6}$	0.2007	$< 10^{-13}$
Bitcoin	0.123	$< 10^{-6}$	0.2391	$< 10^{-18}$
Tether	0.1646	0.0008	0.384	$< 10^{-16}$
Ethereum Classic	0.1143	0.0009	0.2594	$< 10^{-14}$
Stellar	0.0856	0.0038	0.1856	$< 10^{-10}$
Zcash	0.101	0.0038	0.1674	$< 10^{-6}$
Ripple	0.0714	0.0099	0.2207	$< 10^{-16}$
Litecoin	0.0661	0.0171	0.2011	$< 10^{-13}$
NEO	0.1175	0.0198	0.1759	0.0005
Dogecoin	0.06	0.0305	0.1861	$< 10^{-11}$
Bitcoin Gold	0.1047	0.0383	0.1555	0.002
Ethereum	0.0459	0.1026	0.2256	$< 10^{-16}$
EOS	0.0732	0.1723	0.2267	$< 10^{-5}$
Cardano	-0.0638	0.2501	0.176	0.0014
OmiseGO	0.0332	0.5006	0.0837	0.0885
Bitcoin Cash	0.0292	0.5245	0.2193	$< 10^{-6}$

Table C.6: **Granger Causality between Wikipedia views and edits and price for each cryptocurrency.** The table reports the values for the Granger causality tests. F_{vp} and P_{vp} are the F statistic and P values related to the hypothesis that changes in Wikipedia views do not cause changes in price. F_{pv} and P_{pv} are the F statistic and P values for the hypothesis that changes in a cryptocurrency price do not cause changes in Wikipedia views. F_{ep} and P_{ep} are the F statistic and P values for the hypothesis that changes in a cryptocurrency Wikipedia edits do not cause changes in its price. F_{pe} and P_{pe} are the F statistic and P values for the hypothesis that changes in a cryptocurrency price do not cause changes in its Wikipedia edits. Values are sorted in ascending order according to the value of P_{vp}

Cryptocurrency	F_{vp}	P_{vp}	F_{pv}	P_{pv}	F_{ep}	P_{ep}	F_{pe}	P_{pe}
Ethereum Classic	9.3007	0.0024	8.6553	0.0034	3.142	0.0767	0.9523	0.3294
Tether	9.244	0.0025	6.1783	0.0133	1.5401	0.2153	1.3332	0.2489
Bitcoin	5.6925	0.0172	19.442	0.0	1.2476	0.2642	6.2227	0.0127
Monero	4.8285	0.0282	20.9264	0.0	2.6284	0.1052	8.2711	0.0041
Stellar	4.1569	0.0417	25.1997	0.0	0.943	0.3317	0.3191	0.5723
Cardano	2.7315	0.0994	2.5546	0.1109	0.3401	0.5601	0.2431	0.6223
Dash	2.2928	0.1302	24.1867	0.0	0.3828	0.5362	1.5701	0.2104
Dogecoin	1.6981	0.1928	0.0023	0.9617	0.3467	0.5561	0.4643	0.4957
Zcash	1.6162	0.204	2.02	0.1556	0.5985	0.4394	0.4751	0.4908
NEO	1.515	0.2191	2.9372	0.0874	5.097	0.0245	0.5766	0.4481
Litecoin	1.1518	0.2834	9.0573	0.0027	0.386	0.5345	2.9388	0.0867
Ethereum	1.0931	0.296	25.2834	0.0	0.5141	0.4735	2.6115	0.1063
Bitcoin Cash	0.3604	0.5486	0.1015	0.7501	0.9594	0.3278	0.3197	0.5721
Bitcoin Gold	0.29	0.5906	3.7785	0.0526	0.0625	0.8028	0.0487	0.8254
Ripple	0.1754	0.6754	6.6337	0.0101	0.5783	0.4471	0.2186	0.6402
OmiseGO	0.0199	0.8879	1.6613	0.1982	1.2272	0.2686	0.0219	0.8825
EOS	0.0048	0.9449	6.9566	0.0087	0.0152	0.9021	0.2413	0.6235

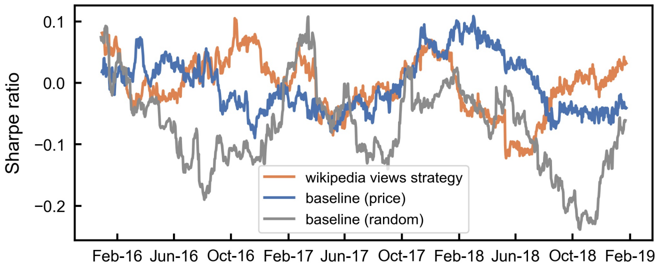


FIGURE C.8: **Sharpe ratio performance with time.** The sharpe ratio of the returns obtained using the Wikipedia-based strategy (orange line), the baseline strategy based on prices (blue line) and the random strategy (grey line). The Sharpe ratio calculated using a 6-months rolling window. Data for the random strategy is obtained from 1000 independent realizations. Results are shown for investments between July 2015 and January 2019 for all cryptocurrencies which can be traded marginally combined.

D Appendix to chapter 8

D.1 Parameter optimisation

In Figure D.1, we show the optimisation of the parameters w (A,C) and n (B,D) for the baseline strategy. In Figure D.2, we show the optimisation of the parameters w (A,D), $W_{training}$ (B,E), and n (C,F) for Method 1. In Figure D.3, we show the optimisation of the parameters w (A,D), $W_{training}$ (B,E), and n (C,F) for Method 2. In Figure D.4, we show the median squared error obtained under different training window choices (A), number of epochs (B) and number of neurons (C), for Ethereum, Bitcoin and Ripple. In Figure D.5, we show the optimisation of the parameter n (C,F) for Method 3.

D.2 Return under full knowledge of the market evolution.

In D.6, we show the cumulative return obtained by investing every day in the top currency, supposing one knows the prices of currencies on the following day.

D.3 Return obtained paying transaction fees.

In this section, we present the results obtained including transaction fees between 0.1% and 1% [282]. In general, one can not trade a given currency with any given other. Hence, we consider that each day we trade twice: We sell altcoins to buy Bitcoin, and we buy new altcoins using Bitcoin. The mean return obtained between Jan. 2016 and Apr. 2018 is larger than 1 for all methods, for fees up to 0.2% (see Table D.1). In this period, Method 3

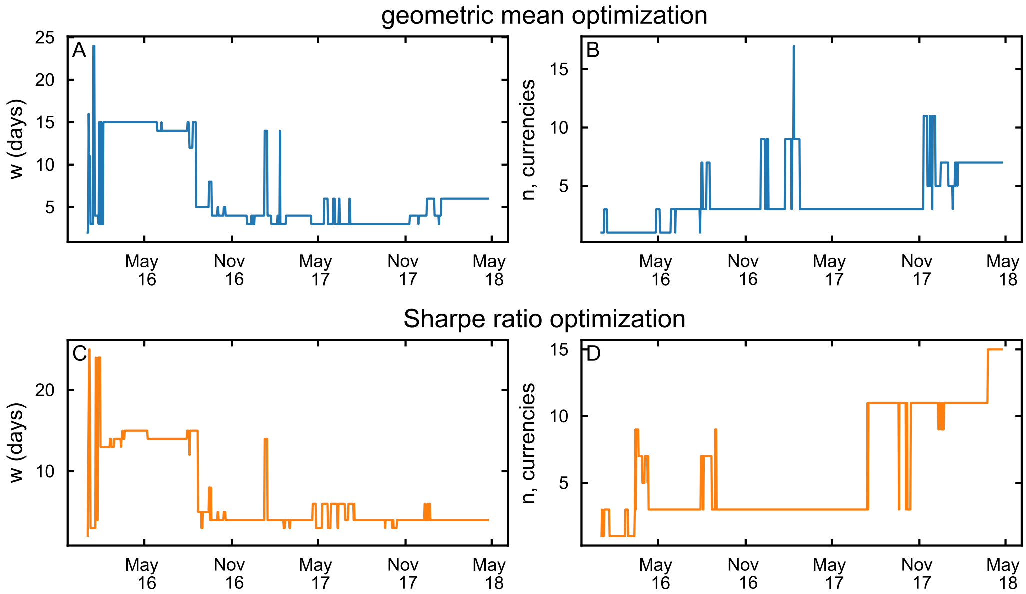


FIGURE D.1: **Baseline strategy: parameters optimisation.** The sliding window w (A,C) and the number of currencies n (B,D) chosen over time under the geometric mean (A,B) and the Sharpe Ratio optimisation (C,D). Analyses are performed considering prices in BTC.

achieves positive returns for fees up to 1%. The returns obtained with a 0.1% (see Figure D.7) and 0.2% (see Figure D.8) fee during arbitrary periods confirm that, in general, one obtains positive gains with our methods if fees are small enough.

TABLE D.1: **Daily geometric mean return for different transaction fees.** Results are obtained considering the period between Jan. 2016 and Apr. 2018.

	no fee	0.1%	0.2%	0.3%	0.5%	1%
Baseline	1.005	1.003	1.001	0.999	0.995	0.985
Method 1	1.008	1.006	1.004	1.002	0.998	0.988
Method 2	1.005	1.003	1.001	0.999	0.995	0.985
Method 3	1.025	1.023	1.021	1.019	1.015	1.005

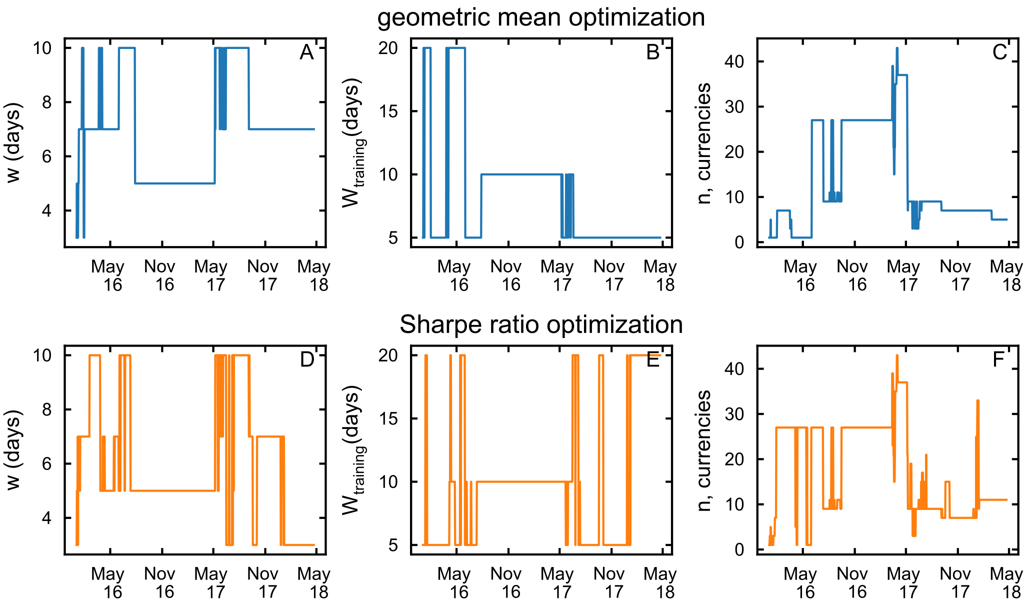


FIGURE D.2: **Method 1: Parameters optimisation.** The sliding window w (A,D), the training window $W_{training}$ (B,E) and the number of currencies n (C,F) chosen over time under the geometric mean (A,B,C) and the Sharpe Ratio optimisation (D,E,F). Analyses are performed considering prices in BTC.

D.4 Results in USD

In this section, we show results obtained considering prices in USD. The price of Bitcoin in USD has considerably increased in the period considered. Hence, gains in USD (Figure D.9) are higher than those in Bitcoin (Figure 8.5). Note that, in Figure D.9, we have made predictions and computed portfolios considering prices in Bitcoin. Then, gains have been converted to USD (without transaction fees). In Table D.2, we show instead the gains obtained running predictions considering directly all prices in USD. We find that, in most cases, better results are obtained from prices in BTC.

D.5 Geometric mean optimisation

In Figure D.10, we show the geometric mean return obtained by between two arbitrary points in time under geometric mean return optimisation for

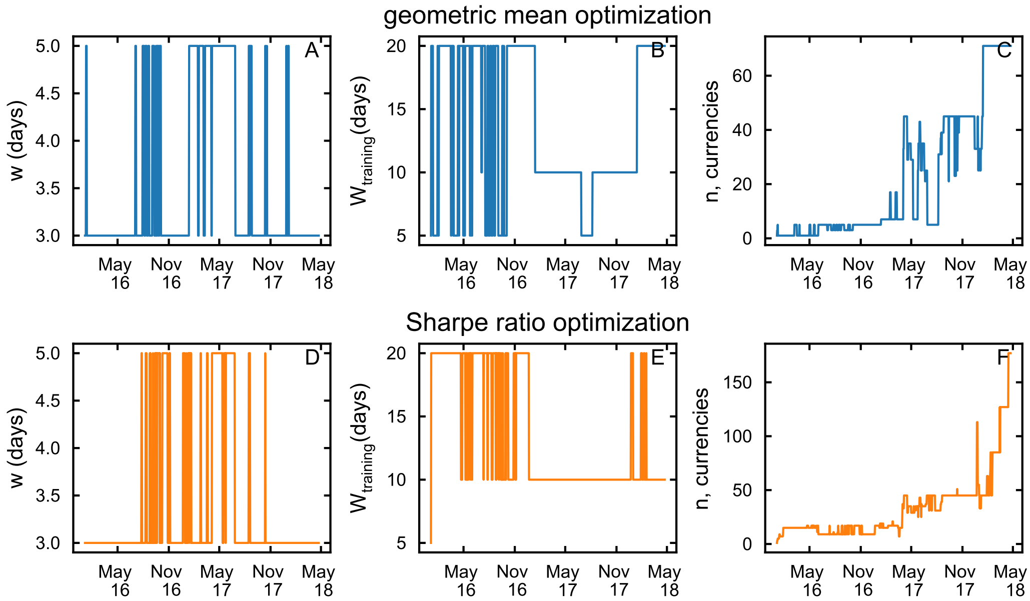


FIGURE D.3: **Method 2: Parameters optimisation.** The sliding window w (A,D), the training window $W_{training}$ (B,E) and the number of currencies n (C,F) chosen over time under the geometric mean (A,B,C) and the Sharpe Ratio optimisation (D,E,F). Analyses are performed considering prices in BTC.

the baseline (D.10-A), Method 1 (Figure D.10-B), Method 2 (Figure D.10C), and Method 3 (Figure D.10D).

TABLE D.2: **Geometric mean returns in USD.** Results are obtained for the various methods by running the algorithms considering prices in BTC (left column) and USD (right column).

	Geometric mean in USD (from BTC prices)	Geometric mean in USD (from USD prices)
Baseline	1.0086	1.0141
Method1	1.0121	1.0085
Method2	1.0091	1.0086
Method3	1.0289	1.0134

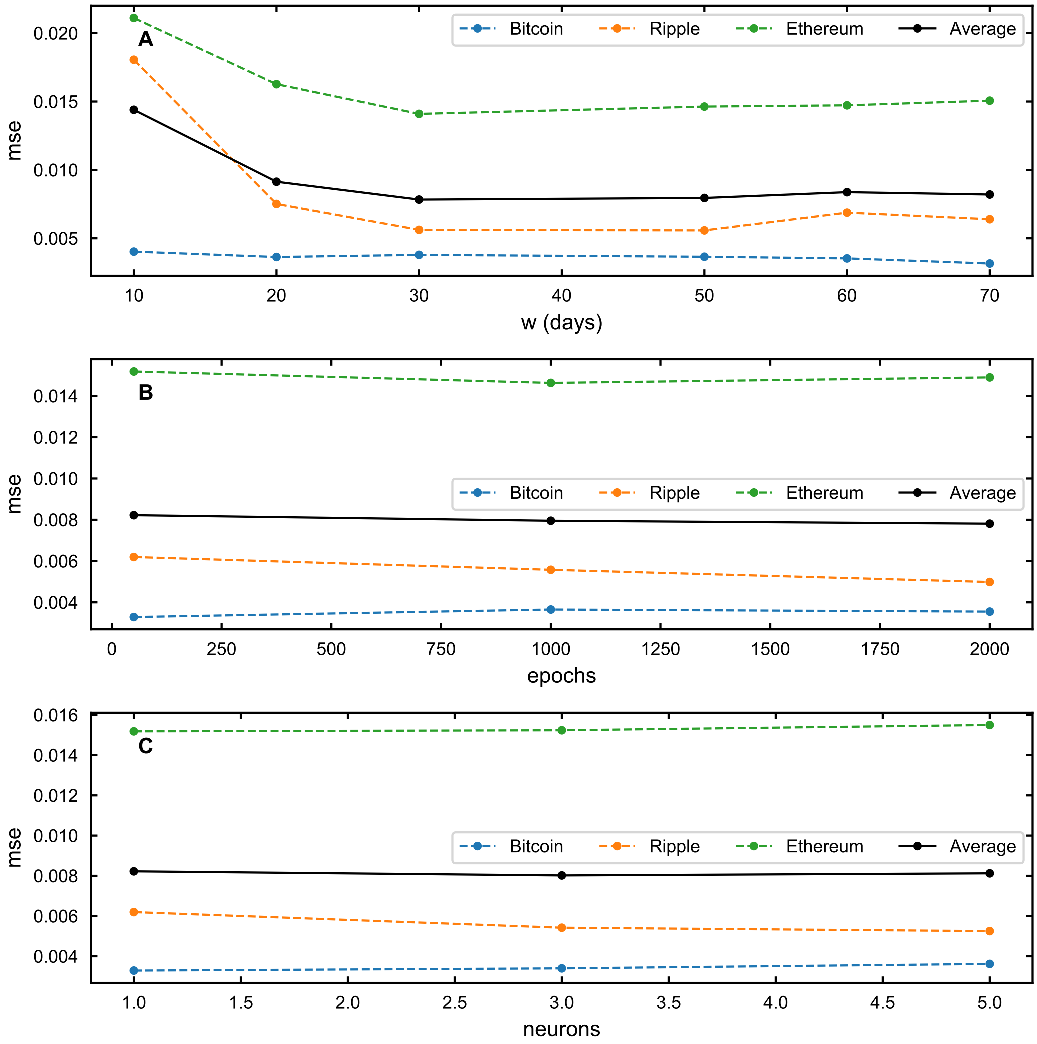


FIGURE D.4: **Method 3: Parameters optimisation.** The median squared error of the ROI as a function of the window size (A), the number of epochs (B) and the number of neurons (C). Results are shown considering prices in Bitcoin.

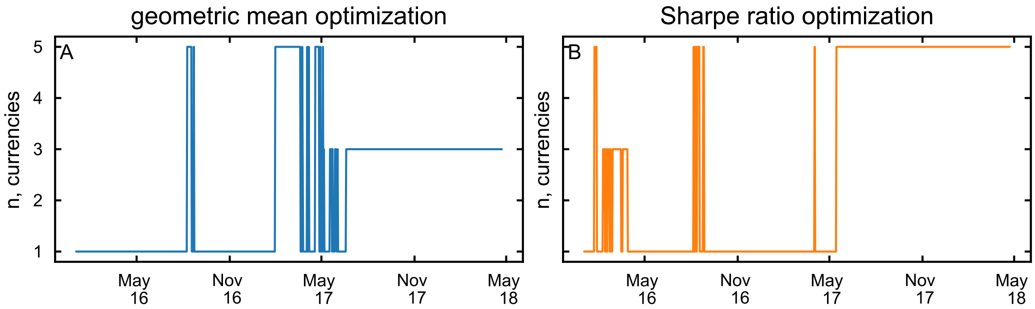


FIGURE D.5: **Method 3: Parameters optimisation.** The number of currencies n chosen over time under the geometric mean (A) and the Sharpe Ratio optimisation (B). Analyses are performed considering prices in BTC.

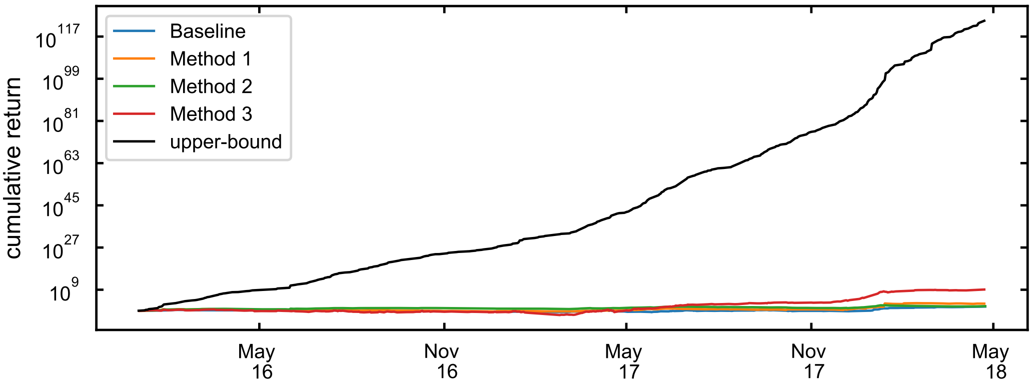


FIGURE D.6: **Upper-bound for the cumulative return.** The cumulative return obtained by investing every day in the currency with highest return on the following day (black line). The cumulative return obtained with the baseline (blue line), Method 1 (orange line), Method 2 (green line), and Method 3 (red line). Results are shown in Bitcoin.

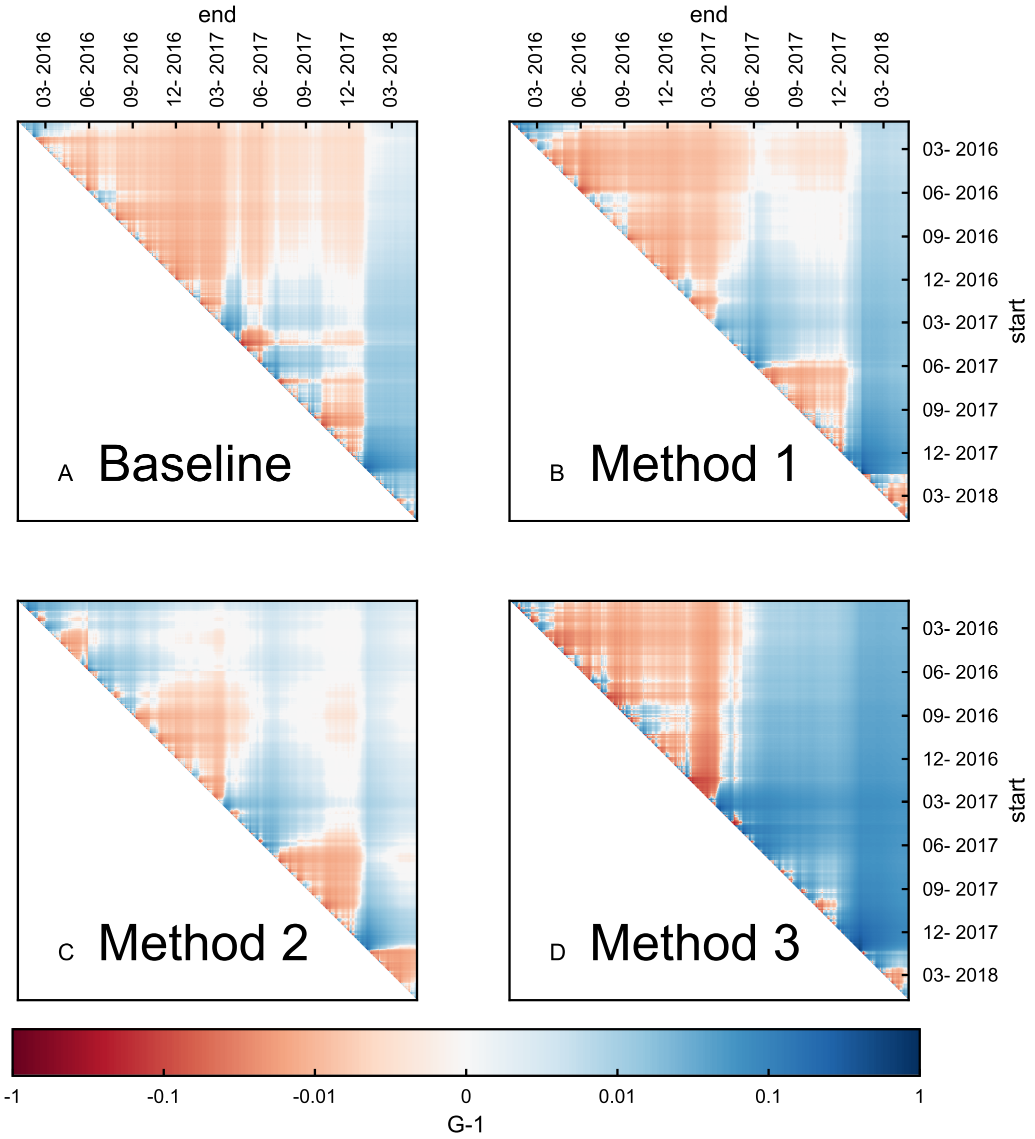


FIGURE D.7: **Daily geometric mean return obtained under transaction fees of** 0.1%**.** The geometric mean return computed between time ”start” and ”end” using the Sharpe ratio optimisation for the baseline (A), Method 1 (B), Method 2 (C) and Method 3 (D). Note that, for visualization purposes, the figure shows the translated geometric mean return G-1. Shades of red refers to negative returns and shades of blue to positive ones (see colour bar).

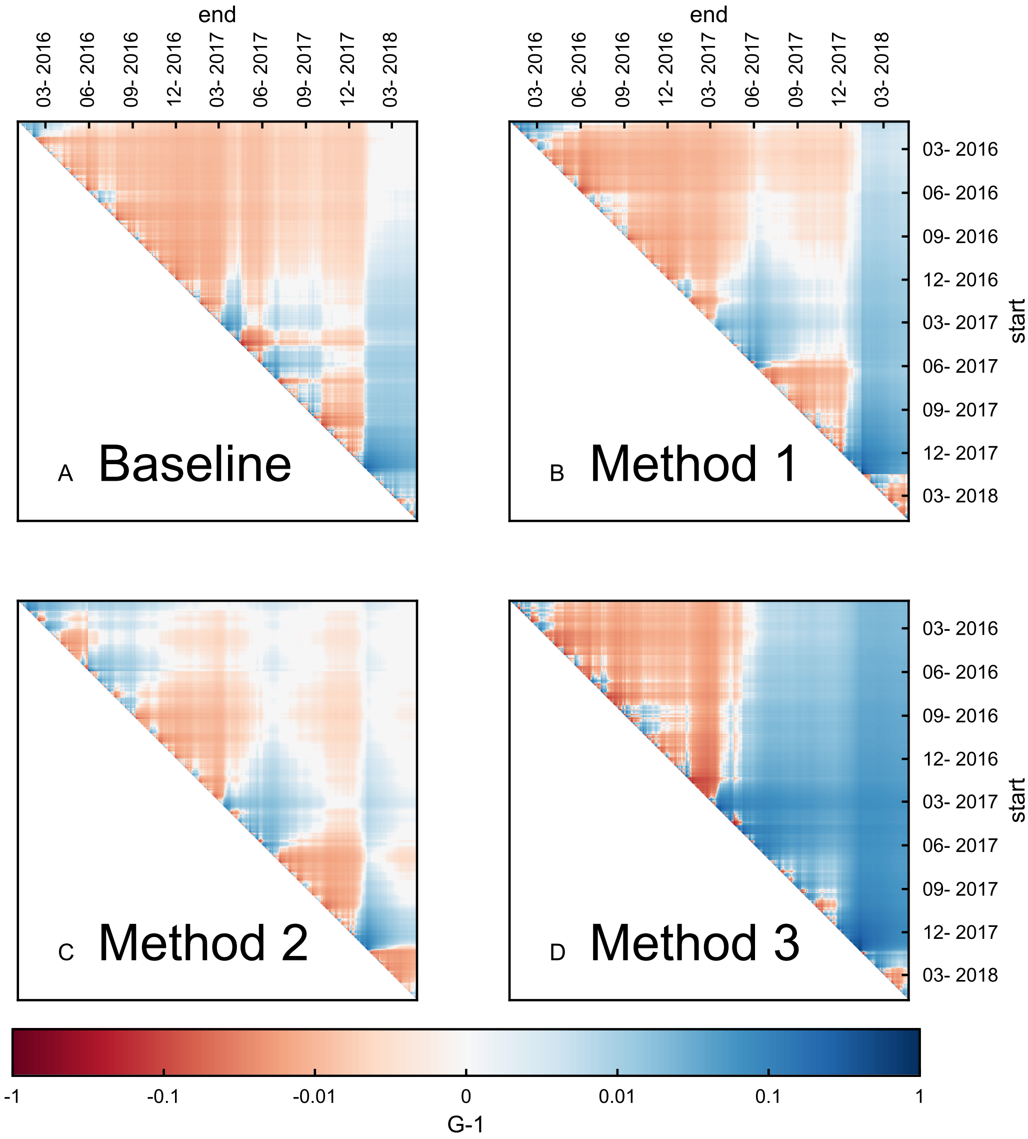


FIGURE D.8: **Daily geometric mean return obtained under transaction fees of** 0.2%**.** The geometric mean return computed between time ”start” and ”end” using the Sharpe ratio optimisation for the baseline (A), Method 1 (B), Method 2 (C) and Method 3 (D). Note that, for visualization purposes, the figure shows the translated geometric mean return G-1. Shades of red refers to negative returns and shades of blue to positive ones (see colour bar).

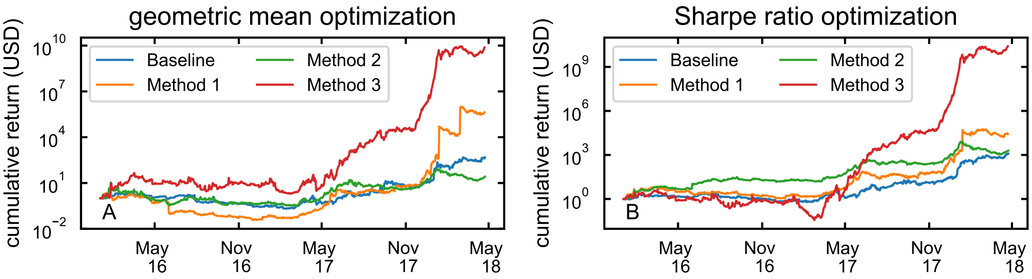


FIGURE D.9: **Cumulative returns in USD.** The cumulative returns obtained under the Sharpe Ratio optimisation (A) and the geometric mean optimisation (B) for the baseline (blue line), Method 1 (orange line), Method 2 (green line) and Method 3 (red line). Analyses are performed considering prices in BTC.

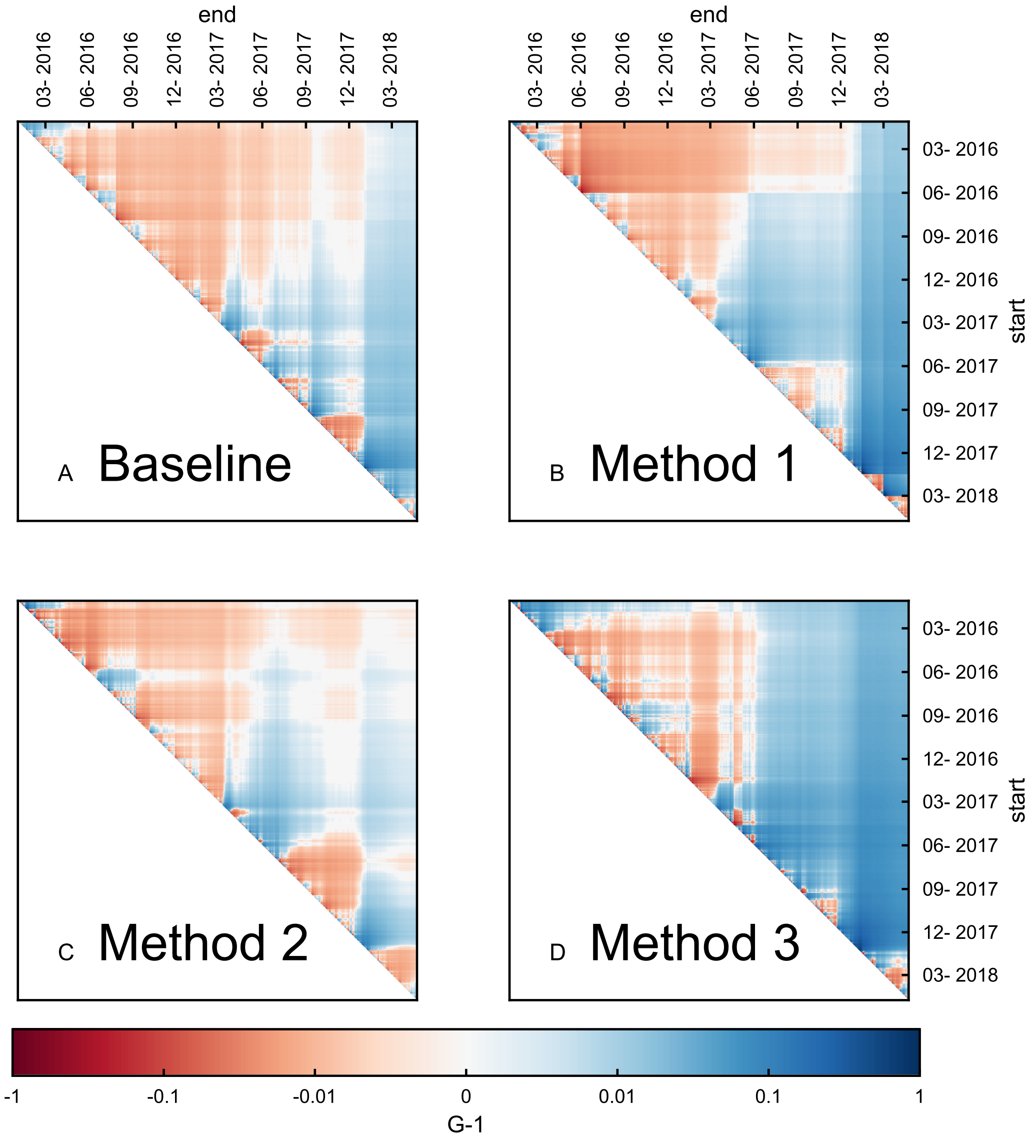


FIGURE D.10: **Geometric mean return obtained within different periods of time.** The geometric mean return computed between time ”start” and ”end” using the Sharpe ratio optimisation for the baseline (A), Method 1 (B), Method 2 (C) and Method 3 (D). Note that, for visualization purposes, the figure shows the translated geometric mean return G-1. Shades of red refers to negative returns and shades of blue to positive ones (see colour bar).

Bibliography

[1] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008. working paper.

[2] Nicolas Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd international conference on World Wide Web*, pages 213–224. Acm, 2013.

[3] Catherine England and Craig Fratrik. Where to bitcoin? *Journal of Private Enterprise*, 33(1), 2018.

[4] coinmarketcap. *coinmarketcap.com*, 2013. Accessed: 28 Ocober 2019.

[5] Shehu M Sarkintudu, Huda H Ibrahim, and Alawiyah Abd Wahab. Cryptocurrency platform ecosystem: a systematic literature review from information systems perspective. *International Journal of Advanced Computer Research*, 9(44):308–315, 2019.

[6] Rainer Böhme, Nicolas Christin, Benjamin Edelman, and Tyler Moore. Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives*, 29(2):213–38, 2015.

[7] Florian Glaser, Kai Zimmermann, Martin Haferkorn, Moritz Weber, and Michael Siering. Bitcoin-asset or currency? revealing users' hidden intentions. *Revealing Users' Hidden Intentions (April 15, 2014). ECIS*, 2014.

[8] David Garcia and Frank Schweitzer. Social signals and algorithmic trading of bitcoin. *Open Science*, 2(9):150288, 2015.

[9] Martina Matta, Ilaria Lunesu, and Michele Marchesi. Bitcoin spread prediction using social and web search media. In *UMAP Workshops*, 2015.

[10] Jermain Kaminski. Nowcasting the bitcoin market with twitter signals. *arXiv preprint arXiv:1406.7577*, 2014.

[11] Ladislav Kristoufek. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports*, 3:3415, 2013.

[12] Ladislav Kristoufek. What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PloS one*, 10(4):e0123923, 2015.

[13] Stuart Colianni, Stephanie Rosales, and Michael Signorotti. Algorithmic trading of cryptocurrency based on twitter sentiment analysis. *CS229 Project*, 2015.

[14] Zhengyao Jiang and Jinjun Liang. Cryptocurrency portfolio management with deep reinforcement learning. *arXiv preprint arXiv:1612.01277*, 2016.

[15] Neil Gandal and Hanna Halaburda. Can we predict the winner in a market with network effects? competition in cryptocurrency market. *Games*, 7(3):16, 2016.

[16] Kyle Soska and Nicolas Christin. Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. In *SEC'15 Proceedings of the 24th USENIX Conference on Security Symposium*, volume 15 of *Sec'15*, pages 33–48, August 2015.

[17] David Garcia, Claudio J Tessone, Pavlin Mavrodiev, and Nicolas Perony. The digital traces of bubbles: feedback cycles between socio-economic signals in the bitcoin economy. *Journal of the Royal Society Interface*, 11(99):20140623, 2014.

[18] Chia-Yen Tan, You-Beng Koh, and Kok-Haur Ng. Structural change analysis of active cryptocurrency market. *arXiv preprint arXiv:1909.10679*, 2019.

[19] Ke Wu, Spencer Wheatley, and Didier Sornette. Classification of cryptocurrency coins and tokens by the dynamics of their market capitalizations. *Royal Society open science*, 5(9):180381, 2018.

[20] Guglielmo Maria Caporale, Luis Gil-Alana, and Alex Plastun. Persistence in the cryptocurrency market. *Research in International Business and Finance*, 46:141–148, 2018.

[21] Peter M Krafft, Nicolás Della Penna, and Alex Pentland. An experimental study of cryptocurrency market dynamics. *arXiv preprint arXiv:1801.05831*, page 605, 2018.

[22] United Nations Office on Drugs and Crime. World Drug Report. Technical report, United Nations, 2019.

[23] European Monitoring Centre for Drugs and Drug Addiction. *Drugs And The Darknet: Perspectives For Enforcement, Research And Policy*. European Monitoring Centre for Drugs and Drug Addiction, 2017.

[24] United Nations Office on Drugs and Crime. *World drug report 2018*. United Nations Publications, 2018.

[25] Robleh Ali, John Barrdear, Roger Clews, and James Southgate. The economics of digital currencies. *Bank of England Quarterly Bulletin*, page Q3, 2014.

[26] Danton Bryans. Bitcoin and money laundering: mining for an effective solution. *Ind. LJ*, 89:441, 2014.

[27] Danny Yuxing Huang, Maxwell Matthaios Aliapoulios, Vector Guo Li, Luca Invernizzi, Elie Bursztein, Kylie McRoberts, Jonathan Levin, Kirill Levchenko, Alex C Snoeren, and Damon McCoy. Tracking ransomware end-to-end. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 618–631. Ieee, 2018.

[28] United Nations Office on Drugs and Crime. *World drug report 2019*. United Nations Publications, 2019.

[29] Paola Ceruleo. Bitcoin: a rival to fiat money or a speculative financial asset? *The Libera Università Internazionale degli Studi Sociali*, 2014.

[30] Angela Rogojanu, Liana Badea, et al. The issue of competing currencies. case study–bitcoin. *Theoretical and Applied Economics*, 21(1):103–114, 2014.

[31] Paolo Tasca. The dual nature of bitcoin as payment network and money. In *VI Chapter SUERF Conference Proceedings*, volume 1, 2016.

[32] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder. *Bitcoin and cryptocurrency technologies: a comprehensive introduction*. Princeton University Press, 2016.

[33] Stuart Haber and W Stornetta. How to time-stamp a digital document, crypto?90, lncs 537, 1991.

[34] J Lawrence Carter and Mark N Wegman. Universal classes of hash functions. *Journal of computer and system sciences*, 18(2):143–154, 1979.

[35] Qin Lv, Pei Cao, Edith Cohen, Kai Li, and Scott Shenker. Search and replication in unstructured peer-to-peer networks. In *Proceedings of the 16th international conference on Supercomputing*, pages 84–95. Acm, 2002.

[36] Bip-34. `https://github.com/bitcoin/bips/blob/master/bip-0034.mediawiki`, 2019. Accessed: 1 October 2019.

[37] Joseph Bonneau, Andrew Miller, Jeremy Clark, Arvind Narayanan, Joshua A Kroll, and Edward W Felten. Sok: Research perspectives and challenges for bitcoin and cryptocurrencies. In *2015 IEEE Symposium on Security and Privacy*, pages 104–121. Ieee, 2015.

[38] Arthur Gervais, Vedran Capkun, Srdjan Capkun, and Ghassan O Karame. Is bitcoin a decentralized currency? *IEEE security & privacy*, 12(3):54–60, 2014.

[39] Juan Garay, Aggelos Kiayias, and Nikos Leonardos. The bitcoin backbone protocol: Analysis and applications. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 281–310. Springer, 2015.

[40] Jordi Herrera-Joancomartí. Research and challenges on bitcoin anonymity. In *Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance*, pages 3–16. Springer, 2014.

[41] Lear Bahack. Theoretical bitcoin attacks with less than half of the computational power (draft). *arXiv preprint arXiv:1312.7013*, 2013.

[42] Ittay Eyal and Emin Gün Sirer. Majority is not enough: Bitcoin mining is vulnerable. *Communications of the ACM*, 61(7):95–102, 2018.

[43] crypto exchange giant binance report a hack of 7000 bitcoin. `https://www.bloomberg.com/news/articles/2019-05-08/crypto-exchange-giant-binance-reports-a-hack-of-7-000-bitcoin`, 2019. Accessed: 11 Septemeber 2019.

[44] Ayana Aspembitova, Ling Feng, Valentin Melnikov, and Lock Yue Chew. Fitness preferential attachment as a driving mechanism in bitcoin transaction network. *PloS one*, 14(8):e0219346, 2019.

[45] Dániel Kondor, Márton Pósfai, István Csabai, and Gábor Vattay. Do the rich get richer? an empirical analysis of the bitcoin transaction network. *PloS one*, 9(2):e86197, 2014.

[46] whale-alert.io/alerts. https://whale-alert.io/alerts . Accessed: 15 October 2019.

[47] Dorit Ron and Adi Shamir. Quantitative analysis of the full bitcoin transaction graph. In *International Conference on Financial Cryptography and Data Security*, pages 6–24. Springer, 2013.

[48] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M Voelker, and Stefan Savage. A fistful of bitcoins: characterizing payments among men with no names. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 127–140. Acm, 2013.

[49] Elli Androulaki, Ghassan O Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Capkun. Evaluating user privacy in bitcoin. In *International Conference on Financial Cryptography and Data Security*, pages 34–51. Springer, 2013.

[50] coinmetrics.io. http://coinmetrics.io, 2017. Accessed: 15 October 2019.

[51] Matthias Lischke and Benjamin Fabian. Analyzing the bitcoin network: The first four years. *Future Internet*, 8(1):7, 2016.

[52] VISA. Visa annual report. Technical report, VISA, 2018.

[53] The blockchain scalability problem & the race for visa-like transaction speed. https://hackernoon.com/the-blockchain-scalability-problem-the-race-for-visa-like-transaction-speed-5cce48f9d44, 2019. Accessed: 1 October 2019.

[54] Shoji Kasahara and Jun Kawahara. Effect of bitcoin fee on transaction-confirmation process. *arXiv preprint arXiv:1604.00103*, 2016.

[55] Karl J O'Dwyer and David Malone. Bitcoin mining and its energy footprint. 2014.

[56] Sean Foley, Jonathan R Karlsen, and Tālis J Putniņš. Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies*, 32(5):1798–1853, 2019.

[57] Gwern. gwern.net/dnm-survival. https://www.gwern.net/DNM-survival, 2019. Accessed: 3 July 2019.

[58] Frank Wehinger. The dark net: Self-regulation dynamics of illegal online markets for identities and related services. In *2011 European Intelligence and Security Informatics Conference*, pages 209–213. Ieee, 2011.

[59] Coindesk. www.coindesk.com/sheep-marketplace-track-stolen-bitcoins. https://www.coindesk.com/sheep-marketplace-track-stolen-bitcoins, 2019. Accessed: 28 August 2019.

[60] Bassam Zantout, Ramzi Haraty, et al. I2p data communication system. In *Proceedings of ICN*, pages 401–409. Citeseer, 2011.

[61] Vitalik Buterin. Bitcoin multisig wallet: the future of bitcoin. *Bitcoin Magazine URL: https://bitcoinmagazine. com/11108/multisig-future-bitcoin*, 2014.

[62] Julia Buxton and Tim Bingham. The rise and challenge of dark net drug markets. *Policy brief*, 7:1–24, 2015.

[63] Kyle Soska and Nicolas Christin. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pages 33–48, 2015.

[64] David Décary-Hétu and Luca Giommoni. Do police crackdowns disrupt drug cryptomarkets? a longitudinal analysis of the effects of operation onymous. *Crime, Law and Social Change*, 67(1):55–75, 2017.

[65] Judith Aldridge and David Décary-Hétu. Not an'ebay for drugs': the cryptomarket'silk road'as a paradigm shifting criminal innovation. *Available at SSRN 2436643*, 2014.

[66] Diana S Dolliver. Evaluating drug trafficking on the tor network: Silk road 2, the sequel. *International Journal of Drug Policy*, 26(11):1113–1123, 2015.

[67] Martin Dittus, Joss Wright, and Mark Graham. Platform criminalism: The'last-mile'geography of the darknet market supply chain. In *Proceedings of the 2018 World Wide Web Conference*, pages 277–286. International World Wide Web Conferences Steering Committee, 2018.

[68] Scott W Duxbury and Dana L Haynie. Building them up, breaking them down: Topology, vendor selection patterns, and a digital drug market?s robustness to disruption. *Social Networks*, 52:238–250, 2018.

[69] Monica J Barratt, Jason A Ferris, and Adam R Winstock. Use of silk road, the online drug marketplace, in the united kingdom, a ustralia and the united states. *Addiction*, 109(5):774–783, 2014.

[70] Marie Claire Van Hout and Tim Bingham. Silk road, the virtual drug marketplace: A single case study of user experiences. *International Journal of Drug Policy*, 24(5):385–391, 2013.

[71] Joe Van Buskirk, Amanda Roxburgh, Michael Farrell, and Lucy Burns. The closure of the silk road: what has this meant for online drug trading? *Addiction*, 109(4):517–518, 2014.

[72] Malte Möser, Rainer Böhme, and Dominic Breuker. An inquiry into money laundering tools in the bitcoin ecosystem. In *2013 APWG eCrime Researchers Summit*, pages 1–14. Ieee, 2013.

[73] Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles E Leiserson. Anti-money

laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*, 2019.

[74] Masarah Paquet-Clouston, Bernhard Haslhofer, and Benoit Dupont. Ransomware payments in the bitcoin ecosystem. *Journal of Cybersecurity*, 5(1):tyz003, 2019.

[75] Mauro Conti, Ankit Gangwal, and Sushmita Ruj. On the economic significance of ransomware campaigns: A bitcoin transactions perspective. *Computers & Security*, 79:162–189, 2018.

[76] Segwit, explained. `https://cointelegraph.com/explained/segwit-explained`, 2019. Accessed: 1 October 2019.

[77] Joseph Poon and Thaddeus Dryja. The bitcoin lightning network: Scalable off-chain instant payments, 2016.

[78] Lightning network developers warn of bug that could cause loss of bitcoin. https://www.coindesk.com/lightning-network-developers-warn-of-bug-that-could-cause-loss-of-bitcoin, 2019. Accessed: 1 October 2019.

[79] What is bitcoin block size debate and why does it matter. http://www.coindesk.com/what-is-the-bitcoin-block-size-debate-and-why-does-it-matter/, 2015. Accessed: 5 June 2017.

[80] Cathy Barrera and Stephanie Hurder. Blockchain upgrade as a coordination game. 2018.

[81] Litecoin. `https://litecoin.org`, 2019. Accessed: 1 October 2019.

[82] Dogecoin. `https://dogecoin.com`, 2019. Accessed: 1 October 2019.

[83] Freicoin. `http://freico.in`, 2019. Accessed: 1 October 2019.

[84] Primecoin. `http://primecoin.io`, 2019. Accessed: 1 October 2019.

[85] Ethereum. `https://www.ethereum.org`, 2019. Accessed: 1 October 2019.

[86] Neo. `https://neo.org`, 2019. Accessed: 1 October 2019.

[87] Peercoin. `https://www.peercoin.net`, 2019. Accessed: 1 October 2019.

[88] Ben Laurie. An efficient distributed currency. *Practice*, 100, 2011.

[89] Libra. `https://libra.org/en-US/`, 2019. Accessed: 1 October 2019.

[90] JP Koning. Fedcoin: a central bank-issued cryptocurrency. *R3 Report*, 15, 2016.

[91] Bank of england. `https://www.bankofengland.co.uk/research/digital-currencies`, 2019. Accessed: 1 October 2019.

[92] J.p. morgan creates digital coin for payments. `https://www.jpmorgan.com/global/news/digital-coin-payments`, 2019. Accessed: 1 October 2019.

[93] Solve. `https://solve.care`, 2019. Accessed: 1 October 2019.

[94] Basic attention. `https://basicattentiontoken.org`, 2019. Accessed: 1 October 2019.

[95] Steem. `https://steem.com`, 2019. Accessed: 1 October 2019.

[96] Blockchain for peer review. `https://www.blockchainpeerreview.org`, 2019. Accessed: 1 October 2019.

[97] Garrick Hileman and Michel Rauchs. *Global Cryptocurrency Benchmarking Study*. Cambridge Centre for Alternative Finance, 2017.

[98] Igor Makarov and Antoinette Schoar. Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 2019.

[99] Library of Congress. Regulation of Cryptocurrency Around the World. https://www.loc.gov/law/help/cryptocurrency/world-survey.php, 2018. Accessed: 15 October 2019.

[100] Bitcoin Mining Pools. `https://www.buybitcoinworldwide.com/mining/pools/`, 2019. Accessed: 11 Septemeber 2019.

[101] Cryptocurrency exchanges list. `https://en.bitcoinwiki.org/wiki/Cryptocurrency_exchanges_list`, 2019. Accessed: 11 Septemeber 2019.

[102] A Huge List of Cryptocurrency Thefts. `https://hackernoon.com/a-huge-list-of-cryptocurrency-thefts-16d6bf246389`, 2019. Accessed: 11 Septemeber 2019.

[103] Gerald P Dwyer. The economics of bitcoin and similar private digital currencies. *Journal of Financial Stability*, 17:81–91, 2015.

[104] Anne Haubo Dyhrberg. Bitcoin, gold and the dollar–a garch volatility analysis. *Finance Research Letters*, 16:85–92, 2016.

[105] Anne Haubo Dyhrberg. Hedging capabilities of bitcoin. is it the virtual gold? *Finance Research Letters*, 16:139–144, 2016.

[106] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y Kenett, H Eugene Stanley, and Tobias Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific reports*, 3:1801, 2013.

[107] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3:srep01684, 2013.

[108] Johan Bollen and Huina Mao. Twitter mood as a stock market predictor. *Computer*, 44(10):91–94, 2011.

[109] Chester Curme, Tobias Preis, H Eugene Stanley, and Helen Susannah Moat. Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*, 111(32):11600–11605, 2014.

[110] Young Bin Kim, Jurim Lee, Nuri Park, Jaegul Choo, Jong-Hyun Kim, and Chang Hun Kim. When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PloS one*, 12(5):e0177630, 2017.

[111] *Bitcoin froum*, 2016. Accessed: 19 February 2019.

[112] coindesk.com. http://coindesk.com, 2017. Accessed: 15 October 2019.

[113] Alexander Dickerson. Algorithmic trading of bitcoin using wikipedia and google search volume. 2018.

[114] Isaac Madan, Shaurya Saluja, and Aojia Zhao. Automated bitcoin trading via machine learning algorithms. *URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan*, 20, 2015.

[115] Huisu Jang and Jaewook Lee. An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *IEEE Access*, 6:5427–5437, 2018.

[116] Sean McNally. *Predicting the price of Bitcoin using Machine Learning*. PhD thesis, Dublin, National College of Ireland, 2016.

[117] KAREEM HEGAZY and SAMUEL MUMFORD. Comparitive automated bitcoin trading strategies.

[118] Tian Guo and Nino Antulov-Fantulin. Predicting short-term bitcoin price fluctuations from buy and sell orders. *arXiv preprint arXiv:1802.04065*, 2018.

[119] *Ethereum froum*, 2016. Accessed: 19 February 2019.

[120] *Rippl chat*, 2016. Accessed: 19 February 2019.

[121] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, 11(8):e0161197, 2016.

[122] *Bitcoin reddit*, 2016. Accessed: 19 February 2019.

[123] Ross C Phillips and Denise Gorse. Mutual-excitation of cryptocurrency market returns and social media topics. In *Proceedings of the 4th International Conference on Frontiers of Educational Technologies*, pages 80–86. Acm, 2018.

[124] Matthew Purver, Thomas L Griffiths, Konrad P Körding, and Joshua B Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 17–24. Association for Computational Linguistics, 2006.

[125] Ross C Phillips and Denise Gorse. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pages 1–7. Ieee, 2017.

[126] Paul S Addison. *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.

[127] Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998.

[128] Lukas Vacha and Jozef Barunik. Co-movement of energy commodities revisited: Evidence from wavelet coherence analysis. *Energy Economics*, 34(1):241–247, 2012.

[129] Chaker Aloui and Besma Hkiri. Co-movements of gcc emerging stock markets: New evidence from wavelet coherence analysis. *Economic Modelling*, 36:421–431, 2014.

[130] Aslak Grinsted, John C Moore, and Svetlana Jevrejeva. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear processes in geophysics*, 11(5/6):561–566, 2004.

[131] Jean-Philippe Lachaux, Antoine Lutz, David Rudrauf, Diego Cosmelli, Michel Le Van Quyen, Jacques Martinerie, and Francisco Varela. Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Neurophysiologie Clinique/Clinical Neurophysiology*, 32(3):157–174, 2002.

[132] Ross C Phillips and Denise Gorse. Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PloS one*, 13(4):e0195200, 2018.

[133] *reddit*, 2016. Accessed: 19 February 2019.

[134] cryptocurrency trading robot. `https://github.com/owocki/pytrader`, 2017. Accessed: 5 December 2017.

[135] *Cryptobot package*, 2017. Accessed: 19 February 2019.

[136] *CryptoCurrencyTrader package*, 2017. Accessed: 5 December 2017.

[137] *Btctrading package*, 2017. Accessed: 19 February 2019.

[138] *Bitpredict package*, 2017. Accessed: 19 February 2019.

[139] *Bitcoin bubble burst*, 2017. Accessed: 19 February 2019.

[140] Eaman Jahani, Peter M Krafft, Yoshihiko Suhara, Esteban Moro, and Alex Sandy Pentland. Scamcoins, s*** posters, and the search for the next bitcoin tm: Collective sensemaking in cryptocurrency discussions. *Proceedings of the ACM on Human-Computer Interaction*, 2(Cscw):79, 2018.

[141] Maryam Zamani, Fereshteh Rabbani, Attila Horicsányi, Anna Zafeiris, and Tamas Vicsek. Differences in structure and dynamics of networks retrieved from dark and public web forums. *Physica A: Statistical Mechanics and its Applications*, 525:326–336, 2019.

[142] Johannes Beck, Roberta Huang, David Lindner, Tian Guo, Zhang Ce, Dirk Helbing, and Nino Antulov-Fantulin. Sensing social media signals for cryptocurrency news. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1051–1054. Acm, 2019.

[143] Mike Seiferling, Abeer ElBahrawy, Tales Padilha, and Keith Chan. *Cryptocurrencies and the future of money*. IE Business School, 2019. to appear.

[144] Blockchain explorer. `https://www.blockchain.com/explorer`, 2019. Accessed: 1 October 2019.

[145] walletexplorer.com. https://www.walletexplorer.com/, 2009. Accessed: 11 April 2019.

[146] Chainalysis. `https://www.chainalysis.com`, 2019. Accessed: 1 October 2019.

[147] Elliptic. `https://www.elliptic.co`, 2019. Accessed: 1 October 2019.

[148] longhash.com. https://www.longhash.com , 2017. Accessed: 15 October 2019.

[149] Carol Alexander and Michael Dakos. A critical investigation of cryptocurrency data and analysis. *Available at SSRN 3382828*, 2019.

[150] Dormant Bitcoin. `https://bitinfocharts.com/top-100-dormant_8y-bitcoin-addresses.html`, 2019. Accessed: 1 October 2019.

[151] DATA Ultimatum: CoinMarketCap Requests More Information From Exchanges to Make Market More Transparent. https://cointelegraph.com/news/data-ultimatum-coinmarketcap-requests-more-information-from-exchanges-to-make-market-more-transparent, 2019. Accessed: 1 October 2019.

[152] *www.mediawiki.org/wiki/API:Main_page*, 2016. Accessed: 19 February 2019.

[153] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462. Acm, 2007.

[154] Ronald Rivest. The md5 message-digest algorithm. Technical report, 1992.

[155] *xtools.wmflabs.org*, 2016. Accessed: 19 February 2019.

[156] Wikipedia. Wikipedia API: Main Page, 2019.

[157] Malte Möser and Rainer Böhme. Trends, tips, tolls: A longitudinal study of bitcoin transaction fees. In *International Conference on Financial Cryptography and Data Security*, pages 19–33. Springer, 2015.

[158] Ernie GS Teo. Emergence, growth, and sustainability of bitcoin: the network economics perspective. In *Handbook of digital currency*, pages 191–200. Elsevier, 2015.

[159] blockchain.com/explorer. https://www.blockchain.com/explorer, 2009. Accessed: 11 April 2016.

[160] Julia Heidemann, Mathias Klier, and Florian Probst. Identifying key users in online social networks: A pagerank based approach. 2010.

[161] Michael Fleder, Michael S Kester, and Sudeep Pillai. Bitcoin transaction graph analysis. *arXiv preprint arXiv:1502.01657*, 2015.

[162] Wikipedia Now Accepts Bitcoin Donations. `https://www.coindesk.com/wikipedia-now-accepts-bitcoin-donations`, 2019. Accessed: 1 October 2019.

[163] YoonJae Chung. Cracking the code: How the us government tracks bitcoin transactions. *Analysis Of Applied Mathematics*, page 152, 2019.

[164] Peter V Marsden. Egocentric and sociocentric measures of network centrality. *Social networks*, 24(4):407–422, 2002.

[165] United states of america vs. ross william ulbricht. https://www.justice.gov/sites/default/files/usao-sdny/legacy/2015/03/25/US%20v.%20Ross%20Ulbricht%20Indictment.pdf, 2019. Accessed: 10 July 2019.

[166] Yin-Wong Cheung and Kon S. Lai. Lag Order and Critical Values of the Augmented Dickey Fuller Test. *Journal of Business & Economic Statistics*, 13(3):277–280, July 1995.

[167] Fergal Reid and Martin Harrigan. An analysis of anonymity in the bitcoin system. In *Security and privacy in social networks*, pages 197–223. Springer, 2013.

[168] Tobias Bamert, Christian Decker, Lennart Elsen, Roger Wattenhofer, and Samuel Welten. Have a snack, pay with bitcoins. In *Peer-to-Peer Computing (P2P), 2013 IEEE Thirteenth International Conference on*, pages 1–5. Ieee, 2013.

[169] Christian Decker and Roger Wattenhofer. Information propagation in the bitcoin network. In *Peer-to-Peer Computing (P2P), 2013 IEEE Thirteenth International Conference on*, pages 1–10. Ieee, 2013.

[170] Sha Wang and Jean-Philippe Vergne. Buzz factor or innovation potential: What explains cryptocurrencies returns? *PloS one*, 12(1):e0169556, 2017.

[171] Paul Vigna and Michael J Casey. *The age of cryptocurrency: How bitcoin and digital money are challenging the global economic order*. St. Martin's Press, 2015.

[172] Michael J Casey and Paul Vigna. Bitcoin and the digital-currency revolution. *The Wall Street Journal*, 23, 2015.

[173] Simon Trimborn and Wolfgang K Härdle. Crix an index for blockchain based currencies. 2016.

[174] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[175] SP Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography*. Princton University Press, 2001.

[176] Neil Gandal and Hanna Halaburda. Competition in the cryptocurrency market. 2014.

[177] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[178] Lada A Adamic and Bernardo A Huberman. Zipf's law and the internet. *Glottometrics*, 3(1):143–150, 2002.

[179] R Alexander Bentley, Carl P Lipo, Harold A Herzog, and Matthew W Hahn. Regular rates of popular culture change reflect random copying. *Evolution and Human Behavior*, 28(3):151–158, 2007.

[180] Warren J Ewens. *Mathematical population genetics 1: theoretical introduction*, volume 27. Springer Science & Business Media, 2012.

[181] Brian J McGill, Brian A Maurer, and Michael D Weiser. Empirical evaluation of neutral theory. *Ecology*, 87(6):1411–1423, 2006.

[182] Motoo Kimura et al. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[183] Fraser D Neiman. Stylistic variation in evolutionary perspective: inferences from decorative diversity and interassemblage distance in illinois woodland ceramic assemblages. *American Antiquity*, pages 7–36, 1995.

[184] Damian Ruck, R Alexander Bentley, Alberto Acerbi, Philip Garnett, and Daniel J Hruschka. Neutral evolution and turnover over centuries of english word popularity. *arXiv preprint arXiv:1703.10698*, 2017.

[185] R Alexander Bentley, Matthew W Hahn, and Stephen J Shennan. Random drift and culture change. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1547):1443–1450, 2004.

[186] David Alonso, Rampal S Etienne, and Alan J McKane. The merits of neutral theory. *Trends in ecology & evolution*, 21(8):451–457, 2006.

[187] James P O'Dwyer and Anne Kandler. Inferring processes of cultural transmission: the critical role of rare variants in distinguishing neutrality from novelty biases. *arXiv preprint arXiv:1702.08506*, 2017.

[188] Matthew W Hahn and R Alexander Bentley. Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 1):S120–s123, 2003.

[189] Tiziana Di Matteo, Tomaso Aste, and Michel M Dacorogna. Scaling behaviors in differently developed markets. *Physica A: Statistical Mechanics and its Applications*, 324(1-2):183–188, 2003.

[190] Tiziana Di Matteo, Tomaso Aste, and Michel M Dacorogna. Long-term memories of developed and emerging markets: Using the scaling analysis to characterize their stage of development. *Journal of Banking & Finance*, 29(4):827–851, 2005.

[191] Ryan Flanagan and Lucas Lacasa. Irreversibility of financial time series: a graph-theoretical approach. *Physics Letters A*, 380(20):1689–1697, 2016.

[192] Moreno Bonaventura, Valerio Ciotti, Pietro Panzarasa, Silvia Liverani, Lucas Lacasa, and Vito Latora. Predicting success in the worldwide start-up network. *arXiv preprint arXiv:1904.08171*, 2019.

[193] Tobias Preis, Johannes J Schneider, and H Eugene Stanley. Switching processes in financial markets. *Proceedings of the National Academy of Sciences*, 108(19):7674–7678, 2011.

[194] Federico Botta, Helen Susannah Moat, H Eugene Stanley, and Tobias Preis. Quantifying stock return distributions in financial markets. *PloS one*, 10(9):e0135600, 2015.

[195] Didier Sornette and Anders Johansen. Significance of log-periodic precursors to financial crashes. *Quantitative Finance*, 1(4):452–471, 2001.

[196] Neil Johnson, Guannan Zhao, Eric Hunsader, Hong Qi, Nicholas Johnson, Jing Meng, and Brian Tivnan. Abrupt rise of new machine ecology beyond human response time. *Scientific reports*, 3:srep02627, 2013.

[197] Bitcoin Soars to Record High Due to High Demand in Japan. http://time.com/4763114/bitcoin-record-high-japan/, 2017. Accessed: 5 June 2017.

[198] *Bitcoin drops by $100 as China?s central bank corrals the market*, 2017. Accessed: 19 February 2019.

[199] Andrea Baronchelli. The emergence of consensus. *arXiv preprint arXiv:1704.07767*, 2017.

[200] David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 1969.

[201] Mitsuo Yoshida, Yuki Arase, Takaaki Tsunoda, and Mikio Yamamoto. Wikipedia page view reflects web search trend. In *Proceedings of the ACM Web Science Conference*, page 65. Acm, 2015.

[202] Evita Stenqvist and Jacob Lönnö. *Predicting Bitcoin price fluctuation with Twitter sentiment analysis*, 2017. Accessed: 19 February 2019.

[203] Maria Glenski, Emily Saldanha, and Svitlana Volkova. Characterizing speed and scale of cryptocurrency discussion spread on reddit. In *The World Wide Web Conference*, pages 560–570. Acm, 2019.

[204] blog.coinmarketcap. *https://blog.coinmarketcap.com/2018/07/19*, 2017. Accessed: 19 February 2019.

[205] *alexa.com/topsites*, n.d. Accessed: 19 February 2019.

[206] Abeer ElBahrawy, Laura Alessandretti, Anne Kandler, Romualdo Pastor-Satorras, and Andrea Baronchelli. Evolutionary dynamics of the cryptocurrency market. *Royal Society Open Science*, 4(11), 2017.

[207] Besiki Stvilia, Michael B Twidale, Linda C Smith, and Les Gasser. Assessing information quality of a community-based encyclopedia. In *Iq*, 2005.

[208] Fernanda B Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. Acm, 2004.

[209] *stats.wikimedia*, 2016. Accessed: 19 February 2019.

[210] Grace Caffyn. What is the bitcoin block size debate and why does it matter. *URL: http://www.coindesk. com/what-is-the-bitcoin-block-size-debate-and-why-does-it-matter/(visited on 27/11/2015)*, 2015.

[211] James M Heilman and Andrew G West. Wikipedia and medicine: quantifying readership, editors, and the significance of natural language. *Journal of medical Internet research*, 17(3), 2015.

[212] Lev Muchnik, Sen Pei, Lucas C Parra, Saulo DS Reis, José S Andrade Jr, Shlomo Havlin, and Hernán A Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3:1783, 2013.

[213] Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19, 2007.

[214] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: a study of power editors on wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 51–60. Acm, 2009.

[215] Amos Tversky and Daniel Kahneman. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics*, 106(4):1039–1061, 1991.

[216] Alexander Brauneis and Roland Mestel. Price discovery of cryptocurrencies: Bitcoin and beyond. *Economics Letters*, 165:58–61, 2018.

[217] Eugene F Fama. Perfect competition and optimal production decisions under uncertainty. *The Bell Journal of Economics and Management Science*, pages 509–530, 1972.

[218] Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, and Andrea Baronchelli. Anticipating cryptocurrency prices using machine learning. *Complexity*, 2018, 2018.

[219] Robert S Hudson and Andros Gregoriou. Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. *International Review of Financial Analysis*, 38:151–162, 2015.

[220] David Garcia and Frank Schweitzer. Social signals and algorithmic trading of bitcoin. *Royal Society Open Science*, 2(9), 2015.

[221] Manlio De Domenico and Andrea Baronchelli. The fragility of decentralised trustless socio-technical systems. *EPJ Data Science*, 8(1):2, 2019.

[222] Bitcoin. bitcoin-core. https://bitcoin.org/en/bitcoin-core/, 2019. Accessed: 3 July 2019.

[223] Blockchain. blockchain.com. www.blockchain.com, 2019. Accessed: 3 July 2019.

[224] Martin Dittus. www.oii.ox.ac.uk/blog/a-distributed-resilience-among-darknet-markets. https://www.oii.ox.ac.uk/blog/a-distributed-resilience-among-darknet-markets/, 2019. Accessed: 28 August 2019.

[225] Matthew A Napierala. What is the bonferroni correction. *AAOS Now*, 6(4):40, 2012.

[226] National Institute on Drug Abuse. Overdose Death Rates, January 2019.

[227] Jakob Demant, Rasmus Munksgaard, David Decary-Hetu, and Judith Aldridge. Going Local on a Global Platform: A Critical Analysis of the Transformative Potential of Cryptomarkets for Organized Illicit Drug

Crime. *International Criminal Justice Review*, 28(3):255–274, September 2018.

[228] Scott Higham, Sari Horwitz, and Katie Zezima. Obama officials failed to focus as fentanyl burned its way across America - Washington Post, March 2019.

[229] Silas W. Smith and Fiona M. Garlich. Chapter 3 - Availability and Supply of Novel Psychoactive Substances. In Paul I. Dargan and David M. Wood, editors, *Novel Psychoactive Substances*, pages 55–77. Academic Press, Boston, January 2013.

[230] United Nations Office on Drugs and Crime. Global Synthetic Drugs Assessment. Technical report, United Nations, 2017.

[231] Emcdda. European Drug Report 2019: Trends and Developments. Technical report, European Monitoring Centre for Drugs and Drug Addiction, 2019.

[232] Monica J. Barratt and Judith Aldridge. Everything you always wanted to know about drug cryptomarkets* (*but were afraid to ask). *International Journal of Drug Policy*, 35:1–6, September 2016.

[233] Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, 2008.

[234] Yan Carriere-Swallow and Felipe Labbe. Nowcasting with Google Trends in an Emerging Market: Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting*, 32(4):289–298, July 2013.

[235] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, March 2011.

[236] Zhi Da, Joseph Engelberg, and Pengjie Gao. In Search of Attention. *The Journal of Finance*, 66(5):1461–1499, October 2011.

[237] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014,

February 2009.

[238] T. Preis and H. S. Moat. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, 1(2), October 2014.

[239] Donald R. Olson, Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Computational Biology*, 9(10):e1003256, October 2013.

[240] Maimuna S. Majumder, Mauricio Santillana, Sumiko R. Mekaru, Denise P. McGinnis, Kamran Khan, and John S. Brownstein. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak. *JMIR public health and surveillance*, 2(30), 2016.

[241] David Molnar, Serge Egelman, and Nicolas Christin. This is our data on drugs: lessons computer security can learn from the drug war. In *Proceedings of the 2010 workshop on New security paradigms - NSPW '10*, page 143, Concord, Massachusetts, USA, 2010. ACM Press.

[242] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. User Intent, Behaviour, and Perceived Satisfaction in Product Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 547–555, 2018.

[243] Rachel S. Wightman, Jeanmarie Perrone, and Lewis S. Nelson. Comparative Analysis of Opioid Queries on Erowid.org: An Opportunity to Advance Harm Reduction. *Substance Use & Misuse*, 52(10):1315–1319, August 2017.

[244] Mitsuo Yoshida, Yuki Arase, Takaaki Tsunoda, and Mikio Yamamoto. Wikipedia Page Views Reflect Web Search Trends. In *Proceedings of the ACM Web Science Conference - WebSci '15*, pages 1–2, Oxford, United Kingdom, 2015. ACM Press.

[245] Robert Todd Perdue, James Hawdon, and Kelly M. Thames. Can Big Data Predict the Rise of Novel Drug Abuse? *Journal of Drug Issues*, 48(4):508–518, October 2018.

[246] Duilio Balsamo, Paolo Bajardi, and Andr Panisson. Firsthand opiates abuse on social media: Monitoring geospatial patterns of interest through a digital cohort. In *The World Wide Web Conference*, pages 2572–2579. Acm, 2019.

[247] Jinhui Zhao, Tim Stockwell, and Scott Macdonald. Non-response bias in alcohol and drug population surveys: Non-response bias in surveys. *Drug and Alcohol Review*, 28(6):648–657, May 2009.

[248] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports*, 3(1), December 2013.

[249] Marton Mestyan, Taha Yasseri, and Janos Kertesz. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLoS ONE*, 8(8):e71226, August 2013.

[250] Abeer ElBahrawy, Laura Alessandretti, and Andrea Baronchelli. Wikipedia and Digital Currencies: Interplay Between Collective Attention and Market Performance. *arXiv:1902.04517 [physics, q-fin]*, February 2019. arXiv: 1902.04517.

[251] Ladislav Kristoufek. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3(1), December 2013.

[252] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. Global Disease Monitoring and Forecasting with Wikipedia. *PLoS Computational Biology*, 10(11):e1003892, November 2014.

[253] Martin Dittus, Joss Wright, and Mark Graham. Platform Criminalism: The 'Last-Mile' Geography of the Darknet Market Supply Chain. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*,

pages 277–286, 2018. arXiv: 1712.10068.

[254] David Tsurel, Dan Pelleg, Ido Guy, and Dafna Shahaf. Fun Facts: Automatic Trivia Fact Extraction from Wikipedia. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, Wsdm '17, pages 345–354, New York, NY, USA, 2017. Acm. event-place: Cambridge, United Kingdom.

[255] Connor McMahon, Isaac L. Johnson, and Brent J. Hecht. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *Icwsm*, volume 11. AAAI Publications, 2017.

[256] Mark Graham, Stefano De Sabbata, and Matthew A. Zook. Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, 2(1):88–105, June 2015.

[257] Kristy Kruithof, Judith Aldridge, David Decary Hetu, Megan Sim, Elma Dujso, and Stijn Hoorens. Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands. Product Page, RAND Corporation, 2016.

[258] Binance.com. Binance, 2017.

[259] Dunamu. Upbit, 2017.

[260] Payward. Inc. Kraken, 2012.

[261] Sean Foley, Jonathan Karlsen, and Tlis J. Putni. Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *Ssrn*, 32(5):https://ssrn.com/abstract=3102645, 2018.

[262] Pavel Ciaian, Miroslava Rajcaniova, and dõArtis Kancs. The economics of bitcoin price formation. *Applied Economics*, 48(19):1799–1815, 2016.

[263] Gabriel Gajardo, Werner D Kristjanpoller, and Marcel Minutolo. Does bitcoin exhibit the same asymmetric multifractal cross-correlations with crude oil, gold and djia as the euro, great british pound and yen? *Chaos, Solitons & Fractals*, 109:195–205, 2018.

[264] Hermann Elendner, Simon Trimborn, Bobby Ong, Teik Ming Lee, et al. The cross-section of crypto-currencies as financial assets: An overview. Technical report, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, 2016.

[265] David Enke and Suraphan Thawornwong. The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, 29(4):927–940, 2005.

[266] Wei Huang, Yoshiteru Nakamori, and Shou-Yang Wang. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10):2513–2522, 2005.

[267] Phichhang Ou and Hengshan Wang. Prediction of stock market index movement by ten data mining techniques. *Modern Applied Science*, 3(12):28, 2009.

[268] Martin Gavrilov, Dragomir Anguelov, Piotr Indyk, and Rajeev Motwani. Mining the stock market (extended abstract): which measure is best? In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 487–496. Acm, 2000.

[269] K Senthamarai Kannan, P Sailapathi Sekar, M Mohamed Sathik, and P Arumugam. Financial stock market forecast using data mining techniques. In *Proceedings of the International Multiconference of Engineers and computer scientists*, volume 1, page 4, 2010.

[270] Alaa F Sheta, Sara Elsir M Ahmed, and Hossam Faris. A comparison between regression, artificial neural networks and support vector machines for predicting stock market index. *Soft Computing*, 7:8, 2015.

[271] Pei-Chann Chang, Chen-Hao Liu, Chin-Yuan Fan, Jun-Lin Lin, and Chih-Ming Lai. An ensemble of neural networks for stock trading decision making. *Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence*, pages 1–10, 2009.

[272] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[273] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[274] William Brock, Josef Lakonishok, and Blake LeBaron. Simple technical trading rules and the stochastic properties of stock returns. *The Journal of finance*, 47(5):1731–1764, 1992.

[275] Thomas Kilgallen. Testing the simple moving average across commodities, global stock indices, and currencies. *The Journal of Wealth Management*, 15(1):82, 2012.

[276] Blake LeBaron. The stability of moving average technical trading rules on the dow jones index. *Deriv. Use Trad. Regul*, 5:324–338, 2000.

[277] Craig A Ellis and Simon A Parbery. Is smarter better? a comparison of adaptive, and simple moving average trading strategies. *Research in International Business and Finance*, 19(3):399–411, 2005.

[278] George T Friedlob and Franklin J Plewa Jr. *Understanding return on investment*. John Wiley & Sons, 1996.

[279] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. Acm, 2016.

[280] Kaggle Inc. Kaggle, 2018.

[281] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[282] Bitcoin Wiki. Comparison of exchanges., 2017.

[283] Dániel Kondor, István Csabai, János Szüle, Márton Pósfai, and Gábor Vattay. Inferring the interplay between network structure and market effects in bitcoin. *New Journal of Physics*, 16(12):125003, 2014.

[284] Andrew Urquhart. What causes the attention of bitcoin? *Economics Letters*, 166:40–44, 2018.

[285] Tianyu Ray Li, Anup S Chamrajnagar, Xander R Fong, Nicholas R Rizik, and Feng Fu. Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. *arXiv preprint arXiv:1805.00558*, 2018.

[286] Pedro Franco. *Understanding Bitcoin: Cryptography, engineering and economics*. John Wiley & Sons, 2014.

[287] Bitcoin miners ditch ghash.io pool over fears of 51% attack. http://www.coindesk.com/bitcoin-miners-ditch-ghash-io-pool-51-attack/, 2014. Accessed: 5 June 2017.

[288] Bitcoin energy consumption index. http://digiconomist.net/bitcoin-energy-consumption, 2017. Accessed: 15 May 2017.

[289] Map of coins. https://www.mapofcoins.com, 2013. Accessed: 5 June 2017.