Complexity science approach to study decentralized financial systems using tools from statistical physics and machine learning

Aspembitova, Ayana T.

2020

Aspembitova, A. T. (2020). Complexity science approach to study decentralized financial systems using tools from statistical physics and machine learning. Doctoral thesis, Nanyang Technological University, Singapore. <https://hdl.handle.net/10356/148631>

<https://hdl.handle.net/10356/148631>

<https://doi.org/10.32657/10356/148631>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 21:53:49 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**COMPLEXITY SCIENCE APPROACH TO
STUDY DECENTRALIZED FINANCIAL
SYSTEMS USING TOOLS FROM STATISTICAL
PHYSICS AND MACHINE LEARNING**

AYANA T. ASPEMBITOVA

**SCHOOL OF PHYSICAL AND MATHEMATICAL
SCIENCES**

2020

**COMPLEXITY SCIENCE APPROACH TO
STUDY DECENTRALIZED FINANCIAL
SYSTEMS USING TOOLS FROM STATISTICAL
PHYSICS AND MACHINE LEARNING**

AYANA T. ASPEMBITOVA

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

A thesis submitted to the Nanyang Technological University in
partial fulfilment of the requirement for the degree of Doctor of
Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

18 Nov 2020

.....
Date

.....
Aspembitova Ayana

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

18 Nov 2020

.....
Date

.....
Assoc. Prof. Chew Lock Yue

Authorship Attribution Statement

Please select one of the following; *delete as appropriate:

~~*(A) This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.~~

*(B) This thesis contains material from 2 papers published in the following peer-reviewed journal.

Chapter 5 is published as A. Aspembitova, L. Feng, V. Melnikov and L.Y. Chew (2019). Fitness preferential attachment as a driving mechanism of bitcoin transaction network. PLOS One 14:1-20. <https://doi.org/10.1371/journal.pone.0219346>

Chapter 6 is published as A. Aspembitova, L. Feng, and L.Y. Chew (2021). Behavioral structure of users in cryptocurrency market. PLOS One 16:1-19. <https://doi.org/10.1371/journal.pone.0242600>

The contributions of the co-authors are as follows:

- Assoc Prof Chew and Dr Feng provided the initial project direction and edited the manuscript drafts.
- I co-designed the study with Assoc Prof Chew and Dr Feng and performed all the simulation and analytical work at the School of Physical and Mathematical Sciences and at the Institute of High Performance Computing.
- I prepared the manuscript drafts. The manuscript was revised by Assoc Prof Chew and Dr Feng.
- V. Melnikov processed the data for research.

18 Nov 2020

.....
Date

.....
Aspembitova Ayana

Abstract

Decentralized Finance is the new socioeconomic system growing with an extremely fast pace and changing the way financial interactions are being conducted. In comparison with the growing importance of digital assets and blockchain technology, there is still little understanding of Decentralized Finance as a *system*. In this thesis we analyze transaction datasets from Bitcoin and Ethereum blockchains to obtain a comprehensive understanding of digital assets - from studying the behaviour of each part to investigating the whole structure and deriving the relations between micro and macro properties of the cryptocurrency systems. Using the Complex Networks approach we explained the system's overall structure and dynamics, and uncovered the mechanism behind network formation. It was found that there is fitness preferential attachment among nodes in the bitcoin network that leads the system to scale-free behaviour. We proposed the quantifiable definition of fitness and supported our finding by simulating a synthetic network and reproducing the main properties of the bitcoin network. After having a good understanding about the structure of the system, we zoom in into its parts by studying the behavioral patterns among the system's users (people). We develop the methodology based on Machine Learning models to define distinct behavioral types in the cryptocurrency systems and find that despite differences between the bitcoin and ethereum systems, there are *four* common strategies that users follow in both markets. Based on our finding, we model the dynamics of people's behaviour in market as an Absorbing Markov Chain. This approach allowed us to present the behavioral switches in a comprehensive and intuitive way. Moreover, we were able to obtain the predictions on the longevity of users in the system according to their behaviour. Finally, we use the Granger causality test to derive the relations between all system characteristics. We attempt to explain the effect of behavioral switches on the structural properties and price; we find that indeed, switches of users from certain behavioral groups causes a change in price which affects the size of the network as well. We hope that the work and results presented in this thesis will advance the understanding of the new field of Decentralized Finance and expect that the research approach and methodologies developed for this study will be helpful to investigate various complex systems as well.

Acknowledgments

My PhD time has been an exciting journey, where I was surrounded by vibrant and warm community. This thesis reflects the help and support from many people, and most importantly, from my supervisors – Dr. Feng Ling and Prof Chew Lock Yue. Their invaluable guidance and kind personalities made this journey fruitful and pleasant.

Ling, first of all thanks for your exceptional scientific work that inspired me to start my PhD and work my full potential. For guiding me through my PhD with care and patience, while giving enough freedom and independence. I have learnt a lot from you!

Lock Yue, thanks for your guidance, patience and constant encouragement. For being the kindest supervisor ever, for generously sharing your knowledge and for many insightful discussions. Your support and valuable advises helped me a lot in my professional and personal development.

I would like to acknowledge the Agency of Science, Technology and Research (A*STAR) Graduate Academy for providing me with the scholarship for my PhD studies. And I would also like to acknowledge the Institute of High Performance Computing for the fruitful research attachment.

Special thanks to Dr. Christopher Monterola for his help and guidance in the beginning of my PhD journey. Thanks to my past and present groupmates in Nonlinear and Complex Systems Lab – Cheryl Abundo, Quek Wei Liang, Khi Pin, He Chong, Luca Vismara, Matthew Ho Shu Hui, Suryadi, Andri Pradana, Ritvik Dobriyal, Vee-Liem Saw and Lian Jie for the fruitful discussions and for creating warm and friendly atmosphere in our little windowless office.

Apart from my supervisors and colleagues, there are many people who have contributed to this thesis by sharing their knowledge and providing the feedback. I am very grateful for: My friend and collaborator Valentin R. Melnikov for his help with processing huge datasets, for his support and valuable comments. My TAC members – Prof Lai Choy Heng and Prof Yong Ee Hou for their time, critical comments, sharing their ideas and advises. My thesis benefited a lot from our TAC meetings. Prof Cheong Siew Ann for his review of my study in the early stage and his feedback, Prof Jose F. Mendes for very interesting discussions on the network part of my thesis, Prof Peter M. Slood for inspiring conversations and discussions, Prof Yaneer Bar-Yam and Prof Dan Braha for interesting lectures and their time to provide insightful feedback on my research.

As with any achievement in my life, a special thanks always go to my beloved professors from Tomsk – Valentin G. Barannik and Natalia F. Barannik. Thanks for nurturing me from my early teens and always having unwavering faith in me. Thanks to my Master's thesis supervisor Elena V. Istigecheva for her help with every aspect in my life and for creating many opportunities to help me to grow both professionally and personally. To my

professor Victor V. Kryssanov from Ritsumeikan University for inspiring and encouraging me to realize my potential, for his wise advises and his help to start this journey. To my friend and mentor - professor Tatiana Pashinskaya for teaching me a lot about financial modelling, for her support and help.

This thesis would not have been possible without my big and supportive family. I would like to say thanks to the most important people in my life: To my father, Dr. Tleukan E. Aspembitov for his truly unconditional love and for supporting my every single decision in life. Papa, thanks for everything you do for me, for everyday making sure that I am feeling happy, eating healthy and exercising regularly (and properly). You are my best friend, my adventure buddy, my fitness coach, my life teacher and the best father in the world. To my mother, Ekaterina S. Aspembitova for her love and endless support, for being an inspirational example of beautiful, intelligent, strong and feminine woman. To my sister Ayagul T. Batyrova for her kindness, silliness and for random parcels with beautiful presents. To my strong and protective brother Dosymzhan Aspembitov for always being my pillar of strength and his wife Asel for her support. To my kids – Rishat, Dilyara, Madina, Azamat, Ameliya and Rihanna for the joy and laughter they bring to my life every day. To my friends – Elena Emelyanenko, Sean Krupp, Victor Yurov, Parastoo Fahimi, Luca Pignata and Palina Tolmach for their help and support.

Эта докторская диссертация посвящается моим родителям - моему любимому папочке Аспембитову Тлеукану и моей дорогой мамочке Аспембитовой Екатерине

Contents

1	Introduction	2
1.1	Overview of the current state of Financial Systems Studies	5
1.2	Complexity Science approach and Methodology	7
2	Blockchain and Cryptocurrencies	9
2.1	Decentralized Financial Systems and Blockchain technology	9
2.1.1	Blockchain 1.0 and Bitcoin	14
2.1.2	Blockchain 2.0 and Ethereum	16
2.2	Obtaining Blockchain Data	18
2.2.1	Data Extraction and Processing	18
2.2.2	Data Description	20
2.3	Conclusions	20
3	Stylized Facts in Decentralized Financial Markets	22
3.1	Introduction	22
3.2	Absence of Autocorrelation in returns	23
3.3	Volatility clustering	24
3.4	Power-law probability distribution of returns	26
3.5	Power-law probability distribution of trading volume and number of trades	26
3.6	Conclusions	27
4	Network Analysis of Cryptocurrency Systems	28
4.1	Empirical studies of Networks	29
4.2	Static Network Fundamentals and Main Properties of Cryptocurrency Net- works	33
4.2.1	Static Network Representation	34
4.2.2	Network Visualisation	35
4.2.3	Main Properties of Cryptocurrency Networks	38
4.2.4	Robustness in Cryptocurrency Networks	43
4.3	Temporal Networks	44
4.3.1	Definitions and Properties of Temporal Networks	45
4.3.2	Temporal properties of cryptocurrency transaction networks	50

4.4	Conclusions	53
5	Mechanism of Bitcoin Network Formation and Evolution	55
5.1	Models of Network Formation	55
5.2	Mechanism behind the Bitcoin transaction network evolution	58
5.2.1	Degree Attachment Mechanism	60
5.2.2	Wealth Attachment Mechanism	62
5.2.3	Fitness Distribution of Network Nodes	63
5.3	Generative algorithm for the Bitcoin Transaction Network	66
5.3.1	Algorithmic model for bitcoin network generation	66
5.3.2	Synthetic Network simulations results	67
5.4	Conclusions	70
6	Behavioural Structure of Cryptocurrency System	72
6.1	Background on Modelling People's Behaviour in Financial Systems	72
6.2	Pre-analysis for Behavioural Classification	75
6.2.1	Periods for Analysis	75
6.2.2	Feature Extraction and Analysis	76
6.2.3	Review of Methods	80
6.3	Methodology and Results	82
6.3.1	Methodology	82
6.3.2	Results and Discussion	84
6.4	Dynamics of Users' Behaviour as Absorbing Markov Chain	90
6.4.1	Daily Behavioral Composition and Switching Mechanism	90
6.4.2	Modelling the Dynamics of Users' Behaviour	92
6.5	Conclusions	95
7	Causal Relations between Market Effects, Users' Behavior and Network Properties	97
7.1	Time Series of System's Evolution and Correlation Analysis	97
7.2	Background of Granger Causality Test	101
7.3	Causality Analysis Results	102
7.4	Conclusions and Discussions	104
8	Conclusions and Future Work	106

List of Figures

2.1	Comparative Review of Traditional and Decentralized Finance	10
2.2	Schematic illustration of bitcoin block structure.	15
2.3	Schematic illustration of ethereum block structure.	17
3.1	Autocorrelations of BTC and ETH log returns of price.	24
3.2	Evolution of logarithmic returns of BTC and ETH.	25
3.3	Evolution of absolute returns of BTC and ETH in a weekly time scale. . .	25
3.4	Probability distributions of logarithmic returns of BTC and ETH.	26
3.5	Probability distributions of trades' volumes in BTC and ETH.	27
4.1	Network structure of bitcoin transactions in the early period of system development (2009-2010).	36
4.2	Network structure of bitcoin transactions without isolated nodes in the early period of system development (2009-2010).	36
4.3	Visualization of bitcoin transactions network structure in the mid period of its development (January 2011).	37
4.4	Network structure of ether transactions in the early period of system development. One month (Sep 2015) is visualised.	37
4.5	Visualisation of one day ether transactions network structure in the late period of its development (January 1, 2017).	38
4.6	Close-up visualization of one day ether transactions network structure in the late period of its development (January 1, 2017).	38
4.7	Degree distributions empirically calculated for bitcoin separately for the two periods and shown as blue lines in the form of complementary cumulative distribution functions (CCDFs). The red lines show a power law function with an exponent of -2.	39
4.8	Degree distributions empirically calculated for ethereum separately for the two periods and shown as blue lines in the form of complementary cumulative distribution functions (CCDFs). The red lines show a power law function with an exponent of -2.	40
4.9	Degree assortativity mixing in BTC and ETH networks.	41
4.10	Wealth assortativity mixing in BTC and ETH networks.	42
4.11	Monthly assortativity coefficients.	51
4.12	Monthly clustering coefficients for bitcoin and ether networks.	51
4.13	Evolution of the monthly network average degree.	52

4.14	The daily evolution of the total number of nodes and links and the history of the bitcoin market price.	52
4.15	The daily evolution of the total number of nodes and links in ether network and the history of the ether market price.	53
4.16	Longevity distribution for nodes in bitcoin and ethereum.	54
5.1	Relation between the number of nodes and links for daily data.	59
5.2	Illustration of a toy network.	60
5.3	Illustration of the growth of toy network based on various attachment mechanisms.	61
5.4	Normalized degree attachment probabilities to new and old (existing) nodes in both periods.	62
5.5	Normalized wealth attachment probabilities for attachment to new and old (existing) nodes in both periods.	63
5.6	The fitness distributions computed for all nodes month by month in the second (stable) period of bitcoin transactions and shown in CCDF (blue line). The Red line shows the power-law function with an exponent of -2.	64
5.7	Evolution of the fitness coefficient for all nodes.	65
5.8	Evolution of the network growth rate R	67
5.9	A comparison of the complimentary cumulative degree distribution between the empirical and synthetic network for both periods. All distributions are observed to display power-law behavior. The red line shows the power law function with an exponent of -2.	69
5.10	A comparison on the clustering coefficient distribution between the empirical and synthetic network for both periods. All distributions are observed to display power-law behavior. The red line shows the power law function with an exponent of -2.	70
6.1	Proportion of “noisy” users in transaction networks.	78
6.2	Correlation between all features in ETH.	79
6.3	Correlation between all features in BTC.	79
6.4	The algorithmic steps to classify users into behavioral groups.	84
6.5	Number of clusters in ETH data found using the elbow method.	85
6.6	Number of clusters in BTC data using the elbow method.	85
6.7	Silhouette coefficient calculated for filtered ethereum data set.	87
6.8	Silhouette coefficient calculated for filtered bitcoin data set.	87
6.9	The percentage of users in different groups in the in bitcoin and ethereum during different periods of systems’ evolution.	89
6.10	The example of daily percentage of users in different groups in the bitcoin and ethereum	90
6.11	State transition diagram.	91
6.12	Percentage of absorbed users predicted correctly by the fundamental matrix.	95
7.1	Systematic Overview of obtained time series.	98

7.2	All time series that describe the Ethereum system's dynamics. All properties are calculated in daily time resolution	100
-----	--	-----

List of Tables

2.1	An example of bitcoin's dataset obtained using BitIodine software.	20
2.2	An example of ether's dataset.	20
4.1	Properties of different networks with scale-free behaviour.	33
4.2	Calculated values for α for the in, out, and overall degree distributions for the first (1P) and second (2P) periods.	39
4.3	Various approaches of calculating main properties in Temporal Networks in comparison with the traditional Static Network.	47
5.1	Correlation between the fitness, degree, and age for all the nodes in the network by the end of the second (stable) period.	65
5.2	A comparison of basic network properties between the synthetic and empirical bitcoin network.	69
6.1	Periods when bitcoin and ethereum systems show significant and non-significant price changes.	75
6.2	Network properties calculated for bitcoin (BTC) and ethereum (ETH) networks for each period under analysis.	76
6.3	Distinct properties of the majority of nodes in each cluster	86
6.4	The percentage of misclassified points shown in comparison with various feature sets used.	88
6.5	Table of matrix notations	92
7.1	Summary of all time series	99
7.2	Table of p-values for Granger causality test	104

Chapter 1

Introduction

Decentralized Finance is a new socioeconomic system that has been developing at an extremely fast pace starting from the introduction of bitcoin cryptocurrency a decade ago [1]. Originally, there was a big stigma around cryptocurrencies and blockchain technology being used mostly for suspicious activities. Slowly, the revolutionary idea behind the blockchain technology was realised - it utilises advanced cryptography algorithms to solve the consensus problem without trust (Byzantine Generals' Problem) that makes any transactions independent from intermediary organizations. Bitcoin and many other cryptocurrencies using the same technology are solely used for financial transactions and represented as an alternative monetary system. An ambitious application of blockchain technology was proposed by Vitalik Buterin who introduced Ethereum [2] - Turing-complete blockchain that executes smart-contracts (codes) to solve any reasonable computational problem. The Ethereum blockchain can be used not only for monetary transactions but also for creating and trading various decentralized financial assets and moreover, it has several non financial applications as well. Nowadays, the field of Decentralized Finance is extremely popular with thousands of various assets being introduced and attracting attention of financial institutions, investors and regulatory entities, and spurring research interest from scientists from different fields.

Overall, there are several research directions related to the study of Decentralized Finance, Blockchain and cryptocurrency markets. This field heavily relies on technological and computational advances and currently blockchain technology has many problems that were revealed by the Bitcoin blockchain - scalability issues, increasing transaction fees and most importantly, the ecological consequences of mining due to the consensus algorithm called "*proof-of-work*". New ways and paradigms are currently being actively explored to solve these problems [3–5]. Also there are privacy and security issues [6–9]

and other problems related to the applications and protocols [10, 11]. Another research direction is related to the study of the economic nature of cryptocurrency systems, since currently blockchain is mostly used as an instrument for financial applications. With the increasing popularity and credibility of decentralized financial instruments, bitcoin has spurred great interest from scientists from the economic and finance fields, and their contributions are focused on financial risks [12–14] and the descriptive statistics [15–17]. Also, there are many studies dedicated to the trading and price prediction problems using various methods [18–24].

In this thesis I present the physics and complexity science perspective of studying the new field of Decentralized Finance. My first goal is to provide a comprehensive and systematic understanding of the new socioeconomic system. While security, technological efficiency, price prediction, financial risks and trading in cryptocurrency systems are quite popular research topics, the systematic study of cryptocurrency markets is still scarce. In [25–29] the Complex Network approach was used to investigate the overall structure of the bitcoin transaction network and derive important statistics with scale-free behaviour being the most prominent one. Comprehensive analysis of the whole cryptocurrency market has been done in the study of [30], where the authors looked at various market time series and proposed a model for the cryptocurrency ecology. The experimental study of cryptocurrency market dynamics has been done in [31] by placing random orders in an exchange platform and studying the susceptibility of users to peer influence. The dynamics of all financial instruments within the Ethereum blockchain have been analyzed in [32] and it has been found that the decentralized market functions as a coherent one. In the study [33] authors aim to investigate the disorder and complexity of the whole cryptocurrency market using the complexity-entropy causality plane. Authors have found that different cryptocurrencies have very different behaviour over time from being deterministic to having high entropy and zero complexity (unpredictable). The systematic investigation of cryptocurrencies’ price evolution has been also studied from various perspectives - in [34] authors aimed to detect the chaos, randomness and temporal correlations in bitcoin price movements. They found that chaos is present in both, stable and turbulent periods, while multi-fractality was found to be the consequence of fat-tailed return distributions. In [35–39] authors studied various statistical features of bitcoin price from physics perspective. Early warning signals for price fluctuations were studied in [40–42]. In the study [40] the properties derived from the transaction network were found to be useful to predict the extreme price volatility. In [41] the critical transition was found to be associated with the jumps in standard deviation of price residual time series.

Overall, there is a growing understanding about the structure of decentralized financial markets, how they function and some insights about the relation between various decentralized financial instruments. However, there are still many open problems related to the lack of systematic understanding - why do we observe the same structural properties in cryptocurrency system as in other, completely different, real-world systems; and is that a good thing for financial system to have these striking similarities? How do people behave in the system and how do their behavioural patterns affect the overall system's dynamics? What is the relation between behavioral patterns and the system's structure and evolution? What is the cause and what are the consequences? Along with these, there are still many more questions which remain open. As the field of Decentralized Finance grows at a very fast pace, attracting more and more participants from individual and corporate levels, there is a high possibility for us to slowly switch to the new blockchain paradigm. Therefore, it is crucial not only to solve current technological problems, but also it is equally important to have a clear understanding of the systematic properties in order to create the safe financial system. Moreover, it needs to be emphasized that Decentralized Finance is a *self-organized* system with no external power and therefore, understanding how parts of the system (people) react to their environment and how they behave is the key point to obtain a comprehensive picture of the system. In this thesis we aim to contribute to the systematic understanding of cryptocurrency systems and provide answers to the questions listed above.

Another goal of this study would be to contribute to the understanding of financial systems in general. Although financial systems are one of the most extensively studied areas with many models and theories developed over the centuries, we are still struggling to explain various phenomena. The lack of information availability related to financial interactions was one of the biggest challenges to study financial systems as this kind of data is either proprietary or considered to be highly sensitive. Blockchain technology provided an opportunity to freely access information about financial transactions. Therefore, we would like to use this opportunity to provide more insight into a better understanding of financial markets by drawing a parallel between traditional and decentralized financial markets.

To achieve our research goals, we choose to use datasets from the two largest blockchains - Bitcoin blockchain and Ethereum blockchain, since the first is the most popular and the latter represents the latest advance of the Decentralized Finance field. We apply the *Complexity Science* approach to obtain a systematic and comprehensive understanding of the system and rely on the variety of tools and knowledge from Statistical Physics, Data

Science, Econophysics, Complex Networks and Machine Learning.

The thesis starts from a discussion of what is the blockchain and what are cryptocurrencies. In Chapter 2 I provide a thorough explanation of these technologies and introduce the dataset used for this research. Understanding of the terms and nature of the dataset will be important to follow Chapters 4, 5 and 6. Then in Chapter 3, I will describe the similar properties found in both traditional and decentralized finance. This similarity would be one of the motivations to study behavioral composition described in Chapter 6. The next step of the analysis is to obtain the structural properties of a system - we use Complex Networks tools to look at the transaction networks and derive various statistics in Chapter 4. Interesting insights gained after Network analysis inspired us to investigate the structural properties deeper and understand the mechanism behind the system's formation - described in Chapter 5. After we have a great understanding of the system's macro properties, we zoom in into the parts of system (people) by investigating their behavioral patterns - in Chapter 6 we propose a methodology based on Machine Learning models to derive distinct behavioral types and we construct a Markov Chain model to describe the dynamics of people's behaviour. Finally, after understanding both micro and macro properties of the system, we aim to derive causal relations between them - Chapter 7 shows the preliminary results of the analysis.

The rest of this Chapter consists of two sections - the first one is dedicated to providing an overview of the current state of understanding of Financial Systems in general; and second section explains the underlying research approach and the choice of used methods.

1.1 Overview of the current state of Financial Systems Studies

Financial systems are extremely complex ones with diverse types of participants (people, banks, firms etc.) that constantly interact with each other, forming different patterns and with hidden inter-dependencies between social and technological developments. Researchers from various fields have extensively studied financial systems from different perspectives both qualitatively and quantitatively. The physics approach has provided some unique perspectives in the development of financial modelling - the first application of physics concepts to finance goes as far back as 1900 when Louis Bachelier in his PhD thesis introduced the stochastic process applied to the valuation of stock options [43]. Later, mathematician Andrey Markov introduced the prototypical memory-less stochastic

process [44]. Based on these studies, many important findings were made like the Wiener process (a continuous type of Markov stochastic process) and the Ito process (a generalized Wiener process). Insights based on stochastic processes gave rise to the development of quantitative models for finance, starting from the revolutionary Black-Scholes-Merton model that was introduced to price the European type of options. Finally, the field of *Quantitative Finance* was formed that uses advanced mathematics to develop models for pricing various financial instruments and their derivatives. As more complicated mathematical models were used to forecast profits and risks, there were no attempts made to understand the financial system as a whole. As a result, the complexity of instruments and models grew as a matryoshka doll and led to financial crises. Later, physics concepts and approaches to study economic and financial systems formed a separate scientific field - *Econophysics*, which can be considered as the first attempt to assemble the pieces of a financial system together and understand the whole behaviour of the system. One of the greatest contributions of Econophysics is the discovery of *stylized facts* (universal properties) in financial time series like power-law distributions of returns, number of trades and trading volumes, volatility clustering, long memory in absolute and squared returns [45–48] etc. Many models were developed by econophysicists to explain these properties, assuming that the interaction between the system’s parts (or agents) are the main cause of them (*agent-based modelling*) - there are many models developed starting from the idea of fundamentalist-chartist interactions [49, 50] to modeling the effect of herding behaviour [51–54] and incorporating agent-based modelling to the traditional stochastic models [55] etc. These models are based on the concepts of Statistical Physics (when micro effects lead to large consequences), while parts of the system (agents) were modelled as heterogeneous and complex ones. Overall, Econophysics and agent-based modelling had a positive contribution towards explaining many interesting phenomena observed in financial markets. However, there are shortcomings of this approach and the main one would be that models are based on hypothetical assumptions that a person’s behavior can be subjected to confounding interpretations. There are also methodological problems related with the difficulty to calibrate and validate agent-based models [56, 57].

Nowadays, with the development of blockchain technology that allows us to observe all financial interactions freely and in great details, we can move to the next step of understanding financial systems - based on data-driven empirical modelling.

1.2 Complexity Science approach and Methodology

Blockchain Technology opened an opportunity to look at the detailed information about the interactions of market participants in very high time resolution. However, only having information is not sufficient, a powerful research approach is needed to extract meaningful patterns and derive an understanding of this interconnected system - we choose to use the Complexity Science approach as a framework for this study. Complexity Science is the relatively new and fast growing field that aims to understand properties of complex systems and derive universal laws for them. “Complex system [is a system] made up of a large number of parts that interact in a nonsimple way” - the famous definition given by Herbert Simon in his paper in 1962 [58]. There are core concepts of Complex systems - emergence and self-organization. *Emergence* is about the system’s properties at different scales, about the relationships between the details of the system and larger picture. It does not state that macroscopic properties are more important or vice versa, it is concerned with the relationships between the two. Another key idea of Complex systems is *self-organization* - it is when the system spontaneously, without any external design organizes itself and produces nontrivial behaviour with time. Complexity Science tries to understand the relationships between components, how these relationships give rise to the collective behaviour of system and how self-organizing processes take place. To achieve this goal, Complex Systems Science utilizes tools from various disciplines and develops new conceptual, mathematical and computational models to describe the behaviour and properties of a system. This approach has contributed to the understanding of how various complex systems work with examples from biology and ecology [59–61], economical and social systems [62–67], transportation systems [68–70] etc. Complexity Science has also played a huge role in solving various practical problems, where many examples of which can be found in the book of Bar-Yam [71].

As one of the goals of Complexity Science is to understand the overall structure and dynamics of a system, the *Complex Networks* is a great tool as many complex systems can be represented as networks under the condition that information is available about the parts of the system and their interaction. This modelling approach allows one to zoom out from the micro details and obtain a big picture of the system by representing parts of the system as dots (nodes) and any connection or interaction between the parts as lines (links). Although this is a very simplified explanation, when large systems are modelled as networks, various useful statistics can be derived that describes the system at a macroscopic level [72–78]. Network theory is fundamentally different from the tradi-

tional dynamical models (e.g. cellular automata, continuous field models) Components of system does not have to be connected uniformly and regularly (like in cellular automata). This non-homogeneous connectivity makes it more challenging to use math to analyze the system’s properties. However, Network theory allows to represent the connections in system more realistically by specifying in details which component connects to which and how they are connected. This makes the implementation of Network theory quite data-intensive. To manage and manipulate this information in efficient way requires new tools and models, I will explain in fuller details existing methods in Chapter 4. Also, number of components may increase or decrease over time and change the structure of the network. To capture this property of real world networks, the theory of *Temporal Network* has been developed. It is the more recently developed approach that provides a great opportunity to study the dynamics of complex systems by adding time stamp information to the links. Although the theoretical base is still under development and there are many open questions [79–83], the overall attitude of modelling systems proposed by this field has already helped solve various practical problems [27, 84, 85].

As it was discussed earlier, to understand the collective behaviour of all parts is crucial, but it is equally important to understand the behaviour of each part and the effect of their interaction on the system’s evolution. While in a simple system the behaviour of each part (like atoms or molecules) is well-known and all units of the system show similar behaviour, in a complex system each unit is complex by itself and there could be many diverse types of units within the system. Having detailed information about each user of a decentralized financial system makes it possible to distinguish different behavioural types in the system and this is similar to the *classification* problem in the Machine Learning field. This approach to investigate the behavioral structure of a system is different from agent-based modelling, where the heterogeneity of users is assumed by the modeller, and from the experimental studies of financial markets [31, 86, 87], which has the downside of sample size limitation. Using Machine Learning methods on the blockchain transaction data allows us to comprehensively investigate the *real* heterogeneity of market participants and observe the patterns of their behavior across different times of the system’s evolution.

In summary, this thesis contributes to the understanding of the structural properties of cryptocurrency systems and the behavioral properties of its users. As this is the first time obtaining such information in financial systems, we believe this thesis will overall advance financial modelling as well.

Chapter 2

Blockchain and Cryptocurrencies

Blockchain is a technological invention that became extremely popular just in few years. Initially, this technology was applied for the decentralization of currency and payments, and the first currency based on this technology was the Bitcoin. Although it has been a decade since the successful start of the blockchain era, there is still some confusion when it comes to the terminology – Bitcoin and blockchain words sometimes used interchangeably in general public to refer to the underlying technology, to the actual cryptocurrency (money) or also to the whole concept of cryptocurrencies. To make it less confusing for a reader, I will first provide brief comparative review between the traditional centralized and decentralized financial systems based on the blockchain. Then, I will explain thoroughly about the technology itself, which in fact goes beyond of its most popular application - cryptocurrencies. Finally, I will discuss its evolution and the opportunities blockchain provides at each step of its development.

2.1 Decentralized Financial Systems and Blockchain technology

Although the blockchain technology can go far beyond of financial applications, the Decentralized Finance is still one of its most important use cases. Blockchain was able to provide the strong alternative to the traditional finance by providing more attractive and fair conditions for system users. The Figure 2.1 shows the main differences between the traditional and decentralized finance. In Decentralized Finance all users have trust in technology, while in traditional one people trust their funds to the banks and institutions. As a result of this fundamental difference, many other distinct properties arise in

Decentralized Finance - users manage their funds independently, without relying on any custody, they do not need permission to enter the system and moreover, every user has an opportunity to contribute to the field by creating new financial products, which makes the field to innovate in a very fast pace.

Figure 2.1. Comparative Review of Traditional and Decentralized Finance

To show how these attractive properties are implemented from the technical viewpoint, I will dedicate this Chapter to explain how blockchain works, what's the difference between Bitcoin and Ethereum blockchains and also provide brief explanations on cryptography behind the system.

Overall, blockchain can be seen as a huge database with the specially designed storage structure. This database is open for public and anybody can contribute to it. Let's imagine there is some information that needs to be recorded - it can be any deal, agreement or transaction between two parties. In traditional system, we would need a third party to approve it (some related entity, like company or government). The innovative idea behind the blockchain technology is that it does not depend on a third party and also, it does not require trust from participants as it operates as peer-to-peer network of users connected through the internet. It is a *Distributed Network* where each user participates in creating and verifying new records. It is almost impossible to tamper the information once it is stored in blockchain. All these attractive features make the blockchain technol-

ogy applicable to many various fields like financial transactions, smart contracts, public services, Internet of Things, reputation and security services.

Each new record in blockchain lists all necessary details about the transaction. After record is made, it goes through the check by the whole network of users. Transparency of transaction history is another difference from the traditional system - every user can observe and download the previous transaction records in blockchain system. This helps to perform the validation of new records - all nodes (users of system) check the details of new record and approve it. After the record has been approved, it can be added to the block. Each block contains its unique identifier (hash), information about the new records and also the information about the previous block in form of the hash. These hash codes connect blocks together forming a chain of blocks.

This is a simplified explanation of what is the blockchain and how it operates. I would like to discuss the main concepts step by step in greater details, explain the cryptography behind it and also briefly discuss the cost and current problems of the technology. Overall, blockchain is the combination of existing technologies and ideas formed into the technology with the innovative functions.

Digital Signature Traditionally, transaction is when two parties make an agreement on something and sign it for verification. In blockchain environment, the logic of transaction is the same, but the process fully relies on cryptographic algorithms - first, information about the transaction goes through the hashing algorithm and then parties digitally sign the agreement.

Transaction is the record that shows identities of both sender and receiver and the value to be redistributed. Cryptographic *Hash function* (function that maps the data of any size to a bit array of fixed size and it is practically impossible to go back to the original text) reverts this input information with transaction details to an unique binary output using some cryptographic algorithm.

Then, each transaction needs to be digitally signed by both parties. The purpose of the Digital Signatures is the same as of the traditional signatures - to prove someone's identity. The difference is that the digital signature uses math (cryptography) to do so and therefore, it is more secure. Blockchain uses the asymmetric Digital Signature Cryptography Algorithm (DSA) - sender needs a public key of the recipient in order to send the money, and the recipient uses his private key to receive the transaction. Key generation has two phases. The first phase is a choice of algorithm parameters which may be shared between different users of the system, while the second phase computes public and private keys for a single user. Public key can be seen as an analogy to the

bank account number, so that everyone who knows account number can send money to it. Private key is like a card PIN number, only owner can have an access to his wallet.

This Digital Signature Algorithm ensures:

1. *Authentication* - the recipient can verify that the message came from the sender.
2. *Non-repudiation* - the sender can not deny sending the message.
3. *Integrity* - the message can not be tampered with.

Blockchains typically use the Elliptic Curve Digital Signature Algorithm (ECDSA) [88]. ECDSA consists of combining the math behind finite fields and elliptic curves to create one way equations to ensure that funds can be spent only by their owners. The private key in blockchain is the arbitrarily chosen number between 1 and the order above. Public key can be easily derived from the private key: $PublicKey = PrivateKey \times P$, where P is the base point. It should be noted that Blockchain uses extremely large numbers in hexadecimal points as a base. Once private key is generated, it is multiplied by the base P to obtain the public key. Properties of elliptic curves makes the Digital Signature algorithm more secure, making it computationally harder to work backwards to obtain the private key from the known public key.

Block structure After the transaction has been made and both parties signed it, the record goes to the block. Each block can consist many records of transactions. To reduce the on-chain storage, the *Merkle tree* cryptographic data structure is used. Merkle tree represents the set of transactions as a binary tree to store them in block. Each leaf in the tree is the hashcode of a transaction, the root is the hash value of all leafs (transactions) in this block. Such organization makes records lightweight and therefore, it is easier to validate the transactions on later step. Different blockchains use different variations of the Merkle tree (binary, hexary etc.).

Apart from the Merkle tree root hash, each block also contains the information about the timestamp, the parent block hash (256 bit value that points to the previous hash) and a nonce. Nonce is a random number, starts from 0 and increases for every hash calculation. Blocks may contain more information, depending on the application it is used for [89] . In later sections, different blockchain architectures will be discussed in greater details.

Consensus in Blockchain The unique structure of data storage - having a hash of the previous block as a part of current block, makes it difficult to tamper the information, since it will lead to the necessity to change the information in all blocks. However, it is still possible to do using powerful modern computers. To make system even more secure, there is an additional important step that is made before adding the new block with

transaction records into the blockchain. It is called consensus in the blockchain network, or *mining*. Cryptographically, mining consists in finding a string of 256 bit length using certain cryptographic algorithm, based on the all input information given. Overall, the role of hash-functions is to enforce the discipline in writing transaction records in the blockchain. Hashing also ensures:

1. *Pre-image resistance* - given a hash h it is difficult to find a message m such that $hash(m) = h$.
2. *Second pre-image resistance* - given message m_1 it is difficult to find a different message m_2 such that $hash(m_1) = hash(m_2)$.
3. *Collision resistance* - It is difficult to find two different messages m_1 and m_2 such that $hash(m_1) = hash(m_2)$.

Mining is the agreement among the existing nodes in network on the common view of blockchain. Reaching this consensus is a complicated problem and there are few existing approaches on how to solve it. In a decentralised network, there should be some node that is responsible to publish the block of transactions into the blockchain. Such node can not be selected randomly as it might lead to the attack if the selected node is not honest. I will briefly review few methods on how to select the node that will create new block and how the consensus can be reached.

Proof of Work

This consensus strategy was proposed by Nakamoto in [1] and the idea is that nodes compete for finding a hash by exhaustively querying a hash code to satisfy certain condition. I would like to explain the Proof of Work (PoW) algorithm by using the definition given by W. Wang et. al. [90]:

"Given an adjustable hardness condition parameter h , the process of PoW puzzle solution aims to search for a solution string, nonce, such that for a given string x assembled based on the candidate block data, the hashcode (i.e, the target block header bh) of the concatenation of x and nonce is smaller than a target value $D(h)$:

$$bh = H(x \parallel nonce) \leq D(h) \tag{2.1}$$

where for some fixed length of bits L $D(h) = 2^{L-h}$ and H is the hash function."

To approve broadcasted transactions, nodes (miners) need to solve the puzzle by guessing the secret value (nonce, shown in Fig 2.2 and Fig 2.3) before adding it into the block. All the information inside the block header (guessed nonce, information about the

transactions, timestamp, hash of the previous block) will be then converted to an SHA-256 function. If the output of this function is below the given difficulty threshold, then block will be added to the chain. Otherwise, the node will have to make another guess. The difficulty value h is adjusted after every 2016 blocks are added to the chain, making the average speed of block mining to be around 10 minutes. Miners have to do many resource intensive calculations in PoW algorithm. On one hand, due to the high cost it prevents malicious attacks, on the other hand, it is becoming economically impractical for a single node to perform mining, since it is becoming more computationally expensive. This leads to the creation of mining pools which defeats the purpose of blockchain - that is supposed to be decentralised system.

Proof of Stake

This is the less resource intensive consensus algorithm, alternative of PoW. Before competing for finding a new block, miners need to prove their ownership of the amount of currency. People with higher balances are more likely to win the competition. This algorithm is based on the assumption that richer nodes are less likely to be malicious. Although this algorithm uses less resources, it might lead to a situation where the richest node becomes dominant and it is also contradicts the idea of blockchain to be decentralised.

There are also other algorithms as Practical Byzantine Fault Tolerance (PBFT) [3], Ripple [4] and Tendermint [5] and all of them have certain advantages and disadvantages. For this research data from PoW and PoS types of blockchains were used, therefore I will not discuss in greater details about the other types.

2.1.1 Blockchain 1.0 and Bitcoin

The term Blockchain 1.0 is used for the decentralization of money and payments. Nowadays there are many currency systems based on the blockchain, first of which was the bitcoin. During the first year after the system went online in 2009, the bitcoin was used only by a few people and was not considered as a real value currency. However, in 2011 the popularity of bitcoin grew dramatically - many people and companies joined the system. Nowadays, the cryptocurrency concept is quite strong and bitcoin is one of the most popular ones. It should be noted that *Bitcoin* refers to the blockchain, while *bitcoin* is the native currency of the Bitcoin blockchain.

The idea of “untraceable cash” based on the cryptography appeared quite a long time ago. Interestingly, the mathematical foundations have been proposed long before the bitcoin has been introduced by [91]-“untraceable cash” that prevents double-spending

and ensures users' anonymity. It was thought that generating electronic cash would be feasible only for banks and therefore, it is still a centralised system.

Though the idea of digital money based on the cryptography was innovative at the time, it did not attract much attention and only later with the invention of bitcoin, it became popular. Introduced after the global financial crisis, when trust to banks was undermined, the idea of decentralised cash looked very promising and the popularity of bitcoin grew dramatically just in few years.

The principle of work of bitcoin (as well as many other cryptocurrencies) is quite novel, comparing to the traditional currencies and financial instruments as it fully relies on the blockchain technology, which was described earlier.

Bitcoin blockchain uses the Elliptic Curve Digital Signature Algorithm for signing transactions and *binary* Merkle tree to compress them. For hashing Bitcoin system uses the *Secure Hashing Algorithm-256* (SHA-2).

Figure 2.1 illustrates the block structure:

Figure 2.2. Schematic illustration of bitcoin block structure.

Each block has its unique hash, information about the hash of the previous block, timestamp, nonce and information about the all transaction records in the form of Merkle tree. Consensus in bitcoin blockchain is reached by using the Proof of Work algorithm and finding a hash string looks as following:

$$H_i = SHA - 2(SHA - 2(H_{i-1} \parallel T_i \parallel TX_i \parallel d_i \parallel N_i \parallel)) < f(d_i) \quad (2.2)$$

where:

H_{i-1} - hash from previous block; T_i - timestamp in sec.; N_i - 32-bit nonce; TX_i - hash of the transaction data; d_i - difficulty parameter. This equation is same as the equation 2.1, but in a different, more comprehensive form. The understanding of this equation is as follows - to find the hash of block H_i we need to solve the puzzle of finding the nonce N_i , while other parameters (hash of previous block H_{i-1} , timestamp T_i and hash of the transaction data TX_i) are known. The output of function when nonce N_i is guessed has to be less than difficulty parameter d_i . Currently the difficulty of finding a new block (see equation 2.2 is $f(d_i) < 2^{192}$ (have its 64 most significant bits equal to zero and this is an extremely rare event).

As a reward for finding such a hash and producing a new block, miners are rewarded by bitcoins. In the beginning, this amount was of 50 bitcoins, then it halves approximately every four years after 210,000 new blocks added. Anonymity of users are ensured by providing the possibility to create as many wallets (addresses) with different IDs as needed (it is recommended to create new wallet for every transaction).

2.1.2 Blockchain 2.0 and Ethereum

Blockchain 2.0 is for the decentralization of different kinds of assets beyond the currency - smart contracts, smart properties etc.

Originally, Satoshi Nakamoto planned to have a three-step technology when developing the Bitcoin: 1. The blockchain (public ledger) 2. Bitcoin protocol (the transaction system) and 3. Turing-complete robust scripting system. As for bitcoin and other currency and payment systems (Blockchain 1.0) first two steps are enough, but to perform more complicated functions, such as recording and transacting with the complex assets, the third step would be necessary.

Ethereum is a good example of the next generation of the blockchain technology. If Bitcoin is the protocol running over the blockchain, Ethereum is the fundamental platform that can run all blockchains and protocols using its own Turing-complete programming language (Solidity) proposed by Vitalik Buterin [2]. Ethereum has two types of accounts - externally owned (same as in bitcoin, controlled by the private key) and contract accounts (controlled by the code). The later one is for the smart contracts. Smart contracts are combination of Ethereum account and the code running on that account. These codes are written in Solidity language for the Ethereum Virtual Machine (EVM). Transactions in Ethereum system are performed not only to transfer coins between addresses, but also to

create new contracts and invoke functions of contracts. Differently from Bitcoin, where user's balance is the amount of UTXO (change), Ethereum also handles users and contract balances. It should be noted that *Ethereum* is the name of blockchain, while *ether* is the name of the native currency of Ethereum.

Ethereum uses the *Ethash* Proof of Work function, which is based on Keccak hash-function standartized to SHA-3. However, Ethereum is transitioning from Proof of Work to the Proof of Stake consensus algorithm. For compressing the transaction information, Ethereum uses the *hexary* Merkle Patricia trees. In Ethereum blockchain *gas* is the cost necessary to perform a transaction. Miners set the gas fee depending on the supply and demand in computational power. This information is also stored in blocks. Overall, each block in Ethereum stores more information, in comparison with the Bitcoin. The figure 2.2 illustrates the block architecture in Ethereum:

Figure 2.3. Schematic illustration of ethereum block structure.

Each block has its unique hash, information about the hash of the previous block and also the hash of the sibling blocks (ommer hash). There is also data about the gas used and gas limit, timestamp of block creation, nonce and information about the all transaction records in the form of Merkle tree. Opposite to the bitcoin blockchain, ethereum blocks stores more data directly in blocks like account balances and global states.

Similar to Bitcoin, miners in Ethereum solve puzzles and broadcast new valid blocks to the network. Ethereum has higher mining rate than Bitcoin - difficulty of puzzle updates dynamically and takes on average 12 seconds for each block, while in Bitcoin it takes approximately 10 minutes. Miners in Ethereum are paid for their service by a *gas*. Each

transaction specifies the limit of gas that user is willing to pay and the price per unit of gas. Therefore, the higher the unit price, the higher the chance that miners will execute the transaction.

2.2 Obtaining Blockchain Data

In this section I will provide the thorough description about obtaining and processing the blockchain data. Working with the Bitcoin blockchain was particularly challenging due to its anonymization concept. As it was discussed in Section 2.1.1, Bitcoin users can own many wallets. Therefore, deanonymization was the additional step we made before being able to start using this data for analysis. Deanonymization of bitcoin users is quite popular research topic in Computer Science and there are few methods that have been already proposed. In this section a brief review of existing attacks is provided with the detailed explanation of methods used for the research.

2.2.1 Data Extraction and Processing

It is quite a big challenge to obtain data from a blockchain and transform it into a human-readable, analyzable format. When obtaining the data for this research, the size was approximately 250 Gigabytes for the Bitcoin blockchain and 181 Gigabytes for the Ethereum blockchain.

Ethereum blockchain data has been already processed by E. Medvedev and uploaded to the Google Big Query website [92,93]. We have queried the data from the website and processed it into the format suitable for this research. More details about this data set will be given in the next section.

Bitcoin transaction data was obtained from scratch with the help of our collaborator Valentin R. Melnikov from Complexity Institute at NTU by downloading the full Bitcoin blockchain, starting from the genesis block (dated 3 January 2009) up to block 560,000 (dated 25 January 2019). Next step was the deanonymization of bitcoin users. Although users are supposed to be anonymous, due to the public availability of transaction history, there is still a chance to link wallets (pseudonyms) to the certain user. This possibility was theoretically discussed in the original paper of bitcoin. Later more papers empirically showed that it is, indeed, possible [6,94]. Fair deanonymization of bitcoin users is quite important step for my research, since results are directly affected by the quality of the obtained data. Therefore, the review and comparison of existing attack methods has been

done before processing the data.

Constructing Transaction and User Networks

Reid and Harrigan [6] demonstrated the possibility to uncover users' identities by constructing two networks - transaction network and user network. Transaction Network shows the flow of bitcoins between transactions. A node in this network represents a transaction and a directed link between the source and target represents the output of the source node and input of the target node. User network shows the flow of bitcoins between the public keys of users. Each node in this network is the public key and the link shows the direction of transaction between the source and the target. Due to the concept of "change" or Unspent Transaction Output (UTXO) in Bitcoin system it becomes possible to partially reveal users identities. For example, if user *A* has an input of 4 bitcoins and wants to make a transaction that needs only 3, then amount of 1 bitcoin remains unused. For this, Bitcoin creates an additional address and output of this address will be marked as UTXO (change). Reid and Harrigan method uses the concept of "change" and combines the transaction and user networks made for users identification.

This method has been widely used and the implementation codes can be found in [95] and [96] as well as processed data [97]. We have used the data set processed with the code of Ivan Brugere (from 2009 till 2013) for some part of this research.

Clustering Bitcoin Addresses

Spagnuolo introduced the whole framework on user identification called BitIodine [98]. It consists of several modules - the scraper, the block parser and the clusteriser. Scraper crawls the web and finds the bitcoin addresses of users that have already publicly posted them. This scraper is somewhat similar to what has been proposed by [99]. Parserer stores all bitcoin blocks into a database. Wallet clusteriser groups addresses based on the two heuristics. (1) several addresses transacting to one account are considered to belong to one user; and (2) the address of a transaction which appears to be an UTXO "change" transaction is considered to belong to the sender of the transaction. Based on the processed data, the temporal network of interactions of bitcoin users was estimated. Since the clustering algorithm is heuristics-based, it does not guarantee that all the wallets in the network are clustered to corresponding users. Therefore, we might expect a certain fraction of non- or poorly clustered wallets, but still this algorithm results in a significant improvement of network representation of financial interactions in bitcoin.

We have use the BitIodine software to perform the wallet clustering for whole Bitcoin blockchain from 2009 till 2019.

2.2.2 Data Description

After finishing the processing part, we have obtained data sets for bitcoin and ether in the format shown in tables 2.1 and 2.2. It contains the information about the sender, the receiver, the amount transacted and time of transaction.

Sender key	Receiver key	Date	amount
501194	2645	20101105221244	0.90
1EHwci1gVKrs	1986	20101105221244	0.10
834628	834630	20100718123114	2.00
834628	C8x2hqqgE2b	20100718123114	0.05
713610	5188	20110218223432	50.00

Table 2.1. An example of bitcoin’s dataset obtained using BitIodine software.

The data obtained using BitIodine software [98] contains information on the users’ ID (both sender and receiver), exact date and time of the transaction, and the amount of transaction in bitcoins. Every wallet is reflected as a hash in the blockchain. For users with only one wallet, their user ID is the same as the hash of their wallet. In the case when a user has more than one wallet, all his/her wallets are clustered into one unit and a random number is assigned to the cluster. The user ID in this case is the ID number of the wallet cluster.

Sender key	Receiver key	Date	amount
0xea674fdde7	0x52bc44d537	20180105221244	5.70
0x209c4784ab	0x2a65aca4d5	20180105221244	0.30
0x61c808d82a	0x7ed1e469fc	20180018123114	1.00
0x209c4784ab	0x52bc44d537	20180018123114	3.05
0xb2930b358f	0xabbb6bebf	20180018223432	2.00

Table 2.2. An example of ether’s dataset.

Similar to bitcoin, the data obtained from [92] contains information on users’ ID (both sender and receiver), exact data and time of the transaction, and the amount of transaction in ethereum. Since deanonymization and wallet clustering is not required for ethereum’s data, user IDs are the original hashes of the users’ wallets.

2.3 Conclusions

In this chapter I provided the detailed technical explanation of what is the blockchain, how it operates and I also discussed the stages of its development and applications. For

this research, two data sets were used - from Blockchain 1.0 (bitcoin) and from Blockchain 2.0 (ether). Differences between them were discussed, as well as their functionalities.

Overall, the goal of this Chapter is to provide reader with the understanding of the nature of data sets, explain the concepts of blockchain and cryptocurrencies and demonstrate the data used for the research. In the next Chapter we start our study of cryptocurrency systems by investigating whether they also have *stylized facts* that found to be universal in the traditional financial markets.

Chapter 3

Stylized Facts in Decentralized Financial Markets

3.1 Introduction

Traditionally, the role of financial institutions was to connect and establish trust between transacting parties. Nowadays, with the development of financial technologies the role of institutions is being reduced to some extent. The blockchain technology created a new paradigm of Decentralized Finance and made it possible to eliminate the need of intermediaries and ensure the trust using cryptography algorithms. The paradigm of the Decentralized Finance is drastically different from the traditional one. It is an open system with distributed trust and decentralized platform that reduces transaction costs. Although, the development of the Decentralized Finance is still at the early stage, it has a great potential to become an alternative financial system.

As for traditional financial instruments, their properties and evolution were extensively studied and many interesting patterns have been found that are universal across a variety of markets and instruments. These universal properties are called “stylized facts” in literature and mostly related to the statistical properties of assets’ prices and can be the consequence of people’s interaction in market. The first universal feature of financial time series was discovered by Mandelbrot in 1963 [100]. Later new common patterns were found - related to the scaling properties [45–48, 101, 102] and to long and short memory of returns [103, 104].

As Decentralized Finance is conceptually different from the traditional one, we wonder if the instruments would show the similar properties. The similarity would be interesting and even desirable - on the one hand, it would allow us to use the existing traditional

finance theories to model Decentralized Financial instruments as well. On the other hand, having detailed information about the transactions in blockchain systems would allow us to empirically derive models and improve/enrich existing theories of financial modelling. This Chapter is dedicated to the discussion of common properties of traditional financial instruments and analysis of cryptocurrencies' financial time series.

3.2 Absence of Autocorrelation in returns

It is well known that stable predictions of price change in short time scale is almost impossible due to the arbitrage - if there is significant correlation, simple strategy will be used by people to get the profit and therefore, reduce the correlation. The dependence between changes of price is normally investigated using the autocorrelation function of time variations of price logarithms:

$$\varrho_l = \frac{\langle (r_{t+l} - \mu)(r_t - \mu) \rangle}{\sigma^2} \quad (3.1)$$

where μ is the mean value of returns, and σ^2 is the variance of returns, notion $\langle \dots \rangle$ refers to the average.

If ϱ_l is close to zero, it means that there is no predictability on returns. This property found in various financial markets is often cited as a support for “efficient market” hypothesis [105, 106]. Also, it should be noted that the autocorrelation of returns is a fast decaying function and when the time scale is increased (weekly or monthly), returns do show the high correlation, which will be discussed in later section. To check if systems of decentralized finance also have this property, we calculated the price return of both bitcoin and ether as logarithmic function:

$$r_t = \ln(S_t/S_{t-\tau}) \quad (3.2)$$

Figure 3.1 shows that the linear correlations in intraday returns are quite small for both cryptocurrencies, which is similar to other financial instruments.

Dotted line marks the noise level in Figure 3.1, it can be seen that autocorrelation values for both instruments are quite small, less than 0.1%. Having small autocorrelation function shows that price movements in both bitcoin and ethereum is difficult to predict in short time scales that is similar to any other financial asset.

Figure 3.1. Autocorrelations of BTC and ETH log returns of price.

3.3 Volatility clustering

Although financial time series are not autocorrelated in short time scales, the interesting property of *long memory* have been discovered for various instruments. Time series of returns of financial instruments often show the *clustering of volatility* - when large fluctuations tend to be followed by large fluctuations, while small changes also cluster together [103, 104, 107]. As it was shown earlier, the returns themselves are linearly uncorrelated, however absolute and squared returns $|r_t|$ show the positive and slowly decaying autocorrelation function, $\text{corr}(|r_t|, |r_{t+\tau}|) > 0$, with τ that could range from few minutes to several weeks.

First, we plot the evolution of *logarithmic returns* to visually observe if there are clusters of price fluctuations. Figure 3.2 shows that fluctuations are clustered and it is more pronounced in bitcoin system.

Clusters of volatility can be observed in both instruments when big fluctuations are followed by big fluctuations and small fluctuations - by small ones. Then we look at the autocorrelation function of absolute returns in a weekly time scale that is defined as following:

$$\varrho_a = \frac{\langle (|r_{t+l}| - \mu_{|r|})(|r_t| - \mu_{|r|}) \rangle}{\sigma_{|r|}^2} \quad (3.3)$$

where $\mu_{|r|}$ is the mean value of absolute returns, and $\sigma_{|r|}^2$ is the variance of absolute returns.

Figure 3.2. Evolution of logarithmic returns of BTC and ETH.

Figure 3.3. Evolution of absolute returns of BTC and ETH in a weekly time scale.

Figure 3.3 shows that there is a well pronounced long memory in bitcoin returns, while in ether only short lags show high autocorrelation. As for ethereum it still shows high autocorrelation but only on short lags with decrease of ACF value as number of lags increased. However, it still stays higher than autocorrelation in Figure 3.1, meaning that in longer time scale there is higher chance of historical price repeating the trend.

3.4 Power-law probability distribution of returns

We then check the scaling properties of cryptocurrencies' time series - first, we look at the distribution of returns, as for traditional instruments it was found to follow the power-law function. In early 1960s Mandelbrot suggested that normal distribution is not sufficient to model the real asset returns [100]. Later, many studies have found that high-frequency returns of various financial instruments indeed have fat tails, which means that the chance of big price fluctuations is more likely than what the Gaussian distribution assumes. This phenomena has been empirically observed across different markets during different periods and it has been found that their returns are power-law distributed [46, 101, 102].

We investigate if in instruments of decentralized finance also show the same property and plot logarithmic returns of bitcoin and ethereum as complimentary cumulative distribution function in log scale. Figure 3.4 shows that both of instruments follow the power law distribution in returns:

Figure 3.4. Probability distributions of logarithmic returns of BTC and ETH.

The dotted blue line shows the power-law function with an exponent of -2 and solid blue line is the empirical distribution. It can be seen that both distributions follow the same power law. Having the power-law distribution of returns means that in both cryptocurrencies chance of having large price fluctuations is quite high and comparable to many traditional instruments.

3.5 Power-law probability distribution of trading volume and number of trades

It has been found that not only return distributions follow the power law function, but also other quantities of financial markets like number of trades [47], trading volumes [108]

etc. also follow the same distribution.

Having precise information about the volumes traded in cryptocurrency systems, we calculated the total transacted volume for each user of system as the sum of incoming and outgoing transactions made by the user.

Figure 3.5. Probability distributions of trades' volumes in BTC and ETH.

As it can be seen from the Figure 3.5, trades' volume also follow the power-law function. As for the number of trades, we found that it indeed also follows the power-law - please refer to the Figure 4.7 and Figure 4.8 that show the *degree distribution*, where degree is the number of trades (transactions).

3.6 Conclusions

In this Chapter we showed that two largest decentralized currencies' financial time series have similar statistical properties to the traditional financial instruments'. More stylized facts of other cryptocurrencies has been also researched in [109]. These findings are quite important since drawing the parallel between the traditional and decentralized financial instruments could be beneficial for the understanding of financial markets in general. Abundance of information of financial transactions provided by the blockchain can be beneficial for improving the modelling of traditional instruments, while rich toolbox of traditional finance can be used as a base to develop the analytical methods and models for decentralized financial instruments.

Now, after we showed the similarities between the traditional and decentralized financial instruments, we continue the research by looking closer to the transaction datasets described in Chapter 2 and the first step would be to investigate into the structural characteristics by constructing *complex networks* from the transaction datasets.

Chapter 4

Network Analysis of Cryptocurrency Systems

When it comes to the understanding of Complex Systems, Networks representation becomes a powerful tool. Thinking about large systems as a collection of points and lines leads to useful insights, allows to observe general trends by zooming out from the small details. Field of Complex Networks has been a domain of a branch of discrete mathematics - *graph theory* that began in 1736 when Swiss mathematician Leonhard Euler was solving the problem of bridges in the city of Königsberg (now Kaliningrad). Since then, graph theory has been developed a lot and many practical problems has been solved using this mathematical tool. Networks as a science started its development from the *Social Network Analysis* in the early 1920s and mostly focused on the relations between members in social groups and social entities.

Nowadays Complex Networks is a big field, thoroughly studied both empirically and theoretically. Although, the real systems can be modelled by the Graph Theory, the analytical tool it offers is not sufficient to describe the properties of the real complex systems. Studies on the structural properties, Temporal Networks, Multiplex Networks, Dynamics on and of Networks, Studies on Controllability, Representation and Link Prediction, Studies on the Applications of Networks etc. - popularity of this research areas shows how useful is Network Theory to understand Complex Systems. There are review articles, books and theses that have provided the great systematic review of the Network Science [110–113]. Complex Networks theory applied to the cryptocurrency systems has already provided some interesting insights - Network analysis on bitcoin dataset has been done in [7, 25–27], where the main properties were calculated. In [114, 115] authors analysed the ethereum transaction network using Complex Networks approach.

In this Chapter we aim to build upon on the growing research on complex networks applied to cryptocurrency systems. First, we will give a brief review of the Network Theory and most importantly, show the place of the new cryptocurrency network in this field, how similar/different it is to other networks and also demonstrate how tools and methods from Network Science can be applied to the new socioeconomic system of cryptocurrencies. We will start by using the traditional Static Network approach to analyse bitcoin and ethereum systems and some results would confirm the findings obtained earlier in literature. On the top of existing knowledge we will additionally perform the comparative analysis of the cryptocurrency systems with the other complex systems based on their network properties. Also, we will review the new field of Temporal Networks and apply its concepts and tools to the bitcoin and ether transaction data. Under the Temporal Network approach we represent the network as a sequence of smaller static networks (multilayer approach), calculate the properties as for a static network, and observe the evolution of properties over time. Datasets of ether and bitcoin transaction were used for analysis presented in this Chapter. In Chapter 2, section 2.2 the procedure of obtaining and processing data was explained in details for the reader's reference.

4.1 Empirical studies of Networks

In literature, complex networks are often divided into categories - technological, biological, information and social ones. This was made mostly for the organizational purposes. Moreover, due to the very applicable nature of the network field, this division is also based on the purpose of application or scientific branch to which the network belongs. When we talk about the networks as power grids, network of cells or telephone networks, it is clear and fair to classify them as technological or biological types. These networks have quite distinct difference. However, as we noticed, together with the technology development, it becomes more complicated to bring the network to the certain category. And cryptocurrency networks are the great example - they are technological peer-to-peer networks (if you think of them as a vehicle for the distributed databases), at the same time, they are financial networks, they also can be considered as information networks, like internet (packed-switched data network with no central authority), they can even be considered as social network, since it is still about the interactions between people. Same can be said about the Twitter, Facebook, and there are plenty of other examples when the classification becomes quite fuzzy.

Nevertheless, in this section I will follow this traditional classification of networks

and provide the overview of empirical studies on real-world systems and show how similar/different cryptocurrency network to others. I will limit myself with the overview of social, technological, financial and economical networks only, since these type of networks are close to the cryptocurrency network structurally, by their nature and the methodology used for analysis.

Social Networks is perhaps the oldest area of the applied graph theory. Jacob Moreno, Romanian psychiatrist considered as the founder of the field of Social Network Analysis. In his book “Who Shall Survive?” [116] he introduced the first hand-drawn representation of system (schoolchildren) as a network, which he called the sociogram. Though the analysis he made on the system was not rigorous comparing with the nowadays standards, his work contains the main idea of how to reveal the important patterns in system by visualising it in a certain way. Other remarkable early applications of graph theory to the social systems has been done in works of Davis [117] and Roethlisberger [118]. Since that time, huge amount of research has been done applied to different problems and communities - from romance among adolescences [119] to business contacts [120] and scientists collaborations [121, 122], from dating patterns [123] to the fictional characters in Marvel comic books [124].

Social Network Analysis field has developed greatly since the start of social networking sites like MySpace, Facebook, Twitter, LinkedIn etc. Opportunity to access the data from such online communities helped us to get more understanding about the network structure in social systems [74, 125–127]. One of the new insights is finding influential nodes in social network, it became quite popular topic with applications for the viral marketing and opinion spreading [63, 128, 129] Moreover, interesting facts about the people’s behaviour were revealed by performing experiments on social networking websites - for example, Kramer et al. [130] showed that emotional contagion do not need in-person interaction and long-lasting moods (happiness, depression etc.) can be easily influenced by others’ in our News Feed. Another article also published in PNAS showed that easily accessible digital traces of human behaviour can accurately predict the sensitive personal attributes like sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender [131]. One of the most scandalous implemenation of the Network Theory together with the Machine Learning to predict human behaviour was done in the research by Cambridge Analytica [132] where they used the data of 87 millions of Facebook users to cluster personality traits into distinctions.

Questions asked in social network analysis have much in common with complex network studies in general - what's the structure of the network, how dynamical processes are affected by the network structure (for example, for promotion of adds and opinions or for disease spreading).

Technological and Infrastructural Networks. The distinct feature of these networks is that their edges and vertices are physical objects - power grids, railroads, wiring of the internet, etc. The interesting fact about these type of networks is that even though they are engineered and created by people, we still had a little understanding of how exactly they work in a global scale and what is their structure. Early studies on railway and airport networks showed that they have a small-world properties [68, 133–135]. Studies on power grids investigated their structure [136, 137] by looking at the spatial distribution. Also, studies showed the power-law behaviour in the size of power outages [69, 138]. Number of authors applied the network analysis to various technological and infrastructural networks and successfully gained meaningful insights about systems and solved practical problems [70, 139–143].

Economic and Financial Networks. Network analysis is appropriate for many economical and financial systems. Early studies on the trade networks [144] represented the node to be the party (organization, individual or company) and link to be the trade relation (one node buys something from another). Another example is networks of stock companies, where the node is the company and the link between them appears if their share values are correlated [145–148]. There was not much research done on the financial networks, since modelling money flows between accounts was impossible due to the lack of such data.

Later, after the financial crisis of 2008 and data revolution network analysis of financial systems gained more popularity. It has been realized that network approach would help to get better understanding of economic phenomena [149]. Overall, financial system can be well modelled as complex network - various entities can be represented as nodes and links can represent contracts or any other type of relations. One of the main interests of researchers was to understand the systemic risk arising from the inter-bank interactions. In studies [150–154] the contagion and systemic risks in banking system were analyzed by looking at the inter-bank network. In the work of Battiston et. al. [155] the multilayer network of credit contracts was constructed among many financial institutions. Analysis of this network led to the conclusion that the higher interdependence between banks, the higher the possibility of incorrect pricing of debt instruments. The overview of financial networks has been done in the review study of Bardoscia et. al. [156] where authors dis-

cussed how various types of financial systems can be modelled by single-layer, multiplex, high-order, correlation and similarity-based networks.

Cryptocurrency Networks. There have been a few studies on cryptocurrency transaction data using tools from network science. One of the earliest studies on the bitcoin transaction graph is by Ron and Shamir [7] where the authors analysed the statistical properties of the graph as the number of links, distribution of addresses and the flows of the largest transactions. This work mainly focuses on understanding the structure of the bitcoin network and detecting behaviours as long chains, forks and binary tree-like distributions. Another exploratory study of the bitcoin network has been performed in [26] to show the basic statistics and dynamics of its network properties. The authors have reported the correlation between the user activity, transaction volume and the exchange rate of bitcoin. As for the network properties, the authors have found a scale-free degree distribution and a very large average path length. This work gave a preliminary understanding of the bitcoin graph in the initial stage of its development. Later research provided more insights into the system - in [25] the network characteristics together with their evolution over time were investigated. The authors have measured the wealth and degree distributions over time and found that the network is highly heterogeneous and assumed that nonlinear degree preferential attachment [157] might be the reason for these observed characteristics. However, no direct investigation on their proposed mechanism has been made.

Several studies have contributed to the understanding of bitcoin network's structure and evolution. There is, however, a dearth of research done on the ethereum network. Nonetheless, recent studies [114, 115] have shown similarities of its network properties with that of bitcoin.

In Table 4.1 we show the general characteristics of several real networks - technological, biological, social and cryptocurrency. The purpose of this table is to show how similar/different is the cryptocurrency networks (bitcoin and ether) comparing to the others, more information about the each type of network shown in the Table can be found in references provided below. For each network we show the number of nodes N , the average degree k , the average path length L , clustering coefficient C and r - assortativity coefficient and α is the value of power-law exponent for the networks.

Ether network was divided into 3 subnetworks according to the [114] - Money Flow Graph (MFG), Smart Contract Creation Graph (CCG) and Smart Contract Invocation Graph (CIG). However, here we omit the CCG due to its very specific properties. It can be seen that bitcoin and ether (MFG) networks, as well as any other, despite of the

Network	type	N	k	L	C	r	α	ref
Internet	tech	2×10^8	7.5	16	0.11	-0.065	2.12	[158]
Routers	tech	228,263	2.8	9.5	0.03	-0.189	2.1	[159]
Protein	bio	2,115	6.8	2.12	0.07	-0.156	3.1	[160]
Metabolic	bio	778	3.2	7.4	0.7	-0.195	2.8	[161]
Math1999	social	57,516	5.00	8.46	0.15	0.125	≈ 3.1	[110]
e-mail	social	59,812	2.88	4.95	0.03	-0.12	3.5	[162]
Facebook	social	721×10^6 (in 2011)	190	4.7	0.14	0.226	2	[126]
bitcoin	crypto	4,892,708 (in 2012)	1.68	2.43	0.12	-0.1	2	[27]
ether (MFG)	crypto	5,267,627 (in 2018)	5.2		0.17	-0.12	1.5	[114]
Ethereum (CIG)	crypto	1,281,500	7.1		0.004	-0.2	3.3	[114]

Table 4.1. Properties of different networks with scale-free behaviour.

large size, have relatively small average degree and average path length is also not so big. Although in ether transaction network (MFG), nodes are tend to transact more than in bitcoin - average degree of 5 and in smart contract network (CIG) each contract makes on average 7 transactions. As for the clustering coefficient, bitcoin and ether networks are closer for internet and citation networks (Math 1999), while technological (routers, email) and biological (protein) networks do not tend to be highly clustered. Bitcoin and ether networks have a slightly dissasortative behavior as well as all of technological and biological networks, only social network has a positive assortative coefficient. Overall, comparison of bitcoin and ether (MFG) networks' main properties with the other network types shows that it is more similar to the technological networks, rather than social or biological.

4.2 Static Network Fundamentals and Main Properties of Cryptocurrency Networks

Mathematical framework for the complex networks is a graph theory [163, 164], and all the definitions and notations are quite close in both graph theory and complex networks. Traditionally, in graph theory literature notation for graph is $G = (V, E)$, where $V \equiv (v_1, v_2, \dots, v_N)$ is a set of vertices and $E \equiv (e_1, e_2, \dots, e_K)$ is a set of ordered (unordered) pairs of nodes from set V . In Network Science, the term *vertex* is usually replaced by the term *node*, while the *edge* is often called *link*. Currently, in Networks literature these terms are used interchangeably - nodes and edges, vertices and links etc. In this thesis the terms *nodes* and *links* will be used.

4.2.1 Static Network Representation

The first step when modelling the Complex System as a Network is to find the best way to represent the system as a set of nodes and links. There are many types of Network representations - undirected, directed, weighted, matrix form, multilayer and multiplex networks, bipartiate networks etc. Depending on the nature of the system under analysis, the certain representation would help to get the most meaningful insights.

1. *Graphical Representation.* Simple visual representation is drawing lines between points, lines can be *directed*, *undirected* or have some information (weight) on the links (*weighted*). In the directed network, nodes connected by arrows, indicating the direction of each link. In the weighted network, links have additional information (weight) $w_{i,j}$ of each link.

2. *Matrix representation.* The fundamental mathematical representation of a network is the *adjacency matrix* $A_{i,j}$ with $n \times n$ dimensions, where:

$$A_{i,j} = \begin{cases} 1, & \text{if there is an edge between nodes } i \text{ and } j. \\ 0, & \text{otherwise.} \end{cases} \quad (4.1)$$

Adjacency matrix for directed network has elements:

$$A_{i,j} = \begin{cases} 1, & \text{if there is an edge from } i \text{ to } j. \\ 0, & \text{otherwise.} \end{cases} \quad (4.2)$$

It should be noted that diagonal elements are zero if there are no self-loops and that undirected network's adjacency matrix is symmetric, while undirected network's is asymmetric.

Adjacency matrix captures the entire structure of the network and its matrix properties can provide us with a lot of useful information. Another matrix representation of a network is the *Graph Laplacian* - widely used method when working with the large networks.

Cryptocurrency system is fast-evolving one, many transactions are happening every day, new nodes entering the system and some nodes are leaving at the same time. It is a directed network - there is a sender and receiver in transactions, and every link has an information about the amount of transaction. Therefore, we represent the network as *directed and weighted* one. In this research we represent it as both, static and temporal to gain more insights about the system. Temporal properties and visualisation will be

discussed in later section.

4.2.2 Network Visualisation

After deciding which way would be the best to represent the network, visualisation of network helps to get the basic understanding of system. There are quite a few methods to visualise complex networks - graph-tool and networkx libraries in Python [165], Gephi software [166], Pajek [167], etc. I used the Gephi software to visualise the cryptocurrency networks, since it has a rich set of layout algorithms. It is important to choose the appropriate layout to get the meaningful visualisation. ForceAtlas2 layout [168] in Gephi was chosen - this layout was specially designed for scale-free networks and allows to place a node depending on its connections and other nodes, which is important when plotting transaction network.

First, the bitcoin network topology was visualized for different periods - the beginning of system (till the end of 2010) and in the later period (from 2011). And the structure was found to be completely different. In the early period (Figure 4.1) we observe a lot of isolated nodes, which are not connected to the whole system, also we see the big clusters of nodes. The physical meaning of isolated nodes could be the following - mining without using the mined coins for further transactions, new users trying the system by sending bitcoins to their own wallet, making transaction once with another user of system who never transacted with anyone else. Figure 4.2 shows the same network structure as in Figure 4.1, but without isolated nodes. It can be seen that there are large groups of nodes that tend to transact only with each other. In the late period (Figure 4.3) network topology looks quite uniform - there are no isolated nodes and a few hubs. However, it should be noted that the topology for the late period is shown for the one month data (January 2013). Due to the big data size, it was impossible to visualize the whole period in Gephi software. Nevertheless, after dividing period by chunks we were able to make a movie of network growth, that will be discussed in later part of this Chapter.

Ether transaction network was visualised in the beginning of system (September 2015). Figure 4.4 shows the structure of network. It looks very different from the bitcoin early stage network - not many isolates and large separate groups of nodes. Instead, we see few hubs - nodes with high degrees and many nodes with small degrees, like in the typical scale-free network. Then we visualise the ether network for later period - due to the large size of data, only one day of transactions was visualised (January 1, 2017). In Figure 4.5 we observe larger network, but structure is quite similar to the one in earlier period - few big hubs and many nodes with small degree. Close-up visualisation of the largest hub in

Figure 4.1. Network structure of bitcoin transactions in the early period of system development (2009-2010).

Figure 4.2. Network structure of bitcoin transactions without isolated nodes in the early period of system development (2009-2010).

Figure 4.6 shows that most of nodes in network connected to the hub.

The differences in early period network structures between bitcoin and ether can be explained by the fact that bitcoin was the first instrument for decentralized finance and there were many inexperienced users and people using for fraudulent transactions that led to the structure shown in the Figure 4.1. By 2015, when Ethereum was introduced, blockchain users overall were more experienced and the structure of transaction network

Figure 4.3. Visualization of bitcoin transactions network structure in the mid period of its development (January 2011).

Figure 4.4. Network structure of ether transactions in the early period of system development. One month (Sep 2015) is visualised.

looks similar to the one in bitcoin later period - scale free with few hubs and many nodes with few connections.

However, it should be noted that cryptocurrency networks are very big and their structure changes every second. Although the static representation gives us some qualitative understanding of the system, there is still a lot of information missing and more sophisticated tools needed in order to fairly understand these systems.

Figure 4.5. Visualisation of one day ether transactions network structure in the late period of its development (January 1, 2017).

Figure 4.6. Close-up visualization of one day ether transactions network structure in the late period of its development (January 1, 2017).

4.2.3 Main Properties of Cryptocurrency Networks

Degree and Degree distributions The basic definition in network theory is a *degree* of node - number of links connected to the node. For directed graph, the degree of a node has two components - number of outgoing links (outdegree) and number of incoming links (indegree). In cryptocurrency network, in degree is the number of transactions received

by user, out degree is the number of transactions sent by user. Average degree of a node in cryptocurrency network shows how many transactions on average each user makes.

One of the most important topological characterizations in network - *degree distribution*, has attracted a lot of attention since the discovery made by Adamic et. al. [169], Barabasi [170] et al. and Faloutsos brothers [158] that many real-world networks follow a power-law distribution. For bitcoin and ethereum networks, degree distribution was calculated and plotted for different periods. Figures 4.7 and 4.8 show that the in-degree, out-degree, and overall degree distributions resemble power-law functions.

Figure 4.7. Degree distributions empirically calculated for bitcoin separately for the two periods and shown as blue lines in the form of complementary cumulative distribution functions (CCDFs). The red lines show a power law function with an exponent of -2.

Table 4.2 shows the calculated value of the power law exponent. Fittings and exponent calculations were made using the “powerlaw” python package [171], which implements the existing statistical methods of [76, 77, 172]. In Table 4.2 1P stands for the first (early) period of bitcoin and 2P stands for the second (late) period.

	In (1P)	In (2P)	Out (1P)	Out (2P)	All (1P)	All (2P)
α	1.87	2.2	1.8	2	1.87	2.07

Table 4.2. Calculated values for α for the in, out, and overall degree distributions for the first (1P) and second (2P) periods.

It has been found that network of ether transactions also follow a power-law distribution [114] with the exponent around 1.5. The similar degree distribution in bitcoin and ether transaction networks is not surprising - as it was mentioned in Chapter 3, in traditional finance the number of trades also converge to the power-law distribution. In

Figure 4.8. Degree distributions empirically calculated for ethereum separately for the two periods and shown as blue lines in the form of complementary cumulative distribution functions (CCDFs). The red lines show a power law function with an exponent of -2.

the next Chapter I will discuss this property in greater details, review existing scale-free network formation models and empirically show the formation process in cryptocurrency networks that results in power-law degree distribution.

Assortativity. Another important property related with the degree of a node is *assortativity* - the correlation between nodes' degrees. To characterize degree correlations we calculate the Pearson correlation coefficient of the out and in degrees in pair of nodes as it was proposed in [75]:

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j / 2m)k_i k_j}{\sum_{ij}(k_i \sigma_{ij} - k_i k_j / 2m)k_i k_j}, \quad (4.3)$$

where j_i^{out} is the outdegree of a node, k_i^{in} is the in-degree of a node.

Sometimes real networks are correlated - probability that node with degree k will connect to another node with degree k' depends on the degree k . The correlated networks are called *assortative*. In *disassortative* networks nodes with low degree are most likely to connect to the nodes with higher degree.

Degree is not the only measure to calculate assortativity - other node characteristics can also be used (for example, the age of a node). Assortative mixing by degree is a special case in network science since the degree itself is a network property and it often leads to another interesting features when assortativity dictates another property. For example, core-periphery structure is often found in assortative social networks [173].

Also, calculating the assortativity coefficient as it is shown in 4.3, not the only way to measure degree mixing. One can plot the *correlation profile* - two-dimensional histogram of degrees. This gives very precise picture of the network shaping process. This method has been used by Maslov for protein and Internet networks [174,175]. We have also looked at the correlation profile of the cryptocurrency networks to understand better where the disassortativity in degrees comes from. Also, we looked at the assortativity mixing profile of the wealth in both bitcoin and ether networks.

Figure 4.9. Degree assortativity mixing in BTC and ETH networks.

In Figure 4.9 x-axis is the degree range that receiver belongs to (0-814, 814-1632 etc.) y-axis is the degree range of a sender. The color of heatmap changes from light pink to dark purple according to the number of transactions made between two degree ranges. It can be seen that disassortativity in ETH mostly comes from the medium degree nodes - they tend to connect to nodes from other, mostly higher degree ranges. While small degree nodes (0 to 159) and high degree nodes tend to connect to similar range nodes. Overall, this mixing profile leads to the average assortativity coefficient of -0.055. In bitcoin network the disassortative behaviour is mostly among the small and medium degree nodes, while high degree nodes show the preference to connect to the high degree nodes. Overall assortativity coefficient for this network is -0.008 (close to random network coefficient). It should be noted that the number of nodes in ranges with higher degree (that show highly assortative behaviour) is quite small (it can be seen from the

degree distribution plots 4.7, 4.8), while number of nodes showing random attachment and disassortative behaviour is much higher. Therefore, it takes the overall assortativity coefficient to -0.055 for ethereum, which is more disassortative than bitcoin one (-0.008). From Fig 4.9 it can be seen that this difference in assortativity comes from the nodes with medium degree range - they tend to connect to nodes with higher degree in ETH, while in BTC they show random preference, resulting in coefficient closer to 0.

Figure 4.10. Wealth assortativity mixing in BTC and ETH networks.

Similar to the degree profile, x-axis is the balance range of receiver, y-axis is the balance range of sender. Balance in wallets was calculated as $value_{in} - value_{out}$, since the initial balance is unknown, some wallets are shown as having a negative balance. Topology of wealth mixing profile is quite different from the degree one - nodes with small wealth tend to receive more from all nodes with various balances. This topology looks more disassortative than the degree one.

Same as in Fig 4.10, x-axis is the receiver balance range and y-axis is the sender balance range. Topology of wealth mixing is quite different from the ether one - small balance nodes receive most of the coins from users with various balance ranges. Users of different balance range are quite uniform in both sending and receiving preferences.

Clustering, (also known as *transitivity*) tells the fraction of a neighbours which are connected to each other. Most of authors use the definition of clustering coefficient

proposed by Watts and Strogatz [176]:

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \quad (4.4)$$

Formula 2.4 is a local clustering coefficient, to compute the clustering coefficient for the whole graph, we need to take an average of C_i over all the nodes in graph:

$$C = \frac{1}{N} \sum C_i \quad (4.5)$$

Basic properties calculated for the static representations of bitcoin and ether networks are shown in table 4.1 in comparison with other empirical networks.

4.2.4 Robustness in Cryptocurrency Networks

We have studied the structure of bitcoin and ethereum network in previous sections - having a good understanding about the network’s structure helps to answer many important questions related to the future behaviour and resilience of the system. Unfortunately, there is still no a comprehensive theory that will connect the structure and function of networks. Nevertheless, studies on percolation and robustness of networks played a key role in estimating the resilience of various biological [177], social [178] and technological [179] systems. Overall, robustness is the “ability of the network to avoid malfunctioning when a fraction of its constituents is damaged” [110]. This topic is mostly studied due to the practical reasons and currently can be encountered in two variants: *static* robustness and *dynamic* robustness. Static robustness refers to the removal of nodes from network without considering the re-distribution of quantities that is being transported by the network, while dynamic robustness is when after each removal redistribution of flows need to be taken into account. Static robustness is used for social systems, while dynamic suits better for technological networks like Internet. Static robustness can be analytically treated using the percolation theory [180], but estimation of the dynamic robustness is much harder and one often has to rely on numerical simulations.

As for robustness in cryptocurrency networks, we need to first decide on what can be considered as “failure” of the network. As it was discussed in the earlier sections, cryptocurrencies can be treated as technological, financial or information networks. In this thesis we study the cryptocurrency systems as *financial* transaction networks, rather than studying from the technological or information flow points of view. Of course, the

network will be considered as failed when nodes will not be able to transact with each other, but this situation is likely when there are issues with the underlying blockchain technology, analysis of which is out of scope of this thesis. Overall, using the standard tools from percolation theory to check the robustness of cryptocurrency networks will not give the clear picture of the system's resilience - removing transacting people might change the structure of network, but will not lead to the failure of system as financial transaction system, as long as there are at least two transacting parties. Removing nodes (especially hubs) could lead to the instability from the economic point of view by causing price volatility and re-distribution of tokens. The economic effects of removing nodes with various properties (degree, betweenness, weight etc.) is an interesting and quite hard topic to study, where dynamics of token flows need to be carefully simulated. Moreover, it would need the clear understanding of relation between the market effects and token distributions among users. Also, when assessing the stability of cryptocurrency networks, the temporal nature of these systems also needs to be taken into account - network changes its structure with every block mined and single node can change its degree (as well as the other properties) from very small to the very high within a short period. The robustness analysis on temporal networks is still an open research question. Overall, considering the temporal nature of cryptocurrency networks together with the difficulties with simulating the dynamics of token flow, we leave the robustness check of cryptocurrency systems as an open research question. However, we would like to shed more understanding into the temporal properties of bitcoin and ethereum system in the next section.

4.3 Temporal Networks

As it was shown in the previous section, traditional Network Theory (Static Networks) is a developed field with strong mathematical foundation and it was successfully applied to many different systems and helped to understand and solve variety of problems. However, many real systems are not static and different approach is needed to model these systems. In empirical studies of systems with the temporal information it has been showed that indeed, we observe the different behaviour and neglecting the time stamp leads to the crucial loss of information. The importance of temporal information was discussed in [181]. Good example was shown by D. Braha - in empirical social network the degree and centrality measures change with time and static network is not able to capture this behaviour [182, 183].

Currently, the field of Temporal Networks is still understudied and there is no com-

monly agreed set and systematic understanding of Temporal Networks and how they affect the whole network structure. There is also little understanding on which features are important in Temporal Networks and if they are universal for various systems. Systematizing the field and proposing the framework for temporal network analysis has been done in [79, 80, 184]. As for the visualisation of temporal networks, movies are the best way to capture the changing temporal structure of networks. There is still no dedicated software tool to visualise networks with the time information. Nevertheless, we were able to visualise the bitcoin network by making a movie of multilayer networks (each layer is the one day network). This movie can be found in the Github page [185].

In this section I will review these frameworks, show differences, discuss advantages and disadvantages. Also, I will provide an overview of existing models of temporal network formation and demonstrate how temporal information of network was used in cryptocurrencies networks.

4.3.1 Definitions and Properties of Temporal Networks

When one tries to model the temporal information of network, there could be few ways to represent it. Temporal Networks can be shown as a contact sequences [123] – two nodes involved in interaction and the time of interaction; as a sequence of static network – each time step is a snapshot of static network; as a dynamic evolving static network – new nodes come to system, new links created over time [170]; as a memory network [80] – creation of interaction is a move, and one move is independent of others (Markov process).

Depending on the way of representing, there are different approaches how to model system as a Temporal Network. Overall, I would divide the existing approaches to model Temporal Networks into 3 categories:

1. **Extension/modification of Static Networks.** When researchers divide the system into temporal layers of certain length - multilayer representation of the original temporal network. In this case tools from Static Networks can be used for analysis. Overall, it is a very powerful approach, since the field of Static Networks is quite developed. But if we have a system with high time resolution, this approach might lead to the misleading results.
 - Time-aggregated network - splitting time into slices and building a graph (snapshot) for each time slice. By sliding the time slices one observes the dynamics of network properties.

- Weighted time-aggregated network – same as time-aggregated network, but each link is given a weight – number of times link appeared during the time slice. Problem of these two methods – time slices should be big enough, so that they capture significant information, but not too large, so that they won't lose the temporal information.
 - Dynamic networks – models that show how Static Network grows and evolves. Pioneering work of Barabasi and Albert [170], followed by many modifications and generalizations. This approach is not so much to capture the temporal nature of systems, but more to explain the emergent properties like power-law degree distribution that will be discussed in later section.
 - Building time series of contacts on a static graph – ($2 \times M$ matrix of the nodes of every link in the network). Allows to apply all the tools from SN, but for large network becomes too complicated for further analysis.
2. **Memory networks** (based on Markov process) – when interaction between two nodes is considered as a state, and next interaction is a possible state to which two already interacted nodes can go with certain probability. This approach transforms the set of links into the set of time-respecting paths on the network's nodes. [186–190]
 3. **Graph streams and Stream Links** – generalization of the classic Graph Theory to a language that deals with interactions over time. Instead of trying to model Temporal Networks as a static with additional modifications, this approach takes the classic Graph Theory which is a mathematical framework for static networks and generalizes the concepts. Stream graphs allow to directly model interactions over time just like Graph Theory models interactions between nodes. There is no dynamics in stream graphs, and therefore, it is equivalent to the classical graphs and it is possible to calculate all the properties as we do in Graph Theory [184].

Different approaches to model Temporal Networks propose different definitions of how to calculate the network properties. The table 4.3 summarises the main approaches of calculating properties in Temporal Networks:

Table 4.3. Various approaches of calculating main properties in Temporal Networks in comparison with the traditional Static Network.

Properties	Static Network	Temporal Network by [123]	High-Order Temporal Network [80]	Stream Graphs and Link Streams [184]	Comments
Network Representation	$G = (V, E)$, V - set of nodes, E - set of links. $E \subseteq V \otimes V$		$G^T = (V, E^T)$ V - set of nodes, E - set of links. $E \subseteq V \times V \times [0, T]$, $(v, w; t) \in E^T for [0, T]$	$S = (T, V, W, E)$, V - set of nodes, T - set of time instants, W - set of temporal nodes, $W \subseteq T \times V$. E - set of temporal links $E \subseteq T \times V \otimes V$	In HOTN, the temporal nodes are not taken into account, only links have time steps
Number of Nodes				$n = \sum_{v \in V} n_v = \frac{ W }{ T }$	Number of nodes under the most of approaches are calculated as for SN. Stream Graphs propose to calculate the contribution of node, since some nodes in network might be present longer/shorter time than others
Number of Links				$m = \sum_{uv \in V \otimes V} m_{uv} = \frac{ E }{ T }$	Most researchers calculate the links as for SN, stream graphs approach calculates the contribution of a link between pair of nodes

Table 4.3 – continued from previous page

Properties	Static Network	Temporal Network by [123]	High-Order Temporal Network [80]	Stream Graphs and Link Streams [184]	Comments
Local clustering coefficient	$C(G) = \frac{2m}{n(n-1)}$	Topological overlap of neighbourhood: $C_i(t_m, t_{m+1}) = \frac{\sum_j a_{ij}(t_m) a_{ij}(t_{m+1})}{\sqrt{[\sum_j a_{ij}(t_m)][\sum_j a_{ij}(t_{m+1})]}}$ average topological overlap – local clustering coefficient for TM: $C_i = \frac{1}{M-1} \sum_{m=1}^{M-1} C_i(t_m, t_{m+1})$		$C(S) = \frac{\sum_{uv \in V \otimes V} T_{uv} }{\sum_{uv \in V \otimes V} T_u \cap T_v },$ $cc(v) = \frac{\sum_{uw \in V \otimes V} T_{vu} \cap T_{vw} \cap T_{uw} }{\sum_{uw \in V \otimes V} T_{vu} \cap T_{vw} }$	
Global clustering coefficient	$tr(G) = \frac{ \nabla }{ \nabla },$ where ∇ - all triangles in G and $\vee$ - all triplets in G			$cc(V) = \frac{1}{n} \cdot \sum_{v \in V} n_v \cdot cc(v)$ $= \sum_{v \in V} \frac{ T_v }{ W } \cdot cc(v)$	
Degree and Average Degree	Degree - $k_i = \sum_{j=1}^n A_{ij}$, average degree - $c = \frac{1}{n} \sum_{i=1}^n k_i$			$d(v) = \frac{ N(v) }{ T } = \frac{ T_{uv} }{ T } - \text{degree}$ $d(V) = \frac{1}{n} \cdot \sum_{v \in V} n_v \cdot d(v)$ $= \sum_{v \in V} \frac{ T_v }{ W } \cdot d(v) - \text{average degree}$	

Table 4.3 – continued from previous page

Properties	Static Network	Temporal Network by [123]	High-Order Temporal Network [80]	Stream Graphs and Link Streams [184]	Comments
Paths and Distances	<p>Average path (directed network): $\langle d \rangle = \frac{1}{2L_{max}} \sum_{i,j \neq i} d_{ij}$</p> <p>Undirected: $\langle d \rangle = \frac{1}{L_{max}} \sum_{i,j > i} d_{ij}$, where $L_{max} = \frac{N(N-1)}{2}$</p>	<p>Temporal distance d_{ij} – length of temporal shortest path</p> <p>Characteristic shortest path length: $L = \frac{1}{N(N-1)} \sum_{ij} d_{ij}$</p>	<p>$dist^{temp}(u, v) := \min dist_{t_0}^{temp}(u, v, t_0)$</p> <p>- distance</p> <p>$P^{temp}(u, v)$ $dist^{temp}(u, v)$</p> <p>- path</p>	<p>Path P from $(\alpha, u) \in W$ to $(w, v) \in W$ is a sequence $(t_0, u_0, v_0), (t_1, u_1, v_1), \dots, (t_k, u_k, v_k)$ of elements of $T \times V \times V$</p>	All of the approaches define the path in TN as a time-dependent path, where ordering of links is crucial.
Closeness centrality	$C(v) = \sum_{u \neq v} \frac{1}{\partial(v, u)}$	$C_c(i, t) = \frac{1}{N-1} \sum_{j \neq i} \frac{1}{\tau(i, j, t)}$	$CC^{temp}(v) = \sum_{u \neq v} \frac{1}{dist^{temp}(u, v)}$	$C_t(v) = \sum_{u \in V} \int_{(s, u) \neq (t, v)} \frac{1}{c_t(v, (s, u))} ds$	
Betweenness centrality	$B(v) = \sum_{u \in V, w \in V} \frac{\sigma(u, v, w)}{\sigma(u, w)}$	$C_i^B(t_m) = \frac{1}{(N-1)(N-2)} \sum_{j \neq i} \sum_{k \neq j, k \neq i} \frac{U(i, t_m, j, k)}{\sigma_{jk}}$ σ_{jk} - number of shortest paths from j to k	$BC^{temp}(v) = \sum_{u \neq v \neq w} P^{temp}(u, w; v) $	$B(t, v) = \sum_{u \in V, w \in W} \int_{\substack{i \in T \\ u, j \in T_w}} \frac{\sigma((i, u), (j, w), (t, v))}{\sigma((i, u), (j, w))} di dj$	

Overall, centrality measures for Temporal Networks are comparatively easy to define by adapting the concepts from Static Networks. As it was shown in Holme’s approach [123] – in centrality measures the distance can be replaced by temporal distance or latency. As for other properties (how to define average degree in network, how to calculate the clustering coefficient etc.), it becomes more difficult, since there is no well-defined understanding of what is the temporal network – whether it is dynamic static network, sequence of contacts etc. Stream Graph approach was able to give the exact mathematical definition of what is Temporal Network and accordingly, derive all the properties. However, it should be noted, that it is a mathematical treatment that has not yet been applied to real systems. As for High-Order Temporal Networks and methods summarised by Holme, they were derived from and tested on empirical data sets.

4.3.2 Temporal properties of cryptocurrency transaction networks

In the previous section bitcoin and ethereum networks were visualised and analysed using the traditional Static Network approach. However, taking into account the temporal nature of these systems, we expanded the analysis to include the time information of link creation. We represented the network as a sequence of smaller static networks (multilayer approach), calculated the properties as for a static network, and observed the evolution of properties over time. We also calculated various properties using daily and monthly time windows. For properties such as assortativity, average degree, and clustering coefficient, a one-month time window allowed us to observe clear trends, while using a daily time window was too sparse to provide many meaningful insights. Fig 4.11 shows the evolution of the mixing patterns in the bitcoin and ethereum networks, Fig 4.12 shows how the clustering changes over time. As for the dynamics of the average degree (see Fig 4.13), we see its value grow steadily over time for both bitcoin and ether with some fluctuations at the later periods.

From Figure 4.11 it can be seen that the bitcoin network is highly assortative in the first period and stabilizes to slightly disassortative behavior (-0.1) in the second period. As for ethereum, there is an opposite behaviour - from highly disassortative at the start to slightly disassortative as the system evolves. As for bitcoin network, the changing from highly assortative to disassortative can be explained by the overall evolution of network: the early period was characterized by the many self-looped nodes, big hubs etc. After bitcoin system become developed and more and more people started to use it for the real

Figure 4.11. Monthly assortativity coefficients.

transactions, rather than for speculations. Mining pools (represented as the nodes with very high degree) are selling the bitcoins for the system users - higher degree node is tend to connect to the lower degree nodes. People use bitcoins to transact with the various companies and entities - it also explains the preference of low degree nodes to connect to the high degree one. Overall, as both systems evolve, the preference to transact becomes more random - assortativity coefficient is close to zero in both bitcoin and ether.

Figure 4.12. Monthly clustering coefficients for bitcoin and ether networks.

Figure 4.12 shows that in the beginning, the bitcoin network was not clustered. Then, wild fluctuations of the clustering coefficient are observed followed by the small fluctuations around 0.1-0.15. As for the ether network, clustering coefficient also had big fluctuations in the beginning of the system's development and also stabilized around 0.1-0.15. Overall, as systems evolve, both cryptocurrencies show lower and more stable clustering, which can indicate the start of using these systems for the actual monetary transactions. High clustering coefficient means that node's transacted parties (neighbours in network)

are connected with each other and this kind of behaviour signifies of the immature financial system. With time, both bitcoin's and ether's clustering coefficient becomes small and does not fluctuate much, which coincides with the period when systems were recognized by general public to be used for monetary transaction (around 2014 for bitcoin and around 2017 for ether).

Figure 4.13. Evolution of the monthly network average degree.

We observe that the average degree of both networks grows steadily with time. In bitcoin users transact more than in ether (average number of transactions 3 for bitcoin and 2 for ether at the end of period under analysis).

For the evolution of nodes and links, the monthly network shows a general trend of steady growth, while the daily time window allows us to see that the network size is directly related to its price dynamics. (Figs 4.14 and 4.15 illustrate this relationship).

Figure 4.14. The daily evolution of the total number of nodes and links and the history of the bitcoin market price.

Interestingly, the evolution of the total number of nodes and links shares some sim-

ilarity with the evolution of the bitcoin market price with a peak in 2011. Correlation for nodes number/price and links number/price was calculated and is 91% and 92 % respectively.

Figure 4.15. The daily evolution of the total number of nodes and links in ether network and the history of the ether market price.

Same as in the bitcoin market, price evolution and number of nodes/links is highly correlated - is 88% and 89 % respectively.

In addition to the analysis of active and inactive nodes, we also calculated the distribution of nodes' longevity. We define the longevity of a node as the duration that it stayed active in the network (creating new transactions) before the last transaction. More specifically, in the our networks we distinguish three states of nodes - active (transacted in a given time period), inactive (did not transact in a given period, but resurrected later) and dead nodes (disappeared from the network and never transacted again). Longevity for a node i was calculated as the total number of months when it was active (skipping the inactive months) before it becomes dead. Fig 4.16 illustrates that most of the nodes do not live long (stay active) in both systems.

To calculate the longevity the period from 2009 till 2015 was considered for bitcoin and 2015-2018 for ethereum. It can be seen that in both systems most of the nodes stay in the network for a short period, while only a few nodes remain active for long period.

4.4 Conclusions

In this Chapter a detailed network analysis was performed for the bitcoin and ether transaction networks. Static network analysis provided some insights into the network structure, revealing some interesting properties. We found that the in-degree, out-degree, and the overall degree distributions both in bitcoin and ether display a scale-free behavior.

Figure 4.16. Longevity distribution for nodes in bitcoin and ethereum.

It is difficult to assign the cryptocurrency networks to one of the previous defined category, but according to the main network characteristics they are closer to the technological one. Overall, the static network approach provides a good descriptive analysis of the system.

Anyway, the temporal network approach appeared to be a more informative - we were able to observe the dynamics of different properties and learn more about the network. One of the findings was the strong relation between the network growth (number of nodes and links) and the market price of bitcoin and ether - there is almost 90% correlation between the number of nodes (edges) and the price of bitcoin and ether. Also, for bitcoin network, it has been found that some network characteristics change their behavior before there is a change in system - stabilizing of assortativity and clustering coefficients before the bitcoin system entered the second period.

Overall, we obtained the good understanding of the overall structure of bitcoin and ethereum systems using the Complex Networks approach. However, there is still a big question of why does the system forms the scale-free structure? Next Chapter will be dedicated to the further investigation of the mechanism of the system formation.

Chapter 5

Mechanism of Bitcoin Network Formation and Evolution

Scale-free behaviour discovered in both bitcoin and ethereum systems was also found in many real-world systems and led to the development of various theoretical models aiming to explain its nature. In this Chapter we aim to uncover the mechanism behind the cryptocurrency network formation - we limit the investigation with the bitcoin transaction dataset only, since bitcoin is used for monetary transactions and reflects better the interactions between individuals. First, we provide an overview of the existing theories explaining the power-law behaviour and then show the results of the empirical investigation.

5.1 Models of Network Formation

Study of empirical networks has showed that features of real-world networks are quite different from those studied in graph theory. This discovery led to the necessity to develop new models that would be able to capture the properties of real networks.

The first attempt to model real systems as networks was under the assumption that the link between the nodes in real networks appears randomly, and it seemed to be a fair approach to model the systems as random graphs. The study of random graphs was initiated by Erdos and Renyi in 1959 [191], they proposed a model for random graph generation with L links and N nodes - Erdos Renyi random graph model ($G_{N,L}^{ER}$). The model takes N isolated nodes and makes connections between randomly selected nodes until the number of edges equals to L . When generating the random network it is useful to provide the parameters p (the probability that L attempts to connect $N(N-1)/2$ pairs

of nodes resulted in a link) and N - number of nodes. The probability that network has L links is then a binomial distribution. From here it is possible to calculate statistics as expected number of links Eq. 5.1, the average degree in random network Eq. 5.2 as well as the second moment $\langle k^2 \rangle$ and variance σ_k .

$$L_{pL} = p \frac{N(N-1)}{2} \quad (5.1)$$

$$\langle k \rangle = \frac{2 \langle L \rangle}{N} = p(N-1) \quad (5.2)$$

The Erdos-Renyi model was later modified by Gilbert by fixing the probability of connections instead of fixing the number of links ($G_{N,p}$) [192], which makes the model more realistic since in real networks the number of connections is rarely fixed. However, random graphs can not describe the behavior of real systems and adequately model their properties. There are a few modifications of random graph models based on the idea for degree not to follow the Poisson distribution. Bender and Canfield [193], and later Newman [78] proposed the models that allow to sample network with a given degree distribution. Later Chung and Lu took a power-law distribution as a given degree distribution to include into the random graph, so that it makes the random graph models useful for analyzing realistic networks [194].

Later the phenomena of small world was found in real networks - when despite of the network large size, the path length between any two random nodes is quite small. Clearly, the random graphs are not able to reproduce both of these features and a new model was proposed by Watts and Strogatz to capture the small world property and high clustering coefficient [195], [176]. This model interpolates between the regular lattice and random graph, so that it could have a high clustering coefficient as a regular lattice, but incorporate the small-world property which can be fitted into random graph.

Discovery of the power-law degree distribution in empirical networks led to the new theoretical models of network formation. The previous models were not capturing some important properties of real systems with the growth and preferential attachment being two of them. This inspired the creation of simple generative model - Barabási-Albert model or *scale-free* models. We call a network scale-free if its degree distribution follows a power-law, or in other words, when probability for node i to have a degree k exhibits a

power-law distribution:

$$P(k_i) = ck_i^{-\alpha}, \quad (5.3)$$

where c is some constant coefficient, and k_i is the degree of node i .

The Barabási-Albert model is the algorithmic model and defined as follows:

- **Growth.** At each time step add new node with $m(\leq m_0)$ links that connect the new node to m nodes that already exist in the network.
- **Preferential attachment.** the probability that a new node will connect to another node i is:

$$p_i = \frac{k_i}{\sum_l k_l}. \quad (5.4)$$

This is an algorithmic model that where nodes with higher degree have higher chance to attract links from new nodes. These two simple mechanisms are able to produce the scale-free behaviour of real systems [196]. Modifications of this model have been able to capture many realistic features of complex networks - growth [197], rewiring [198], aging [199], initial attractiveness of the node [200], and node's fitness [201]. The Dorogovtsev-Mendes-Samukhin model generalizes the Barabási-Albert model by adding a parameter of the initial node attractiveness [200], while Krapivsky added a nonlinear term to define an attachment probability of the new node [157]. Bianconi and Barabási have given a definition of fitness that captures the overall ability of a node to attract new links. Their model combines the impact of both degree and fitness to drive network connectivity [201].

Caldarelli replaced the preferential attachment rule based on degree, assuming that every node in a network has a fitness coefficient x , which is a random number taken from a given probability distribution. This model is somewhat close to the one proposed by Goh for static networks [202]. The fitness of a node represents its importance [203]. In [204], it was analytically proven that a power-law fitness distribution can be generated from any fitness distribution, with the simplest realization manifesting from the power law distributed fitness values. The probability of new node to connect to another node is

as follows:

$$p_i = \frac{x_i}{\sum_l x_l}, \quad (5.5)$$

where x is the fitness of a node. This is also an algorithmic model where fitter nodes are more likely to become hubs. Caldarelli model can be understood as the generalization of the Barabási-Albert model, where $x = k$ in particular cases.

The alternative theory of preferential attachment was proposed in [205–207]. Sample space reducing process assumes that over time complex systems reduce their number of possible states (options). This leads to the emergence of power-law behaviour, sample space reducing model allows to tune parameters so it is possible to reconstruct the network with any exponent. Dorogovtsev and Mendes showed that accelerated growth of links in a network (when the number of links grows faster than the number of nodes) may produce scale-free behavior [72]. The nonlinear growth has been observed in the WWW [73], Internet [208] and collaboration [209] networks.

Though the simulations based on these models were able to reproduce the scale-free behaviour, few studies have been done on empirical investigation of whether real-world systems follow these rules. Scale-free behaviour is one of the common features of many real-world complex networks and it has been observed in various complex systems - social, communication, biological, etc. Many studies have proposed generative mechanisms for scale-free degree distributions in empirical networks. Here we investigate some of the major mechanisms using our dataset so as to uncover the real underlying mechanism that leads to a power-law degree distribution of the bitcoin transaction network.

5.2 Mechanism behind the Bitcoin transaction network evolution

Network analysis revealed some interesting properties of the bitcoin system such as a power-law degree distribution and varying degree correlations.

First, we investigate if the bitcoin network possesses an accelerated growth [72] which could be the cause of the observed power-law degree distribution. Fig 5.1 shows that the number of links grows roughly linearly in relation to the number of nodes. Therefore, in the bitcoin network, we do not observe the accelerated growth that occurs in the WWW,

internet, or in collaboration networks.

Figure 5.1. Relation between the number of nodes and links for daily data.

From Figure 5.1 it can be seen that the growth of links in the bitcoin network is approximately linear. Next we check if the scale-free behaviour results from the mechanism of preferential attachment. Nodes in different systems might have various preferences to connect to other nodes. Degree preferential attachment proposed by Barabasi and Albert [196] has assumed that new nodes tend to connect to existing nodes with higher degree rather than to those with less connections Eq. 5.5.

Later generalizations of this model [197], [198], [199], [200], [201] add more parameters but still assume that a node's degree plays an important role in link creation.

For some complex systems, degree preferential attachment might be a reasonable explanation for the scale-free behaviour, while for other systems the information about the node's degree is either not available or is not an important factor for other nodes to connect to. In the bitcoin network, it is also reasonable to assume that richer nodes could attract more connections. Therefore, we might observe the mechanism of wealth preferential attachment, where the probability that new node will connect to node i is:

$$p_i = \frac{w_i}{\sum_l w_l}, \quad (5.6)$$

where w_i is the amount of bitcoins in the wallet of node i .

Another possible factor to create new connections might be the type of user in the bitcoin network. For example, exchanges and businesses would have more chance to become hubs, while common users' accounts would have less potential to have many transactions. For the intrinsic potential of a node to attract new links in a network, we define the characteristic of the *fitness* of a node. The fitness coefficient depends on the

type of user in the bitcoin network and therefore, does not change significantly over time. Fig 5.2 demonstrates a toy network, where each node has three properties - degree k , wealth w and fitness x .

Figure 5.2. Illustration of a toy network.

Each existing node in the network (nodes A, B, C, D) has properties such as degree k (how many transactions were made), wealth w (amount of bitcoins in wallet), and fitness x (the potential of node i to create new connections, where different types of users in the system have different intrinsic fitness). A new node E might choose to connect to any existing node depending on its preference - based on degree, wealth or fitness. Figure 5.2 shows how a toy network evolves based on different attachment mechanisms. For simplicity, one new node is added at each time step with other links remain static and unchanged.

5.2.1 Degree Attachment Mechanism

To see if there is degree preferential attachment in our network, we empirically check how a node's degree affects its ability to create connections. Having precise time steps of interactions in our dataset allowed us to derive the statistics of attachment preferences. We checked transaction by transaction if there were new nodes (users) that entered the system. For every new node we determined the degree of the node to which it was attached. We look at the two periods of bitcoin development separately (early period 2009-2011 and late period 2011-2013). For each period we check for every new node entering the system to which node it was attached (transacted with) - for degree attachment check, we look at the degree statistics of the node to which new node was attached. Then, we plot

Figure 5.3. Illustration of the growth of toy network based on various attachment mechanisms.

the distribution of obtained statistics. All new nodes were considered for both periods in order to obtain the comprehensive understanding of the attachment mechanism.

As we surmise that the attachment mechanism of the new nodes to the existing ones might differ from the mechanism of how old nodes connect with each other, we have examined the attachment mechanism separately for new and old nodes for both periods. For a fair comparison, we have normalized the overall distribution of degree attachment statistics.

Figure 5.4. Normalized degree attachment probabilities to new and old (existing) nodes in both periods.

Figure 5.4 shows the results of our analysis, which indicates that there is no preference for nodes to attach to high-degree nodes. Moreover, we found that in the second period, the preference to connect to nodes with smaller degree is more pronounced than to higher degrees. Based on this finding, we conclude that in the bitcoin network, degree preferential attachment is not the generative mechanism that drives the bitcoin network towards a scale-free degree distribution.

5.2.2 Wealth Attachment Mechanism

Given the fact that we are working with the financial network, it would be also fair to assume that *wealth preferential attachment* might be the mechanism leading to the scale-free behaviour. Since the bitcoin network is directed and links are weighted, it is possible

to track the amount of bitcoins in each wallet. For an additional analysis, we checked if attachment preference depends on the node's wealth.

Figure 5.5. Normalized wealth attachment probabilities for attachment to new and old (existing) nodes in both periods.

The analysis was normalized by the total amount of bitcoins transacted in each period. From Fig 5.5 we can see the clear preference in both periods that nodes do not tend to transact with other nodes that have higher amounts of bitcoins in their wallets.

5.2.3 Fitness Distribution of Network Nodes

As it was shown, there is no clear dependency between the degree (or wealth) of a node and its ability to create new links in the bitcoin network. We then empirically calculate the ability of a node to attract new connections at each time step and call it nodes' *fitness*:

$$x_{i,n} = \frac{k_{i,t=n} - k_{i,t=n-1}}{\sum_{j=1}^m (k_{j,t=n} - k_{j,t=n-1})}, \quad (5.7)$$

where $k_{i,t}$ is the degree of node i at time t , $n=1, 2, \dots, 23$ is the month of the given period, and m is the number of nodes in the network.

We found that there are two distinct patterns in node fitness: first, some fit nodes do not live long in the network - they create relatively more transactions in a short period and then disappear. We would attribute this behaviour pattern to the speculative users of the system. Another behaviour type observed in nodes with high fitness is high longevity

and constant creation of new links. This type of node might be exchanges, businesses that accept bitcoin payments, etc.

Figure 5.6. The fitness distributions computed for all nodes month by month in the second (stable) period of bitcoin transactions and shown in CCDF (blue line). The Red line shows the power-law function with an exponent of -2.

Considering the limitation of space, we show the fitness distributions for every 3 to 4 months. It can be seen that the distribution has a power-law shape with some deviation from month 4 until month 13. We have calculated fitness values for all the nodes for each month over the second period, as shown in Figure 5.6. Due to limited space, we have only shown the results for each month at an interval of three to four months. Figure 5.6 reveals that the distribution of the nodes' fitness has essentially the shape of a power-law except for the period from month 4 to month 13. The observed deviation from power-law for these months arises from a hike in the transaction activities of bitcoin with a corresponding rise in the number of nodes and links, and price fluctuations (see Fig 4.14). Once this transient spurt of interest is over, the fitness distribution of the nodes was found to settle back into a power-law distribution. Figure 5.6 displays the complementary cumulative distribution function of the nodes' fitness (blue line) plotted with a power-law fit (red line). The exponent of the power-law has a value of -2. Note that nodes with zero fitness were removed from the analysis because of their inactivity during the given months.

To understand if the fitness coefficient of a node remains approximately constant during its lifetime (or if the coefficient changes significantly instead), we have looked into the evolution of all nodes. We have divided nodes according to their initial fitness coefficient into three groups - with low, medium and high fitness coefficients. Figure 5.7 shows the evolution of the average fitness coefficient for nodes, where we observe a level

of fitness that remains relatively constant over time. This finding supports our definition of fitness (intrinsic potential to create new connections that depends on the type of user) in the bitcoin network due to the nature of the nodes - exchanges remain to be exchanges while common users do not become traders or businesses over time. The standard error for all groups was close to 0 across all months under analysis.

Figure 5.7. Evolution of the fitness coefficient for all nodes.

The age of a node is an indicator of the temporal nature of a network. In a static network, nodes are immortal. On the other hand, the age of a node is finite in a temporal network. As a result, we are interested in how the degree and the fitness coefficient of a node are related to its age. One might expect that a long-lived node would have a higher degree, or a larger fitness coefficient. However, we found a low correlation between the age of a node and both its degree and fitness throughout the period of study (see Table 5.1). The correlation between fitness and degree is also lower than expected for a network with links created by fitness preferential attachment. These lowered correlation can be explained by the fact that while high fitness nodes create many links in a short span of time, these nodes become inactive quickly and do not have a high degree by the end of the period. This dynamical and topological behavior of the nodes is not surprising in a fast evolving financial network of transactions.

	Fitness	Degree	Age
Fitness	-	0.47	0.03
Degree	0.47	-	0.02
Age	0.03	0.02	-

Table 5.1. Correlation between the fitness, degree, and age for all the nodes in the network by the end of the second (stable) period.

We found the correlation between degree and age, and between fitness and age, to be small.

5.3 Generative algorithm for the Bitcoin Transaction Network

In the previous section, we have investigated the driving mechanism of the bitcoin network that could potentially lead to the observed scale-free behaviour (with a power-law degree distribution), but we found that the degree or wealth of a node is not the cause of link creation in the bitcoin network. On the other hand, we have uncovered that nodes possess intrinsic fitness which is power-law distributed, and we expect this to bring about a power-law degree distribution. In the study of Caldarelli [203, 204], it was shown that scale-free behavior might arise from any fitness distribution with the simplest realization being fitness coefficients following a power-law. The approach of Caldarelli has inspired our creation of a generative algorithm for the bitcoin transaction network, with the generated network yielding both the static and temporal properties of the real-world system.

5.3.1 Algorithmic model for bitcoin network generation

The algorithmic steps of the model are as follows:

1. Create a large set of seed nodes with links to initiate the synthetic network. For every node i , assign a fitness coefficient x_i randomly from a power-law distribution with the exponent 2. The value of x_i reflects the node's ability to create new connections.
2. Simulate the entrance of new nodes to the network. This is the birth process.
3. Randomly create links between new-to-new nodes (all new nodes have an equal chance to create the link with the another new node). On the other hand, links between new-to-old nodes and old-to-old (existing) nodes are created with old nodes being assigned a number of links proportional to their level of fitness.
4. Assign fitness coefficients for the new nodes so that the fitness distribution remains power-law distributed.
5. Define the fraction of nodes that will be inactive for the next time step. Select these inactive nodes randomly according to the given fraction and assign a fitness coefficient of zero to them. This is the death process.
6. Repeat the algorithm from step 2.

Note that steps 2 and 5 emulate a fundamental property of temporal networks - the networks do not only grow with time, they can also shrink. Nodes are not constantly active, they might be temporarily inactive or even leave the network. For nodes that have become inactive but return to an active state later, we call them resurrected nodes. If the

inactive nodes never turn active again, we consider the nodes to be dead. Also, assigning the fitness coefficient at step 4 instead of calculating at step 3 is important so that new nodes also have chance to obtain the higher coefficient (like in real system), instead of calculating it at the step (3) and keeping constant.

A feature of our generative algorithm is its incorporation of our empirical findings, where links are created according to fitness preferential attachment, and the fitness coefficients of nodes are power-law distributed (which do not change significantly over time). Also, we have implemented the birth and death processes by capturing the evolution of nodes in the bitcoin system.

5.3.2 Synthetic Network simulations results

We have created synthetic temporal networks based on the proposed algorithm. During the generation process, we used the initial network from the empirical data at the start of each given period (first and second) as seeds to start our simulations. The next step of network generation is to include new nodes. The number of nodes to be added at each time step is defined according to the empirical growth rate as follows:

$$R = \frac{n_{new}}{n_{total}}, \quad (5.8)$$

where n_{new} is the number of new nodes entering the system at a particular month, while n_{total} is the total number of nodes in the same month. Fig 5.8 shows the evolution of this network growth rate R over four years.

Figure 5.8. Evolution of the network growth rate R .

Figure 5.8 shows the fraction of new nodes entering the system at a given month over the total number of active nodes in that month for both periods. Sharp decline in July

2011 can be explained by the shock event, when the largest exchange at the moment Mt. Gox was robbed in June 2011 and the trust to bitcoin was disrupted.

As described in 5.2.3, we calculated the statistics of nodes to which new ones were attached. From there we observed that in the real bitcoin network approximately $2/3$ of new nodes are attached to the other new ones (degree of zero at the moment of attachment), while $1/3$ of new nodes are attached to the existing ones (degree greater than zero).

We followed this same *rule* in our network generation process. Furthermore, the number of new links we created at each time step has been set to be consistent with the average degree of the empirical dataset – thus for each new node created, we add 1.57 links for the first period, and 2.44 links for the second period.

We then assigned fitness coefficients randomly to the new nodes of the synthetic network according to an empirical fitness distribution . After which, we randomly set a fraction f of the network nodes to the inactive state according to the empirical findings. We repeated the algorithm, and after running for 24 time steps, the synthetic network was found to evolve towards a structure with a power-law degree distribution. We calculated the nodes' degrees of our synthetic network by aggregating the in- and out-degrees, thereby treating the network as undirected. A comparison of the degree distribution of our synthetic network with the empirical bitcoin network shows good correspondence in scale-free behaviour (see Figure 5.9). In particular, it is worth noting that the exponent of the degree distribution around -2 is close to that of the fitness distribution. This is expected since the fitness coefficients directly determine new link creation for the node i . The similar exponent value with those observed in the empirical network thus validates the assumptions of our model. It should be noted that there is a difference in maximum degree between the real and synthetic networks, this can be explained by the existence of large entities (exchanges etc.) in real system, that the design of synthetic network simulation was not able to reproduce.

Subsequently, we generated an ensemble of 10 synthetic temporal networks for the evaluation of their average network structural properties. Table 5.2 compares these network properties between the synthetic and the bitcoin transaction network. The scale-free behaviour is well captured by the model, while the clustering coefficient differs between the real and model networks. A higher clustering coefficient in the real network might arise from the social aspects - recommendations, following the choice of neighbours, etc. Our model simulates connections without capturing these social aspects and therefore, we observe the model network to be less clustered. For the assortativity coefficient, our

Figure 5.9. A comparison of the complimentary cumulative degree distribution between the empirical and synthetic network for both periods. All distributions are observed to display power-law behavior. The red line shows the power law function with an exponent of -2.

simulation was able to capture the decreasing trend from first to second period, although the value slightly differs.

Property	Model	Empirical
Number of nodes (1st Period)	193,012	193,012
Number of nodes (2nd Period)	4,689,069	4,689,069
Number of links (1st Period)	302,985	302,116
Number of links (2nd Period)	11,441,315	11,432,100
Clustering coefficient (1st Period)	0.08 ± 0.02	0.116
Clustering coefficient (2nd Period)	0.12 ± 0.03	0.24
Average Degree (1st Period)	1.57	1.57
Average Degree (2nd Period)	2.44	2.44
Assortativity coefficient (1st Period)	-0.05 ± 0.01	0
Assortativity coefficient (2nd Period)	-0.08 ± 0.01	- 0.1
Power law exponent (1st Period)	1.94.	1.98
Power law exponent (2nd Period)	2.16	2

Table 5.2. A comparison of basic network properties between the synthetic and empirical bitcoin network.

We observe that while our model was able to reproduce the scale-free behaviour and disassortativity of the empirical network (for the 2 periods), the clustering coefficients still differ. We relate it to the social aspects of the bitcoin system which were not implemented in the current model. To further validate our model, we have performed analysis on the other topological properties of the synthetic network to see how close they are to the empirical network. For this, Fig 5.10 illustrates our results on the complementary

cumulative distribution of the clustering coefficient $c(k)$ versus node degrees. Note that $c(k)$ was calculated by averaging over the clustering coefficient of all nodes in the network that possess the same degree k . We found that the clustering coefficient is power-law distributed for both periods, and for both the empirical and the synthetic networks. In the first period, we observed that $c(k) \propto k^{-1.68}$ for the synthetic network, and $c(k) \propto k^{-2.2}$ for the empirical network. On the other hand, during the second period, we found that $c(k) \propto k^{-1.71}$ and $c(k) \propto k^{-2.32}$ for the synthetic and the empirical network respectively.

Figure 5.10. A comparison on the clustering coefficient distribution between the empirical and synthetic network for both periods. All distributions are observed to display power-law behavior. The red line shows the power law function with an exponent of -2.

5.4 Conclusions

In this Chapter, our detailed network analysis of the bitcoin transaction network has provided insights into its network structural properties and the underlying evolution mechanisms. In the previous Chapter we found that the in-degree, out-degree, and the overall degree distributions during the first and second periods display scale-free behavior. In this Chapter we have uncovered the dynamics of different structural properties which yielded a change in the set of network characteristics over the two periods. In particular, the assortativity and clustering coefficients were observed to stabilize before the second period, while deviation from the power-law of the fitness distribution happened a few months before the sudden increase in transaction activities and market price. Our empirical study on the attachment mechanism of the bitcoin transaction network showed that there is neither degree preferential attachment of nodes nor wealth preferential attachment in the system, although certain nodes are intrinsically better at attracting links due to

properties not observed in their local network properties. We surmise that the power-law behavior observed in the degree distributions results from the nodes' intrinsic fitness, i.e., its intrinsic ability to create new connections. This could be related to the types of users - exchanges are more attractive to connections than active traders, who are then more attractive than a common adopter of bitcoin. Our empirical analysis on the distribution of node's fitness of the bitcoin network indicates a power-law distribution that is relatively stable over time. By generating synthetic temporal networks based on this fitness distribution and further incorporating the birth/death of nodes, we have demonstrated that the scale-free behavior of the degree distribution, as well as other topological properties, are consequences of the mechanism of fitness preferential attachment.

Chapter 6

Behavioural Structure of Cryptocurrency System

In the Chapters 4 and 5 we looked at the big picture of the cryptocurrency systems, calculated the statistical properties of the overall structure and uncovered the mechanism behind the network formation. While we were closely looking at the *links* of networks, we have treated all *nodes* as particles in physical system disregarding their possible difference between nodes. However, for the complex, self-organized system like bitcoin and ethereum, behaviour of nodes (people in our systems) is quite important to take into account. In this Chapter we focus on the understanding of the system's parts, we develop the methodology to derive distinct behavioral types in system, and our approach is novel and very different from models used traditionally in financial systems. First, we provide an overview how people's behaviour was modelled in traditional financial systems and discuss advantages and shortcomings of existing approaches. Next, we propose the methodology based on the Machine Learning models to derive the real behavioral structure in cryptocurrency markets. Then, we model the dynamics of obtained behavioral composition as Absorbing Markov Chain and discuss obtained results.

6.1 Background on Modelling People's Behaviour in Financial Systems

It is well-known that financial systems are complex and their evolution depends heavily on the behavior of their agents (users). This realization can be traced back to the times of Adam Smith in the late 1700s. Subsequently, there are many theories trying to model

the complexity of financial systems based on historical patterns (Keynesian economics), with the eventual emergence of the paradigm of “rational expectations” in the twentieth century. Rational expectation assumes that people have access to all the information, act rationally, and adapt fast to new conditions. This is an idealized theory within which many useful models were developed and it provided insights into key aspects of market behavior. However, under this paradigm many real situations observed in market were not able to be characterized and explained. It was soon realised that human behavior is more heterogeneous and complex than what efficient market theory has assumed. It spurred the invention of many agent-based models which simulate diverse users (agents) interacting according to a set of prescribed rules. Agent-based modelling proposed by physicists in analogy with the interacting particles that is well studied in Statistical Mechanics. These models are able to explain many stylized facts in financial time series that previous models have failed to reproduce. Although these models perform well in explaining facts observed in financial data, they are based on hypothetical assumptions and they differ from modeller to modeller. Due to the sensitivity of financial data, there was little opportunity to construct behavioral models on the basis of empirical findings. However, researchers have found ways to gain experimental evidence by conducting laboratory studies [86, 87, 210]. The downside of the experimental approach is the limitation of sample size which prevents generalization of its findings to real financial market.

The invention of blockchain and cryptocurrencies has overcome this problem by opening up large data-sets of financial transactions for close examination. The study of user behavior in the cryptocurrency market has also mainly been conducted from the point of view of anomaly detection [211–214]. In [215] it was shown that outliers in in-degree distributions in bitcoin network are due to the artificial behaviour of users rather than normal economic interactions. Research on strategies and users behavior in cryptocurrency markets are few. Cocco [216, 217] has proposed an agent-based model to explain price movements in bitcoin. They assumed that there are two types of behavior in the bitcoin system: chartists and random traders. The authors then prescribed behavioral rules to the agents according to their type and observed how they affected the market price of bitcoin. As for experimental research to understand the behavior of users of the cryptocurrency system, interesting work has been done by Krafft [31]. These researchers have conducted online experiments to study how users are susceptible to peer influence in cryptocurrency markets. By placing experimental orders in the Cryptsy exchange and then observing users’ behavior afterwards, they assessed the strength of peer influence on the users. This study has shed light on the understanding of causal impact of individual

opinion in large cryptocurrency markets. However, there is still no understanding of the real behaviour of people in market.

The use of machine learning methods in blockchain and cryptocurrency data sets is not new and has already been implemented for various purposes. The most popular task is to use machine learning to detect anomalous user behavior. The authors in [211] have analysed bitcoin transaction network data for the four years (2009-2013) with the goal of detecting suspicious users. They used Local Outlier Factor (LOF) to first detect outliers in the dataset, and then employed k-means clustering to calculate the relative distances between the cluster centroids and the detected outliers. This enabled them to estimate the performance of LOF – if the computed distance is small, LOF has performed poorly. Overall, the authors were able to detect anomalous transactions using this approach. Another study on the detection of anomalous user behavior has been conducted done by Monamo [212]. They used trimmed k-means clustering to detect outliers, i.e. data points that are farthest away from the cluster centroids. In [213], the authors trained a supervised machine learning algorithm to predict the category of the unidentified users. Identified users (a sample of 957 of the 385 million transactions) were used as a training set for the Gradient Boosting algorithm, and classifiers were built to differentiate users among 12 categories - exchange, mining pool, personal wallet, scam, darknet, ransomware, hosted wallet, gambling, mixing, stolen coins, merchant services and others. Interestingly, they are able to predict the category that the user belongs to with an accuracy of 80%.

Overall, the detection of user types in cryptocurrency systems is mainly to address the security and privacy issues. Furthermore, most of recent research focuses on the bitcoin transaction network, with less work being performed to understand the cryptocurrency system as a whole [218]. So far, lots of work have been done to understand the behavior of financial markets' participants both at the theoretical and experimental levels. These approaches have been successfully applied to cryptocurrency markets as well. However, there is still a lack of studies that derive users behavior in financial markets from empirical evidences. The successful implementation of machine learning methods in identifying anomalies and user categories has inspired us to employ them for the identification of behavioral patterns in cryptocurrency system which would contribute to the understanding of human behavior in financial markets [219].

6.2 Pre-analysis for Behavioural Classification

6.2.1 Periods for Analysis

Our interest is to elucidate the behavioral composition of the users at different periods of the bitcoin and ethereum systems. Overall, cycles in financial systems can be summarised as price drops (bear market), price increase (bull market) and periods of stable price. In order to get the robust behavioral composition, we take into account various periods of price movements. To define the trend of price movement, the average relative daily return and the total relative return in one month for both cryptocurrencies are calculated. Table 6.1 shows the chosen periods with the return values. For bitcoin, the distinct price increase was detected in June 2011, followed by bear market in July 2011. For ether, May 2017 showed the highest price increase, followed by sudden decrease in July 2017. We take these distinct periods of price change in short period for analysis, together with the periods when the price change was minimal (stable period) - January 2014 for bitcoin and April 2017 for ether.

Period	Average return	Total return	Trend
BTC Jun 2011	7.2%	68%	Increase
BTC Jul 2011	-0.5%	-30%	Decrease
BTC Jan 2014	0%	-0.03%	Stable
ETH May 2017	4%	202%	Increase
ETH Jul 2017	-0.5%	-30%	Decrease
ETH Apr 2017	1%	5.2%	Stable

Table 6.1. Periods when bitcoin and ethereum systems show significant and non-significant price changes.

The average and total returns are calculated for each period, and based on the evaluated value, we define the trend of the price (price increase, price decrease, or stable price).

We additionally analyze user composition in both bitcoin and ethereum during the occurrence of extreme events that affect the entire cryptocurrency system: (a) December 2017 - January 2018 (aka Crypto Bubble), (b) the period after this event, and (c) the shock event at the end of 2018 known as Crypto Winter. Overall, our choice of periods for analysis is quite diverse - we take into account three different stages of price movements for each currency and moreover, we look into the behavioral composition during shock events.

I design a temporal, weighted, and directed network for each of the defined periods

where each link (i, j, w, t) is a transaction between two nodes (users) i and j at time t with the amount of coins w . The main properties of each network are shown in Table 6.2.

Period	V number	E number	Average Degree	Clustering	Assortativity
BTC Jun 2011	342,722	663,642	1.93	0.041	-0.1
BTC Jul 2011	373,014	609,790	1.63	0.04	-0.1
BTC Jan 2014	1,092,008	2,891,972	2.65	0.12	-0.07
ETH Apr 2017	364,504	725,205	1.99	0.104	-0.115
ETH May 2017	774,119	1,567,020	2.02	0.096	-0.105
ETH Jul 2017	1,387,519	2,890,970	2.08	0.091	-0.169
BTC Dec 2017	12,593,945	22,716,559	1.8	0.101	-0.1
BTC Feb 2018	6,194,508	10,777,423	1.74	0.05	-0.08
BTC Nov 2018	8,629,863	14,184,283	1.64	0.04	-0.04
ETH Jan 2018	8,788,002	18,748,342	2.13	0.065	-0.055
ETH Mar 2018	4,278,371	7,656,868	1.79	0.097	-0.085
ETH Dec 2018	3,055,070	5,430,566	1.78	0.153	-0.122

Table 6.2. Network properties calculated for bitcoin (BTC) and ethereum (ETH) networks for each period under analysis.

Overall, we can see that all the constructed networks are large, with millions of nodes (V) and links (E) for most cases. The networks are also quite sparse with an average degree of about 2 and a small clustering coefficient. Both bitcoin and ethereum networks are slightly disassortative, with a coefficient of around -0.1 .

6.2.2 Feature Extraction and Analysis

From the constructed networks, we calculate properties (features) of each user. Our goal is to yield the various strategies employed in the cryptocurrency market. According to the literature [220–222], agents in the financial markets tend to adopt one of the following strategies: buy, sell, trade or hold. To see if this behavioral composition holds in the cryptocurrency market, we need the following information from each user: (i) frequency and amount of transactions, (ii) total accumulated/spent coins, and (iii) number of outgoing and incoming transactions. Based on the constructed networks with properties shown in Table 6.2, it is possible to obtain the required information.

The following are features that we shall use to define the type of strategy employed in the cryptocurrency market:

- Total degree of a node i at time interval t . (This feature corresponds to the total number of interactions of node i at time interval t , which reflects the transaction frequency of a user.)

- In-degree of a node i at time interval t . (This feature corresponds to the total number of interactions at time interval t with the node being a receiver of the transaction.)
- Out-degree of a node i at time interval t . (This feature corresponds to the total number of interactions at time interval t with the node being a sender of the transaction. These two features: in- and out-degrees of a node, help us understand how active a node is in sending/receiving coins.)
- Outgoing value. (This is the total amount of eth/btc sent in the time interval t . It is the sum of the weights w of a node i when it is a sender of the transaction.)
- Incoming value. (This is the total amount of eth/btc received in the time interval t . It is the sum of the weights w of a node i when it is a receiver of the transaction.)
- Total balance. (This is the net number of coins in the account balance of node i in the time interval t . These three features: incoming value, outgoing value, and balance of a node show the wealth of a node and its preference to accumulate or spend coins.)
- Total transacted value. (This is the sum of outgoing and incoming values of node i in the time interval t .)

Due to the novel nature of cryptocurrencies, there are many users both in bitcoin and ethereum that are experimenting with system by placing random orders and leaving the system quickly. We consider these users as noise and omit them from the analysis, since they do not contribute to the overall understanding of market behavioral structure. Moreover, they cause confusion to our analysis by the machine learning algorithms as we try to define the existing behavioural pattern in the market. We define “noisy” user as follows - total degree for whole transaction history is less than 2, and traded value is less than 1% of all transacted values at that time step. However, it should be noted that not all “noisy” users in bitcoin can be considered as experimenting users - some of them might be the consequence of BitIodine clustering limitations, as discussed in the Chapter 2. Overall, the percentage of “noisy” users during different periods shown in Fig. 6.1:

More than 50% of users in each period of ETH have degree less than 2, small transaction value and appear only once. In bitcoin vast majority are “noisy” users but surprisingly, in the period of Crypto Winter and after Crypto Bubble the proportion of “noisy” users is only 36-38%. “Noisy” users are removed for the further analysis.

Figure 6.1. Proportion of “noisy” users in transaction networks.

Then features are calculated for all users under analysis in each network (see Table 6.2). In [25, 27] it has been shown that degree and wealth in the bitcoin network are power-law distributed with the exponent around 2. We found that indeed, in, out and total degree are power-law distributed with the exponent around 2-2.3, while in, out and total value are also power-law distributed with the exponent around 1.7-1.9. Heavy tailed degree and wealth distributions in various financial markets is well-known fact and has been extensively researched [47, 48, 108, 223]. Cryptocurrency market also shows this property similar to other markets as it was shown in Chapter 3.

We then select features for the machine learning model from those listed above. Overall, unsupervised feature selection methods can be categorized as filter and wrapper approaches [224]. Filter approach selects the most relevant features based on certain criteria (correlation, entropy etc.), while wrapper approach first defines subsets of features and then evaluates them based on the result of the certain clustering algorithm. In our research we first calculate correlation between various features, then define a few set of features for further analysis. The final choice of the most optimal feature set will be based on the clustering result.

Fig 6.2 and Fig 6.3 show correlations between features for different periods in bitcoin and ethereum:

From Figure 6.2 it can be seen that total degree and total value are redundant features as they are always highly correlated with others. In later periods, value in and value out also show very high correlation.

Figure 6.3 shows that in, out and total degrees are always highly correlated, same as

Figure 6.2. Correlation between all features in ETH.

Figure 6.3. Correlation between all features in BTC.

in, out and total value.

After understanding relationships between users' properties, we define sets of features for each cryptocurrency - *Set 1* includes all features listed above, *Set 2* omits total degree and total value as they showed the highest correlation with other features, *Set 3* includes only features with low correlation, this set is different for etherem (in degree, out degree, total value and balance) and bitcoin (total degree, total value and balance). Later we will use all sets of features and show which set provides better clustering.

6.2.3 Review of Methods

There are many unsupervised clustering algorithms that can be broadly divided into two categories: hierarchical and partitional techniques [225]. Hierarchical clustering produces a dendrogram (a hierarchy of nested clusters) by iteratively merging smaller clusters to larger ones [225, 226]. Although this method provides detailed information on how clusters are related to each other, final output on structural information can not be clearly visualised when performed on large number of points. This might result to a wrong interpretation of patterns similarity in dataset. Moreover, hierarchical clustering is computationally expensive method due to the necessity to calculate the dissimilarity matrix on each step to compare each pattern to others. In the study of [227] the performance of various unsupervised clustering algorithms was evaluated - it has been found that k -means clustering outperforms other partitional methods like DBSCAN [228, 229] or OPTICS [230] in both one and multi-dimensional feature sets. Authors have also found that spectral methods have higher accuracy than k -means clustering when there are more than 10 feature sets. However, with the less number of features, their performance was compatible. Therefore, we decided to use k -means clustering method that have both methodological and computational advantages.

K -means algorithm is a partitional squared-error clustering method. It is broadly used for unsupervised clustering due to its computational effectiveness and easy implementation. The method comprises the following steps [231]:

1. Initialize the cluster centroids $\mu_1, \mu_2, \dots, \mu_k$.
2. Segment the data into k groups. Assign each data point to the closest centroid and change the centroid to the average of its assigned points so that the distortion function:

$$J(c, \mu) = \sum_{i=1}^m \|x^{(i)} - \mu_{c(i)}\|^2 \quad (6.1)$$

converges. Note that $x^{(i)}$ is one of the m data points and $\mu_{c^{(i)}}$ is the cluster centroid assigned to the i -th data point, i.e. $c^{(i)} \in \{1, 2, \dots, k\}$.

Despite the efficiency of k -means clustering, there is a preliminary step where the optimal number of clusters has to first be carefully identified. Various statistical methods offer different ways to calculate this number from the dataset [232–234]. We use the elbow method in the next section to determine the optimal number of clusters for the bitcoin and ethereum cryptocurrency systems. Elbow method runs k -means clustering on the data set for a range of values (clusters) and calculates the average score for each value (in our case, we calculate the sum of square distances from each point to its assigned cluster center). When all values are plotted, it is possible to decide on the optimal number of clusters - there is a strong inflection point that makes the plot look like elbow - hence the name.

Support Vector Machines (SVM) is a supervised machine learning method that has gone from being largely unnoticed [235] to a famous method [236]. It is widely applied in the field of digital recognition [237], computer vision [238], and text classification [239]. SVM shows comparable results with neural-network-based algorithms when applied to classification and pattern recognition problems [240]. When used for classification purposes, SVM divides the training dataset with the hyper-plane which is the most distant from the data points. Assume that we have a training labelled dataset

$$S = (\mathbf{x}^1, y^1), \dots (\mathbf{x}^m, y^m) \quad (6.2)$$

of size m , where $\mathbf{x}^i$ is the data point and y^i is its label. The SVM algorithm finds a hyper-plane $(\mathbf{w}, b)$ such that γ in the following equation is maximised:

$$\gamma = \min_i y^i \{ \langle \mathbf{w}, \phi(\mathbf{x}^i) \rangle - b \}. \quad (6.3)$$

Note that the quantity $\{ \langle \mathbf{w}, \phi(\mathbf{x}^i) \rangle - b \}$ corresponds to the distance between the point $\mathbf{x}^i$ and the decision boundary, where $\mathbf{w}$ is the weight vector, $\mathbf{x}^i$ is the input vector and b is the bias vector. When multiplied by the label y^i it gives a positive value for correct classifications and a negative for wrong ones. Minimum of this quantity over the whole data set is positive if data is linearly separable, and γ is called the margin. When data is not linearly separable, SVM can be improved with kernels that realize the non-linear

mapping to the feature space.

Originally, SVM was designed for binary classification. It has now been extended to cases where there is more than two classes by considering the multi-class problem as a series of binary ones (one-against-one, or one-against-all strategies) [241]. In section 6.3.2, we will show how these methods are used to identify user groups with different strategies within the cryptocurrency systems.

6.3 Methodology and Results

6.3.1 Methodology

In consequence of the power-law distribution of features, the preliminary step is to normalize and scale the features for proper performance of the machine learning methods. Otherwise, the common machine learning algorithms can lead to biased results due to the data in the fat tail of the distributions. Normalization was performed by taking the logarithm of the original value of the feature: $x' = \log(x)$ so the power-law behaviour becomes a linear one.

Since our dataset is unlabelled, we do not have information on the existing strategies employed in the cryptocurrency markets. Hence, we use k -means clustering to cluster users into groups. But before using k -means clustering, we need to identify the optimal number of clusters. We use the elbow method [232] for this purpose and calculate the number of clusters for all datasets and all feature sets. After clustering data using k -means we examine the distribution of the original properties of each group of users and analyse how distinct they are from those in the other groups. Based on that, in section 6.3.2 we define rules for each cluster and provide the characteristic descriptions of their properties.

We then validate statistically how well our data points fit into each cluster by calculating their silhouette score [234] for all feature sets and all periods as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \quad (6.4)$$

where $a(i)$ is the mean distance between i and all other data points in the same cluster, while $b(i)$ is the smallest mean distance from i to all points in the other clusters that i does not belong to.

Then we use more advanced technique of Support Vector Machine to cluster users

more precisely according to the defined rules.

Overall, our algorithm is as follows:

1. Use elbow method for all feature sets to find the optimal number of clusters
2. Perform k -means clustering on the unlabelled feature matrices A_c and obtain the vector of labels V_c :

$$A_c = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} \\ a_{21} & a_{22} & \dots & a_{2j} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nj} \end{bmatrix}, \quad V_c = \begin{Bmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{Bmatrix}.$$

Note that A_c is the feature set for the calculation of the optimal number of clusters for the division of users into distinct groups. Also, the size of A_c is $n \times 7$ for Set 1, $n \times 5$ for Set 2 and $n \times 4$ (ETH) $n \times 3$ (BTC) for Set 3, where n is the number of nodes (users) and the columns correspond to the defined features. V_c gives the set of labels to the n nodes (users) based on the clustering.

3. Analytical validation - examine clusters qualitatively and find the distinct properties in each group of clustered users.
4. Perform statistical validation of clusters by calculating the silhouette score.
5. According to the distinct properties in clusters, define a rule that majority of data points follow in cluster. Adjust A_c and V_c accordingly by removing data points that do not follow the rule.
6. Use adjusted A_c and V_c as a training set for the SVM model.
7. Use the trained SVM model to label the sets of features $A_1, A_2, \dots, A_m$ from other datasets. Each matrix $A_1, A_2, \dots, A_m$ represents feature sets in each period shown in Table 6.2.

Fig 6.4 illustrates the algorithm described above.

Due to the complexity of the feature sets, we found that solely applying unsupervised method is insufficient. Our approach is to first group the set of features A_c that k -means clustering has good performance. From which, we obtain the set of labels V_c . Subsequently, we use the clustered output as training data for the more advanced supervised

Figure 6.4. The algorithmic steps to classify users into behavioral groups.

algorithm SVM. The final step is to employ the trained SVM model as classifier for the rest of the feature sets $A_1, A_2, \dots, A_m$ to obtain their corresponding labels $V_1, V_2, \dots, V_m$.

In section 6.3.2, we will discuss in greater detail the results obtained in each step of our algorithm. In summary, our algorithm has allowed us to perform user classification more precisely and it has also increased the speed of the clustering process.

6.3.2 Results and Discussion

First, we explore how many clusters can be distinguished in dataset - for this purpose, we use the elbow method for all defined feature sets. For ethereum, it has been found that Set 1 and Set 2 produce sharp elbow at 4 clusters (Fig 6.5, while Set 3 has failed to find optimal k . For bitcoin, the best results were obtained using Set 2 and 3 (Fig 6.6, however for periods of global events, there is no distinct optimal number of clusters regardless of feature set used).

Figure 6.5 shows that for all periods number of clusters 4 appeared to be the optimal.

From Figure 6.6 it is clear that for local events, optimal number of clusters is 4. However, for global events there is no “sharp” elbow.

K -means clustering with $k = 4$ then allows us to obtain the vector of labels V_c for feature matrices A_c where defining the optimal number of clusters was possible. To understand better about each cluster, we examine the behavior of users in each group and analyse their distinctive features. In order to do this properly, we map the obtained label vector V_c to the original feature matrix A_c before the process of normalization and

Figure 6.5. Number of clusters in ETH data found using the elbow method.

Figure 6.6. Number of clusters in BTC data using the elbow method.

the scaling of feature values. The Table 6.3 shows the main characteristic properties for each group that are universal for both bitcoin and ethereum. These characteristics were obtained by observing the distributions of all properties in each group in both cryptocurrencies. Note that the negative balance displayed in the table should not be interpreted literally. This is because blockchain technology does not allow users to double-spend or overdraw. A negative balance in our results implies that during the period under consideration, outgoing transactions exceed the incoming ones. Since we are not aware of the

initial balance of the users, we cannot estimate exactly the amount each user holds at the moment. Therefore, in our data it is displayed as negative balances.

	In degree k_{in}	Out degree k_{out}	In value v_{in}	Out value v_{out}	Balance b
Group 1	$k_{in} > 0$	$k_{out} = 0$	$v_{in} > 0$	$v_{out} = 0$	$b > 0$
Group 2	$k_{in} = 0$	$k_{out} > 0$	$v_{in} = 0$	$v_{out} > 0$	$b < 0$
Group 3	$k_{in} > 0$	$k_{out} > 0$	$v_{in} > 0$	$v_{out} > 0$	$b > 0$
Group 4	$k_{in} > 0$	$k_{out} > 0$	$v_{in} > 0$	$v_{out} > 0$	$b < 0$

Table 6.3. Distinct properties of the majority of nodes in each cluster

Based on their behavioral strategies as shown in Table 6.3, we name each group as follows:

- Optimists (Group 1) - Users who invest in currency. Their persistent strategy for a period of one month is to buy and accumulate coins.
- Pessimists (Group 2) - Users who sell the currency. They do not buy coins and their balance is negative at the period under consideration.
- Positive Traders (Group 3) - Users who alternate between buy and sell, but have positive balance that shows their preference to accumulate coins.
- Negative Traders (Group 4) - Users who alternate between buy and sell, but have negative balance that demonstrates their pessimistic attitude.

We next statistically validate if the data is well clustered by calculating the silhouette score for each data point based on Eq. (6.4). Fig. 5.1 and Fig. 5.2 demonstrate silhouette scores calculated for filtered bitcoin and ethereum datasets.

For local periods all data points have positive scores, while small number of data points for global events are not well matched with their cluster (have negative score).

In all periods we see amount of data points with negative score and average silhouette coefficient for all is 0.4.

We have found that there are data points in each cluster that have negative silhouette score. These users appeared to have different properties than it is shown in Table 6.3 and we consider them as misclassified points. Table 6.4 shows the percentage of misclassified points in each period for various feature sets. Although the silhouette analysis for bitcoin global events shows that most of data points have positive score, users in each cluster do not share similar properties. Therefore, we consider this case as fail for k -means to define meaningful clusters. Overall, we consider the classification to be failed if 1) after

Figure 6.7. Silhouette coefficient calculated for filtered ethereum data set.

Figure 6.8. Silhouette coefficient calculated for filtered bitcoin data set.

using elbow method we do not see the optimal number of clusters (“sharp elbow”) 2) after performing k -means clustering on the dataset and analysing each clusters’ properties, we see distributions of all properties to be similar with no distinction among clusters.

From Table 6.4 we can conclude that feature set 2 shows the best performance.

We take the properties listed in Table 6.3 as an ideal case (rule) for defining the behavioral clusters in cryptocurrency system. Then we use the labelled datasets A_c and V_c (all labelled sets for local periods were merged to increase training set for better

	Set 1	Set 2	Set 3
Local ETH	$\approx 0.001\%$	$\approx 0.001\%$	fail
Global ETH	$0.05 - 0.1\%$	$0.012 - 0.016\%$	fail
Local BTC	fail	$0.6 - 7\%$	$3 - 12\%$
Global BTC	fail	fail	fail

Table 6.4. The percentage of misclassified points shown in comparison with various feature sets used.

performance) and adjust them by removing misclassified points. The method of SVM is then used by treating adjusted A_c and V_c as a training set, the linear kernel for SVM was used and the C parameter is equal to 1. We classify datasets for global events using trained SVM and compare properties of users in each group with the ideal case (rule) shown in the Table 6.3. All points for both bitcoin and ethereum global events were classified properly according to the rule. We would also like to mention about the variability among users in same group, as well as the variability of group properties in different periods. Note that rules described in Table 6.3 aim to classify users with certain attitude (behaviour), while they can be of any type - from independent users to large entities. Therefore, properties (e.g. balance, degree) may vary from very small value to very large. Group properties also vary across different periods - although general rules described in Table 5 would hold true for all groups across all periods, average values of those properties (as well as minimum and maximum) may vary. There also can be users with slightly negative (or slightly positive balance) that would be classified by our defined rule as group 4 (or group 3), though their properties are not much different. We checked on the number of users with the slightly negative (or positive) balance (balance that is under the 1st percentile of all balances in system at that period). We found that there is a small percentage in ETH (less than 0.5%) of users in group 3 whose balance is slightly positive. Number of users with slightly negative balance in group 4 is about 1- 7%. In BTC the number of users are also less than 0.5% (for group 3) and from 6-15% (for group 4).

We then look at the evolution of behavioral types in the cryptocurrency system at different periods: Fig 6.9 shows the evolution of users behavior in bitcoin and ethereum. Overall, it can be seen that the user composition is more stable in the ethereum system, with the decreased number of positive traders during the shock events - 20%-40% smaller comparing with the periods of local events. Consequently, populations of negative traders, pessimists and optimists increased (50%, 10%-80% and 20%-30% relatively). As for bitcoin, people's behavior is more volatile depending on the price movement. Price decline

and stable price periods show the similar behavioral composition, but the growth of price leads to the change in users behaviour - there are no pessimists and number of positive traders increased up to 35%, while number of optimists and negative traders remains the same. As for the systematic events, bitcoin's users behaviour changes dramatically during the Crypto Bubble with an increase of optimists' population up to 45%, comparing with the local events. Periods after Bubble and Crypto Winter show very different behavioral composition with the majority of users (up to 70% from total number) being a positive traders. The difference between the two dominant cryptocurrencies could be due to their distinctive nature. Although both are considered 'currencies' by many, ethereum has direct utility as 'gas' payment that enables computation of smart contracts. This is in contrary to bitcoin, which is mostly considered as a store of value (and sometimes ledger) that many people tend to profit upon its price fluctuations, contributing to the change in user composition during periods of price increase and decrease. Systemic events affect users' strategies in both currencies compared with the local events, however their strategy choices were quite different. Users in bitcoin appeared to be more optimistic during the Crypto Bubble. The increased number of users in group 3 during the Crypto Winter and after Bubble period might also mean the lack of persistency in optimistic and pessimistic strategies - people switch strategies often and our algorithm averages them out as a group of traders.

Figure 6.9. The percentage of users in different groups in the in bitcoin and ethereum during different periods of systems' evolution.

6.4 Dynamics of Users' Behaviour as Absorbing Markov Chain

Studying switching mechanisms between different strategies and overall dynamics of people's behaviour is also quite important as this information was not available before and it can possibly explain some phenomena in financial systems [242, 243]. We derive the switches between different strategies among users of cryptocurrency system and use the *Absorbing Markov Chain* approach to model these behavioral switches - this allows us to demonstrate the behavioral dynamics in comprehensive and intuitive way and moreover, allows easy computation of various statistic for both, single users and overall system's evolution. This approach have been used in several other systems to model the dynamics of various components [244–246] and allowed to accurately estimate the system's evolution - e.g. resources needed for technological and ecological systems.

6.4.1 Daily Behavioral Composition and Switching Mechanism

First, we apply the Machine Learning model to obtain the *daily* behavioral composition for both cryptocurrencies. Figure 6.10 shows the example of the daily behavioral composition in bitcoin and ethereum.

Figure 6.10. The example of daily percentage of users in different groups in the bitcoin and ethereum

After we obtain the composition of users in different groups in intraday level, we then look at the evolution of each user's strategic preferences - how often they change

their strategy and which group they prefer to switch. This information of the behavioral dynamics can be modelled as *Absorbing Markov Chain* - for any system with n states parameters of the Absorbing Markov Chain are the probabilities of transitioning from one state to another, parameters are contained in the $n \times n$ matrix M . Sum of the i th row in matrix is equal to 1 and there is at least one absorbing state - likelihood of transitioning from absorbing state to any other equals to 0 (once entered, can not be exited). In our Absorbing Markov Chain each state is the certain strategy and the transition probabilities are the percentage of users switching from one state to another. Overall, as it was found earlier, there are *three* main strategies users follow - optimist, pessimist and trader (we merge the positive and negative traders to one group for the convenience of modelling). Moreover, each user can be inactive for some time - we define it as *holder* state. Also there are new users that can start their activity from the certain strategy and there are users that leave the system. In total, in our system there are six states - 5 of them are transient and 1 is absorbing.

Figure 6.11 illustrates the possible transition probabilities in the system:

Figure 6.11. State transition diagram.

New node can become either optimist, pessimist or trader. There is no transitions to the new state. Optimist, pessimist, trader and holder states can transit to either state (apart from new) and also can transit to the same state. State “dead” absorbs nodes from optimist, pessimist and trader states and there is no more transitions from this state.

The example of the transition matrix M for one day is shown below, where the last row is the absorbing state and fifth row shows transitions for a new node to other transient states.

$$M_1 = \begin{bmatrix} 0.714 & 0.124 & 0.02 & 0.138 & 0.0 & 0.0 \\ 0.129 & 0.765 & 0.006 & 0.058 & 0.0 & 0.04 \\ 0.064 & 0.004 & 0.868 & 0.011 & 0.0 & 0.05 \\ 0.115 & 0.029 & 0.005 & 0.814 & 0.0 & 0.034 \\ 0.0 & 0.486 & 0.09 & 0.421 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1 \end{bmatrix}$$

Transition matrices M_1 to M_n were obtained for each day for both cryptocurrencies. These matrices are able to comprehensively describe the evolution of users' behaviour and provide a systematized description of strategy switching among all users in system.

6.4.2 Modelling the Dynamics of Users' Behaviour

Modelling the evolution of users' behaviour as Absorbing Markov Chain enables us to obtain useful insights by calculating various statistics as proposed in [247]. First, we would like to provide the table with notations for convenience:

Notation	Meaning
s	Number of transient states
r	Number of absorbing states
Q	Matrix of transition probabilities between transient states
R	Matrix of transition probabilities from transient states to absorbing states
I	Identity Matrix
N	Matrix of expected number of times in transient state j , given starting state i
A	Matrix of probabilities of absorption in absorbing state j , given starting state i
T	Vector of row sums of fundamental matrix N

Table 6.5. Table of matrix notations

We then re-arrange the transition matrix M as shown in Eq. 6.5:

$$M = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix} \quad (6.5)$$

where Q is the $s \times s$ matrix of transient states, R is the $s \times r$ matrix of transition probabilities from transient to absorbing states and I is the $r \times r$ identity matrix.

From the example of matrix M_1 shown above, matrices Q and R will be as following

(since we have only one absorbing state, matrix R is the vector):

$$Q = \begin{bmatrix} 0.714 & 0.124 & 0.02 & 0.138 & 0.0 \\ 0.129 & 0.765 & 0.006 & 0.058 & 0.0 \\ 0.064 & 0.004 & 0.868 & 0.011 & 0.0 \\ 0.115 & 0.029 & 0.005 & 0.814 & 0.0 \\ 0.0 & 0.486 & 0.09 & 0.421 & 0.0 \end{bmatrix}$$

$$R = \begin{Bmatrix} 0.0 \\ 0.4 \\ 0.05 \\ 0.034 \\ 0.0 \end{Bmatrix}$$

From here, we can calculate the *fundamental matrix* N of absorbing chain as following:

$$N = I + Q + Q^2 + \dots = (I - Q)^{-1} \quad (6.6)$$

Fundamental matrix N is quite interesting for our analysis as it shows the expected number of times transient state j moves to the state i before it gets absorbed. Fundamental matrix calculated for the example matrix M is as following:

$$N = \begin{bmatrix} 12.785 & 8.303 & 2.778 & 12.239 & 0 \\ 9.572 & 10.649 & 2.334 & 10.560 & 0 \\ 7.288 & 4.926 & 9.187 & 7.487 & 0 \\ 9.593 & 6.927 & 2.328 & 14.791 & 0 \\ 9.347 & 8.535 & 2.942 & 12.033 & 1 \end{bmatrix}$$

From this matrix, we can conclude that on the day when matrix M was calculated, users with the strategy “hold” will on average transit to the state “hold” around 13 times before they get absorbed (leave the system), and around 8 times they will shift to the state “optimist” before leaving the system. Although, this information can be quite interesting and useful, it is not so intuitive as it also incorporates the probability of not reaching the state. More intuitive way would be to calculate the expected number of steps from each

state before the absorption. This can be done by calculating the vector T - the row sum of matrix N . As for our system, vector T can indicate the expected *longevity* of the user according to his behavioral pattern.

$$T = \begin{pmatrix} 36.105 \\ 33.115 \\ 28.888 \\ 33.639 \\ 33.857 \end{pmatrix}$$

The interpretation of vector T is quite straightforward - users in state “hold” at the given date are expected to stay in system around 36 more days, users in state “optimist” - 33, “pessimist” - 29, “trader” - 34 and new nodes - 34. However, as fundamental matrix assumes the infinite number of transitions, we can expect the certain number of errors when applying to our system.

To check how accurate the predictions based on the fundamental matrix, we select the random set of 30 transition matrices and perform all calculations as described earlier. Then, we calculate the percentage of users from each behavioral group that were absorbed after the certain number of steps predicted by the vector T . Overall, there are around 75,000 of users in our test sample. Figure 6.12 demonstrates the results of predictions based on the fundamental matrix.

The bottom of each box in figure is the third quantile of data, top - the first quantile. Green line shows the median value, the lowest black line is the minimum value and the upper black line is the maximum value. Dots below the minimum value and above the maximum value are outliers. Group 0 is the state “holder” and on average more than 80% of users from this group were absorbed after the number of steps calculated by the vector T with the lowest predicted number of 71%. Group 1 (state “optimist”) predictions are not so accurate with only 50% of absorbed users predicted correctly similar to Group 3 (state “trader”). Group 2 (state “pessimist”) and Group 4 (state “new”) also showed great accuracy with more than 85% of users longevity predicted correctly. Although, there are couple outliers on lower side for Group 4.

Overall, modelling behavioral patterns as Absorbing Markov Chain is a powerful approach that allowed us to describe the dynamics of cryptocurrency system users’ behaviour over time and answer to the number of relevant questions. Under Markov Chain model approach we were able to understand the switching mechanism between different strate-

Figure 6.12. Percentage of absorbed users predicted correctly by the fundamental matrix.

gies and estimate the longevity of users. Results shown in Figure 6.12 also can be a cross-validation for the Machine Learning algorithm developed in the section 6.3 - the high accuracy of calculations compared with the empirical results on absorbed users in 3 groups based *only* the behavioral patterns indicate the adequate performance of the methodology.

6.5 Conclusions

In conclusion, we have constructed empirical human behavioral analysis in financial system during different periods of its evolution. Our developed methodology to classify different strategies that exist in the cryptocurrency market, using a combination of unsupervised machine learning method (k -means clustering) and supervised learning method (Support Vector Machine), has allowed us to derive distinct and robust clusters of users having different behavior. This methodology has been applied to the two largest cryptocurrencies - bitcoin and ethereum, during periods of local price changes and also during large systemic events. Our obtained results show that there are four distinct behavioral types in the cryptocurrency systems: optimists, pessimists, positive traders and negative traders. These four types of behaviour appeared to be common for both systems across various periods of their development. We have analysed the behavioral composition of bitcoin and ethereum users during the periods under consideration and we found ethereum

users' behavior to be more stable. We infer this behavior to result from the long-term view that the ethereum users have on the market relative to that of the bitcoin users. As for systemic events, users' behavior changes in both currencies with very different strategy preferences.

Although both bitcoin and ethereum are digital tokens that serve as decentralised currency based on blockchain technology, there are crucial differences between them. While bitcoin has positioned itself as an alternative monetary system in the financial market, ethereum has mostly focused on monetising smart contracts. Also, being the first cryptocurrency, bitcoin has been widely used for speculative purposes. These traits are reflected in the user composition as shown in Fig 6.9, where the behavior of ethereum users is observed to be more stable as these users are more optimistic of the market. In contrast, the behavior of the bitcoin users tend to fluctuate according to the trend of the market, with a loss of optimism when the market goes down.

We extended the study on behavioral composition by studying the dynamics of people's behaviour in cryptocurrency systems. We developed the Absorbing Markov Chain model consisting of 4 transient states (holder, optimist, pessimist and trader), and 1 absorbing state. Transitions between states were calculated empirically on the whole available dataset as the percentage of users changing their strategy on each day. This modelling approach appeared to be very powerful and helped us to represent the dynamics of behavioral changes in system in intuitive and comprehensive way. More importantly, the framework developed for Absorbing Markov Chains allowed us to calculate various statistics for the system's dynamics. Interestingly, it was found that it is possible to predict how long each user will stay in system based on its behaviour with high accuracy for certain groups of users.

Chapter 7

Causal Relations between Market Effects, Users' Behavior and Network Properties

By now we have a great understanding of the bitcoin and ethereum cryptocurrency systems - we studied the properties of price movements in Chapter 3, looked at the whole transaction networks and calculated various statistics (average degree, assortativity coefficient, clustering coefficient etc.) in a granular level using the Complex Networks approach in Chapter 4 and Chapter 5. We have also obtained information about users' behaviour and how it evolves over time in Chapter 6. Now we would like to understand the relations between all the various characteristics of systems obtained previously - does price evolution depends on the behavioral switches? Does the change of different behavioral group sizes affect the structural properties etc.? In this Chapter we focus on deriving the relations between all properties we have obtained earlier. First, we present all systems' characteristics in the form of time series and perform simple correlation analysis. Then we explore the causal relations between all time series using the Granger causality test and discuss our findings.

7.1 Time Series of System's Evolution and Correlation Analysis

To answer the questions above we first obtain time series of all information available in daily time resolution - time series of price, logarithmic returns of price, number of nodes,

number of edges, assortativity and clustering coefficients, sizes of all behavioral groups and percentage of users switching from one strategy to another. Overall, we have a *very comprehensive* description of the cryptocurrency system reflected in the 34 time series. It should be noted that in this Chapter we focus only on *Ethereum* blockchain, since in the previous Chapters we have found that it shows more stable behaviour and we would like to start the analysis from the more stable system. In future work we plan to expand the analysis to the Bitcoin blockchain as well.

The Figure 7.1 provides a systematic review how this information was obtained.

Figure 7.1. Systematic Overview of obtained time series.

There are four time series about the network structure - evolution of number of nodes and edges, clustering and assortativity coefficients. Information about users was obtained based on the properties of each user (in degree, out degree, in value, out value, total degree and balance) that were fed to the Machine Learning (ML) model described in Chapter 6 and for each user we obtain the *label* that reflects its behaviour. From the users' label information we obtain four time series of behavioral group sizes. After modelling behavioral switches as Markov Chain, we get 22 time series of switches (hold-hold, hold-pessimist etc.). Behavioral transitions were derived from transitions matrices, where each value in time series is the probability to transit from one state to another (hold-hold, hold-pessimist etc.) in a daily interval. Since we have 3 states (optimist, pessimist and

trader) with 5 possible transitions (they can go back to their previous state or transit to any other state but new), 1 state (hold) with 4 possible transitions (any but new and dead states), and 1 state (new) that can transit to 3 other states (any but hold and dead). All transitions from various states sum up to 22. Lastly, we obtain the information about the price evolution - price itself, logarithmic return of price and volatility. Overall, 33 time series were obtained to describe the system's dynamics.

Abbreviation	Full Name	How it was obtained	Reference
hold_hold	hold to hold	From transition matrices $M_1...M_n$	Section 6.4.2
hold_opt	hold to optimist	From transition matrices $M_1...M_n$	Section 6.4.2
hold_pes	hold to pessimist	From transition matrices $M_1...M_n$	Section 6.4.2
hold_trd	hold to trader	From transition matrices $M_1...M_n$	Section 6.4.2
new_opt	new to optimist	From transition matrices $M_1...M_n$	Section 6.4.2
new_pes	new to pessimist	From transition matrices $M_1...M_n$	Section 6.4.2
new_trd	new to trader	From transition matrices $M_1...M_n$	Section 6.4.2
opt_death	optimist to dead	From transition matrices $M_1...M_n$	Section 6.4.2
opt_hold	optimist to hold	From transition matrices $M_1...M_n$	Section 6.4.2
opt_opt	optimist to optimist	From transition matrices $M_1...M_n$	Section 6.4.2
opt_pes	optimist to pessimist	From transition matrices $M_1...M_n$	Section 6.4.2
opt_trd	optimist to trader	From transition matrices $M_1...M_n$	Section 6.4.2
pes_death	pessimist to dead	From transition matrices $M_1...M_n$	Section 6.4.2
pes_hold	pessimist to hold	From transition matrices $M_1...M_n$	Section 6.4.2
pes_opt	pessimist to optimist	From transition matrices $M_1...M_n$	Section 6.4.2
pes_pes	pessimist to pessimist	From transition matrices $M_1...M_n$	Section 6.4.2
pes_trd	pessimist to trader	From transition matrices $M_1...M_n$	Section 6.4.2
trd_death	trader to dead	From transition matrices $M_1...M_n$	Section 6.4.2
trd_hold	trader to hold	From transition matrices $M_1...M_n$	Section 6.4.2
trd_opt	trader to optimist	From transition matrices $M_1...M_n$	Section 6.4.2
trd_pes	trader to pessimist	From transition matrices $M_1...M_n$	Section 6.4.2
trd_trd	trader to trader	From transition matrices $M_1...M_n$	Section 6.4.2
nodes	number of nodes	From network analysis	Section 4.2.3
edges	number of edges	From network analysis	Section 4.2.3
price	value of price	From price statistics	Section 3.2
assort	assortativity	From network analysis	Section 4.2.3
clustering	clustering	From network analysis	Section 4.2.3
volatility	volatility	From price statistics	Section 3.2
hold	number of holders	From behavioral analysis	Section 6.3.2
opt	number of optimists	From behavioral analysis	Section 6.3.2
pes	number of pessimists	From behavioral analysis	Section 6.3.2
trd	number of traders	From behavioral analysis	Section 6.3.2

Table 7.1. Summary of all time series

Figure 7.2 shows all obtained time series, we skip the plotting of returns time series, since its evolution is well-captured by the volatility series:

Figure 7.2. All time series that describe the Ethereum system's dynamics. All properties are calculated in daily time resolution

To understand the relation between all time series, we first perform the simple correlation analysis by calculating the sample Pearson correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (7.1)$$

where n is the sample size, x_i and y_i are the data points in time series where correlation is calculated.

Figure ?? shows the correlation coefficients between various properties. The table shows only significant correlations between time series, where p-value of t-test was higher than 0.05. Price is highly correlated with the number of nodes and edges in network, this relation has been already shown in Chapter 4, surprising relation is the high correlation between the assortativity coefficient and price - 68%. As for relations between behavioral

switches and price dynamics, switches from optimist and trader strategies to any other shows the highest positive and negative correlations with the price.

7.2 Background of Granger Causality Test

Simple correlation analysis provided some interesting insights into the relations between various properties of the system. To investigate more into the causal relations we use the Granger causality test [248] - statistical hypothesis test used to define whether certain time series is relevant to prediction of another time series. Granger causality test was originally introduced to derive causalities in economic time series and it is based on two premises - 1) cause appears before the effect 2) predictions of effect can be improved by knowing the cause. In other words, time series $\{x_{i,t}\}$ is Granger causal of time series $\{x_{j,t}\}$ if inclusion of the history x_i to the history of x_j itself improves the prediction of $x_{j,t}$:

$$VAR(x_{j,t} - \hat{x}_{j,t}|x_{j,0:t-1}) > VAR(x_{j,t} - \hat{x}_{j,t}|x_{j,0:t-1}, x_{i,0:t-1}) \quad (7.2)$$

where VAR is the vector autoregressive model and $\hat{x}_{j,t}$ is the error variance.

To investigate the causality between $\{x_{i,t}\}$ and $\{x_{j,t}\}$ let the Eq. 7.3 be the definition of full $VAR(P)$ of all time series components:

$$\begin{bmatrix} x_{j,t} \\ x_{i,t} \end{bmatrix} = \sum_{p=1}^P \begin{bmatrix} A_{j,j}^f(p) & A_{j,i}^f(p) \\ A_{i,j}^f(p) & A_{i,i}^f(p) \end{bmatrix} \begin{bmatrix} x_{j,t-p} \\ x_{i,t-p} \end{bmatrix} + \begin{bmatrix} w_{j,t}^f \\ w_{i,t}^f \end{bmatrix} \quad (7.3)$$

where f means the full model (by full we mean both cause and effect components are being included), $\{w_{j,t}^f\}$ and $\{w_{i,t}^f\}$ are the noise processes with zero mean.

One step predictor for full $VAR(p)$ then would be:

$$\hat{x}_{j,t}^f = \sum_{p=1}^P \sum_{m \in i,j} A_{j,m}^f(p) x_{m,t-p} \quad (7.4)$$

There are many assumptions and limitations of this test about the system to be linear, time-invariant and stationary that might lead to the inconsistent results. As these limitations were realized to cause misinterpretations of relations between time series, several generalizations were proposed to include non-linearity, multi-step predictions

etc. [249, 250]. Shortcomings of Granger causality test when applied to the analysis of neuroscience time series was greatly discussed in [251] which also holds true for many other systems as well. Nevertheless, Granger causality test has been widely applied to various economic and financial time series [252–254] and provided meaningful analysis. Also many other fields (particularly, neuroscience applied this method a lot) used this method for the analysis of time series [255–258].

7.3 Causality Analysis Results

Granger causality test is the test of zero hypothesis H_o that time series $\{x_{i,t}\}$ *does not* Granger cause time series $\{x_{j,t}\}$. This test can be performed only on stationary time series, therefore the first step of our analysis would be to check all time series for the stationarity using the Augmented Dickey-Fuller (ADF) test. For each time series x_t ADF derives the autoregressive representation and the test is applied to the following model:

$$\Delta x_t = \mu + \gamma t + \alpha x_{t-1} + \sum_{j=1}^{k-1} \beta_j \Delta x_{t-j} + u_t \quad (7.5)$$

where Δ is the difference operator, u_t is the white noise, μ is the constant, γ is the time trend coefficient. ADF examines the negativity of the parameter α - the more negative is α , the stronger the rejection of the hypothesis that there is a unit root in time series and therefore, time series is stationary.

We have performed the Augmented Dickey-Fuller test on all time series and found that the following time series are not stationary: *number of nodes, number of edges, assortativity coefficient, price, optimist to hold, new to pessimist, optimist to optimist and trade to hold switches*. For these time series the differencing $x_t - x_{t-1}$ was made to make them stationary. Differencing is the common technique applied to non-stationary time series to remove the trend and seasonality, and transform series to stationary [259]. After the first differencing, ADF test was performed again and it was found that all time series are stationary.

Next step is to construct Vector Autoregressive (VAR) models only for effect time series $\{x_{j,t}\}$ for the certain number of lags:

$$x_{j,t} = a_0 + a_1 y_{t-1} + a_2 y_{t-2} \dots + a_m y_{t-m} + \epsilon \quad (7.6)$$

where m is the number of lags and ϵ is the error.

Then, we construct another Vector Autoregressive model for two variables that includes the lagged values of cause time series $\{x_{i,t}\}$ as well:

$$x_{j,t} = a_0 + a_1y_{t-1} + a_2y_{t-2}\dots + a_my_{t-m} + b_0 + b_1y_{t-1} + b_2y_{t-2}\dots + b_my_{t-m}\epsilon \quad (7.7)$$

where m is the number of lags and ϵ is the error.

The number of lags m was selected the following way - we increase the number of lags until there is no significant autocorrelation in time series. Although there are other ways to assess the optimal number of lags, it was discussed in [260] that relying on classical hypothesis-testing models might not provide good results due to the specifics of data. Therefore, we decide to rely on the absence of autocorrelation as the factor to choose the optimal number of lags and it found to be 10.

On the constructed VAR model we perform the F-test to assess the significance of each lagged cause time series to effect time series. Null hypothesis H_0 of Granger causality test holds true if there is no lagged value of cause time series $\{x_{i,t}\}$ is significant according to the F-test.

We aim to investigate the causal effects to the *price, assortativity, number of nodes and edges*, since these characteristics reflect the system's evolution the best. We use all other time series to understand their influence on the four effect time series. Table below shows the highest p-values of the performed F-tests on all lags of time series. We set the significance level to be 0.99 and red colour highlights the p-values of time series that *Granger cause* the effect time series.

From the Table 7.2 some conclusions can be made about the cause and consequences between various time series - we can see that the number of people in states "holder" and "pessimist" cause the change in price and network size, while the number of people in states "pessimist" and "holder" do not have an impact on the system's evolution. Another observation is that price affects the growth of network and not other way around - it was shown previously in the Figures 4.14 and 4.15, price is highly correlated with the number of nodes and edges and it is interesting to find the causal factor. As for the effect of behavioral switches, we see that switches from optimist strategy to the optimist, dead and hold states causes the change in network structure and does not affect the price movement. Overall, the price evolution is affected by few time series - switches from and to inactive states like hold and dead, and number of optimists and traders in

Cause Time Series	Price	Assortativity	Nodes	Edges
hold to hold	0	0.0003	0	0
hold to optimist	0.06	0.3128	0.2271	0.3232
hold to pessimist	0.312	0.0241	0.1013	0.0346
hold to trader	0.007	0.5268	0.0863	0.0946
new to optimist	0.2192	0.0001	0.0001	0
new to pessimist	0.6509	0.0007	0.131	0.1754
new to trader	0.0874	0.0095	0.0125	0.0013
optimist to dead	0.0796	0.1438	0.0034	0.0006
optimist to hold	0.01	0	0003	0
optimist to optimist	0.0076	0.0011	0.0038	0.0011
optimist to pessimist	0.0002	0.0334	0	0
optimist to trader	0.2277	0.0008	0	0
pessimist to dead	0.1843	0.0001	0.1709	0.0223
pessimist to hold	0.0045	0.0108	0.0277	0.0002
pessimist to optimist	0.0369	0.0236	0	0
pessimist to pessimist	0.6099	0.0741	0.3982	0.3454
pessimist to trader	0.0586	0.0307	0.0546	0.0105
trader to dead	0	0.0002	0	0.0005
trader to hold	0.006	0	0.0013	0
trader to optimist	0.1743	0.0253	0	0
trader to pessimist	0.0732	0.1149	0.0125	0.0059
trader to trader	0.0211	0.0011	0.1527	0.0349
number of nodes	0.4304	0	-	0.0013
number of edges	0.4278	0	0.0005	-
price	-	0.0114	0	0
assortativity	0.0885	-	0.0057	0.0007
number of holders	0.0964	0.059	0.0575	0.0964
number of optimists	0	0.0004	0	0
number of pessimists	0.2393	0.2379	0.0527	0.2393
number of traders	0	0.0005	0	0

Table 7.2. Table of p-values for Granger causality test

system. Network properties are affected greatly by many factors - both changes in price and behavioral composition.

7.4 Conclusions and Discussions

In this Chapter we derived time series to describe the evolution of ethereum system in a comprehensive way, including time series from network analysis, price and behavioral dynamics. Simple correlation analysis showed some interesting dependencies between the

network properties, price changes and behavioral dynamics. We extended the analysis using the Granger test and obtained the information about causal relations between all time series. However, we would like to emphasize the fact that in this Chapter we investigated only linear relations between time series, while there could be more insights obtained when deriving the non-linear relations and extending the causality test by using non-parametric models. Moreover, we would like to perform the causality analysis to the bitcoin transaction dataset as well. Also, it might be helpful to use neural networks for deriving non-linear relations. Overall, more research needs to be done to obtain causal relations between various properties, where the choice of the right methodology would be an important step. Nevertheless, this preliminary analysis showed that there is indeed a relation between all system components - structural properties, behavioral properties of system's parts (users) and the price evolution.

Chapter 8

Conclusions and Future Work

This thesis was inspired by the latest developments in Financial Technologies - introduction of blockchain and cryptocurrencies spurred our interest to understand more about these innovative systems. Although this field has been extensively studied for a decade, most of the research was focused on the technological issues, price prediction and on the economic nature of digital assets. In this thesis we contributed to the systematic understanding of the Decentralized Finance ecosystem by exploring transaction datasets from two biggest cryptocurrencies - bitcoin and ether using the Complexity Science framework. This powerful approach showed promising results in explaining various phenomena in both natural and man-made complex systems and by applying it to the cryptocurrency systems we were able to obtain the systematic understanding and discover new properties and relations in the system.

Overall, our contribution to the systematic understanding of cryptocurrency systems can be summarized in several main points. Firstly, we investigated into the structural properties of transaction networks by constructing temporal, weighted, directed complex networks from the transaction data. We found that cryptocurrency systems share similar properties with other complex systems with the scale-free behaviour being the most prominent one. This inspired us to uncover the mechanism behind the transaction network structure. In earlier studies it was assumed that there is a degree preferential attachment behind the network formation, however we showed empirically some counter-intuitive results where nodes had the opposite behavior than it was assumed earlier. We proposed a quantifiable definition for the *fitness* of a node and showed empirically that this property follows a power-law function and leads the network to the scale-free behavior. We supported our findings by simulating the synthetic network with fitness preferential attachment mechanism and were able to obtain the network structure similar to the bitcoin

network. These findings attracted growing interest in both the network science [261–263] and cryptocurrency research community [264].

Secondly, we are the first at the moment of publication to derive the real behavioral structure of the financial market. We calculate various properties for all users in both systems. Then, we develop the methodology using Machine Learning models to fairly classify all users into the behavioral groups. It has been found that despite of differences in bitcoin and ethereum systems, there are *four* universal distinct types of behavior (or strategies) that users follow. We also investigated how the behavioral composition changes during the different stages of the system development. First we defined 3 periods for the local events (applies only to the particular cryptocurrency) - period during the price growth, price drop and stable price. And also we defined periods for global events (big events to the whole cryptocurrency market) - periods also known as Crypto Bubble, Crypto Winter and also period after Bubble. Behavioral composition of ethereum system’s users appeared to be more stable comparing to the bitcoin.

Another contribution is the modelling the dynamics of people’s behavior in financial market. We presented the changes in people’s behaviour in market as the Absorbing Markov Chain, where each state is the strategy they follow and transitions between states were calculated empirically. This framework not only enabled us to observe the dynamics of people’s strategic preferences over time in a intuitive way, but more importantly, it appeared to be quite accurate in predicting the *longevity* of users in system, based only on their behavioral patterns.

The last contribution of this thesis is the preliminary investigation of causal relations between the market effects (price), behavioral patterns among users and overall system’s structure. After performing the network analysis and deriving behavioral structure, we had the comprehensive information of the cryptocurrency systems that was reflected in the set of time series. We then perform the correlation analysis to understand the linear relations between all time series and find that there is high positive correlation between the price and assortativity coefficient evolution and also switches from the certain strategies are highly correlated with the network size and price. As correlation analysis showed some relations between various time series, we perform the linear causality test to understand the cause and effect of time series relations. We found that price evolution is affected by certain behavioral switches in system, while price causes changes in structural properties of network. It should be noted that findings on causality presented in this thesis are in their preliminary stage and more work needs to be done on uncovering causal relations.

Moreover, we believe that findings made in this thesis would also contribute to the

understanding of financial systems in general. As it has been shown in Chapter 3, Decentralized Financial markets have similar properties as the traditional ones. These properties were extensively researched and many models were developed to explain them under the traditional financial systems studies. For example, prior works in econophysics proposed agent-based models to explain the nature of stylized facts in financial markets. Although these models have strong theoretical background supported by simulations, there was a lack empirical evidence on agents' behaviour. We believe that our studies on behavioral composition and switches would be beneficial to improve existing and create new models to explain stylized facts in financial markets.

As we solve many interesting problems in this thesis, there are still many open questions that we would like to discover in our present and future work. The first topic would be to extend the causality analysis on the time series by studying non-linear relations between them and including non-parametric causality tests. In Chapter 7 we focused only on studying the ethereum time series since it showed more stable behavioral patterns. However, we would also like to investigate whether in bitcoin system there are similar causal relations. Another research direction would be to relate the observed behavioral and structural patterns with the stylized facts. Also, in Chapter 5 we performed the network mechanism check only on the bitcoin data at the early stage, since we worked on this project in 2016, when data on ether network was small and unreliable. In future, the mechanism check of ether network can also be implemented and compared to that of bitcoin.

Bibliography

- [1] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2009.
- [2] V. Buterin. Ethereum: A next-generation smart contract and decentralized application platform, 2014.
- [3] M. Castro and B. Liskov. Practical byzantine fault tolerance and proactive recovery. *ACM Trans. Comput. Syst.*, 20(4):398–461, 2002.
- [4] D. Schwartz, N. Youngs, and A. Britto. The ripple protocol consensus algorithm. 2014.
- [5] J. Kwon. Tendermint : Consensus without mining. 2014.
- [6] F. Reid and M. Harrigan. An analysis of anonymity in the bitcoin system. *Security and Privacy in Social Networks. Springer*, 2011.
- [7] D. Ron and A. Shamir. Quantitative analysis of the full bitcoin transaction graph. In *Financial Cryptography and Data Security, Springer*, number 4, 2012.
- [8] J. Herrera-Joancomartí. Research and challenges on bitcoin anonymity. In *Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance*, pages 3–16, Cham, 2015. Springer International Publishing.
- [9] I. Eyal and E. G. Sirer. Majority is not enough: Bitcoin mining is vulnerable. *Commun. ACM*, 61(7):95–102, June 2018.
- [10] J. Garay, A. Kiayias, and N. Leonardos. The bitcoin backbone protocol: Analysis and applications. In *Advances in Cryptology - EUROCRYPT 2015*, pages 281–310, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.
- [11] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, and E. W. Felten. Sok: Research perspectives and challenges for bitcoin and cryptocurrencies. In *2015 IEEE Symposium on Security and Privacy*, pages 104–121, 2015.
- [12] D. Garcia and F. Schweitzer. Social signals and algorithmic trading of bitcoin. *Royal Society Open Science*, 2(150288), August 2015.
- [13] R. Böhme, N. Christin, B. Edelman, and T. Moore. Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives*, 29(2), 2015.
- [14] D. Garcia and F. Schweitzer. Social signals and algorithmic trading of Bitcoin. Technical Report 1506.01513, June 2015.
- [15] S. Chan, J. Chu, S. Nadarajah, and J. Osterrieder. A Statistical Analysis of Cryptocurrencies. *Journal of Risk and Financial Management*, 10(2), 2017.

- [16] A. K. Tiwari, S. Kumar, and R. Pathak. Modelling the dynamics of bitcoin and litecoin: Garch versus stochastic volatility models. *Applied Economics*, 51(37):4073–4082, 2019.
- [17] S. Nadarajah and J. Chu. On the inefficiency of bitcoin. *Economics Letters*, 150:6–9, 2017.
- [18] J. Kaminski and P. Gloor. Nowcasting the bitcoin market with twitter signals. *ArXiv*, abs/1406.7577, 2014.
- [19] L. Kristoufek. What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PLOS ONE*, 10(4):1–15, 04 2015.
- [20] S. G. Colianni, S. Rosales, and M. Signorotti. Algorithmic trading of cryptocurrency based on twitter sentiment analysis. 2015.
- [21] N. Gandal and H. Halaburda. Can we predict the winner in a market with network effects? competition in cryptocurrency market. *Games*, 7:16, 2016.
- [22] L. Alessandretti, A. ElBahrawy, L.M. Aiello, and A. Baronchelli. Anticipating Cryptocurrency Prices Using Machine Learning. *Complexity*, 2018:8983590, 2018.
- [23] D. Kondor, I. Csabai, J. Szüle, M.Pósfai, and G. Vattay. Inferring the interplay of network structure and market effects in bitcoin. *New Journal of Physics*, 16(125003), 2014.
- [24] S. Wang and J.-P. Vergne. Buzz factor or innovation potential: What explains cryptocurrencies’ returns? *PLOS ONE*, 12(1):1–17, 01 2017.
- [25] D. Kondor, M. Posfai, I. Csabai, and G. Vattay. Do the rich get richer? an empirical analysis of the bitcoin transaction network. *PLoS ONE*, 9(5), February 2014.
- [26] A. Baumann, B. Fabian, and M. Lischke. Exploring the bitcoin network. In *Proceedings of the 10th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST*,, pages 369–374. INSTICC, SciTePress, 2014.
- [27] A. Aspembitova, L. Feng, V. Melnikov, and L. Y. Chew. Fitness preferential attachment as a driving mechanism in bitcoin transaction network. *PLOS ONE*, 14:1–20, 08 2019.
- [28] I. Alqassem, I. Rahwan, and D. Svetinovic. The anti-social system properties: Bitcoin network data analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50(1):21–31, 2020.
- [29] J. Liang, L. Li, and D. Zeng. Evolutionary dynamics of cryptocurrency transaction networks: An empirical study. *PLOS ONE*, 13(8):1–18, 08 2018.
- [30] A. ElBahrawy, L. Alessandretti, Anne Kandler, R. Pastor-Satorras, and A. Baronchelli. Evolutionary dynamics of the cryptocurrency market. *Royal Society Open Science*, 4(11):170623, 2017.
- [31] P. M. Kraff, N. D. Penna, and A. S. Pentland. An experimental study of cryptocurrency market dynamics. *Proceedings of the the 2018 CHI Conference*, 2018.

- [32] S. Somin, Y. Altshuler, G. Gordon, A. Pentland, and E. Shmueli. Network Dynamics of a Financial Ecosystem. *Scientific Reports*, 10(1):4587, 2020.
- [33] D. Stosic, D. Stosic, T. B. Ludermit, and T. Stosic. Exploring disorder and complexity in the cryptocurrency space. *Physica A: Statistical Mechanics and its Applications*, 525:548 – 556, 2019.
- [34] S. Lahmiri and S. Bekiros. Chaos, randomness and multi-fractality in bitcoin market. *Chaos, Solitons and Fractals*, 106:28 – 34, 2018.
- [35] A. F. Bariviera, M.J. Basgall, W. Hasperu, and M. Naiouf. Some stylized facts of the bitcoin market. *Physica A: Statistical Mechanics and its Applications*, 484:82 – 90, 2017.
- [36] J. Alvarez-Ramirez, E. Rodriguez, and C. Ibarra-Valdez. Long-range correlations and asymmetry in the bitcoin market. *Physica A: Statistical Mechanics and its Applications*, 492:948 – 955, 2018.
- [37] S. Begu, Z. Kostanjar, H. E. Stanley, and B. Podobnik. Scaling properties of extreme price fluctuations in bitcoin markets. *Physica A: Statistical Mechanics and its Applications*, 510:400 – 406, 2018.
- [38] D. Stosic, D. Stosic, T. B. Ludermit, and T. Stosic. Collective behavior of cryptocurrency price changes. *Physica A: Statistical Mechanics and its Applications*, 507:499 – 509, 2018.
- [39] D. Stosic, D. Stosic, T. B. Ludermit, and T. Stosic. Multifractal behavior of price and volume changes in the cryptocurrency market. *Physica A: Statistical Mechanics and its Applications*, 520:54 – 61, 2019.
- [40] N. Antulov-Fantulin, D. Tolic, M. Piskorec, Z. Ce, and I. Vodenska. Inferring short-term volatility indicators from the bitcoin blockchain. In *Complex Networks and Their Applications VII*, pages 508–520. Springer International Publishing, 2019.
- [41] C. Tu and Suweis S. D’Odorico, P. Critical slowing down associated with critical transition and risk of collapse in crypto-currency. *Royal Society Open Science*, 7(3):191450, 2020.
- [42] Belinskiy A. Soloviev, V. N. Complex systems theory and crashes of cryptocurrency market. In *Information and Communication Technologies in Education, Research, and Industrial Applications*, pages 276–297, Cham, 2019. Springer International Publishing.
- [43] L. Bachelier. Thorie de la spculation. *Annales Scientifiques De L Ecole Normale Suprieure*, 17, 1900.
- [44] A. Markov. Rasprostranenie zakona bol’shih chisel na velichiny, zavisyaschie drug ot druga (in russian). *Izvestiya matematicheskogo obshestva pri Kazanskom Universitete*, 15, 1906.
- [45] R. N Mantegna and H E. Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376(6535):46–49, 1995.

- [46] T. Lux. The stable paretian hypothesis and the frequency of large returns: an examination of major german stocks. *Applied Financial Economics*, 6(6):463–475, 1996.
- [47] V. Plerou and H. E. Stanley. Tests of scaling and universality of the distributions of trade size and share volume: Evidence from three distinct markets. *Phys. Rev. E*, 76:046109, Oct 2007.
- [48] V. Plerou and H. E. Stanley. Stock return distributions: Tests of scaling and universality from three distinct stock markets. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 77:037101, 03 2008.
- [49] T. Lux and M. Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500, 1999.
- [50] S. Alfarano and T. Lux. A noise trader model as a generator of apparent financial power laws and long memory. *Macroeconomic Dynamics*, 11(S1):80–101, 2007.
- [51] R. Cont and J.-P. Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(2):170–196, 2000.
- [52] D. Stauffer. Percolation models of financial market dynamics. *Advances in Complex Systems*, 04(01):19–27, 2001.
- [53] Stefan Borhholdt. Expectation bubbles in a spin model of markets: Intermittency from frustration across scales. *International Journal of Modern Physics C*, 12(05):667–674, 2001.
- [54] T. Kaizoji, S. Bornholdt, and Y. Fujiwara. Dynamics of price and trading volume in a spin model of stock markets with heterogeneous agents. *Physica A-statistical Mechanics and Its Applications*, 316:441–452, 2002.
- [55] L. Feng, B. Li, B. Podobnik, T. Preis, and H. E. Stanley. Linking agent-based models and stochastic models of financial markets. *Proceedings of the National Academy of Sciences*, 109(22):8388–8393, 2012.
- [56] D. Sornette. Physics and financial economics (1776-2014): puzzles, ising and agent-based models. *Reports on progress in physics. Physical Society*, 77 6, 2014.
- [57] E. Samanidou, E. Zschischang, D. Stauffer, and T. Lux. Agent-based models of financial markets. *Reports on Progress in Physics*, 70:409–450, 2007.
- [58] H. A. Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, pages 467–482, 1962.
- [59] M. Nagy, G. Vásárhelyi, B. Pettit, I. Roberts-Mariani, T. Vicsek, and D. Biro. Context-dependent hierarchies in pigeons. *Proceedings of the National Academy of Sciences*, 110(32):13049–13054, 2013.
- [60] D. Waxman and S. Gavrillets. 20 questions on adaptive dynamics. *Journal of Evolutionary Biology*, 18(5):1139–1154, 2005.
- [61] G. Meszéna, M. Gyllenberg, F. J. Jacobs, and J. A. J. Metz. Link between population dynamics and dynamics of darwinian evolution. *Phys. Rev. Lett.*, 95:078105, Aug 2005.

- [62] W. Fang, J. Ke, J. Wang, and L. Feng. Linking market interaction intensity of 3d ising type financial model with market volatility. *Physica A: Statistical Mechanics and its Applications*, 461:531 – 542, 2016.
- [63] S. Ma, L. Feng, and C.-H. Lai. Mechanistic modelling of viral spreading on empirical social network and popularity prediction. *Scientific Reports*, 8(1):13126, 2018.
- [64] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [65] A. Vázquez, J. G. Oliveira, Z. Dezsö, K. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E*, 73:036127, Mar 2006.
- [66] N. N. Chung, G. S. Jacobs, H. Sudoyo, S. G. Malik, L. Y. Chew, J. S. Lansing, and M. P. Cox. Sex-linked genetic diversity originates from persistent sociocultural processes at microgeographic scales. *Royal Society Open Science*, 6(8):190733, 2019.
- [67] N. N. Chung, L. Y. Chew, W. Chen, R. M. D’Souza, and C. H. Lai. Susceptible individuals drive active social contagion. *Phys. Rev. Research*, 1:033125, Nov 2019.
- [68] W. Li and X. Cai. Statistical analysis of airport network of china. *Phys. Rev. E*, 69:046106, Apr 2004.
- [69] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(2):026103, 2007.
- [70] K. Lee, W.-S. Jung, J. S. Park, and M. Y. Choi. Statistical analysis of the metropolitan seoul subway system: Network structure and passenger flows. *Physica A: Statistical Mechanics and its Applications*, 387(24):6231 – 6234, 2008.
- [71] Y. Bar-Yam, C. Ramalingam, L. Burlingame, and C. Ogata. *Making Things Work: Solving Complex Problems in a Complex World*. NECSI, Knowledge Press, 2004.
- [72] S. N. Dorogovtsev and J. F. F. Mendes. Effect of the accelerating growth of communications networks on their structure,. *Physical Review E*, 63(02101), 2001.
- [73] S. N. Dorogovtsev and J. F. F. Mendes. Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E*, 63(056125), 2001.
- [74] A. Stopczynski, V. Sekara, A. Sapiezynski, P. and Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann. Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978–e95978, apr 2014.
- [75] M. E. J. Newman. Assortative mixing in networks. *Physics Review Letter*, 89(208701), November 2002.
- [76] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *Society for Industrial and Applied Mathematics Review*, 51(4), February 2009.
- [77] M. Newman. Power laws, pareto distributions and zipfs law. *Countemporary Physics*, 46, 2005.

- [78] M.E.J. Newman and S.H. Strogatz. Random graphs with arbitrary degree distributions and their applications. *Physics Review E*, 64(26118), July 2001.
- [79] P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 519(3):97 – 125, 2012. Temporal Networks.
- [80] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. J Tessone, and F. Schweitzer. Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature Communications*, 5(1):5024, 2014.
- [81] T. Takaguchi, Y. Yano, and Y. Yoshida. Coverage centralities for temporal networks. *The European Physical Journal B*, 89(35), August 2016.
- [82] A. Li, S. P. Cornelius, Y.-Y. Liu, L. Wang, and A.-L.Barabási. The fundamental advantages of temporal networks. *Science*, 368(6366):1042–1046, November 2017.
- [83] R.K. Pan and J. Saramaki. Path lengths, correlations, and centrality in temporal networks. *Physics Review E*, 84(016105), 2011.
- [84] N. Masuda and P. Holme. Predicting and controlling infectious disease epidemics using temporal networks. *F1000Prime Reports*, 5(6), 2013.
- [85] W. H. Thompson, P. Brantefors, and P. Fransson. From static to temporal network theory – applications to functional brain connectivity. *Network Neuoscience*, 1(2), 2013.
- [86] M. Cipriani and A. Guarino. Herd behavior in financial markets: An experiment with financial market professionals. *Journal of the European Economic Association*, 7(1), 2009.
- [87] M. Cipriani and A. Guarino. Herd behavior in a laboratory financial market. *American Economic Review*, 2005.
- [88] D. Johnson, A. Menezes, and S. Vanstone. The Elliptic Curve Digital Signature Algorithm (ECDSA). *International Journal of Information Security*, 1(1):36–63, 2001.
- [89] D. Vujicic, D. Jagodic, and S. Randic. Blockchain technology, bitcoin, and ethereum: A brief overview. In *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pages 1–6, 2018.
- [90] W. Wang, Dinh T. H., P. Hu, Z. Xiong, D. Niyato, P. Wang, Y. Wen, and D. Kim. A survey on consensus mechanisms and mining strategy management in blockchain networks. *IEEE Access*, PP:1–1, 01 2019.
- [91] D. Chaum, A. Fiat, and M. Naor. Untraceable electronic cash. In *Proceedings on Advances in Cryptology, CRYPTO '88*, page 319–327, Berlin, Heidelberg, 1990. Springer-Verlag.
- [92] E. Medvedev. Ethereum data extraction, <https://github.com/blockchain-etl/ethereum-etl>.
- [93] E. Medvedev. Ethereum data extraction, <https://github.com/blockchain-etl/ethereum-etl>, 2019.

- [94] M. Spagnuolo, F. Maggi, and S. Zanero. Bitiodine: Extracting intelligence from the bitcoin network. *Proceedings of 18th International Conference on Financial Cryptography and Data Security* pp. 457-468. Springer, 8437, 2014.
- [95] M. Harrigan. Bitcoin tools, <https://github.com/harrigan/bitcointools>.
- [96] I. Brugere. Bitcoin transaction network extraction, <https://github.com/ivanbrugere/bitcoin-transaction-network-extraction>, 2002.
- [97] I. Brugere. Bitcoin transaction network data, <https://dynamics.cs.washington.edu/data.html>.
- [98] M. Spagnuolo, F. Maggi, and S. Zanero. Bitiodine: Extracting intelligence from the bitcoin network. In *Financial Cryptography and Data Security*, pages 457–468. Springer Berlin Heidelberg, 2014.
- [99] A. Biryukov, D. Khovratovich, and I. Pustogarov. Deanonymisation of clients in bitcoin p2p network. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS '14*, page 15–29, New York, NY, USA, 2014. Association for Computing Machinery.
- [100] B. Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36, 1963.
- [101] V Plerou, P Gopikrishnan, LA Nunes Amaral, M Meyer, and HE Stanley. Scaling of the distribution of price fluctuations of individual companies. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 60(6 Pt A):6519–6529, December 1999.
- [102] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423(6937):267–270, May 2003.
- [103] F. Lillo and J. D. Farmer. The long memory of the efficient market. *Studies in Nonlinear Dynamics Econometrics*, 8, 2003.
- [104] Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley. Correlations in economic time series. *Physica A: Statistical Mechanics and its Applications*, 245(3):437–440, 1997.
- [105] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.
- [106] E. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970.
- [107] P. Cizeau, Y. Liu, M. Meyer, C.-K. Peng, and H. E. Stanley. Volatility distribution in the S&P500 stock index. *Physica A: Statistical Mechanics and its Applications*, 245(3):441–445, 1997.
- [108] P. Gopikrishnan, X. Plerou, V. and Gabaix, and H. E. Stanley. Statistical properties of share volume traded in financial markets. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 62, 08 2000.

- [109] W. Zhang, P. Wang, X. Li, and D. Shen. Some stylized facts of the cryptocurrency market. *Applied Economics*, 50(55):5950–5965, 2018.
- [110] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, 2006.
- [111] *The Structure and dynamics of networks*. Princeton University Press, Princeton, NJ, US, 2006.
- [112] P. Commendatore, M. Matilla-García, L.M. Varela, and J.S. Cánovas. *Complex Networks and Dynamics: Social and Economic Interactions*. Lecture Notes in Economics and Mathematical Systems. Springer International Publishing, 2016.
- [113] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, and G. Gabrielli, A. and Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1(1):58–71, 2019.
- [114] T. Chen, Y. Zhu, Z. Li, J. Chen, X. Li, X. Luo, X. Lin, and X. Zhange. Understanding ethereum via graph analysis. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 1484–1492, 2018.
- [115] Q. Bai, C. Zhang, Y. Xu, X. Chen, and X. Wang. Evolution of ethereum: A temporal graph perspective, 2020.
- [116] J.L. Moreno. *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama*. Nervous and Mental Disease Publishing Company, 1934.
- [117] A. Davis, B.B. Gardner, and M.R. Gardner. *Deep South: A Social Anthropological Study of Caste and Class*. University of Chicago Press, 1941.
- [118] F.J. Roethlisberger and W.J. Dickson. *Management and the Worker*. Harvard University Press, 1939.
- [119] P. Bearman, S. Peter, J. Moody, and K. Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. " *American Journal of Sociology*, 110:44–91, 2004.
- [120] P. Mariolis. Interlocking directorates and control of corporations: The theory of bank control. *Social Science Quarterly*, 56(3):425–439, 1975.
- [121] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.
- [122] J.W. Grossman. The evolution of the mathematical research collaboration graph. *Congressus Numerantium*, 158:202–212, 2002.
- [123] P. Holme, C. R. Edling, and F. Liljeros. Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155 – 174, 2004.
- [124] R. Alberich, J. Miro-Julia, and F. Rossello. Marvel universe looks almost like a real social network, 2002.
- [125] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, 30(4):330 – 342, 2008.

- [126] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph, 2011.
- [127] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165 – 4180, 2012.
- [128] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 1029–1038, New York, NY, USA, 2010. Association for Computing Machinery.
- [129] P. Shakarian and D. Paulo. Large social networks can be targeted for viral marketing with small seed sets. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012.
- [130] A.D. I. Kramer, J. E. Guillory, and J.T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.
- [131] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.
- [132] Cambridge analytica — the power of big data and psychographics, presentation, 2016.
- [133] Katherine A. Seaton and Lisa M. Hackett. Stations, trains and small-world networks. *Physica A: Statistical Mechanics and its Applications*, 339(3):635 – 644, 2004.
- [134] V. Latora and M. Marchiori. Is the boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, 314(1):109 – 113, 2002. Horizons in Complex Systems.
- [135] P. Sen, S. Dasgupta, P. A. Chatterjee, A. and Sreeram, G. Mukherjee, and S. S. Manna. Small-world properties of the indian railway network. *Phys. Rev. E*, 67:036106, Mar 2003.
- [136] R. Albert, I. Albert, and G. L. Nakarado. Structural vulnerability of the north american power grid. *Phys. Rev. E*, 69:025103, Feb 2004.
- [137] L. A. N. Amaral, A. Scala, M. Barthélemy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.
- [138] R Kinney, P Crucitti, R Albert, and V Latora. Modeling cascading failures in the north american power grid. *The European Physical Journal B - Condensed Matter and Complex Systems*, 46(1):101–107, 2005.
- [139] R. Guimerà, S. Mossa, A. Turtleschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, 2005.

- [140] R. Gallotti and M. Barthelemy. The multilayer temporal network of public transport in Great Britain. *Scientific Data*, 2(1):140056, 2015.
- [141] A P Masucci, D Smith, A Crooks, and M Batty. Random planar graphs and the London street network. *The European Physical Journal B*, 71(2):259–271, 2009.
- [142] X.Zhu and S. Liu. Analysis of the impact of the mrt system on accessibility in singapore using an integrated gis tool. *Journal of Transport Geography*, 12(2):89 – 101, 2004.
- [143] H. Soh, S. Lim, T. Zhang, X. Fu, G. K. K. Lee, T. G. G. Hung, P.Di, Silvester Prakasam, and L. Wong. Weighted complex network analysis of travel routes on the singapore public transportation system. *Physica A: Statistical Mechanics and its Applications*, 389(24):5852 – 5863, 2010.
- [144] I. Yang, H. Jeong, B. Kahng, and A.-L. Barabási. Emerging behavior in electronic bidding. *Phys. Rev. E*, 68:016102, Jul 2003.
- [145] S. Miccichè, G. Bonanno, F. Lillo, and R. N. Mantegna. Degree stability of a minimum spanning tree of price return and volatility. *Physica A: Statistical Mechanics and its Applications*, 324(1):66 – 73, 2003. Proceedings of the International Econophysics Conference.
- [146] R N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B - Condensed Matter and Complex Systems*, 11(1):193–197, 1999.
- [147] L. Kullmann, J. Kertész, and K. Kaski. Time-dependent cross-correlations between different stock returns: A directed network of influence. *Phys. Rev. E*, 66:026125, Aug 2002.
- [148] G. Bonanno, G. Caldarelli, F. Lillo, and R.N. Mantegna. Topology of correlation-based minimal spanning trees in real and model markets. *Physics Review E*, 68(046130), April 2003.
- [149] F. Allen and Ana Babus. Networks in finance. *History of Finance eJournal*, 2008.
- [150] E. Nier, J. Yang, T. Yorulmazer, and A. Alentorn. Network models and financial stability. *Journal of Economic Dynamics and Control*, 31(6):2033 – 2060, 2007. Tenth Workshop on Economic Heterogeneous Interacting Agents.
- [151] K. Anand, P. Gai, and M. Marsili. Rollover risk, network structure and systemic financial crises. *Journal of Economic Dynamics and Control*, 36(8):1088 – 1100, 2012. Quantifying and Understanding Dysfunctions in Financial Markets.
- [152] M. Bardoscia, S. Battiston, F. Caccioli, and G. Caldarelli. Pathways towards instability in financial networks. *Nature Communications*, 8(1):14416, 2017.
- [153] D. Acemoglu, A. Ozdaglar, and A. Tahbaz-Salehi. Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608, February 2015.
- [154] B. Elliott, M.and Golub and M.O. Jackson. Financial networks and contagion. *American Economic Review*, 104(10):3115–53, October 2014.

- [155] S. Battiston, G. Caldarelli, R. M. May, Tarik Roukny, and Joseph E. Stiglitz. The price of complexity in financial networks. *Proceedings of the National Academy of Sciences*, 113(36):10031–10036, 2016.
- [156] M. Bardoscia, P. Barucca, S. Battiston, F. Caccioli, G. Cimini, D. Garlaschelli, F. Saracco, T. Squartini, and G. Caldarelli. The physics of financial networks.
- [157] P.L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physics Review Letter*, 85(4629), April 2000.
- [158] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Proceedings of the ACM, SIGCOMM [Computer Communication Review]*, 29(251), 1999.
- [159] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, volume 3, pages 1371–1380 vol.3, 2000.
- [160] H. Jeong, S.P. Mason, A-L Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature (London)*, 41(411), 2001.
- [161] H Jeong, B Tombor, R Albert, Z N Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.
- [162] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66(035103), February 2002.
- [163] F. Harrary. *Graph Theory*. Perseus, 1995.
- [164] D.B. West. *Introduction to Graph Theory*. Prentice Hall, 1996.
- [165] A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using networkx. 1 2008.
- [166] *Gephi: an open source software for exploring and manipulating networks*, International AAAI Conference on Weblogs and Social Media, 2009.
- [167] V. Batagelj and A. Mrvar. *Pajek — Analysis and Visualization of Large Networks*. Springer Berlin Heidelberg, 2004.
- [168] M. Jacomy, T. Venturini, and M. Heymann, S.AND Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE*, 9(6):1–12, 06 2014.
- [169] L. A. Adamic and B. A. Huberman. Power-law distribution of the world wide web. *Science*, 287(5461):2115–2115, 2000.
- [170] A.-L. Barabási, R. Albert, and H.Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1):69 – 77, 2000.
- [171] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLOS One*, 9(e85777), January 2014.

- [172] A. Klaus, S. Yu, and D. Plenz. Statistical analyses support power law distributions found in neuronal avalanches. *PLOS One*, 6(e19779.), 2011.
- [173] P. Holme. Core-periphery organization of complex networks. *Phys. Rev. E*, 72:046111, Oct 2005.
- [174] S. Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.
- [175] S. Maslov, K. Sneppen, and A. Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications*, 333:529 – 540, 2004.
- [176] D.J. Watts and S.H. Strogatz. Collective dynamics of small world networks. *Nature*, 393(440), November 1998.
- [177] L. M. Zahn. Robustness of protein networks. *Science*, 355(6325):591–591, 2017.
- [178] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86 14, 2001.
- [179] H. and Barabási A.-L. Albert, R. and Jeong. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.
- [180] D. Stauffer, A. Aharony, and S. Redner. Introduction to percolation theory. 1985.
- [181] A. Li, S. P. Cornelius, Y.-Y. Liu, L. Wang, and A.-L. Barabási. The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046, 2017.
- [182] D. Braha, Y. Bar-Yam, and H. Sayama. *Time-Dependent Complex Networks: Dynamic Centrality, Dynamic Motifs, and Cycles of Social Interactions*, pages 39–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
- [183] S. A. Hill and D. Braha. Dynamic model of time-dependent complex networks. *Phys. Rev. E*, 82:046105, Oct 2010.
- [184] M. Latapy, T. Viard, and C. Magnien. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining*, 8(1):61, 2018.
- [185] A. Aspembitova. Bitcoin network growth. <https://github.com/AspembitovaAyana/AspembitovaAyana.github.io>, 2021.
- [186] J. Xu, T.L. Wickramaratne, and N. V. Chawla. Representing higher-order dependencies in networks. *Science Advances*, 2(5), 2016.
- [187] M. Rosvall, A. V Esquivel, A. Lancichinetti, J. D West, and R. Lambiotte. Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*, 5(1):4630, 2014.
- [188] I. Sholtes, N. Wider, A. Garas, C.J. Tessone, and F. Schweitzer. Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. *The European Physical Journal B*, 89(61), January 2016.

- [189] I. Sholtes, N. Wider, A. Garas R. Pfitzner, C.J. Tessone, and F. Schweitzer. Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nature Communications*, 4(5024), December 2017.
- [190] R. Pfitzner, I. Sholtes, and A. Garas. When is a network a network? multi-order graphical model selection in pathways and temporal networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2, August 2017.
- [191] P. Erdos and A. Renyi. On random graphs. *Nature*, 6(290), 1959.
- [192] E.N. Gillbert. Random graphs. *Annals of Mathematical Statistics*, 30(4), 1959.
- [193] E.A. Bender and E.R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Combinatorial Theory A*, 24(296), May 1978.
- [194] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25), 2002.
- [195] D. J. Watts. *Small worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton, 1999.
- [196] A-L Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439), October 1999.
- [197] R. Albert and A-L. Barabasi. Topology of evolving networks: Local events and universality. *Physics Review Letters*, 85(24), December 2000.
- [198] R. Xulvi-Brunet and I.M. Sokolov. Reshuffling scale-free networks: From random to assortative. *Physical Review E*, 70(066102), December 2004.
- [199] P.L. Krapivsky and S. Redner. Organization of growing random networks. *Physics Review E*, 63(66123), April 2001.
- [200] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks: Exact solution of the barabasi-albert’s model. *Physics Review Letter*, 85(4633), April 2000.
- [201] G. Bianconi and A-L. Barabasi. Competition and multiscaling in evolving networks. *Europhysics Letter*, 54(4), 2001.
- [202] K.I Goh; B. Kahng and D. Kim. Universal behavior of load distribution in scale-free networks. *Physical Review Letters*, 87(278701), December 2001.
- [203] G. Caldarelli, A. Capocci, P. DeLosRios, and M. A. Muñoz. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89(258702), 2002.
- [204] D. Vito, P. Servedio, and G. Caldarelli. Vertex intrinsic fitness: How to produce arbitrary scale-free networks. *Physical Review E*, 70(056126), 2004.
- [205] B. Corominas-Murtra and S. Hanel, R. and Thurner. Understanding scaling through history-dependent processes with collapsing sample space. *Proceedings of the National Academy of Sciences*, 112(17):5348–5353, 2015.

- [206] B. Corominas-Murtra and S. Hanel, R. and Thurner. Sample space reducing cascading processes produce the full spectrum of scaling exponents. *Scientific Reports*, 7(1):11223, 2017.
- [207] B. Corominas-Murtra, R. Hanel, L. Zavojanni, and S. Thurner. How driving rates determine the statistics of driven non-equilibrium systems with stationary distributions. *Scientific Reports*, 8(1):10837, 2018.
- [208] R. Pastor-Satorras, A. Vazquez, and A. Vespignani. Dynamical and correlation properties of the internet,. *Physical Review Letters*, 87(258701), 2002.
- [209] H. Jeong, Z. Neda, and A-L. Barabasi. Measuring preferential attachment for evolving networks. *Europhysics Letters*, 61(4), 2003.
- [210] Li. R. Anderson and C. A. Holt. Information cascades in the laboratory. *The American Economic Review*, 1997.
- [211] T. Pham and S. Lee. Anomaly detection in bitcoin network using unsupervised learning methods. *Unpublished Report*, 2016.
- [212] P. Monamo, Vukosi Marimate, and Bheki Twala. Unsupervised learning for robust bitcoin fraud detection, pp. 129–134. *2016 Information Security for South Africa, ISSA*, 2016.
- [213] H. H. Sun Yin, K. Langenheldt, M. Harlev, R.R. Mukkamala, and R. Vatrapu. Regulating cryptocurrencies: A supervised machine learning approach to de-anonymizing the bitcoin blockchain. *Journal of Management Information Systems*, 2019.
- [214] J. V. Monaco. Identifying Bitcoin users by transaction behavior. In *Biometric and Surveillance Technology for Human and Activity Identification XII*, volume 9457, pages 25 – 39. International Society for Optics and Photonics, SPIE, 2015.
- [215] D. Di Francesco Maesa, A. Marino, and L. Ricci. Detecting artificial behaviours in the bitcoin users graph. *Online Social Networks and Media*, 3-4:63 – 74, 2017.
- [216] L. Cocco, G. Concas, and M. Marchesi. Using an artificial financial market for studying a cryptocurrency market. *Journal of Economic Interaction and Coordination*, Springer, 2015.
- [217] L. Cocco, R. Tonelli, and M. Marchesi. An agent-based artificial market model for studying the bitcoin trading. *IEEE Access*, 7, 2019.
- [218] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander. Where is current research on blockchain technology?—a systematic review. *PLOS One*, 11(10), 2016.
- [219] A. T. Aspembitova, L. Feng, and L. Y. Chew. Behavioral structure of users in cryptocurrency market. *PLOS ONE*, 16(1):1–19, 01 2021.
- [220] Feng L. *Econophysics and Agent-Based Modeling of Financial Market*. PhD thesis, National University of Singapore, 2013.
- [221] B. LeBaron. Agent based computational finance: suggested readings and early research. *Journal of Economic Dynamics and Control*, 24, 2000.

- [222] G. Iori and J. Porter. Agent based modelling for financial markets. *The Oxford Handbook on Computational Economics and Finance*. Oxford University Press, 2018.
- [223] P. Plerou, V. and Gopikrishnan, L. Amaral, X. Gabaix, and H. E. Stanley. Economic fluctuations and anomalous diffusion. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 62:R3023–6, 10 2000.
- [224] S. Solorio-Fernández, J. Carrasco-Ochoa, and J.F. Martínez-Trinidad. A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 01 2019.
- [225] N. Grira, M. Crucianu, and N. Boujemaa. Unsupervised and semi-supervised clustering: A brief survey. *Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (6th Framework Programme)*, 2005.
- [226] F. Murtagh. Structure of hierarchic clusterings: implications for information retrieval and for multivariate data analysis. *Information Processing and Management*, 20(5-6), 1984.
- [227] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. da F. Costa, and F. A. Rodrigues. Clustering algorithms: A comparative approach. *PLOS ONE*, 14(1):1–34, 01 2019.
- [228] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996.
- [229] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.
- [230] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 1999.
- [231] A. Singh, A. Yadav, and A. Rana. k-means with three different distance metrics. *International Journal of Computer Applications*, 67(10), 2013.
- [232] T.M. Kodinariya and P.R. Makwana. Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 6(1), 2013.
- [233] R. Tibshirani. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*, 63(2), 2001.
- [234] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 1987.
- [235] V. N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. (in russian). *Doklady Akademii Nauk USSR*, 181(4), 1968.

- [236] V. N. Vapnik, S. Golovitch, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, 9, 1997.
- [237] V. N. Vapnik. Statistical learning theory. *Wiley*, 1998.
- [238] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. *Proceedings of the International Conference on Computer Vision*, 1998.
- [239] T. Joachims. Text categorization with support vector machines. *Proceedings of the European Conference on Machine Learning*, Springer-Verlag, 1998.
- [240] F. L. Chu and L. Wang. Gene expression data analysis using support vector machines. *Proceedings of the 2003 IEEE International Joint Conference on Neural Networks*, 2003.
- [241] G.H Haldorsson, J.A. Benediktsson, and J.R. Sveinsson. Support vector machines in multisource classification. *Proceedings IGARSS, Toulouse, France*, 2003.
- [242] R. Cont. Long range dependence in financial markets. In *Fractals in Engineering*, pages 159–179, London, 2005. Springer London.
- [243] C.H. Hommes. Financial markets as nonlinear adaptive evolutionary systems. *Quantitative Finance*, 1(1):149–167, 2001.
- [244] V. Madhusudanan Pillai and M.P. Chandrasekharan. An absorbing markov chain model for production systems with rework and scrapping. *Computers and Industrial Engineering*, 55(3):695 – 706, 2008.
- [245] F. Duchin and S. H. Levine. Embodied resource flows and product flows. *Journal of Industrial Ecology*, 14(4):586–597, 2010.
- [246] M. Zais and D.Zhang. A markov chain model of military personnel dynamics. *International Journal of Production Research*, 54(6):1863–1885, 2016.
- [247] J. Kemeny and J. Snell. Finite markov chains. 1976.
- [248] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.
- [249] L. Barnett, A.B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett.*, 103:238701, Dec 2009.
- [250] H. Ltkepohl. New introduction to multiple time series analysis. 2007.
- [251] P. A. Stokes and P. L. Purdon. A study of problems encountered in granger causality analysis from a neuroscience perspective. *Proceedings of the National Academy of Sciences*, 114(34):E7063–E7072, 2017.
- [252] C. Hiemstra and J. D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.
- [253] P. K. Narayan and R. Smyth. Electricity consumption, employment and real income in australia evidence from multivariate granger causality tests. *Energy Policy*, 33(9):1109 – 1116, 2005.

- [254] S. Z. Chiou-Wei, C.-F. Chen, and Z. Zhu. Economic growth and energy consumption revisited — evidence from linear and nonlinear granger causality. *Energy Economics*, 30(6):3063 – 3076, 2008. Technological Change and the Environment.
- [255] A. Roebroeck, E. Formisano, and R. Goebel. Mapping directed influence over the brain using granger causality and fmri. *NeuroImage*, 25(1):230 – 242, 2005.
- [256] A. K. Seth, A. B. Barrett, and L. Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.
- [257] A. Brovelli, A. Ding, M. and Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences*, 101(26):9849–9854, 2004.
- [258] J. R. Freeman. Granger causality and the times series analysis of political relationships. *American Journal of Political Science*, 27(2):327–358, 1983.
- [259] T. Cogley and J. M. Nason. Effects of the hodrick-prescott filter on trend and difference stationary time series implications for business cycle research. *Journal of Economic Dynamics and Control*, 19(1):253–278, 1995.
- [260] D. L. Thornton and D. S. Batten. Lag-length selection and tests of granger causality between money and income. *Journal of Money, Credit and Banking*, 17(2):164–178, 1985.
- [261] S. C. Lera, A. Pentland, and D. Sornette. Prediction and prevention of disproportionately dominant agents in complex networks. *Proceedings of the National Academy of Sciences*, 2020.
- [262] A. Cipriani and A. Fontanari. Dynamical fitness models: evidence of universality classes for preferential attachment graphs, 2019.
- [263] J.-W. Hu, S. Gao, J.-W. Yan, P. Lou, and Y. Yin. Manufacturing enterprise collaboration network: An empirical research and evolutionary model. *Chinese Physics B*, 29(8):088901, aug 2020.
- [264] X. F. Liu, X.-J. Jiang, S.-H. Liu, and C. K. Tse. Knowledge discovery in cryptocurrency transactions: A survey, 2020.