CENTRALISATION IN FINANCIAL, SOCIO-ECONOMIC AND ECOLOGICAL COMPLEX NETWORKS

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
Jian-Hong LIN
(林坚洪)

M.A. USST in System Analysis and Integration

born on March 30, 1988
citizen of P.R. China

accepted on the recommendation of

Prof. Dr. Didier SORNETTE
Prof. Dr. Claudio J. TESSONE

2022

Abstract

This thesis is composed of five research papers I've co-authored with my supervisor, i.e. Professor Didier Sornette and my co-supervisor, i.e. Professor Claudio J. Tessone. While two of the papers constituting this thesis have been published in peer-reviewed journals, the other three are currently under submission. The present thesis focuses on the analysis of centralisation in ecological, economic, financial and social networks.

The first two parts of the thesis are devoted to analyse centralisation in the Bitcoin ecosystem. We consider the Bitcoin Lightning Network (BLN) for 18 months since its launch in January 2018 and analyse its binary and weighted representation at the micro-, meso-, and macro-scale. The results show that the bitcoin distribution in the BLN is strongly uneven: the average Gini coefficient of users' bitcoins is 0.88, reflecting that 10% of the users hold 80% of the bitcoins in the BLN. The increasing unevenness of the bitcoin distribution is further confirmed by the evolution of the Gini coefficient of the centrality measures and by the evidence that the BLN meso-scale structure becomes increasingly compatible with a core-periphery structure.

In the third part of the thesis, we present a novel model for the emergence of collective dynamics in financial markets using an Ising-like model on non-normal networks. Our model has its foundations in the intrinsic asymmetry and hierarchy of social influence that, in turn, can be represented by non-normal networks. The influential nodes in non-normal networks have a large influence on other nodes through directed links. Social imitation and herding that start from the influential nodes' opinions lead to transient dynamics that induce financial bubbles and crashes. Via analytical results, agent-based simulations, and empirical analysis of financial data, we show that financial bubble size is proportional to the Kreiss constant which characterizes the degree of non-normality of the network.

The results of the first three parts of the thesis show that influential nodes play a significant role in the centralisation of the Bitcoin ecosystem as well as in the formation of bubbles in financial systems. Thus, in the fourth part of the thesis, we propose a dynamic Markov process (DMP) to identify influential nodes in complex networks. This method integrates the Markov chain and the spreading dynamics to rank the influence of nodes. Numerical results indicate that the DMP method can accurately evaluate the influence of nodes for both single and multi-spreaders.

In the last part of the thesis, we explore the fitness-complexity algorithm for the nestedness maximization problem. Nestedness refers to a hierarchical network structure where the set of neighbors of a given node is a subset of the neighbors of better-connected nodes. Nestedness maximization aims at sorting the rows and columns of the adjacency matrix to maximize the level of nestedness of the network. By analysing the ecological networks and World Trade country-product networks, we show that the fitness-complexity algorithm is

highly effective to achieve the nestedness maximization task.

Kurzfassung

Diese Arbeit besteht aus fünf Forschungsarbeiten, die ich zusammen mit meinem Betreuer, Prof. Dr. Didier Sornette, und meinem stellvertretender Betreue, Prof. Dr. Claudio Tessone verfasst habe. Zwei der in dieser Dissertation enthaltenen Artikel wurden bereits in von Experten begutachteten Zeitschriften veröffentlicht, während die anderen drei Forschungsarbeiten derzeit eingereicht werden. Die vorliegende Arbeit konzentriert sich auf die Analyse der Zentralisierung in ökologischen, ökonomischen, finanziellen, sowie auch sozialen Netzwerken.

Die ersten beiden Teile der Arbeit widmen sich der Analyse der Zentralisierung im Bitcoin-Ökosystem. Wir betrachten das Bitcoin Lightning-Netzwerk (BLN) während 18 Monaten seit der Einführung im Januar 2018 und analysieren dessen binäre und gewichtete Darstellung auf der Mikro-, Meso-, und Makroebene. Die Ergebnisse zeigen, dass die Bitcoins im BLN stark ungleichmässig verteilt sind: Der durchschnittliche Gini-Koeffizient der Bitcoins pro Benutzer beträgt 0.88, was widerspiegelt, dass 10% der Benutzer rund 80% der Bitcoins im BLN halten. Die zunehmende Ungleichmässigkeit der Bitcoin-Verteilung wird weiter bestätigt durch die Entwicklung des Gini-Koeffizienten der Zentralitätsmasse, sowie durch den Nachweis, dass die mesoskalierte Struktur des BLN zunehmend kompatibel mit einer Kern-Peripherie-Struktur wird.

Im dritten Teil der Arbeit präsentieren wir ein neues Modell für die Entstehung kollektiver Dynamiken in Finanzmärkten unter Verwendung eines Ising-ähnlichen Modells basierend auf nicht-normalen Netzwerken. Unser Modell hat seine Grundlagen in der intrinsischen Asymmetrie und Hierarchie des sozialen Einflusses, der wiederum durch nicht-normale Netzwerke repräsentiert werden kann. Die einflussreichen Knoten in nicht-normalen Netzwerken haben durch gerichtete Verbindungen einen grossen Einfluss auf andere Knoten. Soziale Nachahmung und Herdenverhalten, die von den Meinungen der einflussreichen Knotenpunkte ausgehen, führen zu vorübergehenden Dynamiken, die Finanzblasen und -einbrüche hervorrufen können. Durch analytische Ergebnisse, agentenbasierte Simulationen und empirische Analysen von Finanzdaten zeigen wir, dass die Grösse der Finanzblase proportional zur Kreiss-Konstante ist, die den Grad der Nicht-Normalität des Netzwerks charakterisiert.

Die Ergebnisse der ersten drei Teile der Arbeit zeigen, dass einflussreiche Knoten eine bedeutende Rolle bei der Zentralisierung des Bitcoin-Ökosystems, sowie bei der Bildung von Blasen in Finanzsystemen spielen. Daher schlagen wir im vierten Teil der Arbeit einen dynamischen Markov-Prozess (DMP) vor, um einflussreiche Knoten in komplexen Netzwerken zu identifizieren. Diese Methode integriert die Markov-Kette und die Ausbreitungsdynamik, um den Einfluss von Knoten zu ordnen. Numerische Ergebnisse weisen darauf hin, dass die DMP-Methode den Einfluss von Knoten sowohl für Einzel- als auch für Multi-Spreader genau bewerten kann.

Im letzten Teil der Arbeit erforschen wir den Fitness-Komplexitäts-Algorithmus für das Verschachtelungs-Maximierungs-Problem. Verschachtelung bezieht sich auf eine hierarchische Netzwerkstruktur, bei der die Menge der Nachbarn eines gegebenen Knotens eine Teilmenge der Nachbarn besser verbundener Knoten sind. Die Verschachtelungs-Maximierung zielt darauf ab, die Zeilen und Spalten der Nachbarschafts-Matrix zu sortieren, um den Grad der Verschachtelung eines Netzwerks zu maximieren. Durch die Analyse von ökologischen Netzwerken, sowie Länder-Produkt-Netzwerken des Welthandels zeigen wir, dass der Fitness-Komplexitäts-Algorithmus sehr effektiv ist, um die Aufgabe der Verschachtelungsmaximierung zu erfüllen.

Acknowledgements

This thesis is the results of many efforts during these three years PhD, where my first focus was always improving myself. During this period, many people accompanied me and gave me a lot of support.

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Didier Sornette, for taking me as his PhD student as well as for his continuous support, open-mindedness and patience during my PhD study. His deep insights and detailed feedback had a great impact on my research and pushed our projects to high quality. His knowledge, enthusiasm and optimism influenced me positively in all my academic research. I look forward to continuing working with him in the future.

Next, I am particularly grateful to my co-supervisor Prof. Claudio J. Tessone, for giving me the opportunity to join his team. I am thankful for the excellent discussions and exchanges of ideas with Claudio. Thanks to his continuous support, endless input, ideas, and efforts, I am able to present this thesis. I hope that I still have the opportunity to collaborate also with him in the future.

I thank Prof. Stefano Brusoni for chairing my defense and evaluating my thesis. Special thanks go to Prof. Yi-Cheng Zhang. I have benefited greatly from his tireless support, inspiring ideas and extensive knowledge. I also enjoyed the time playing football and table tennis with him.

I would like to thank my collaborators: Prof. Tiziano Squartini, Prof. Sandro Lera, Dr. Zhao Yang, Dr. Rebecca Westphal and Dr. Manuel Marian. I am grateful to have worked with them. Without their initiative, encouragement and intellectual support, a significant part of this thesis would have not been possible.

Here I want to thank my current and former colleagues: Sumit Kumar Ram, Dongshuai Zhao, Ke Wu, Ran Wei, Ming Chen, Jan-Christian Gerlach, Ali Ayoub, Shengnan Li, Tao Yan, Yu Gao, Yu Zhang, and all the other members from the Chair of Entrepreneurial Risk at ETH Zürich and the Chair of Blockchain and Distributed Ledger Technologies at the University of Zürich. Furthermore, I thank Adriana Schellenbaum-Lenner, Prisca Rohr-Steinmann and Judith Holzheimer for their help during my PhD at ETH Zürich.

I would like to thank Chunping Zeng for having accompanied during my study in Switzer-

land. I want to thank my parents: Zichuan Lin and Shuqin Lin; and the other family members. Without their support, I could have not completed my PhD. I thank my friends: Enrico Maria Fenoaltea, Martin Kindschi, Zhongli Wang, Razan Khattab, Wenyao Zhang, Fanyuan Meng, Izat Bagdatovich Baybusinov, Fujuan Gao, Yuhao Zhou, Ruijie Wang, Tianlong Fan, Junying Cui, Bolun Chen, Tianrong Ding, Xueyu Meng, Weipeng Nie, Alex Mari and Radu Tanase.

Last but not least, I am grateful for the financial support from China Scholarship Council and Prof. Claudio J. Tessone.

Jian-Hong LIN
Zürich, June 2022

Contents

Chapter 1

Introduction

Complex systems involve a variety of research areas, including economics [1], finance [2], society [3], and ecology [4] among many others. Such systems can be described as complex networks, where their elements are abstracted as nodes, and the interactions between them are represented by links. Networks provide a simple and powerful framework for characterising and understanding the properties of complex systems [5]. This thesis focuses on the centralisation in financial, socio-economic, and ecological networks. Centralised networks are those where one or a few nodes are much more important than all the other nodes. The analysis of centralisation is significant for the resilience and security of networked systems [6]. In this thesis, we first show that the emergence of influential nodes gives rise to the centralisation and hierarchy in the Bitcoin ecosystem and the removal of influential nodes leads to the BLN fragmentation into many components [7, 8]. Second, to analyse the influence of influential nodes and hierarchical structure on financial systems, we develop a novel model for the emergence of financial bubbles using an Ising-like model on non-normal networks [9]. Finally, we present a method to identify influential nodes in complex networks [10] and then apply the fitness complexity algorithm on nestedness maximization, where nestedness refers to a hierarchical structure of networks [11]. This thesis comprises five journal papers and each chapter is based on a journal paper [7–11]. In paper [7, 8, 10, 11], I am the sole first author. In paper [9], I am the co-first author. My contributions to this paper are developing the models, performing the analysis, creating most of the figures and writing parts of the paper. In what follows, we provide the motivation and a short summary for each chapter.

In Chapter 2 (based on [7]), we analyse the centralisation of the Bitcoin ecosystem [12]. Bitcoin [13], a decentralised digital currency that can be transferred on a peer-to-peer network, is the world's most adopted cryptocurrency. It has a market capitalization of around 1.17 trillion US dollars at the time of writing and has gained tremendous popularity over the past few years, attracting interests of researchers from diverse disciplines ranging from computer science, to economics and social sciences [14–16]. Such an evidence motivates us to focus on the centralisation of the Bitcoin ecosystem. While the vast majority of the existing papers on the topic focus on the structural analysis of either the Bitcoin user network or the Bitcoin address network [17, 18], we consider the Bitcoin Lightning Network (BLN) representation which has been designed to overcome the major limitation affecting Bitcoin, i.e. its poor scalability [12, 19]. The goal of this chapter is to shed light on the Bitcoin centralisation issue, by answering a simple question: is Bitcoin becoming an increasingly centralised system? To this end, we consider three representations of the BLN (daily, daily-block, and weekly) for 18 months (January 2018-July 2019) on which we compute a set of four centrality measures: degree, closeness, betweenness, and eigenvector centrality. The analysis of the distribution of the aforementioned measures shows that the BLN topology is becoming more and more centralised. Specifically, the BLN is becoming more and more increasingly similar to a combination of star-like sub-graphs, where hubs play the role of channel-switching nodes progressively clustering together into a core. The remaining nodes, on the other hand, constitute the periphery-like part of the network behaving like leaves attached to the core-vertices and being loosely connected among themselves. The tendency to centralisation is also visible by analysing the distribution of strengths: as revealed by the average Gini coefficient: 10% of the nodes hold 80% of the bitcoins at stake in the BLN.

In Chapter 3 (based on [8]), we investigate the centralisation of the weighted BLN. Most of the existing studies have focused on its binary structure [7,20,21] while ignoring the architecture of its weighted counterpart, which has remained largely unexplored. To fill this gap, we consider the weighted BLN for 18 months and analyse its daily-snapshot representation, at the micro-, meso- and macro-scale. We find the presence of fat-tailed degree distributions - often compatible with power-laws - and of weight and strength distributions whose functional form is, instead, compatible with a log-normal; Moreover, we observe disassortative and hierarchical trends. The most remarkable result, however, concerns centralisation: a tendency to centralisation matching the one characterizing the binary BLN in Chapter 2 can be observed. The first evidence is provided by the Nakamoto coefficient, which we have topologically redefined to quantify the percentage of nodes enclosing 51% of the total number of links/total weight. As the size of the BLN increases, the Nakamoto coefficient progressively reduces, indicating that fewer nodes are needed to embody the required percentage. This confirms the appearance of nodes constituting a topological majority. The increasing unevenness of the distribution of the total weight is further confirmed by the evolution of the Gini coefficient of weighted centrality measures and by the evidence that the BLN meso-scale structure becomes increasingly compatible with a core-periphery structure - even from the weighted perspective, with the largest nodes by strength constituting the core of such a network.

In Chapter 4 (based on [9]), we develop a novel model for the growth of transient bubbles on non-normal networks based on the fact that social influences are intrinsically non-symmetric and hierarchically organized [22–24]. Indeed, in financial markets, a famous investor is significantly more influential than others and information does not spread evenly but follows cascading circuits. This is particularly important when modelling bubbles, as they emerge from herding and social imitation of traders. This cascade of opinions can increase the order to buy before its next decline. These insights are finally related to recent bubbles in meme stock trading. The adjacency matrix of the non-normal network should be, in general, asymmetric, and represent a hierarchy of the social influence network. This is illustrated by analyzing Reddit discussion forums of meme stocks. We mimic this dynamical opinion formation using an Ising-like model, in which there are two types of agents, fundamentalists and noise traders, who trade a risky and a risk-free asset. The noise traders are assumed to be on a non-normal network. They are influenced by their in-neighbors when they must decide to invest in the two assets. From the simulated and empirical results, we show that the bubble size is controlled by the Kreiss constant, which is a measure of the degree of non-normality in the network.

In Chapter 5 (based on [10]), we propose a new method to identify important nodes in complex networks. The important nodes are the extraordinary nodes that cause a large influence on the structure and dynamics of networks. For example, the important nodes lead to the centralisation of the Bitcoin ecosystem and give rise to transient bubbles in financial systems. Locating important nodes can help in increasing the spread of news and impact market sentiment collectively [25, 26], preventing the outbreak of the epidemics [27, 28], locating the opinion leaders in social networks and quantifying the influence of scientists and publications [29, 30]. Recent studies have shown that the nodes' influence is determined not only by the network structure but also by the parameters of the dynamical models [31–33]. Indeed, Liu *et al.* [33] show that, when the spreading rate in the susceptible-infected-recovered(SIR) spreading model is small, degree centrality [34, 35] performs better than other centrality measures for ranking nodes' influence . While for large spreading rates, eigenvector centrality [36] is the best one. The key idea of the above methods is to evaluate the fraction of susceptible nodes that have been infected. But, since these methods are linear, they overestimate the spreading influence of nodes as the spreading process is non-linear. To fill this gap, we present a dynamic Markov process (DMP) to evaluate the expected spreading of the outbreak size of the nodes. It overcomes the problem of nonlinear coupling by calculating the probability of the susceptible nodes being infected by their neighbours sequentially and adjusting the state transition matrix during the spreading process. Simulation results in the SIR model show that the DMP method can evaluate the influence of nodes more accurately than the linear methods [31–33] for both single spreader and multi-spreaders.

In Chapter 6 (based on [11]), we apply the fitness-complexity algorithm to the nestedness maximization problem. In Chapters 2 and 3, we observe hierarchical structure in the BLN and in Chapter 4, we show that the hierarchy leads to transient explosive growth. Thus hierarchy plays a crucial role in the network structure and the dynamic process playing on it. Nestedness refers to a hierarchical structure of networks. The concept of nestedness was first coined in biology to characterise the spatial distribution of biotas in isolated, yet spatially-related, landscapes [37]. In structural terms, a perfectly nested pattern is one such that the set of connections of any given node is a subset of the relationships of larger degree ones. The degree of nestedness can be measured by the nestedness temperature. Lower temperatures correspond to more nested topologies. The algorithm to measure the nestedness of the network includes three steps [38]. First, determine a line of perfect nestedness by defining a perfectly nested interaction matrix with the same number of links as in the original matrix. Second, reorder the ranking of rows and columns that produces a ranked matrix of minimal temperature (maximal nestedness). Finally, for a given network and a given ranking of its row-nodes and column-nodes, one calculates the nestedness temperature. In the second step, it is hard to reorder the rows and columns of the adjacency matrix to maximize nestedness. Indeed, there are $N!M!$ possible permutations of rows and columns, where N and M are the number of rows and columns of the adjacency matrix, respectively [39]. The nestedness maximization problem that of determining the ordering of rows and columns of the matrix to maximize the degree of nestedness of a given matrix [39]. Nestedness Temperature Calculator [40] and BINMATNEST (binary matrix nestedness temperature calculator) [38] are the most popular algorithms to quantify the degree of nestedness of a given network. The fitness-complexity algorithm was originally introduced to rank countries and products in the country-product export network [41]. It can sort matrices exhibiting an identifiable "triangular" shape. Thus, we explore the fitness-complexity for the nestedness maximization problem. Our findings on ecological and World Trade data suggest that the fitness-complexity algorithm has the potential to become a standard tool in nestedness analysis.

In Chapter 7, we expose our conclusions and open new research questions associated with each chapter.

Chapter 2

Lightning Network: a second path towards centralisation of the Bitcoin economy

The Bitcoin Lightning Network (BLN), a so-called “second layer” payment protocol, was launched in 2018 to scale up the number of transactions between Bitcoin owners. In this paper, we analyse the structure of the BLN over a period of 18 months, ranging from 14^{th} January 2018 to 13^{th} July 2019, at the end of which the network has reached 8.216 users, 122.517 active channels and 2.732,5 transacted bitcoins. Here, we consider three representations of the BLN: the *daily snapshot* one, the *weekly snapshot* one and the *daily-block snapshot* one. By studying the topological properties of the binary and weighted versions of the three representations above, we find that the total volume of transacted bitcoins approximately grows as the square of the network size; however, despite the huge activity characterising the BLN, the bitcoin distribution is very unequal: the average Gini coefficient of the node strengths (computed across the entire history of the Bitcoin Lightning Network) is, in fact, $\simeq 0.88$ reflecting 10% (50%) of the nodes to hold 80% (99%) of the bitcoins at stake in the BLN (on average, across the entire period). This concentration brings up the question of which minimalist network model allows us to explain the network topological structure. Like for other economic systems, we hypothesise that local properties of nodes, like the degree, ultimately determine part of its characteristics. Therefore, we have tested the goodness of the Undirected Binary Configuration Model (UBCM) in reproducing the structural features of the BLN: the UBCM recovers the disassortative and the hierarchical character of the BLN but underestimates the centrality of nodes; this suggests that the BLN is becoming an increasingly centralised network, more and more compatible with a core-periphery structure. Further inspection of the resilience of the BLN shows that removing hubs leads to the collapse of the network into many components, an evidence suggesting that this network may be a target for the so-called *split attacks*.

Based on Jian-Hong Lin, Kevin Primicerio, Tiziano Squartini, Christian Decker, and Claudio J. Tessone. "Lightning Network: a second path towards centralisation of the Bitcoin economy." *New Journal of Physics* 22, no. 8 (2020): 083022.

2.1 Introduction

The gain of popularity of Bitcoin [13] has made apparent the problems in terms of scalability of the technology upon which it is based: in fact, only a limited amount of transactions per second - whose number is proportional to the size of a block and its release frequency - can be processed by Bitcoin. This shortcoming may prevent the adoption of this payment network at a global scale, especially when considering that classic payment mechanisms (e.g. traditional credit cards) are able to achieve tens of thousands of transactions per second. A naïve (and short term) solution would be represented by an increase of the block size: larger blocks, however, would require larger validation time, storage capability and bandwidth costs, in turn favouring *centralisation*, as fewer entities would become able to validate the new blocks that are appended to the Blockchain; moreover, centralisation in the validation process would make the system less resilient, i.e. more prone to faults and attacks.

The Bitcoin Lightning Network (BLN) [12, 20, 42] aims at breaking the trade-off between block size and centralisation by processing most of the transactions off-chain: it is a "Layer 2" protocol that can operate on top of Blockchain-based cryptocurrencies such as Bitcoin. The origin of the BLN can be traced back to the birth of Bitcoin itself, as an attempt to create *payment channels* across which any two users could exchange money without burdening the entire network with their transaction data - thus allowing for *cheaper* and *faster* transactions (as both the mining fees and the Blockchain confirmation are no longer required). The BLN can, thus, be seen as a solution that does not sacrifice the key feature of Bitcoin, i.e. *decentralisation*, that characterises its *architecture* (i.e. the number of computers constituting the network), its *political organisation* (i.e. the number of individuals controlling the network) and its *wealth distribution* (i.e. the number of individuals owning the actual supply), while enhancing the circulation and the exchange of the native assets.

The BLN has recently raised a lot of interest: Seres [20] argued that the BLN structure can be ameliorated to improve its security; Rohrer [43] showed that the current BLN can be prone to channel exhaustion or attacks aimed at isolating nodes, thus compromising the nodes reachability, the payment success ratio, etc. In this paper, we consider the BLN payment channels across a period of 18 months, i.e. from 14th January 2018 to 13th July 2019, and analyze it at both the daily and the weekly timescale. Our results show that the BLN is characterised by an unequal wealth distribution and by a larger-than-expected centrality of nodes, thus suggesting that the BLN indeed suffers from the aforementioned centralisation issue.

2.2 Methods

Notation. For each time snapshot t, the BLN can be described as a weighted, undirected network with total number of nodes $N^{(t)}$ and represented by the $N^{(t)} \times N^{(t)}$ symmetric matrix $\mathbf{W}^{(t)}$ [44, 45] whose generic entry $w_{ij}^{(t)}$ indicates the total amount of money exchanged between i and j, across all channels, at time t. The total amount of money exchanged by node i, at time t, is $s_i^{(t)} = \sum_{j(\neq i)=1}^{N^{(t)}} w_{ij}^{(t)}$, a quantity that will be also called *capacity*. For the present analysis, we also consider the BLN binary adjacency matrix $\mathbf{A}^{(t)}$, whose generic entry reads $a_{ij}^{(t)} = 1$ if $w_{ij}^{(t)} > 0$ and $a_{ij}^{(t)} = 0$ otherwise. Naturally, the presence of a link between any two nodes i and j, i.e. $a_{ij}^{(t)} = 1$, indicates that one or more payment channels are open, between the same nodes, at time t and the total number of open channels (i.e. links) is simply provided by $L^{(t)} = \sum_{i=1}^{N^{(t)}} \sum_{j=i+1}^{N^{(t)}} a_{ij}^{(t)}$.

Centrality measures. Indices measuring the centrality of a node aim at quantifying the importance of a node in a network, according to some, specific topological property [34,35,46, 47]. Among the measures proposed so far, of particular relevance are the *degree centrality*, the *closeness centrality*, the *betweenness centrality* and the *eigenvector centrality*. Let us briefly describe them:

- the degree centrality [34, 35] of node i coincides with the degree of node i, i.e. the number of its neighbours, normalized by the maximum attainable value, i.e. $N - 1$:

  $$k_i^c = \frac{k_i}{N-1} \tag{2.1}$$

  where $k_i = \sum_{j(\neq i)=1}^{N} a_{ij}$. From the definition above, it follows that the most central node, according to the degree variant, is the one connected to all the other nodes;

- the closeness centrality [34, 35] of node i is defined as

  $$c_i^c = \frac{N-1}{\sum_{j(\neq i)=1}^{N} d_{ij}} \tag{2.2}$$

  where d_{ij} is the topological distance between nodes i and j, i.e. the length of the shortest path(s) connecting them: in a sense, the closeness centrality answers the question "how reachable is a given node?" by measuring the length of the patterns that connect it to the other vertices. From the definition above, it follows that the most central node, according to the closeness variant, is the one lying at distance 1 by each other node;

- the betweenness centrality [34, 48–50] of node i is given by

$$b_i^c = \sum_{s(\neq i)=1}^{N} \sum_{t(\neq i,s)=1}^{N} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{2.3}$$

where σ_{st} is the total number of shortest paths between node s and t and $\sigma_{st}(i)$ is the number of shortest paths between nodes s and t that pass through node i. From the definition above, it follows that the most central node, according to the betweenness variant, is the one lying "between" any two other nodes;

- the eigenvector centrality [33, 34, 50] of node i, e_i^c, is defined as the i-th element of the eigenvector corresponding to the largest eigenvalue of the binary adjacency matrix (whose existence is ensured by the Perron-Frobenius theorem). According to the definition above, a node with large eigenvector centrality is connected to other "well connected" nodes. In this sense, its behavior is similar to the PageRank centrality index.

Gini coefficient. The Gini coefficient has been introduced to quantify the inequality of a country income distribution [51,52]: it ranges between 0 and 1, with a larger Gini coefficient indicating a larger "unevenness" of the income distribution. Here, we apply it to both the distribution of the centrality measures of nodes, i.e.

$$G_c = \frac{\sum_{i=1}^N \sum_{j=1}^N |c_i - c_j|}{2N \sum_{i=1}^N c_i} \tag{2.4}$$

where $c_i = k_i^c, c_i^c, b_i^c, e_i^c$ and to the distribution of the total amount of money exchanged by the nodes of the BLN, i.e.

$$G_s = \frac{\sum_{i=1}^N \sum_{j=1}^N |s_i - s_j|}{2N \sum_{i=1}^N s_i}. \tag{2.5}$$

Centralisation measures. The centrality indices defined above are all normalized between 0 and 1 and provide a rank of the nodes of a network, according to the topological feature chosen for their definition. Sometimes, however, it is useful to compactly describe a certain network structure in its entirety. To this aim, a family of indices has been defined (the so-called *centralisation indices*), encoding the comparison between the structure of a given network and that of the reference network, according to the chosen index. In mathematical terms, any centralisation index reads $C_c = \frac{\sum_{i=1}^N (c^* - c_i)}{\max\{\sum_{i=1}^N (c^* - c_i)\}}$, where $c^* = \max\{c_i\}_{i=1}^N$ represents the maximum value of the chosen centrality measure computed over the network under consideration and the denominator is calculated over the benchmark, defined as the graph providing the maximum attainable value of the quantity $\sum_{i=1}^N (c^* - c_i)$. As it can be proven that the most centralized structure, according to the degree, closeness and betweenness centrality, is the *star graph*, one can define the corresponding centralisation indices:

- the *degree-centralisation* index, as

$$C_{k^c} = \frac{\sum_{i=1}^{N}(k^* - k_i^c)}{(N-2)}; \tag{2.6}$$

- the *closeness-centralisation* index, as

$$C_{c^c} = \frac{\sum_{i=1}^{N}(c^* - c_i^c)}{(N-1)(N-2)/(2N-3)}; \tag{2.7}$$

- the *betweenness-centralisation* index, as

$$C_{b^c} = \frac{\sum_{i=1}^{N}(b^* - b_i^c)}{(N-1)^2(N-2)/2}; \tag{2.8}$$

- the *eigenvector-centralisation* index, as

$$C_{e^c} = \frac{\sum_{i=1}^{N}(e^* - e_i^c)}{(\sqrt{N-1}-1)(N-1)/(\sqrt{N-1}+N-1)}. \tag{2.9}$$

For what concerns the eigenvector index, the star graph does not represent the maximally centralised structure: however, we keep it for the sake of homogeneity with the other quantities.

Benchmarking the observations. Beside providing an empirical analysis of the BLN, in what follows we will also benchmark our observations against a model discounting available information to some extent. Like for other economic and financial systems, we hypothese that local properties of nodes ultimately determine the BLN structure: specifically, we focus on the *degrees* and adopt the the Undirected Binary Configuration Model (UBCM) as a reference model [53,54]. The UBCM captures the idea that the probability for any two nodes to establish a connection depends on their degrees and can be derived within the *constrained entropy maximization* framework, the score function being represented by Shannon entropy

$$S = -\sum_{\mathbf{A}} P(\mathbf{A}) \ln P(\mathbf{A}) \tag{2.10}$$

and the constraints being represented by the degree sequence $\{k_i\}_{i=1}^N$. Upon solving the aforementioned optimization problem [53,54], one derives the probability that any two nodes establish a connection

$$p_{ij} = \frac{x_i x_j}{1 + x_i x_j}, \, \forall \, i < j \tag{2.11}$$

the unknowns $\{x_i\}_{i=1}^N$ representing the so-called Lagrange multipliers enforcing the constraints. In order to numerically determine them, one can invoke the *likelihood maximization principle*, prescribing to search for the maximum of the function

$$\mathcal{L}(\mathbf{x}) = \ln P(\mathbf{A}|\mathbf{x}) = \ln \left[\prod_{i=1}^N \prod_{j=i+1}^N p_{ij}^{a_{ij}} (1 - p_{ij})^{1 - a_{ij}} \right] \tag{2.12}$$

with respect to the vector $\{x_i\}_{i=1}^N$, a procedure leading to the resolution of the following system of equations [53, 54]

$$k_i = \sum_{j(\neq i)=1}^N p_{ij} = \sum_{j(\neq i)=1}^N \frac{x_i x_j}{1 + x_i x_j}, \, \forall \, i. \tag{2.13}$$

Core-periphery detection. Inspecting the evolution of centralisation is useful to understand to what extent the structure of a given network becomes increasingly (dis)similar to that of a star graph; however, although encoding the prototypical centralised structure, carrying out a comparison with such a graph can indeed be too simplistic. Hence, we also check for the presence of the "generalized" star graph structure also known as *core-periphery structure*, composed by a densely-connected core of nodes surrounded by a periphery of loosely-connected vertices. In order to do so, we implement a recently-proposed approach [55], prescribing to minimize the score function known as *bimodular surprise* and reading

$$S_{\parallel} = \sum_{i \geq l_c} \sum_{j \geq l_p} \frac{\binom{C}{i}\binom{P}{j}\binom{V-(C+P)}{L-(i+j)}}{\binom{V}{L}} \tag{2.14}$$

where $V = \frac{N(N-1)}{2}$ is the total number of node pairs, $L = \sum_{i=1}^N \sum_{j=i+1}^N a_{ij}$ is the total number of links, C is the number of node pairs in the core portion of the network, P is the number of node pairs in the periphery portion of the network, l_c is the observed number of links in the core and l_p is the observed number of links in the periphery. From a technical point of view, $S_{\parallel}$ is the p-value of a multivariate hypergeometric distribution [55].

2.3 Data

Since payments in the Bitcoin Lightning Network are *source-routed* and *onion-routed*, the sender must have a reasonably up-to-date view of the network topology, in order to pre-compute the entire payment route. Nodes in the BLN regularly broadcast information about the channels they participate in: each time a channel is opened, or any of its details changes, the two endpoints of the channel announce such changes to the rest of the network. This exchange of information, called *gossip*, allows other nodes to keep their view of the network topology up-to-date, an information that is, then, used to initiate a payment.

The network topology can be visualised by means of the the so-called *routing table*. For this paper, we took regular snapshots of the routing table (every 15 minutes, between January 14^{th} 2018, at blockheight 503816, to July 13^{th} 2019, at blockheight 585844); these snapshots were, then, aggregated into *timespans*, each timespan representing a constant state of a channel from its start to its end. In addition, this information is enriched with data from the Blockchain: since every channel consists of an unspent transaction output on the Bitcoin Blockchain, we can determine the size of a channel and its open and close dates within minutes. Other heuristics can be used to search for potential channels on the Blockchain, without involving the gossip mechanism: this allows us to put a lower bound on the completeness of our measurements.

In the Bitcoin Blockchain, the time between blocks is Poisson distributed with an expected value of 10 minutes between blocks. On a single day, the expected number of new blocks added to the Blockchain is 144. For the sake of simplicity, and without altering in any way the results, we consider this number of blocks our natural timescale (for example, the blocks of the first day range from the 503816^{th} one to the 503959^{th} one while the blocks of the second day range from the 503960^{th} one to the 504103^{rd} one). In this paper, three different representations of the BLN are studied, i.e. the *daily snapshot* one, the *weekly snapshot* one and the *daily-block snapshot* one - even if the results of our analysis will be shown for the daily-block snapshot representation only. A *daily/weekly snapshot* includes all channels that were found to be *active during that day/week*; a *daily-block snapshot* consists of all channels that were found to be *active at the time the first block of the day was released*: hence, the transactions considered for the daily-block representation are a subset of the ones constituting the daily representation.

2.4 Results

Empirical analysis of the BLN binary structure. Figure 2.1 plots the evolution of basic network quantities since launch of the BLN, i.e. the number of nodes, which is a proxy of the number of users, the number of links and the link density. As it can be seen, although the network size increases (for the *daily-block snapshot* N ranges from 2 to 6476 and L ranges from 1 to 55866; in particular, in the last daily snapshot of our dataset we have 6476 nodes and 54440 links), it becomes sparser. However, two different regimes are visible: a first phase

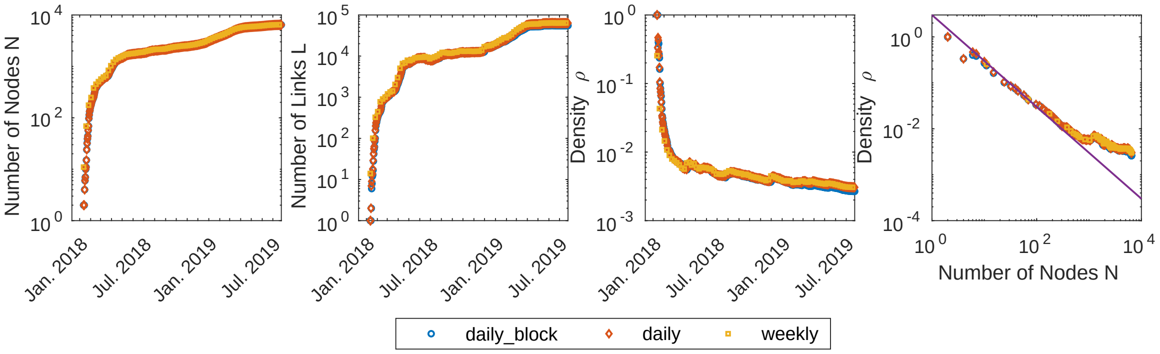

Figure 2.1: (colour online) Evolution of the total number of nodes N, total number of links L and link density $\rho = \frac{2L}{N(N-1)}$ of the BLN. By plotting the link density versus the total number of nodes, further insight can be gained on the functional dependence $\rho = f(N)$: in particular, the position $\rho \sim N^{-1}$ well describes the link density dependence on N for the snapshots satisfying the condition $N \leq 10^3$.

where a steep increase of N and L (descrease of ρ) takes place is followed by a phase during which a much smoother increase (decrease) of the same quantities is observed. Further insight on the BLN evolution can be gained by plotting the link density $\rho = \frac{2L}{N(N-1)}$ versus the total number of nodes N: a trend whose functional form reads $\rho \sim cN^{-\gamma}$, with $\gamma \simeq 1$, clearly appears. However, such a functional form seems to describe quite satisfactorily the BLN evolution up to the period when $N \simeq 10^3$: afterwards, a different functional dependence seems to hold. Notice also that the value of the numerical constant c coincides with the value of the average degree, since $c = \frac{2L}{N-1} = \frac{\sum_{i=1}^{N} k_i}{N-1} \simeq \overline{k}$. By imagining a growth process according to which each new node enters the network by establishing at least one new connection with the existing ones, to ensure that $L^t \geq N^t - 1$, a lower-bound on $c \simeq \overline{k}$ can be deduced: $c \geq 2$ (fig. 2.1 shows the trend $y = 3N^{-1}$ even if the inspection of the evolution of the quantity $c = \frac{2L}{N-1}$ reveals that periods where $c \simeq \overline{k}$ assumes different, constant values can be individuated).

In order to comment on the centrality structure of the BLN, let us explicitly draw it: fig. 2.2 shows the largest connected component of the BLN daily-block snapshot representation on day 16 and on day 34. Several hubs are present (e.g. on day 34, the largest one, having degree $k_{hub}^{34} = 121$, is linked to the 34.3% of nodes): notice that each of them is linked to a plethora of other nodes that, instead, are scarcely linked among themselves. The emergence of structurally-important nodes is further confirmed by plotting the evolution of the Gini index for the distribution of the centrality measures defined in the Methods section (i.e. the degree, the closeness, the betweenness and the eigenvector centrality): fig. 2.3 shows that G_c is increasing for three measures out four, pointing out that the values of centrality are

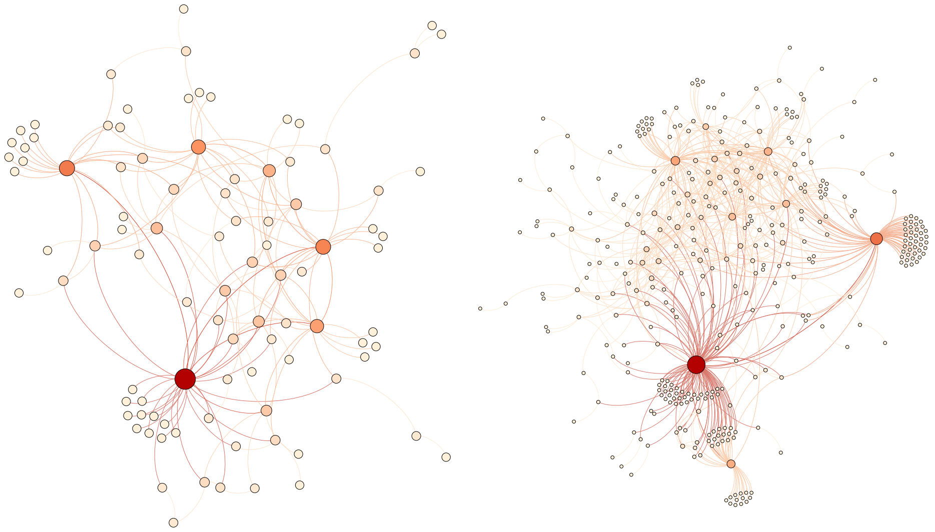

Figure 2.2: (colour online) Comparison between the largest connected component of the BLN (*daily-block snapshot* representation) on day 17 (left - 95 nodes and 155 links are present) and on day 35 (right - 359 nodes and 707 links are present). A visual inspection of the network evolution suggests the presence of a core-periphery structure since its early stages.

more and more unevenly distributed (irrespectively from the chosen indicator). The flat trend characterizing the closeness centrality could be explained by the presence of nodes with large degree ensuring the vast majority of nodes to be reachable quite easily. On the other hand, the evolution of the centralisation indices indicates that the BLN is *not* evolving towards a star graph, although the eigenvector centrality reaches quite large values in the middle stages of the BLN history. As anticipated above, imagining that the picture provided by a star-like structure could provide a good description of the BLN topology is indeed too simplistic.

Benchmarking the observations. Let us now benchmark the observations concerning the centrality and the centralisation indices with the predictions for the same quantities output by the UBCM. More specifically, we have computed the expected value of G_c and C_c (with $c_i = k_i^c, c_i^c, b_i^c, e_i^c, \forall i$) and the corresponding error, by explicitly sampling the ensembles of networks induced by the UBCM. In fig. 2.5 we plot and compare the evolution of the observed and expected values of G_c and C_c, both as functions of N. Such a comparison reveals that the UBCM tends to overestimate the values of the Gini index for the degree, the closeness and the betweenness centrality and to underestimate the values of the Gini index for the eigenvector centrality[1]. These results point out a behavior that is not reproducible

[1] Z-scores, not shown here, confirm that all observations are statistically significant.

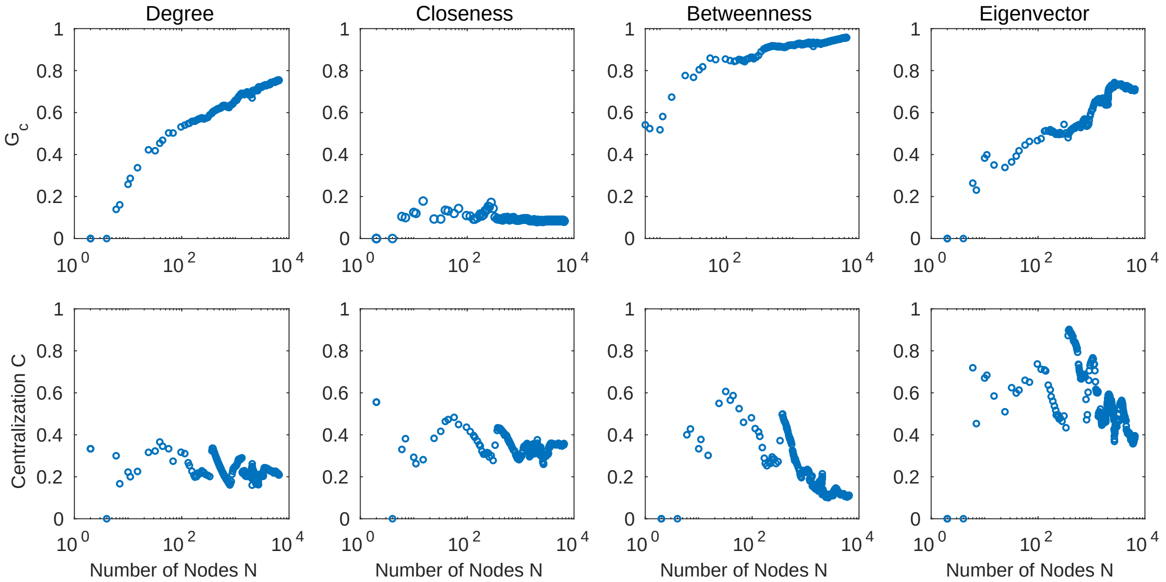

Figure 2.3: (colour online) Top panels: evolution of the Gini index for the degree, closeness, betweenness and eigenvector centrality for the *daily-block snapshot* representation: G_c is characterised by a rising trend, irrespectively from the chosen indicator, pointing out that the values of centrality are increasingly unevenly distributed. Bottom panels: evolution of the degree-, closeness-, betweenness- and eigenvector-centralisation measures: although the eigenvector-centralization index reaches quite large values in the middle stages of the BLN history, the picture provided by a star graph is too simple to faithfully represent the BLN structure.

by just enforcing the degree sequence (irrespectively from the chosen index). The evidence that the UBCM predicts a more-heterogeneous-than observed structure, could be explained starting from the result concerning the eigenvector centrality. The latter, in fact, seems to indicate a non-trivial (i.e. not reproducible by lower-order constraints like the degrees) tendency of well-connected nodes to establish connections among themselves - likely, with nodes having a smaller degree attached to them. Such a *disassortative* structure could explain the less-than-expected level of unevenness characterizing the other centrality measures: in fact, each of the nodes behaving as the "leaves" of the hubs would basically have the same values of degree, closeness and betweenness centrality.

On the other hand, the betweenness- and the eigenvector-centralisation indices suggest that the BLN structure is indeed characterized by some kind of more-than-expected star-likeness: the deviations from the picture provided by such a benchmark, however, could be explained by the co-existence of *multiple* star-like sub-structures (see also fig. 2.2 and the Appendix for a more detailed discussion about this point).

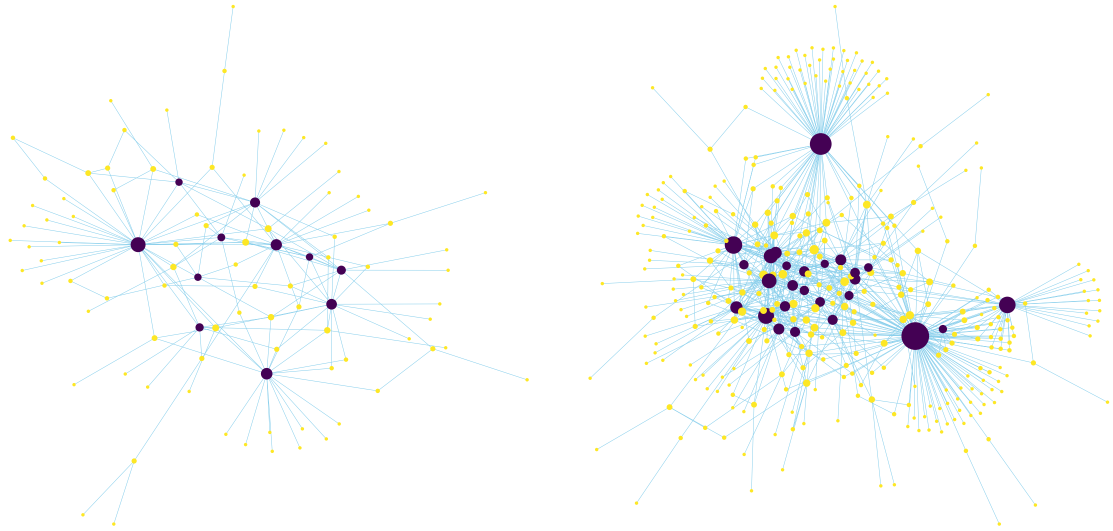

Figure 2.4: (colour online) Core-periphery structure of the BLN *daily-block snapshot* representation on day 17 (left - 95 nodes and 155 links are present) and on day 35 (right - 359 nodes and 707 links are present), with core-nodes drawn in blue and periphery-nodes drawn in yellow.

Core-periphery detection. A clearer picture of the BLN topological structure is provided by the analysis aimed at clarifying the presence of a "core-periphery -like" organization. Inspecting the evolution of the bimodular surprise $S_{||}$ across the entire considered period reveals that the statistical significance of the recovered core-periphery structure increases, a result leading to the conclusion that the description of the BLN structure provided by such a model becomes more and more accurate as the network evolves. As an example, fig. 2.4 shows the detected core-periphery structure on the snapshots depicted in fig. 2.2: the nodes identified as belonging to the core and to the periphery are, respectively, coloured in blue and yellow.

Empirical analysis of the BLN weighted structure. Let us now move to the empirical analysis of the weighted structure of the BLN, by inspecting the evolution of the total capacity W of (i.e. the total number of bitcoins within) the BLN daily-block snapshot representation: fig. 2.6 shows the evolution of W as a function of network size N. The trend shown in the same figure reads $y = aN^b$ with $a = 2 \cdot 10^{-5}$ and $b = 2$. Although the total number of bitcoin rises, inequality rises as well: in fact, the percentage of nodes holding a given percentage of bitcoins at stake in the BLN steadily decreases (on average, across the entire period, about the 10% (50%) of the nodes holds the 80% (99%) of the bitcoins - see the second panel of fig. 2.4). This trend is further confirmed by the evolution the Gini

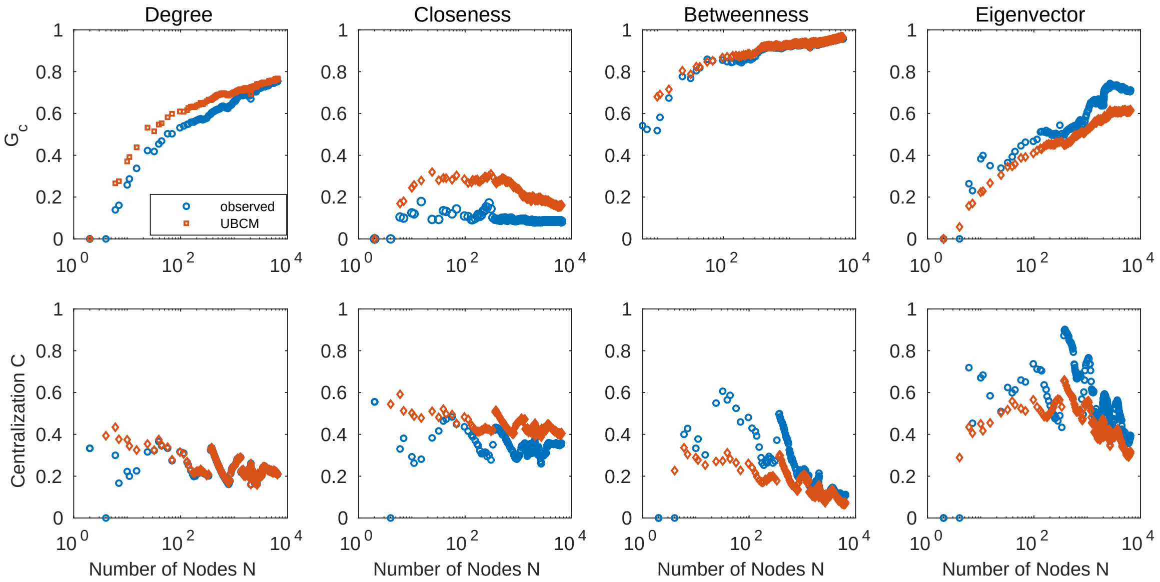

Figure 2.5: (colour online) Top panels: comparison between the observed Gini index for the degree, closeness, betweenness and eigenvector centrality (blue dots) and their expected value, computed under the UBCM (red diamonds) for the *daily-block snapshot* representation. Once the information contained into the degree sequence is properly accounted for, a (residual) tendency to centralisation is still visible. Bottom panels: comparison between the observed degree-, closeness-, betweenness- and eigenvector-centralisation measures and their expected value computed under the UBCM (red diamonds). Once the information contained into the degree sequence is properly accounted for, the emerging picture is that of a network characterized by some kind of more-than-expected star-likeness: deviations from this benchmark, however, are clearly visible and probably due to the co-existence of many star-like sub-structures (see also fig. 2.2).

coefficient G_s, whose value is $\simeq 0.9$ for the last snapshots of our dataset (and whose average value is 0.88 for the daily-block snapshot representation).

2.5 Conclusions

The Bitcoin Lightning Network is a sort of “Layer 2” protocol aimed at speeding up the Blockchain, by enabling fast transactions between nodes. Originally designed to allow for cheaper and faster transactions without sacrificing the key feature of Bitcoin, i.e. its decentralisation, it is evolving towards an increasingly centralised architecture, as our analysis reveals. In particular, its structure seems to become increasingly similar to a core-periphery one, with well-connected nodes clustering together (as revealed by the study of the eigenvector centrality). More precisely, our analysis reveals the presence of many star-like sub-

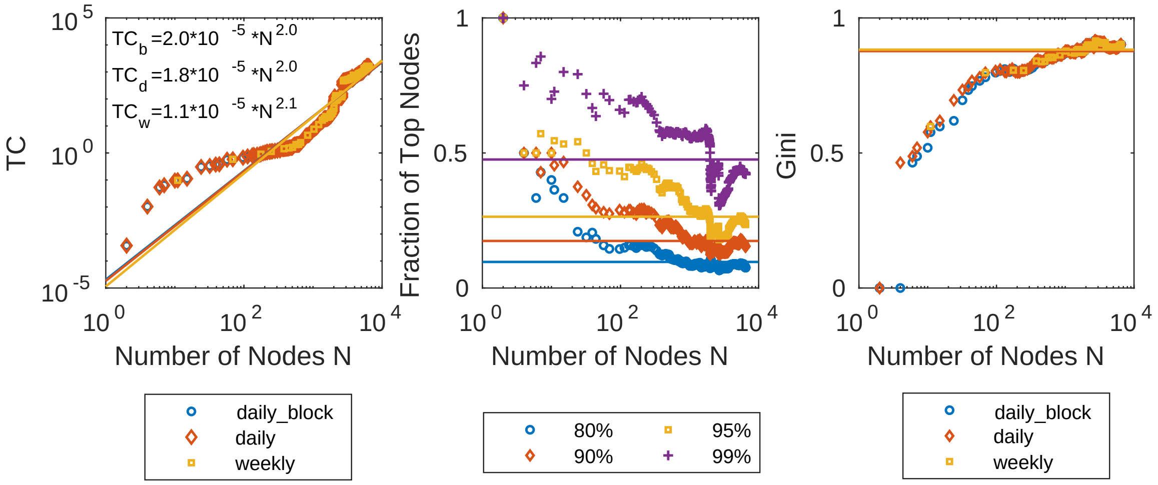

Figure 2.6: (colour online) Evolution of the total capacity of the BLN (left). Percentage of nodes holding the $\simeq 80\%$, $\simeq 90\%$, $\simeq 95\%$ and $\simeq 99\%$ of the total number of bitcoins at stake in the BLN (middle): the former has been computed as the fraction $\frac{n^*}{N}$ of top nodes whose total capacity amounts at $\simeq 80\%$, $\simeq 90\%$, $\simeq 95\%$, $\simeq 99\%$ of the total. Evolution of the Gini coefficient G_s (right): although the total number of bitcoins rises, inequality rises as well.

structures with the role of centers played by the hubs, seemingly acting as channel-switching nodes. Such a tendency seems to be observable even when considering weighted quantities, as only about 10% (50%) of the nodes hold 80% (99%) of the bitcoins at stake in the BLN (on average, across the entire period); moreover, the average Gini coefficient of the nodes strengths is $\simeq 0.88$. These results seems to confirm the tendency for the BLN architecture to become "less distributed", a process having the undesirable consequence of making the BLN increasingly fragile towards attacks and failures.

2.6 Appendix

As anticipated in the main text, the UBCM seems to underestimate the extent to which the topological structure of the BLN is disassortative. Figure 2.7 shows the evolution of the Newman *assortativity coefficient* [56], defined as

$$r = \frac{L\sum_{i=1}^{N}\sum_{j(\neq i)=1}^{N} a_{ij}k_ik_j - \left(\sum_{i=1}^{N} k_i^2\right)^2}{L\sum_{i=1}^{N} k_i^3 - \left(\sum_{i=1}^{N} k_i^2\right)^2}; \tag{2.15}$$

and its expected counterpart under the UBCM: as it is clearly visible, the BLN is more disassortative than expected (i.e. the correlations between degrees are "more negative" than predicted by the UBCM), the reason lying in the presence of the aforementioned star-like sub-structures that, instead, are absent in the model. To further confirm this, we explicitly show two configurations drawn from the UBCM for the snapshots 16 and 34: as fig. 2.7 clearly shows, star-like sub-structures are present to a much lesser extent with respect to the observed counterparts shown in fig. 2.2.

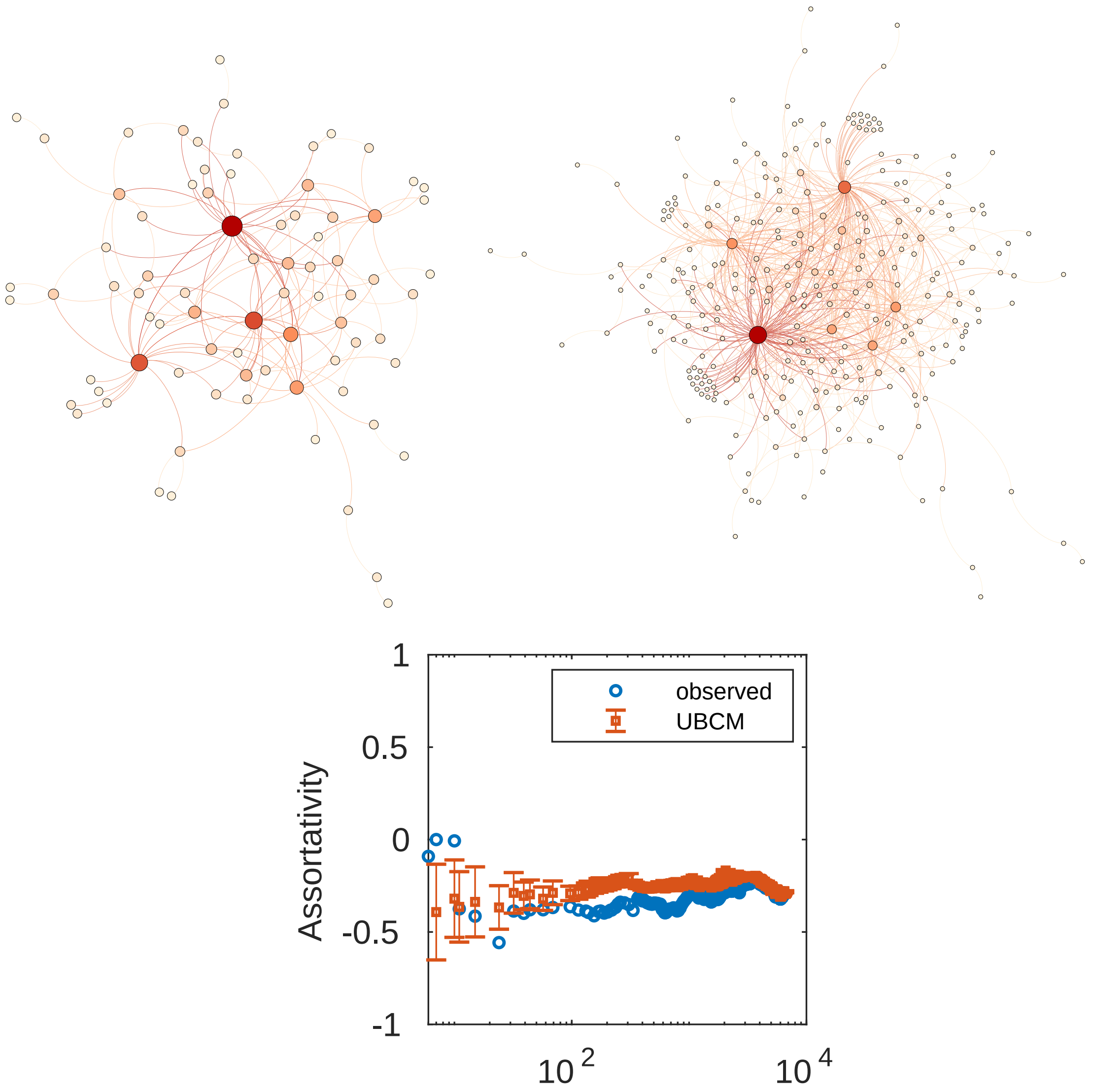

Figure 2.7: (colour online) Top panels: comparison between the largest connected component of the BLN (*daily-block snapshot* representation) generated by the UBCM for the day 17 and the day 35. A visual inspection of these networks confirms that star-like sub-structures are present to a much lesser extent with respect to the observed BLN in the same snapshots. Bottom panel: evolution of the comparison between the empirical assortativity coefficient r (blue dots) and its expected value, computed under the UBCM (red diamonds), for the *daily-block snapshot* representation. The BLN is significantly more disassortative than expected.

Chapter 3

The weighted Bitcoin Lightning Network

The Bitcoin Lightning Network (BLN) was launched in 2018 to scale up the number of transactions between Bitcoin owners. Although several contributions concerning the analysis of the BLN binary structure have recently appeared in the literature, the properties of its weighted counterpart are still largely unknown. The present contribution aims at filling this gap, by considering the Bitcoin Lightning Network over a period of 18 months, ranging from 12th January 2018 to 17th July 2019, and focusing on its weighted, undirected, daily snapshot representation. As the study of the BLN weighted structural properties reveals, it is becoming increasingly 'centralised' at different levels, just as its binary counterpart: 1) the Nakamoto coefficient shows that the percentage of nodes whose degrees/strengths 'enclose' the 51% of the total number of links/total weight is rapidly decreasing; 2) the Gini coefficient confirms that several weighted centrality measures are becoming increasingly unevenly distributed; 3) the weighted BLN topology is becoming increasingly compatible with a core-periphery structure, with the largest nodes 'by strength' constituting the core of such a network, whose size keeps shrinking as the BLN evolves. Further inspection of the resilience of the weighted BLN shows that removing such hubs leads to the network fragmentation into many components, an evidence indicating potential security threats - as the ones represented by the so called 'split attacks'.

Based on Jian-Hong Lin, Emiliano Marchese, Tiziano Squartini and Claudio J. Tessone. "The Weighted Bitcoin Lightning Network." *arXiv*, preprint arXiv:2111.13494.
(submitted to *Chaos, Solitons* & *Fractals*)

3.1 Introduction

The Bitcoin Lightning Network (BLN) represents an attempt to overcome one of the main limitations of the Bitcoin technological design, i.e. *scalability*: at the moment, only a limited amount of transactions per second, whose number is proportional to the size of blocks and their release frequency, can be processed by Bitcoin, a major shortcoming preventing the adoption of this payment system at a global scale - especially when considering that classic payment mechanisms are able to achieve tens of thousands of transactions per second. Increasing the size of the blocks has been proposed as a solution; implementing this choice, however, would require 1) a larger validation time, 2) a larger storage capability and 3) larger bandwidth costs, hence favoring a more centralised validation process: in fact, fewer entities would become able to validate the new blocks, thus making the system as a whole more prone to faults and attacks.

Developers have tried to break the trade-off between block size and centralisation by proposing to process transactions off-chain, i.e. by means of a 'Layer 2' protocol that can operate on top of blockchain-based cryptocurrencies such as Bitcoin: nowadays, such a protocol is known with the name of *Bitcoin Lightning Network* (BLN) and works by creating payment channels across which any two users can exchange money without having the data related to their transactions burdening the entire blockchain.

The BLN has recently raised a lot of interest: Lee et al. [57] showed that the BLN is characterised by a scale-free topology; Lin et al. [7] and Martinazzi et al. [21] analysed the evolution of the BLN topology and found it to have become increasingly centralised at different levels; Seres et al. [20] argued that the BLN structure can be ameliorated to improve its security; the authors of [58] and [43] showed that the current BLN can be prone to channel exhaustion or attacks aimed at isolating nodes, thus compromising their reachability, the payment success ratio, etc. Mizrahi et al. [59] analysed the robustness of the BLN against three different types of attacks: locking channels, disconnecting pairs of nodes and isolating hubs; although their results indicate that the BLN can be disrupted at a relatively low cost, Conoscenti et al. [60] suggested that the BLN is still resilient against the removal of nodes that do not have a significant influence on the probability of success of a payment.

However, most of the aforementioned contributions have just focused on the analysis of the BLN binary structure, leaving its weighted counterpart largely unexplored. With the present paper we aim at filling this gap, by studying the weighted properties of the BLN daily snapshot representation, at the micro-, meso- and macro-scale, across a period of 18 months, i.e. from 12th January 2018 to 17th July 2019.

3.2 Data

Payments in the BLN are *source-routed* and *onion-routed*: hence, in order to pre-compute the entire payment route, the sender must have a reasonably up-to-date view of the network

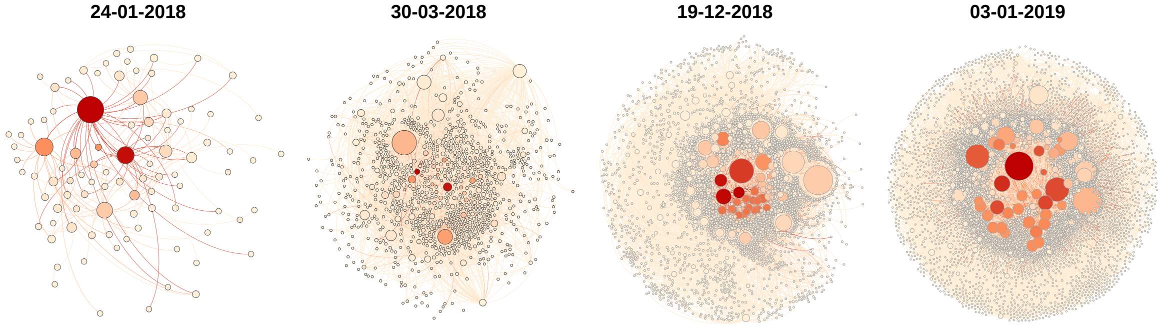

Figure 3.1: Pictorial representation of the four snapshots of the BLN whose LCC is characterised by a number of nodes amounting at 100, 1.000, 3.000, 5.000 and corresponding to the days 24-01-2018, 30-03-2018, 19-12-2018 and 01-03-2019, respectively. The size of each node is proportional to its degree (i.e. the bigger the node, the larger its degree) while the color of each node is proportional to its strength (i.e. the darker the node, the larger its strength).

topology. Nodes in the BLN regularly broadcast information about the channels they participate in: such a mechanism, called *gossip*, allows other nodes to keep their view of the network topology up-to-date.

The BLN topology can be visualised by means of the the so-called *routing table*. For this paper, we took a snapshot of the routing table every 15 minutes, between January 12th 2018, at blockheight 503.816, to July 17th 2019, at blockheight 585.844 [61]: these snapshots were, then, aggregated into *timespans*, each timespan representing a constant state of a channel from its start to its end; for the present analysis, we considered the *daily snapshot* representation of the BLN, including all channels that were found to be active during that day. Importantly, here we do not rest upon estimates of the number of daily blocks - obtainable by considering that the time between the appearance of two subsequent blocks, in the blockchain, is Poisson distributed with an expected value of 10 minutes - but on the exact time our channels have been opened: since every channel consists of an unspent transaction output on the blockchain, we can determine the size of a channel and its opening and closing time within minutes.

3.3 Methods

Notation. On a generic, daily snapshot t, the BLN can be described as a weighted, undirected network with total number of nodes $N^{(t)}$ and represented by an $N^{(t)} \times N^{(t)}$ symmetric matrix $\mathbf{W}^{(t)}$ whose generic entry $w_{ij}^{(t)}$ indicates the total amount of money exchanged between i and j, across all channels established by them, during the snapshot t [44,45]. Consistently, the generic entry of the BLN binary adjacency matrix $\mathbf{A}^{(t)}$ reads $a_{ij}^{(t)} = 1$ if $w_{ij}^{(t)} > 0$ and

$a_{ij}^{(t)} = 0$ otherwise: the presence of a link between any two nodes i and j, i.e. $a_{ij}^{(t)} = 1$, indicates that one or more payment channels have been opened, between the same nodes, during the snapshot t. As a last remark, we will focus on the largest connected component (LCC) of the BLN, throughout its entire history - the percentage of nodes belonging to it being steadily above 90%.

For the sake of illustration, we will plot our results for four snapshots, i.e. the ones whose LCC is characterised by a number of nodes amounting at 100, 1.000, 3.000, 5.000 and corresponding to the days 24-01-2018, 30-03-2018, 19-12-2018 and 01-03-2019, respectively - see fig. 3.1.

Degree and strength distributions. The total number of channels (i.e. *links*) that have been opened during the snapshot t is provided by $L^{(t)} = \sum_{i=1}^{N^{(t)}} \sum_{j=i+1}^{N^{(t)}} a_{ij}^{(t)}$; on the other hand, the total number of channels node i participates in coincides with its *degree*, i.e. $k_i^{(t)} = \sum_{j(\neq i)=1}^{N^{(t)}} a_{ij}^{(t)}$. The weighted counterparts of the notions above coincide with the total weight of the network, i.e. $W^{(t)} = \sum_{i=1}^{N^{(t)}} \sum_{j=i+1}^{N^{(t)}} w_{ij}^{(t)}$, and with the total amount of money exchanged by node i, i.e. $s_i^{(t)} = \sum_{j(\neq i)=1}^{N^{(t)}} w_{ij}^{(t)}$, a quantity often referred to as the node *strength* or the node *capacity*.

While inspecting the functional form of the degree and strength distributions may reveal the presence of hubs, i.e. 'large', single nodes, when dealing with cryptocurrencies it is of interest making a step further and inspecting the presence of 'large subgraphs' of nodes. The meaning of this sentence can be made more precise upon considering the metric designed by Srinivasan et al. [62] to measure the number of addresses required (to collude) for gathering over the 51% of the overall mining power and named *Nakamoto index*: a high Nakamoto coefficient indicates that many miners, or mining pools, need to combine their power to reach the 51% threshold needed to take over the blockchain. Here, we adapt it to quantify a 'topological' kind of majority, by defining

$$N_k = \min\{i \in [1 \dots N] : \sum_i^N f_i \geq 0.51\} \tag{3.1}$$

where $f_i = k_i/2L$ and

$$N_s = \min\{i \in [1 \dots N] : \sum_i^N f_i \geq 0.51\} \tag{3.2}$$

where $f_i = s_i/2W$: the first variant of the Nakamoto index can be calculated by starting from the (node with) largest degree and add them up until the condition above is satisfied; analogously, for the second variant.

Assortativity and hierarchy. In order to gain insight into the higher-order structure of the BLN, we can consider the quantities known as *average nearest neighbors degree*, defined as

$$\text{ANND}_i = \frac{\sum_{j(\neq i)=1}^{N} a_{ij} k_j}{k_i}, \quad \forall\, i \tag{3.3}$$

and *average nearest neighbors strength*, defined as

$$\text{ANNS}_i = \frac{\sum_{j(\neq i)=1}^{N} a_{ij} s_j}{k_i}, \quad \forall\, i; \tag{3.4}$$

while plotting ANND_i versus k_i reveals the (either positive or negative) assortative character of a network, i.e. the presence of (either positive or negative) correlations between degrees, plotting ANNS_i versus k_i reveals the presence of (either positive or negative) correlations between degrees and strengths. On the other hand, the 'cohesiveness' of the neighborhood of each node can be inspected by calculating the *binary clustering coefficient*

$$\text{BCC}_i = \frac{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} a_{ij} a_{jk} a_{ki}}{k_i(k_i - 1)}, \quad \forall\, i \tag{3.5}$$

defined as the percentage of triangles established by any two neighbors of each node and the node itself. Its weighted counterpart reads

$$\text{WCC}_i = \frac{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} w_{ij} w_{jk} w_{ki}}{k_i(k_i - 1)}, \quad \forall\, i \tag{3.6}$$

and is intended to assign a 'weight' to each triangle counted by the BCC, by weighing the connections shaping it. Plotting BCC_i versus k_i reveals the (possibly) hierarchical character of a network, i.e. its organisation in sub-modules; plotting WCC_i versus k_i, instead, provides a hint about the magnitude of the nodes inter-connections as a function of the nodes connectivity.

Disparity. The *disparity index* is defined as

$$Y_i = \sum_{j(\neq i)=1}^{N} \left[\frac{w_{ij}}{s_i}\right]^2 = \frac{\sum_{j(\neq i)=1}^{N} w_{ij}^2}{s_i^2} = \frac{\sum_{j(\neq i)=1}^{N} w_{ij}^2}{\left[\sum_{j(\neq i)=1}^{N} w_{ij}\right]^2}, \quad \forall\, i \tag{3.7}$$

and quantifies the (un)evenness of the distribution of the weights 'constituting' the i-th strength over the k_i links characterising the connectivity of node i. More specifically, the disparity index of node i reads $Y_i = 1/k_i$ in case weights are equally distributed among the connections established by it, i.e. $w_{ij} = a_{ij} s_i / k_i$, $\forall\, j$, any larger value signalling an excess concentration of weight in one or more links.

Centrality. Any index measuring the centrality of a node aims at quantifying its importance in the network, according to some specific topological criterion [34, 35, 46, 47]. While the efforts of researchers have mainly focused on the definition of binary centrality measures, relatively little work has been done on their weighted counterparts. In what follows, we will consider possible extensions of the centrality measures employed in [7], i.e. the *degree*, *closeness*, *betweenness* and *eigenvector* centrality:

- the degree centrality [34, 35] of node i coincides with its degree, normalized by the maximum attainable value, i.e. $\mathrm{DC}_i = k_i/(N-1)$: the strength centrality of node i generalises it by simply replacing the total number of 'node-specific' connections with the total 'node-specific' weight. In what follows we will consider the (simpler) definition

  $$\mathrm{WDC}_i = s_i, \quad \forall\, i \tag{3.8}$$

  from which it follows that the most central node, according to the strength variant, is the one characterised by the largest percentage of weight 'embodied' by (the totality of) its connections;

- the closeness centrality [34,35] of node i is defined as $\mathrm{CC}_i = (N-1)/\sum_{j(\neq i)=1}^{N} d_{ij}$ where d_{ij} is the topological distance between nodes i and j, i.e. the length of any shortest path connecting them. The definition of weighted closeness centrality of node i is based on the redefinition of shortest path length which, in turn, rests upon the redefinition of *weighted distance* between any two nodes, i.e. $d_{ij}^{(w)}$. Possible variants of the latter one read $d_{ij}^{(w)} = \min\{w_{ih} + \dots + w_{hj}\}$ and $d_{ij}^{(w)} = \min\left\{\frac{1}{w_{ih}} + \dots + \frac{1}{w_{hj}}\right\}$ where h indexes the intermediary vertices lying on the path between i and j, $w_{ih} \dots w_{hj}$ are the weights of the corresponding edges and the extremum is taken over all paths between i and j. Naturally, the meaning changes along with the chosen definition: while the first one describes any two nodes as 'closer', the smaller the weights of the intermediate connections, the opposite is true when the second one is considered. Hereby, we opt for the following definition of weighted closeness centrality

  $$\mathrm{WCC}_i = \frac{N-1}{\sum_{j(\neq i)=1}^{N} d_{ij}^{(w)}}, \quad \forall\, i \tag{3.9}$$

  with $d_{ij}^{(w)} = \min\left\{\frac{1}{w_{ih}} + \dots + \frac{1}{w_{hj}}\right\}$. This choice also implies that once the path connecting nodes i and j has been individuated, the WCC is nothing else that the harmonic mean of the weights constituting it;

- the betweenness centrality [34,48–50] of node i is given by $\mathrm{BC}_i = \sum_{s(\neq i)=1}^{N} \sum_{t(\neq i,s)=1}^{N} \frac{\sigma_{st}(i)}{\sigma_{st}}$ where σ_{st} is the total number of shortest paths between node s and t and $\sigma_{st}(i)$ is the number of shortest paths between nodes s and t that pass through node i. The weighted counterpart of it can be defined as

  $$\mathrm{WBC}_i = \sum_{s(\neq i)=1}^{N} \sum_{t(\neq i,s)=1}^{N} \frac{\sigma_{st}^{(w)}(i)}{\sigma_{st}^{(w)}}, \quad \forall\, i \tag{3.10}$$

  where, now, $\sigma_{st}^{(w)}$ is the total number of weighted shortest paths between nodes s and t and $\sigma_{st}^{(w)}(i)$ is the number of weighted shortest paths between nodes s and t that pass through node i;

- the eigenvector centrality [33,34,50] of node i is defined as the i-th element of the eigenvector corresponding to the largest eigenvalue of the binary adjacency matrix - whose existence is guaranteed in case the Perron-Frobenius theorem holds true. According to the definition above, a node with large eigenvector centrality is connected to other 'well connected' nodes. Such a definition can be extended by considering the WEC_i, defined as the i-th element of the eigenvector corresponding to the largest eigenvalue of the weighted adjacency matrix.

$$G_c = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |c_i - c_j|}{2N \sum_{i=1}^{N} c_i}; \tag{3.11}$$

hereby, we apply it to the several definitions of centrality provided above. As a general comment, we would like to stress that a non-normalized centrality measure cannot be employed to compare nodes, across different configurations, in a fully consistent way. However, if our only interest is that of quantifying the (un)evenness of the distribution of our centrality measures, the absence of a normalization term does not make any difference: in fact, the Gini coefficient is not affected by it.

Small-world-ness. The study of the BLN centralisation can be approached from a slightly different perspective by asking if the BLN is (increasingly) becoming a small-world system [63–65]. The usual way of proceeding to answer such a question prescribes to check if

$$\overline{d} = \frac{\sum_{i=1}^{N} \sum_{j(\neq i)=1}^{N} d_{ij}}{N(N-1)} \sim \ln N \tag{3.12}$$

i.e. if the average path length grows logarithmically with the number of nodes and if the average clustering coefficient $\overline{\mathrm{BCC}_i} = \sum_{i=1}^{N} \mathrm{BCC}_i/N$ is larger than the one predicted by an *Undirected Random Graph Model* (URGM) tuned to reproduce the empirical density of links. Recently, however, it has been argued that the same question can be answered by considering the quantity named *global efficiency*, defined as

$$E_g = \frac{\sum_{i=1}^{N} \sum_{j(\neq i)=1}^{N} d_{ij}^{-1}}{N(N-1)}, \tag{3.13}$$

understood as an indicator of the 'traffic capacity' of a network and, quite remarkably, not affected by the analytical problems suffered by the average path length [64] - potentially diverging due to the presence of couples of nodes belonging to disconnected components. Latora et al. [64] have also defined a *local efficiency* as

$$E_l = \frac{1}{N} \sum_{i=1}^{N} E(\mathbf{G}_i), \tag{3.14}$$

a quantity that can be evaluated by, first, calculating the efficiency of the subgraph induced by the nearest neighbors of each node, upon removing it and, then, averaging such numbers. Latora et al. [64] have argued that while E_g plays a role analogous to the inverse of the average path length, E_l plays a role analogous to the average clustering coefficient: hence, small-world networks should have both a large E_g and a large E_l, i.e. should be very efficient in allowing nodes to communicate in both a global and a local fashion.

Core-periphery detection. As it has emerged quite clearly from the binary analysis of the BLN, just inspecting the evolution of centrality measures can return a too simplistic picture of the network under consideration. For this reason, we have checked for the presence of mesoscopic 'centralised' structures such as the *core-periphery* one, composed by a densely-connected subgraph of nodes surrounded by a periphery of loosely-connected vertices. In order to do so, we have implemented the approach recently proposed in [66] and prescribing to minimize the score function

$$\mathscr{W} = \sum_{w_\bullet \geq w_\bullet^*} \sum_{w_\circ \geq w_\circ^*} \frac{\left(\!\binom{V_\bullet}{w_\bullet}\!\right)\left(\!\binom{V_\circ}{w_\circ}\!\right)\left(\!\binom{V-V_\bullet-V_\circ}{W-w_\bullet-w_\circ}\!\right)}{\left(\!\binom{V}{W}\!\right)} \tag{3.15}$$

known as *bimodular surprise*[1]; here, $V = N(N-1)/2$ is the total number of node pairs, $W = \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_{ij}$ is the total weight of the network, $V_\bullet$ is the number of node pairs in the core portion of the network, $V_\circ$ is the number of node pairs in the periphery portion of the network, $w_\bullet^*$ is the observed number of core links and $w_\circ^*$ is the observed number of periphery links. From a technical point of view, $\mathscr{W}$ is the p-value of a multivariate negative hypergeometric

[1] The Python package for surprise optimization, called 'SurpriseMeMore', is freely downloadable at the following URL: https://github.com/EmilianoMarchese/SurpriseMeMore.

distribution and the multiset notation, according to which $\left(\!\binom{V_\bullet}{w_\bullet}\!\right)=\binom{V_\bullet+w_\bullet-1}{w_\bullet}$ allows W to be compactly rewritten in a way that nicely mirrors that of its binary counterpart [55, 66].

Benchmarking the observations. Along the guidelines of the analysis carried out in [7], in what follows we benchmark our observations by employing the recently-proposed null model called CReM$_\text{A}$ - the acronym standing for *Conditional Reconstruction Model A* [67,68] - that allows binary and weighted constraints to be defined in a disentangled fashion. From a purely theoretical point of view, it is defined by the maximisation of the *conditional Shannon entropy*

$$S(W|A)=-\sum_{\mathbf{A}\in\mathbb{A}}P(\mathbf{A})\int_{\mathbb{W}_\mathbf{A}}Q(\mathbf{W}|\mathbf{A})\ln Q(\mathbf{W}|\mathbf{A})d\mathbf{W} \tag{3.16}$$

constrained to reproduce the strengths $\{s_i\}_{i=1}^N$; the (conditional) weighted distribution output by such an optimization procedure reads

$$Q(\mathbf{W}|\mathbf{A})=\frac{e^{-H(\mathbf{W})}}{Z_\mathbf{A}}=\prod_{i=1}^N\prod_{j=i+1}^N q_{ij}(w_{ij}|a_{ij})=\prod_{i=1}^N\prod_{j=i+1}^N(\beta_i+\beta_j)^{a_{ij}}e^{-(\beta_i+\beta_j)w_{ij}}; \tag{3.17}$$

notice the conditional character of the distribution above, embodied by the term a_{ij} at the exponent - as a simple consistency check, the probability that $w_{ij}=0$ in case there is no link is $q(w_{ij}=0|a_{ij}=0)=1$ as it should be. The vector of parameters $\{\beta_i\}_{i=1}^N$ defining the distribution above can be estimated via a (generalized) *likelihood maximisation* procedure [67] that leads to the system of N equations

$$s_i=\sum_{j(\neq i)=1}^N\langle w_{ij}\rangle=\sum_{j(\neq i)=1}^N\frac{p_{ij}}{\beta_i+\beta_j},\quad\forall\, i; \tag{3.18}$$

the coefficients $\{p_{ij}\}_{i,j=1}^N$, instead, are treated as 'prior information' and, as such, left 'untouched' by the estimation procedure above. In a sense, we are free to combine the (conditional) weighted distribution above with the purely binary probability mass function 'best' encoding the available information about the network structure. In what follows, we have considered

- the one defining the *Undirected Binary Configuration Model* (UBCM) and following from the maximisation of the traditional Shannon entropy $S=-\sum_\mathbf{A}P(\mathbf{A})\ln P(\mathbf{A})$ constrained to reproduce the degrees $\{k_i\}_{i=1}^N$: the UBCM captures the idea that the probability for any two nodes to establish a connection (solely) depends on their degrees and can be fully determined by solving the N equations

$$k_i = \sum_{j(\neq i)=1}^{N} p_{ij}^{\text{UBCM}} = \sum_{j(\neq i)=1}^{N} \frac{x_i x_j}{1 + x_i x_j}, \quad \forall\, i; \tag{3.19}$$

- the deterministic recipe $p_{ij} \equiv a_{ij}$, $\forall\, i < j$, accounting for the case in which the prior knowledge concerns the entire network topological structure, now treated as given.

While, in the first case, the generic set of coefficients $\{p_{ij}\}_{i,j=1}^{N}$ is instantiated upon identifying $p_{ij} \equiv p_{ij}^{\text{UBCM}}$, $\forall\, i < j$, in the second one the, identification simply reads $p_{ij} \equiv a_{ij}$, $\forall\, i < j$; in both cases, the resolution of the related system of equations, carried out via the Python package called 'NEMTROPY'[2], leads us to numerically determine the corresponding vector of parameters $\{\beta_i\}_{i=1}^{N}$.

Benchmarking a set of observations ultimately boils down at verifying their 'compatibility' with the predictions output by a chosen null model, by testing their statistical significance against the null model itself. To this aim, one can proceed as follows: first, sampling the ensemble induced by the chosen null model, by generating a 'sufficiently large' number of configurations (in all our experiments, 100); second, calculating the value of any quantity of interest over each configuration; third, deriving the corresponding ensemble CDF. At this point, a p-value remains naturally defined; in what follows, we will employ it to carry out one-tailed tests. Whenever tests of this kind are considered, one may be interested in calculating either the (ensemble) probability $Q(X \geq X^*)$ of observing a value, for the quantity of interest X, that is *larger* than the empirical one, X^*, or the (ensemble) probability $Q(X \leq X^*)$ of observing a value, for the quantity of interest X, that is *smaller* than it; in both cases, if such a probability is found to be smaller than a given threshold, the quantity is deemed as statistically significant, hence incompatible with the description provided by the chosen null model - which (significantly) underestimates or overestimates it, respectively.

3.4 Results

Degree and strength distributions. Giving a look at the four snapshots depicted in fig. 3.1 reveals the presence of a large heterogeneity, with nodes having a large degree/strength co-existing with nodes having a small degree/strength; moreover, while nodes with a large degree also have a large strength (i.e. larger nodes are also darker), small, dark nodes can be observed as well: in other words, an overall positive correlation between degrees and strengths co-exists with a large variability of the strength values - especially for what concerns the nodes with a small connectivity (see fig. 3.2).

[2] The acronym stands for 'Network Entropy Maximization: a Toolbox Running On Python' and the package is freely downloadable at the following URL: https://pypi.org/project/NEMtropy/.

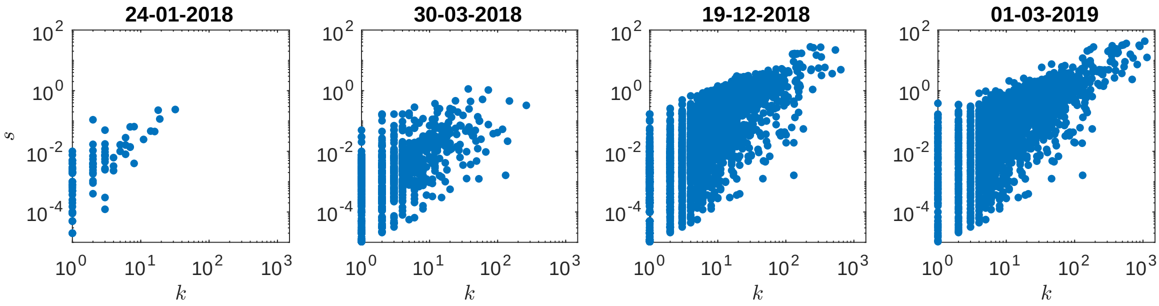

Figure 3.2: Scattering the strength sequence versus the degree sequence reveals the presence of positive correlations between the two sets of quantities: the Pearson coefficient describing them, on our usual four snapshots, amounts at $r = 0.84, 0.42, 0.66, 0.80$, respectively.

As a first empirical analysis, we have inspected the functional form of the degree distribution for four distinct snapshots, i.e. the days 24-01-2018, 30-03-2018, 19-12-2018 and 01-03-2019; to this aim, we have plotted the cumulative density function (CDF), defined as $\text{CDF}(k) = \sum_{h \geq k} f(h)$ where $f(h)$ is the fraction of nodes whose degree is h. As shown in fig. 3.3, the degree distribution becomes broader as the BLN evolves; moreover, running the code released by Clauset et al. [69] to fit the functional form $\text{PDF}(k) = (\alpha - 1)k_{min}^{\alpha-1}k^{-\alpha}$ on the data returns the values $\alpha = 1.9, 2.0, 2.1, 2.2$ and $k_{min} = 1, 3, 14, 26$ while the Kolmogorov-Smirnov test returns the p-values $p = 0.02, 0.03, 0.04, 0.5$. Hence, the null hypothesis that the degrees are distributed according to a power-law is never rejected, at the 1% significance level - while it is, for the first three snapshots, at the 5% significance level. Overall, the null hypothesis that the degrees are distributed according to a power law is not rejected for the 85% of the total number of snapshots, at the 1% significance level, and for the 71% of the total number of snapshots, at the 5% significance level.

As a second empirical analysis, we have calculated the evolution of the CDF of the weights, defined as $\text{CDF}(w) = \sum_{v \geq w} f(v)$. Analogously to what has been observed for the degrees, even the support of the weight distribution has broadened throughout the entire BLN history (see fig. 3.4), although to a lesser extent. Fitting a log-normal distribution, whose functional form reads $\text{PDF}(w) = (w\sigma\sqrt{2\pi})^{-1}e^{-\frac{(\ln w - \mu)^2}{2\sigma^2}}$, on the data reveals that, at both the 1% and the 5% significance levels, the Kolmogorov-Smirnov test does not reject the hypothesis that weights are log-normally distributed when $N < 94$ (i.e. from the fourth day to the twelfth day). For our four snapshots, instead, the hypothesis is rejected - notice that day 24-01-2018 is the thirteenth.

As a third empirical analysis, we have considered the evolution of the CDF of the strengths, defined as $\text{CDF}(s) = \sum_{t \geq s} f(s)$. The support of the distribution is enlarged of a few orders of magnitude during the BLN history (see fig. 3.5). Analogously to the

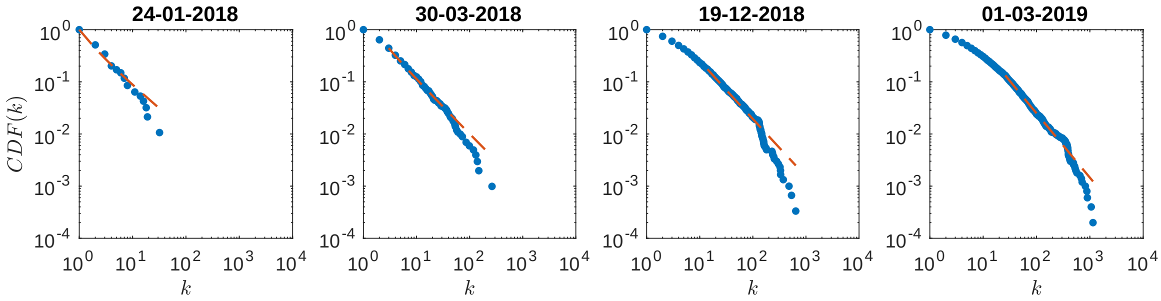

Figure 3.3: Cumulative density function of the degrees, for our usual four snapshots. The support of the distribution has become broader as the BLN has evolved. Fitting a power-law $\text{PDF}(k) = (\alpha-1)k_{min}^{\alpha-1}k^{-\alpha}$ on the data (naturally, for $k \geq k_{min}$), by running the code released by Clauset et al. [69] returns values amounting at $\alpha = 1.9, 2.0, 2.1, 2.2$ and $k_{min} = 1, 3, 14, 26$ while the Kolmogorov-Smirnov test returns the p-values $p = 0.02, 0.03, 0.04, 0.5$. Hence, the null hypothesis that the degrees are distributed according to a power-law is never rejected, at the 1% significance level - while it is, for the first three snapshots, at the 5% significance level. Overall, the null hypothesis that the degrees are distributed according to a power-law is not rejected for the 85% of the total number of snapshots, at the 1% significance level, and for the 71% of the total number of snapshots, at the 5% significance level.

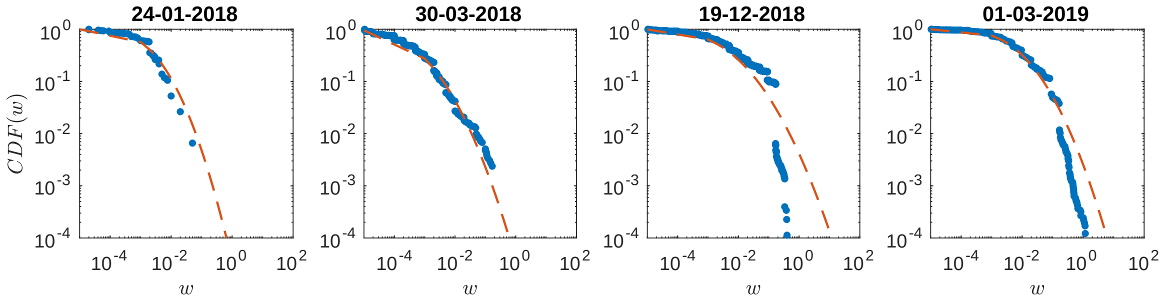

Figure 3.4: Cumulative density function of the weights, for our usual four snapshots. The support of the distribution has become slightly broader as the BLN has evolved. Fitting a log-normal distribution $\text{PDF}(w) = (w\sigma\sqrt{2\pi})^{-1}e^{-\frac{(\ln w-\mu)^2}{2\sigma^2}}$ on the data reveals that, at both the 1% and the 5% significance levels, the Kolmogorov-Smirnov test does not reject the hypothesis that weights are log-normally distributed when $N < 94$ (i.e. from the fourth day to the twelfth day); for our four snapshots, instead, the hypothesis is rejected - notice that day 24-01-2018 is the thirteenth.

case of the weights, we have fitted a log-normal distribution, whose functional form reads $\text{PDF}(s) = (s\sigma\sqrt{2\pi})^{-1}e^{-\frac{(\ln s-\mu)^2}{2\sigma^2}}$, on the data: while the the Kolmogorov-Smirnov test returns the p-values $p = 0.061, 0.006, 4.4 \cdot 10^{-7}$ and $1.6 \cdot 10^{-7}$, hence does not reject the hypothesis that strengths are log-normally distributed, at both significance levels, for the first snapshot

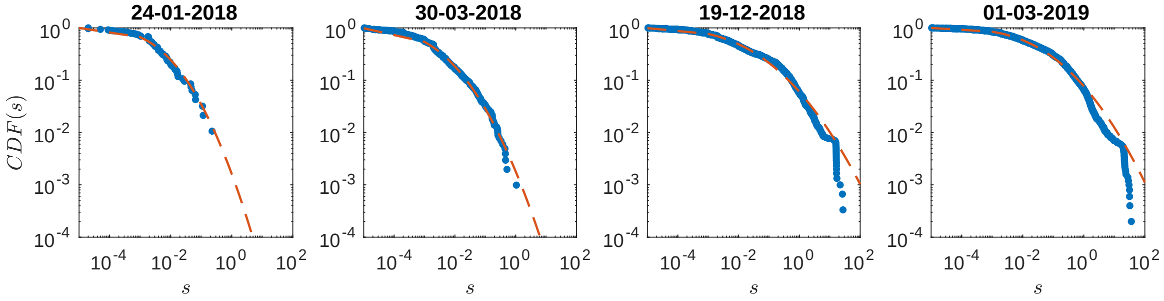

Figure 3.5: Cumulative density function of the strengths, for our usual four snapshots. The support of the distribution is enlarged of a few orders of magnitude during the BLN history. A log-normal distribution $\text{PDF}(s) = (s\sigma\sqrt{2\pi})^{-1}e^{-\frac{(\ln s-\mu)^2}{2\sigma^2}}$, fitted on the data, lets the Kolmogorov-Smirnov test returns the p-values $p = 0.061, 0.006, 4.4 \cdot 10^{-7}$ and $1.6 \cdot 10^{-7}$. Hence, the null hypothesis that strengths are log-normally distributed is not rejected for the first snapshot while it is for the last three ones - at both significance levels. Overall, the null hypothesis that the strengths are distributed according to a log-normal is not rejected for the 16% of the total number of snapshots, at the 1% significance level, and for the 5% of the total number of snapshots, at the 5% significance level.

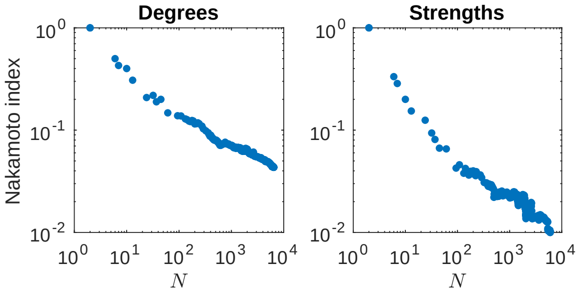

Figure 3.6: Evolution of the Nakamoto index for the degrees and the strengths, plotted versus the total number of nodes: as the size of the system enlarges, the percentage of nodes ‘providing’ the 51% of the total number of links/the total weight progressively reduces, an evidence pointing out that nodes embodying a ‘topological’ kind of majority indeed appear. Moreover, the total weight seems to be distributed less evenly than the total number of connections.

considered here, it does so for the other three ones - an evidence seemingly indicating that, quite early in its history, the BLN has started deviating more and more from the picture provided by the distribution tested here. Overall, the null hypothesis that the strengths are distributed according to a log-normal is not rejected for the 16% of the total number of snapshots, at the 1% significance level, and for the 5% of the total number of snapshots, at the 5% significance level.

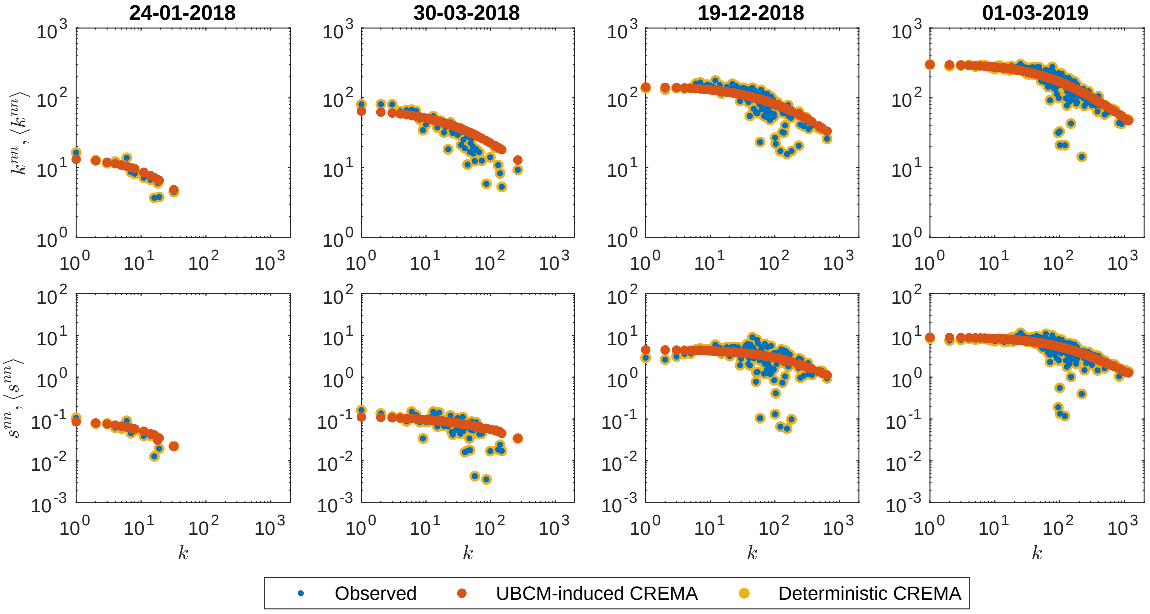

Figure 3.7: ANND_i, $\langle\text{ANND}_i\rangle$ values scattered versus k_i (upper panels) and ANNS_i, $\langle\text{ANNS}_i\rangle$ values scattered versus k_i (bottom panels) for our usual four snapshots (all trends are averaged over the classes of nodes with the same degree). Both trends clearly signal a disassortative behavior, i.e. nodes with a large degree are (preferentially) connected to nodes with a small degree/small strength and viceversa. While the UBCM-induced CReM_A model successfully captures such a disassortative trend, the deterministic CReM_A model reproduces both the ANND and the ANNS values exactly.

The picture provided by the three distributions above can be complemented by the information provided by the Nakamoto index (see fig. 3.6). As its evolution clearly shows, the percentage of nodes 'providing' the 51% of the total number of links/the total weight progressively reduces, as the BLN size enlarges: in particular, the total weight seems to be distributed less evenly than the total number of connections - as fewer nodes are needed to embody the (same) required percentage. This seems to confirm the appearance of nodes constituting the aforementioned 'topological' majority.

Assortativity and hierarchy. Plotting the values of the average nearest neighbors degree versus the degrees reveals the disassortative character of the BLN, i.e. the presence of negative correlations between the degrees: in other words, nodes with a large degree are (preferentially) connected to nodes with a small degree and viceversa. To be noticed that the UBCM-induced CReM_A model successfully captures such a trend, indicating that the information encoded into the degree sequence, leading to

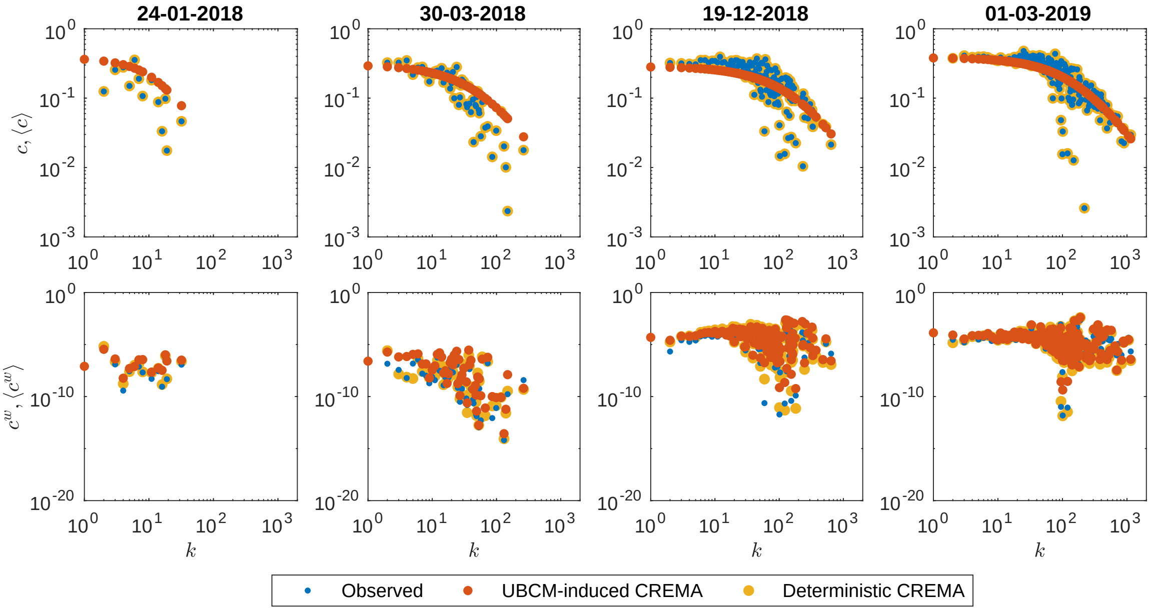

Figure 3.8: BCC_i, $\langle \mathrm{BCC}_i \rangle$ values scattered versus k_i (upper panels) and WCC_i, $\langle \mathrm{WCC}_i \rangle$ values scattered versus k_i (bottom panels) for our usual four snapshots (all trends are averaged over the classes of nodes with the same degree). While the trend of the BCC clearly signals a hierarchical behavior, i.e. the tendency of nodes with a larger degree to participate into a smaller number of connected triples than nodes with a smaller degree and viceversa, this does not seem to be the case for the WCC values when plotted versus the degrees. While the UBCM-induced $\mathrm{CReM_A}$ model successfully captures both trends, the deterministic $\mathrm{CReM_A}$ model reproduces only the BCC values exactly.

$$\langle \mathrm{ANND}_i \rangle \simeq \frac{\sum_{j(\neq i)=1}^{N} \langle a_{ij} \rangle \langle k_j \rangle}{\langle k_i \rangle} = \frac{\sum_{j(\neq i)=1}^{N} p_{ij} k_j}{k_i}, \quad \forall\, i \tag{3.20}$$

with $p_{ij} \equiv p_{ij}^{\mathrm{UBCM}}$, $\forall\, i < j$ and where the symbol $\simeq$ indicates that we have approximated the expected value of a ratio as the ratio of the expected values, is enough to account for the correlations between the degrees as well. An analogous decreasing trend characterises the values of the average nearest neighbors strength when plotted versus the degrees, i.e. nodes with a large degree are (preferentially) connected to nodes with a small strength and viceversa; as for its binary counterpart, the UBCM-induced $\mathrm{CReM_A}$ model successfully reproduces the empirical ANNS values, indicating that the information encoded into the degree and the strength sequences, leading to

$$\langle \mathrm{ANNS}_i \rangle \simeq \frac{\sum_{j(\neq i)=1}^{N} \langle a_{ij} \rangle \langle s_j \rangle}{\langle k_i \rangle} = \frac{\sum_{j(\neq i)=1}^{N} p_{ij} s_j}{k_i}, \quad \forall\, i \tag{3.21}$$

with $p_{ij} \equiv p_{ij}^{\text{UBCM}}$, $\forall\, i < j$ successfully accounts for the correlations between the degrees and the strengths as well (see fig. 3.7). On the other hand, plotting the values of the clustering coefficient versus the degrees reveals the hierarchical character of the BLN: nodes with a larger degree tend to participate into a smaller number of connected triples than nodes with a smaller degree and viceversa; the UBCM-induced CReM$_\text{A}$ model, leading to

$$\langle \text{BCC}_i \rangle \simeq \frac{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} \langle a_{ij} \rangle \langle a_{jk} \rangle \langle a_{ki} \rangle}{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} \langle a_{ij} \rangle \langle a_{ik} \rangle} = \frac{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} p_{ij} p_{jk} p_{ki}}{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} p_{ij} p_{ik}}, \quad \forall\, i \tag{3.22}$$

with $p_{ij} \equiv p_{ij}^{\text{UBCM}}$, $\forall\, i < j$ is able to capture such a trend as well. The same decreasing trend, instead, does not characterise the values of the weighted clustering coefficient when plotted versus the degrees which, instead, appears as rather flat - interestingly, this is no longer true when the weighted clustering coefficient values are plotted versus the strengths: in this case, a clear rising trend is visible, signalling that nodes with a larger strength tend to participate into 'heavier' connected triples of nodes. Again, the UBCM-induced CReM$_\text{A}$ model, predicting

$$\langle \text{WCC}_i \rangle \simeq \frac{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} \langle w_{ij} \rangle \langle w_{jk} \rangle \langle w_{ki} \rangle}{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} \langle a_{ij} \rangle \langle a_{ik} \rangle} = \frac{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} \langle w_{ij} \rangle \langle w_{jk} \rangle \langle w_{ki} \rangle}{\sum_{j(\neq i)=1}^{N} \sum_{k(\neq i,j)=1}^{N} p_{ij} p_{ik}}, \quad \forall\, i \tag{3.23}$$

with $p_{ij} \equiv p_{ij}^{\text{UBCM}}$, $\forall i < j$ successfully reproduces the empirical WCC values, indicating that the information encoded into the degree and the strength sequences successfully accounts for the behavior of third-order properties as well (see fig. 3.8).

Disparity. As anticipated in the paragraph introducing such a quantity, the disparity index of node i reads $Y_i = 1/k_i$ in case weights are equally distributed among the neighbors of node i. Figure 3.9 shows the scatter plot of Y_i as a function of k_i (since it is plotted in a log-log scale, the function $y = -x$ becomes the trend signalling that weights are uniformly distributed among the neighbors of each node): generally speaking, many values lie above the $y = -x$ line, an evidence indicating that some kind of 'excess concentration' of weight (in one or more links) is indeed present - a tendency which is particularly evident for nodes with smaller degree.

Let us now compare the empirical disparity values with the predictions of the null models defined within our CReM$_\text{A}$ framework. To this aim, let us explicitly calculate the expected value of disparity, that reads

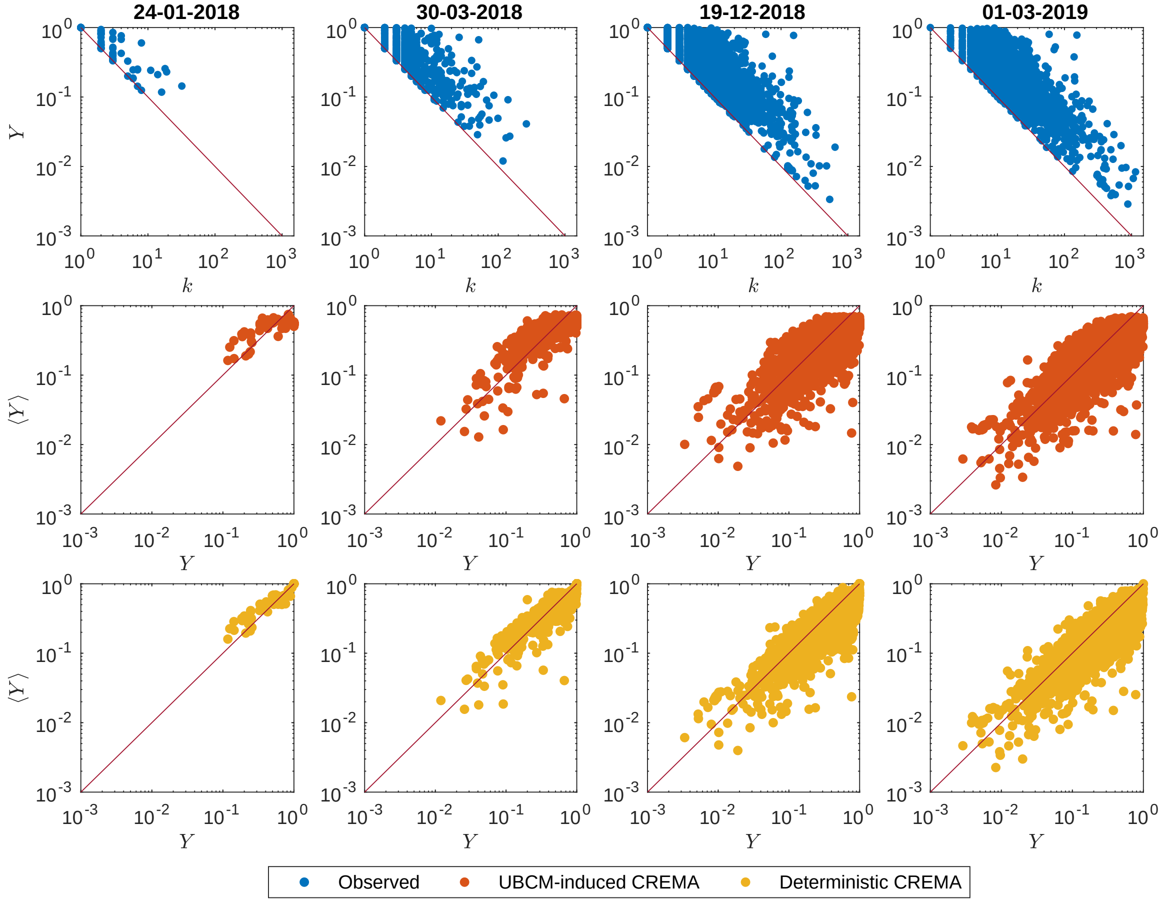

Figure 3.9: Upper panels: empirical disparity values scattered versus the degrees, for our usual four snapshots. As the plots reveal, the vast majority of strength values is not evenly distributed across the connections characterising each node, i.e. $Y_i > 1/k_i$ for the vast majority of nodes. Middle panels: expected disparity values output by the UBCM-induced CReM$_\text{A}$ model scattered versus the empirical disparity values. Bottom panels: expected disparity values output by the deterministic CReM$_\text{A}$ model scattered versus the empirical disparity values. The empirical disparity values are, generally speaking, in agreement with our benchmark models; however, the percentage of nodes for which $Q(Y_i \geq Y_i^*) < 0.05$, for our usual four snapshots, amounts at $0\%, 3.0\%, 9.1\%, 11\%$ for the UBCM-induced CReM$_\text{A}$ model and at $0\%, 5.6\%, 15\%, 17\%$ for the deterministic CReM$_\text{A}$ model: in other words, the percentage of nodes whose empirical disparity is significantly larger than predicted by one of the two null models considered here is rising throughout the entire BLN history - i.e. its vertices increasingly 'favor' some of the links surrounding them.

$$\langle Y_i \rangle \simeq \frac{\sum_{j(\neq i)=1}^{N} \langle w_{ij}^2 \rangle}{\langle s_i^2 \rangle} = \frac{\sum_{j(\neq i)=1}^{N} \left(\text{Var}[w_{ij}] + \langle w_{ij} \rangle^2 \right)}{\text{Var}[s_i] + \langle s_i \rangle^2} = \frac{\text{Var}[s_i] + \sum_{j(\neq i)=1}^{N} \langle w_{ij} \rangle^2}{\text{Var}[s_i] + s_i^2}, \quad \forall\, i \tag{3.24}$$

where

$$\langle w_{ij} \rangle = \frac{p_{ij}}{\beta_i + \beta_j}, \quad \forall\, i < j \tag{3.25}$$

and

$$\text{Var}[s_i] = \sum_{j(\neq i)=1}^{N} \text{Var}[w_{ij}] = \sum_{j(\neq i)=1}^{N} \frac{p_{ij}}{(\beta_i + \beta_j)^2}, \quad \forall\, i \tag{3.26}$$

(naturally, for the the present analysis we have considered both the case $p_{ij} \equiv p_{ij}^{\text{UBCM}}$, $\forall i < j$ and the case $p_{ij} \equiv a_{ij}$, $\forall\, i < j$). As fig. 3.9 shows, disparity is, generally speaking, in agreement with our benchmark models. However, the calculation of the percentage of nodes for which $Q(Y_i \geq Y_i^*) < 0.05$, for our usual four snapshots, reveals it to be $0\%, 3.0\%, 9.1\%, 11\%$ for the UBCM-induced CReM$_\text{A}$ model and $0\%, 5.6\%, 15\%, 17\%$ for the deterministic CReM$_\text{A}$ model: in other words, the percentage of nodes whose empirical disparity is significantly larger than predicted by one of our two null models is rising throughout the entire BLN history. This evidence suggests that, as the BLN evolves, its vertices treat their neighbors less and less equally: indeed, they seem to place weights in a way that increasingly 'favors' some of the links surrounding them - a result that remains true even when a null model constraining the entire topology of the BLN is employed[3].

Centrality. Let us now comment the results concerning the weighted centrality measures considered in the present work. As a general observation, the weighted cases are characterised by trends which are overall similar to the trends characterising the binary cases. As already observed for the purely binary BLN structure, the evolution of the Gini index for most centrality measures points out the latter ones to grow (strongly) unevenly distributed throughout the entire BLN history. While the rise of the Gini coefficient for the weighted degree, betweenness and eigenvector centrality measures suggests the appearance of nodes with 'heavy' connections - further confirmed by the strength distribution, which is a fat-tailed one - likely crossed by many paths and well connected between themselves, the flat trend characterising the evolution of the closeness centrality confirms what has been already observed in the purely binary case, i.e. that the aforementioned 'hubs' ensure the vast majority of nodes to be reachable (hence, to be close to each other) quite easily.

Let us now compare the empirical trends of our four centrality measures with the ones predicted by our two null models. Figure 3.10 reveals that the UBCM-induced CReM$_\text{A}$ model

[3] To be noticed that our null models also underestimate disparity values: however, the corresponding percentages (amounting at $3.2\%, 4.3\%, 5.8\%, 5.2\%$ for the UBCM-induced CReM$_\text{A}$ model and at $12.8\%, 10.8\%, 13.0\%, 13.6\%$ for the deterministic CReM$_\text{A}$ model, for our usual four snapshots), are typically lower and not increasing.

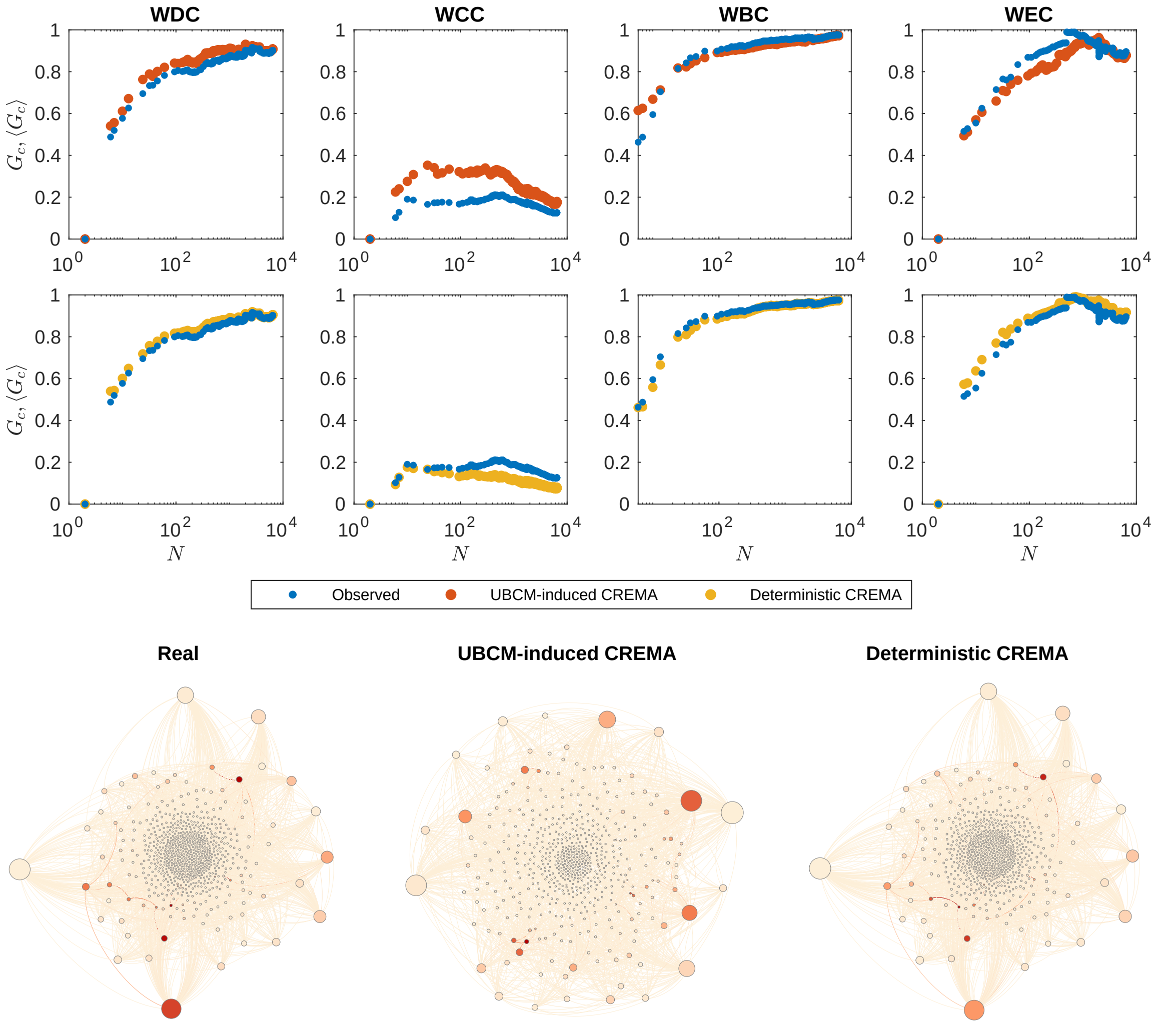

Figure 3.10: Upper and middle panels: evolution of the Gini coefficient of the empirical weighted centrality values (blue dots) and of its expected counterpart, under the UBCM-induced CReM$_{\text{A}}$ model (red dots, upper panels) and the deterministic CReM$_{\text{A}}$ model (yellow dots, bottom panels). The rise of the Gini coefficient for the weighted degree, betweenness and eigenvector centrality points out that the distribution of centrality measures becomes increasingly uneven while the flat trend characterising the evolution of the closeness centrality confirms what has been observed in the purely binary case: the aforementioned 'hubs' ensure the vast majority of nodes to be reachable (hence, to be close to each other) quite easily. Generally speaking, our null models tend to significantly underestimate both the weighted betweenness centrality and the weighted eigenvector centrality, signalling the presence of fewer-nodes-with-heavier-connections than predicted by chance. Bottom panels: comparison between the BLN on day 20-03-2018, a configuration generated by the UBCM-induced CReM$_{\text{A}}$ model and a configuration generated by the deterministic CReM$_{\text{A}}$ model for the same day. The latter one distributes weights more evenly than observed, hence underestimating disparity and letting the strength distribution become wider: this has an interesting consequence, i.e. letting the size of the core become larger than observed, under this model - it amounts at the 12%, the 27% and the 21% of the total number of nodes on the considered day - while still allowing the unevenness of the WEC distribution rise.

tends to overestimate the values of the Gini index for the weighted degree and closeness centrality, i.e. the empirical weighted degree and closeness centrality measures are always significantly lower than their predicted counterparts. For what concerns the weighted betweenness centrality, instead, the percentage of snapshots for which $Q(G_{\text{WBC}} \geq G^*_{\text{WBC}}) < 0.05$ amounts at $\simeq 87\%$, i.e. the UBCM-induced CReM$_{\text{A}}$ model significantly underestimates the weighted betweenness centrality for $\simeq 87\%$ of the total number of snapshots. Analogously, the same null model tends to underestimate the values of the Gini index for the weighted eigenvector centrality roughly one third of the times: in fact, the percentage of snapshots for which $Q(G_{\text{WEC}} \geq G^*_{\text{WEC}}) < 0.05$ amounts at $\simeq 33\%$. Interestingly, the empirical WBC and WEC values are compatible with the predictions output by the UBCM-induced CReM$_{\text{A}}$ model, on the 'remaining' snapshots.

The deterministic CReM$_{\text{A}}$ model, instead, performs slightly better in reproducing the centrality patterns characterising the BLN: in fact, while it still overestimates the Gini index for the weighted degree centrality, the percentage of snapshots for which $Q(G_{\text{WCC}} \geq G^*_{\text{WCC}}) < 0.05$ amounts at $\simeq 96\%$, i.e. the deterministic CReM$_{\text{A}}$ model significantly underestimates the weighted closeness centrality for $\simeq 96\%$ of the total number of snapshots. For what concerns the weighted betweenness centrality, the percentage of snapshots for which $Q(G_{\text{WBC}} \geq G^*_{\text{WBC}}) < 0.05$ amounts at $\simeq 50\%$, i.e. the deterministic CReM$_{\text{A}}$ model significantly underestimates the weighted betweenness centrality roughly half of the times. Lastly, for what concerns the weighted eigenvector centrality, the percentage of snapshots for which $Q(G_{\text{WEC}} \leq G^*_{\text{WEC}}) < 0.05$ amounts at $\simeq 83\%$, i.e. the deterministic CReM$_{\text{A}}$ model significantly overestimates the weighted eigenvector centrality for $\simeq 83\%$ of the total number of snapshots. The deterministic CReM$_{\text{A}}$ model distributes weights more evenly than observed, hence underestimating disparity and letting the strength distribution become wider: this has an interesting consequence, i.e. letting the size of the core under this model become larger than observed - likely, because nodes with relatively low strength become, now, part of the core - while still allowing the unevenness of the WEC distribution rise.

Overall, these results point out a behavior that is not reproducible by just enforcing the degree and the strength sequences - irrespectively from the chosen index: in particular, the behavior of the weighted betweenness centrality points out that both null models - even if to a different extent - predict a more-even-than-observed structure.

Small-world-ness. The evidence that the BLN structure is more-centralised-than-expected rises an interesting question, i.e. if the BLN is small-world or not. From a purely empirical perspective, answering this question amounts at checking the behavior of the average path length, $\overline{d}$, and that of the average clustering coefficient, $\overline{\text{BCC}} = \sum_i \text{BCC}_i/N$ [63–65].

Figure 3.11 shows the results of these two analyses: while the evolution of $\overline{d}$ is described quite accurately by the function $\ln N$ during the first snapshots of the BLN history, its trend has progressively become more and more similar to the smoothest one characterising the function $\ln \ln N$ - which has reached the value $\simeq 3.5$ on the snapshot with 10^4 nodes. For what concerns the average clustering coefficient, one needs to compare it with the value predicted by the URGM, i.e. the null model prescribing to link each pair of nodes with the same probability $p = 2L/N(N-1)$: as fig. 3.11 shows, the URGM significantly underestimates the average clustering coefficient throughout the entire BLN history; taken together, there results indicate that the BLN is indeed small-world. On the other hand, the UBCM overestimates $\overline{\mathrm{BCC}} = \sum_i \mathrm{BCC}_i/N$ during the first half of its history (for $\simeq 40\%$ of the total number of snapshots), thus signalling a tendency of our system to avoid closing paths among triples of nodes.

An alternative way of testing small-world -ness is that of checking the behavior of efficiency. Overall, the global efficiency amounts at $E_g \simeq 0.4$ and it is significantly underestimated by the UBCM throughout the entire history of the BLN. This indicates that the BLN exchanges information more-efficiently-than-predicted by a null model retaining only the information provided by degrees and can be a consequence of the presence of hubs crossed by many paths that shorten the topological distance between (any pair of) nodes.

These results suggest that the BLN has progressively self-organized to 'keep the overall distances low'. What about efficiency from a local point of view? For what concerns the local efficiency, the percentage of nodes for which $Q(E_l \geq E_l^*) < 0.05$ amounts at 75%: hence, the UBCM significantly underestimates it for a large portion of the BLN snapshots - as evident from fig. 3.11, the most recent ones. As the local efficiency $E(\mathbf{G}_i)$ provides information about how efficient the communication between the first neighbors of node i is, upon its removal, our results seem to indicate that the BLN is becoming more and more 'fault tolerant' than its randomised counterpart (interestingly, it appeared to be much more fragile during the first half of its history). This result can be understood by imagining that a larger number of redundant connections has been established, among nodes, in the more recent snapshots of the BLN history - whence the rise of the average clustering coefficient as well.

As an additional exercise, let us inspect the evolution of the BLN global efficiency as nodes are removed either randomly or sequentially, after they have been sorted in decreasing order of weighted degree, closeness, betweenness and eigenvector centrality. The results of our exercise are shown in fig. 3.11. The depicted trends are compatible with a *robust-yet-fragile* architecture, i.e. a topological structure that is robust against a random removal of nodes but fragile against a targeted removal of nodes (e.g. an attack) - or, more correctly, more robust against a random node removal than against a targeted node removal: notice how steeper the decrease of E_g is in the second case; moreover, removing nodes according to

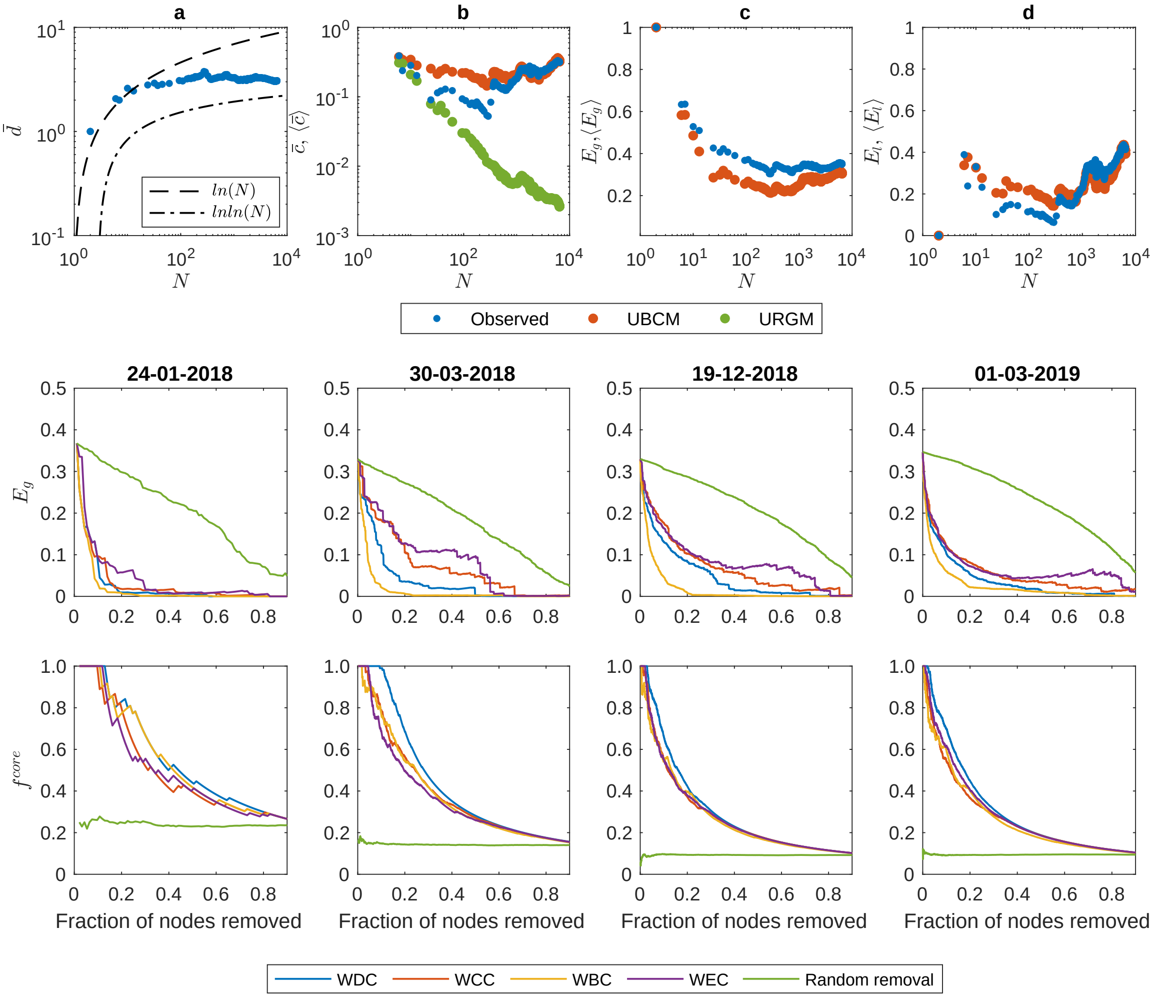

Figure 3.11: Upper panels: (a) evolution of the BLN average path length $\bar{d}$ and of the functions $\ln N$ and $\ln\ln N$; (b) evolution of the empirical average clustering coefficient $\overline{\text{BCC}}$ (blue dots) and of its expected values $\langle\overline{\text{BCC}}\rangle$ under the URGM (green stars) and the UBCM (red squares); (c) evolution of the empirical global efficiency E_g (blue dots) and of its expected values $\langle E_g\rangle$ under the UBCM (red squares); (d) evolution of the empirical local efficiency E_l and of its expected values $\langle E_l\rangle$ under the UBCM (red squares). The BLN is indeed characterised by a small-world structure; moreover, while it has been always more-globally-efficient-than-expected under the UBCM, it has 'recently' become also more-locally-efficient-than-expected under the same null model. Middle panels: evolution of the BLN global efficiency, for our usual four snapshots, when nodes are removed either randomly (green trend) or sequentially, after having been sorted in decreasing order of weighted degree (blue trend), closeness (red trend), betweenness (yellow) and eigenvector (purple) centrality. The trends above characterise a robust-yet-fragile architecture: robust against a random node removal but fragile against a targeted node removal (e.g. an attack). Bottom panels: percentage of core nodes found within the set of nodes removed according to one of the two aforementioned criteria.

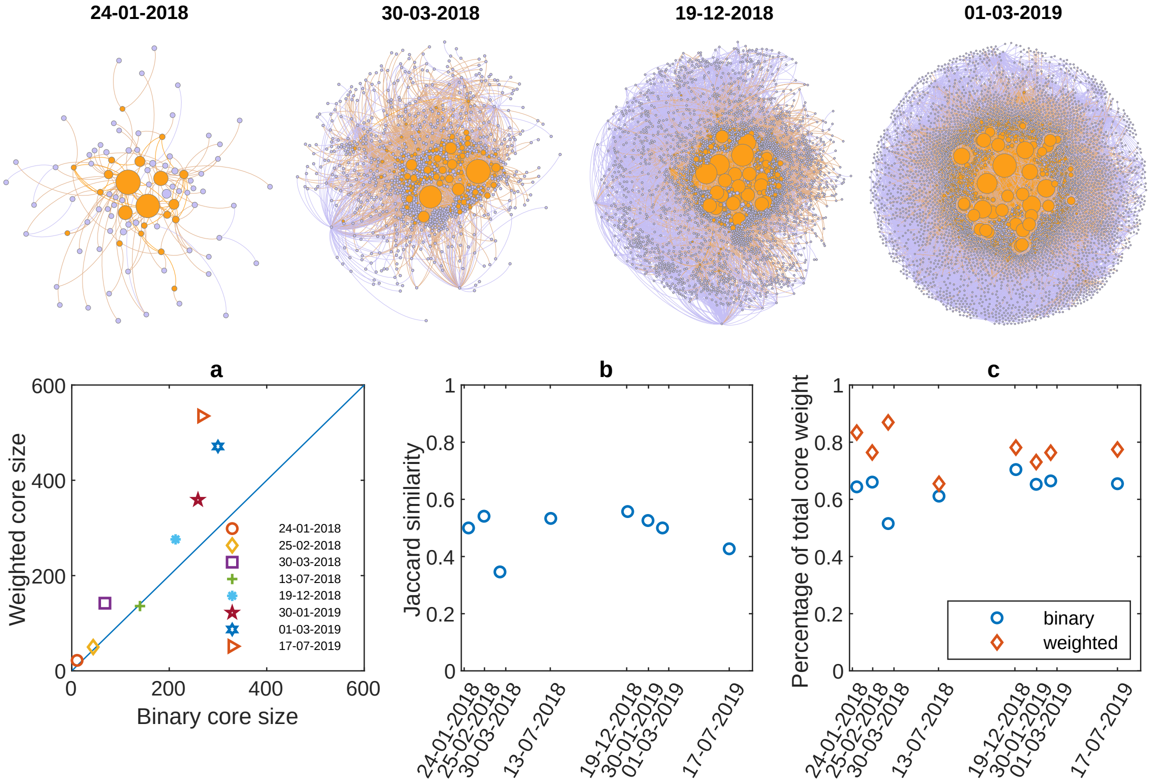

Figure 3.12: Upper panels: core-periphery structure, as revealed by the weighted surprise, W, for our four usual snapshots (core nodes are colored in orange and periphery nodes are colored in purple); the size of the nodes is proportional to their strength: hence, the nodes constituting the core of the network are precisely those with a larger strength. Bottom panels: (a) values of the size of the binary core scattered versus the values of the size of the weighted core; (b) evolution of (the percentage of) the overlap between the set of nodes belonging to the binary core and the set of nodes belonging to the weighted one, estimated via the Jaccard similarity, on a selected subset of snapshots of the BLN; evolution of the percentage of the total network weight embodied by 'core connections', amounting at $\simeq 68\%$ in the binary case (blue circles) and at $\simeq 77\%$ in the weighted one (red diamonds).

their WBC reduces the BLN global efficiency to the largest extent (for the vast majority of snapshots, larger than removing nodes according to their WDC, WCC and WEC).

The same figure also shows that the nodes whose removal brings the most severe damages to the BLN are those belonging to the core (see the next paragraph), whose size shrinks from $\gtrsim 20\%$ to $\simeq 10\%$ of the total number of nodes.

Core-periphery detection. The result concerning the underestimation, by our null models, of the Gini index for the weighted eigenvector centrality - i.e. the presence of well connected nodes which, in turn, are also well connected among them - lets us suppose the BLN to be characterised by a statistically significant core-periphery structure: here, however, we are interested in revealing the presence of a weighted core-periphery structure, i.e. a kind of mesoscale organisation where core nodes are the ones sharing the 'heaviest' connections - and not just those with 'many' connections.

To this aim, we adopt a recently proposed approach, based upon the surprise formalism. In particular, we consider the evolution of the weighted bimodular surprise, W, across the entire BLN history: it reveals that the statistical significance of the recovered core-periphery structure increases, a result leading to the conclusion that the description of the BLN structure provided by such a model becomes more and more accurate as the network evolves. As an example, fig. 3.12 shows the detected core-periphery structure on the snapshots depicted in the same figure: the nodes identified as belonging to the core and to the periphery are, respectively, colored in orange and purple. Notice also that we have drawn the node size proportionally to the node strength: hence, larger nodes, i.e. the ones sharing the 'heaviest' connections, are precisely those constituting the core of the network.

First, let us check the correspondence between the nodes in the core (whose size will be indicated as N_{core}) and vertices with large weighted eigenvector centrality by ranking the nodes in decreasing order of WEC and checking the percentage of top N_{core} nodes that also belong to the core: it amounts at $56\%, 60\%, 57\%, 62\%$, for our usual four snapshots. Then, let us compare the composition of the purely binary core - detected in [7] - with that of the weighted core. As fig. 3.12 shows, a nice correspondence between the size of binary core and that of the weighted one indeed exists although, from a certain moment of the BLN history on, the binary core seems to 'grow slower' than the weighted one which, instead, enlarges to reach a size of $\simeq 600$ nodes: this further confirms that the nodes with a 'large' strength, revealed by surprise as the most central ones, do not necessarily coincide with those having a 'large' degree.

The evolution of (the percentage of) the overlap between the set of nodes belonging to the binary core and the set of nodes belonging to the weighted one further confirms that the two sets do not coincide perfectly, although the Jaccard similarity steadily points out a $\simeq 60\%$ of overlap: in other words, 60% of the nodes belong to both cores - likely, those hubs whose degree *and* strength are large enough to justify their coreness in both senses; similarly, the percentage of the total network weight embodied by 'core connections' amounts at $\simeq 68\%$ in the binary case and at $\simeq 77\%$ in the weighted one.

3.5 Discussion

The analysis of the binary BLN structure carried out in [7] has revealed a system whose topology has become increasingly characterised by star-like structures, whose centers are constituted by 'hubs' to which many nodes having a (much) small(er) degree, in turn, attach. Such a structure - whose disassortativity is confirmed by scattering the ANND values versus the degrees - could explain the more-than-expected level of unevenness characterizing the betweenness and the eigenvector centralisation indices, suggesting them to be due to the emergence of channel-switching nodes - apparently, an unavoidable consequence of the way BLN is designed: on the one hand, as longer routes are more expensive, any two BLN users will search for a short(est) path; on the other, nodes have the incentive to become as central as possible, in order to maximize the transaction fees they may earn.

The tendency to centralisation is observable even when considering weighted quantities, as the percentage of nodes whose connections embody the 51% of the total weight progressively reduces and the Gini coefficient of several (weighted) centrality measures steadily increases throughout the entire BLN history. This clearly points out the co-existence of nodes playing deeply different 'structural' roles, with 'many' peripheral vertices co-existing with 'few' core ones; if, on the one hand, this structure allows the global efficiency to achieve a large value (i.e. hubs facilitate the global exchange of information, being at the origin of another structural BLN peculiarity, i.e. its small-world -ness), on the other it highlights the tendency of the BLN architecture to become increasingly 'less distributed', a process having the undesirable consequence of making it increasingly fragile towards failures and attacks.

Distinguishing between the two is crucial, in order to properly understand the BLN robustness to 'damages'. While resilience towards failures can be tested by looking at how the global efficiency 'reacts' to random node removal, resilience towards attacks can, instead, be quantified by implementing targeted removals of the 'most important' nodes. To this aim, we have ranked nodes in decreasing order of weighted degree, closeness, betweenness and eigenvector centrality and removed them, sequentially: the global efficiency drops rapidly after few (core) nodes are deleted - in fact, for almost all snapshots, removing just *one top node* (according to any of the aforementioned criteria) is enough to disconnect the graph. Moreover, since top nodes are likely to be part of the core - whose size shrinks from $\gtrsim 20\%$ to $\simeq 10\%$ of the total number of nodes - our results indicate that the vertices belonging to it are precisely those whose removal causes the major structural damages. Random failures, instead, cause the decrease of E_g to be much less steep: taken together, the results above seem to indicate that the BLN topology is an example of *robust-yet-fragile* architecture, i.e. a structure that is robust against a random node removal but fragile against a targeted node removal (e.g. an attack).

Chapter 4

Non-normal interactions create socio-economic bubbles

We present a generic new mechanism for the emergence of collective exuberance among interacting agents in a general class of Ising-like models that have a long history in social sciences and economics. The mechanism relies on the recognition that socio-economic networks are intrinsically non-symmetric and hierarchically organized, which is represented as a non-normal adjacency matrix. Such non-normal networks lead to transient explosive growth in a generic domain of control parameters, in particular in the subcritical regime. Contrary to previous models, here the coordination of opinions and actions and the associated global macroscopic order do not require the fine-tuning close to a critical point. This is illustrated in the context of financial markets theoretically, numerically via agent-based simulations and empirically through the analysis of so-called meme stocks. It is shown that the size of the bubble is directly controlled through the Kreiss constant which measures the degree of non-normality in the network. This mapping improves conceptually and operationally on existing methods aimed at anticipating critical phase transitions, which do not take into consideration the ubiquitous non-normality of complex system dynamics. Our mechanism thus provides a general alternative to the previous understanding of instabilities in a large class of complex systems, ranging from ecological systems to social opinion dynamics and financial markets.

Based on Didier Sornette†, Sandro Claudio Lera†, Jian-Hong Lin† and Ke Wu. "Non-Normal Interactions Create Socio-Economic Bubbles." *arXiv*, preprint arXiv:2205.08661.
(†: equal contribution, first authors, submitted to *Science Advances*)

4.1 Introduction

Many complex dynamical systems are characterised by periods of relative stability and "normal" behaviors, interrupted by transient regimes during which the dynamics exhibits an explosive behavior (a “bubble”) or shifts suddenly to another attractor. A large corpus of knowledge and methods have been developed in the last two decades, which are based on the underlying concept of tipping points, wherein a critical threshold is reached at which the system bifurcates to a new state [70–74]. Here, we suggest that this common growing wisdom is incomplete and present a new mechanism for the emergence of large transient instabilities. Based on the mathematics of non-normal dynamical operators represented as non-normal networks, we suggest that this new mechanism is much more general and likely to be often the dominant process at work, because it does not require the fine-tuning close to or sweeping of the system over a critical point. The robust ubiquitous ingredients are (i) asymmetric interactions between elements or agents on the network and (ii) a degree of hierarchy. Together, these two ingredients give rise to networks with non-normal adjacency matrices, whose associated dynamical systems are known to induce transient bursts [75–78]. Interpreted in terms of socio-economic interactions, these transient bursts are responsible for short-lived social contagion even well below any critical threshold. We demonstrate this mechanism in the context of the formation of arguably the largest anomalies of financial markets, namely financial bubbles and their following crashes that lead to enormous economic losses.

A typical ingredient of models of financial markets is the presence of two classes of traders: fundamentalists who maximize their expected utility function and noise traders [79–81]. Noise traders are typically assumed to influence each other according to an Ising-like dynamics, with interaction dependencies captured by an adjacency matrix $\mathbf{A}$ and interaction strength captured by a coupling constant κ. When the imitation strength between noise traders is large enough, collective social behavior can occur, such as polarization of noise traders toward buying (selling), which in turn creates bubbles (crashes) [82–85].

These bubbles and crashes are generally associated with the underlying Ising phase transition separating a disordered opinion regime for low imitation strength κ from an ordered regime where all noise traders tend to be synchronized. In all existing models of this type, bubbles requires the imitation strength κ to be close to or larger than a critical value κ_c associated with the underlying phase transition. In other words, in this class of models, bubbles and crashes are the signatures of the fact that the financial market has entered a “critical regime”, in the technical sense of the emergence of collective order in the decisions of a large fraction of traders. There is an extensive literature on agent-based models and generalized Ising models [86–91]. To the best of our knowledge - in all cases - the abnormal stylized facts, such as excess volatility and transient bubbles and crashes, require the system to be close to, at or above the critical point in the ordered polarizing regime.

In this work, we document a much more general mechanism for the nucleation and growth of transient bubbles. It is based on the fact that social influence is typically directed and hierarchical [22–24]. Indeed, in our example of financial markets, the influence of a famous investor on a retail investor is likely much larger than the other way around. The adjacency matrix $\mathbf{A}$ should thus represent a hierarchy of asymmetric interactions [92, 93]. Mathematically, this is represented by non-normal adjacency matrices $\mathbf{A}$ [94,95]. Analyzing Reddit discussion forums of meme stocks, we show that patterns of influence are highly non-normal, and that the rate of reciprocity is dependent on a user hierarchy. Using a previously validated agent-based model [96,97], we show that non-normal networks give rise to transient bubbles even when the imitation strength is sub-critical. Intuitively, some traders are more influential than others and information does not spread evenly but along cascading circuits. This cascade of opinions can result in an increase of the buy orders (social ordering) before it decays eventually. These insights are finally related to recent financial bubbles in meme stock trading. Our work thus provides a novel angle to substantiate qualitative proposals that financial systems are intrinsically generating crises [98, 99].

Due to the broad applicability of Ising-like interaction models [100], our model is expected to generalize in a straight forward manner to other hierarchical socio-economic systems, and help explain wide-spread phenomena such as social bubbles [101] and herding in opinion dynamics [87].

4.2 Agent Based Price Simulation

Agent-based models (ABMs) have become a popular tool in interdisciplinary research over the last decades [2,102–105], primarily due to their flexibility in accounting for heterogeneous and non-linear interactions. Here, we implement an ABM that simulates a financial market consisting of fundamentalists and noise traders who trade a risky and a risk-free asset [96,97]. The risky asset is a dividend paying stock. The risk-free asset pays a constant return in each time-step and represents a bank account or risk-free government bond. Fundamentalists are rational risk-averse investors who invest by maximizing their expected utility under a constant relative risk aversion utility function. The noise traders invest based on social imitation. Each trader formulates their excess demand for the next time step and the price P_t of the risky asset is calculated as the Walras equilibrium in which supply equals demand. Details are found in SI Appendix 1. Here, we focus our attention on the social imitation mechanism, which is of center importance for the subsequent analysis.

The noise traders' investment strategy is based on an Ising-like social influence model, where they can be modeled as nodes in a network with directed edges. Each node i is considered to be in one of the two possible states, $+1$ (the noise trader holds the risky asset), and -1 (the noise trader holds the risk-free asset). The states are denoted as $s^i = \pm 1$, respectively. The transition probability π that trader i flips its state s_t^i at time t depends on the opinion of their in-neighbours, according to

$$\pi(s_{t+1}^i = -s_t^i) = \frac{p^{\pm}}{2}\left(1 - \kappa \frac{1}{k_i^{in}} s_t^i \sum_j a_{ij} s_t^j\right) \tag{4.2.1}$$

where $p^{\pm}$ controls the average holding time of each asset and the social coupling strength κ determines the noise traders' susceptibility to social imitation. The directed network of N nodes (noise traders) is described by its adjacency matrix $\mathbf{A} = \{a_{ij}\}$, where $a_{ij} = 1$ if there exists a directed edge (influence) from node j to node i, and $a_{ij} = 0$ otherwise. The in-degree of node i is the number of directed edges pointing to node i, which is given by $k_i^{in} = \sum_{j=1}^{N} a_{ij}$.

In the Ising model, if node i switches its state from time step t to time step $t+1$, i.e. $s_{t+1}^i = -s_t^i$, the change of the value of node i's state is $-2s_t^i$. Given the probability of node i to switch its state according to (4.2.1), the average rate of change of the spin starting in the state s_t^i is given by $\Delta s_t^i = -2\ s_t^i\ \pi$ We introduce the n-dimensional state vector $\vec{s}(t) = \left(s_t^1, s_t^2, ..., s_t^N\right)$. Together with (4.2.1), the average rate of state transition can then be written as $\triangle\vec{s}(t) = \vec{s}(t+1) - \vec{s}(t) = p^{\pm}(\kappa\mathbf{\Lambda}\mathbf{A} - \mathbf{I})\vec{s}(t)$ where $\mathbf{\Lambda}$ is an $N \times N$ diagonal matrix with $1\Big/k_i^{in}$ on the i-th diagonal entry and $\mathbf{I}$ is the identity matrix. Introducing the matrix

$$\mathbf{M} \equiv p^{\pm}(\kappa\mathbf{\Lambda}\mathbf{A} - \mathbf{I})\ , \tag{4.2.2a}$$

allows for the simple structure

$$\triangle\vec{s}(t) = \mathbf{M}\vec{s}(t) \tag{4.2.2b}$$

of a general linear stability analysis. At each time step t, the collective opinion ("magnetization" in the Ising language) m_t is defined as

$$m_t = \frac{1}{N}\sum_{i=1}^{N} s_t^i \quad \in [-1, 1]. \tag{4.2.3}$$

This system remains stable as long as the largest eigenvalue of $\mathbf{M}$ remains negative. As is well known, by continuously tuning $\mathbf{M}$, systems whose linear stability is controlled by (4.2.2b) can undergo a bifurcation (or phase transition) from a stable fixed point with zero average change of spin to a state where all spin change state to align to each other (a state described by higher-order terms beyond the linear expansion $\mathbf{M}\vec{s}(t)$). The existence of such states has been related to the emergence of financial bubbles (crashes), diagnosed by the existence of transient super-exponential growth (loss) [2, 85]. In the remainder of this article, we will analyze the sub-critial regime of $\mathbf{M}$, but with $\mathbf{M}$ being non-normal. We will show that this non-normal structure gives rise to transient dynamics that induce bubbles and crashes much

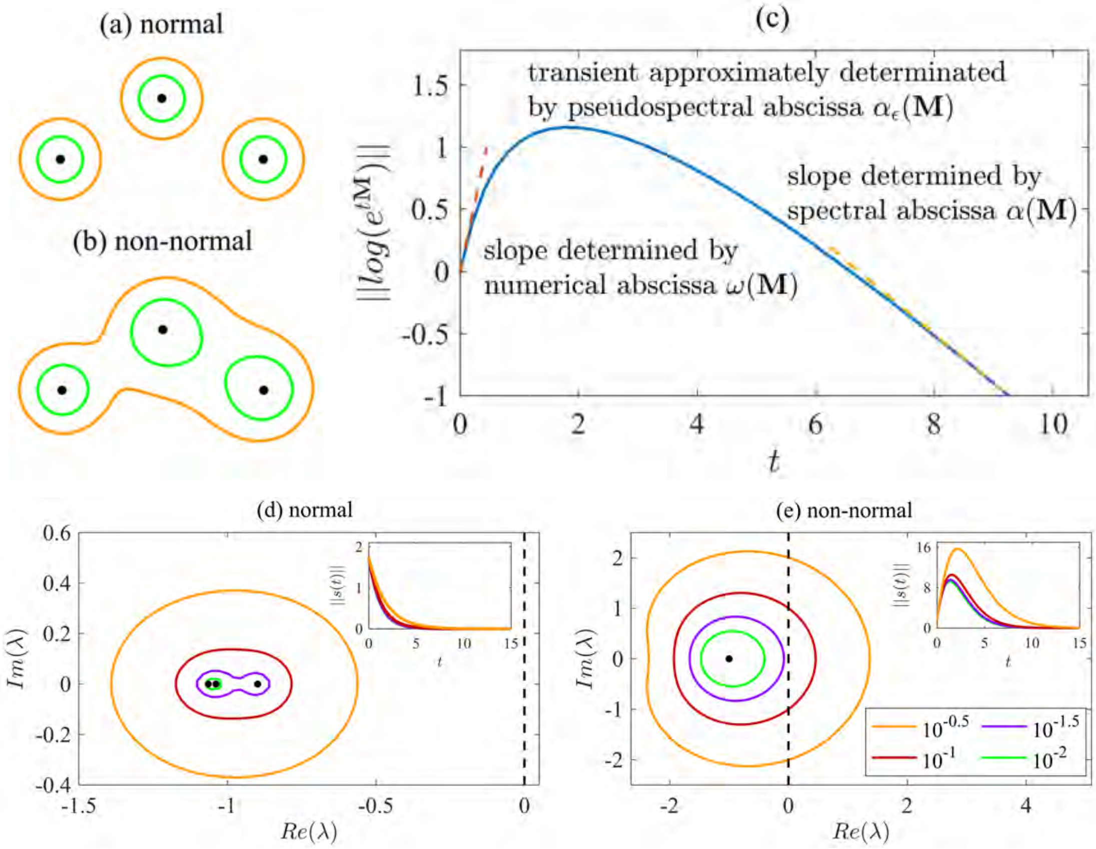

Figure 4.1: (a) and (b) show the geometry of pseudo spectra. In each plot, the contours represent the boundary of $\sigma_\epsilon(\mathbf{M})$ for two values of ϵ. (c) Initial, transient and asymptotic behavior of $||e^{t\mathbf{M}}||$ for a non-normal matrix $\mathbf{M}$. The graph indicates that here $\sigma(\mathbf{M}) < 0$ and hence that the asymptotic behavior is stable. However, the asymptotic behavior is not at all predictive of the transient behavior in case $\mathbf{M}$ is non-normal. Plot (d) shows the eigenvalues (black dots) and some ϵ-pseudospectra (different colors represent different values of ϵ). All eigenvalues as well as the epsilon-spectral lines are confined to the left half plane of $\mathbb{C}$. Accordingly, $||s_t||$ decays exponentially as shown in the inset plot. By contrast, in plot (e), we show a different matrix $\mathbf{M}$. While its eigenvalues are also confined to the left half plane of $\mathbb{C}$, its ϵ spectral lines are not. According to inequality (4.8.27), in the inset plot, we see intermittent transient growth before the asymptotic decay.

like above criticality in normal symmetric networks.

4.3 A Primer on Non-Normality

One of the key properties of non-normal matrices is that their intermediate transient dynamics is significantly different from the long-term asymptotic behavior governed by the largest eigenvalue. Early contributions to the study of non-normal matrices have originated from hydrodynamics, where non-normality plays a role in the emergence of turbulence [75]. Ever since, non-normal matrix theory has helped explain phenomena such as perturbations in ecosystems [106], amplification of neural activities [77], chemical reactions [107] and the formation of Turing patterns [78]. Following a classic textbook [76], we briefly summarize the basic theory behind non-normal matrices (see SI Appendix 2 for details).

Let $\mathbf{M}$ be an $(N \times N)$-matrix. The set of all eigenvalues of $\mathbf{M}$ is called the *spectrum* $\sigma(\mathbf{M})$. A matrix is called *normal* if $\mathbf{M}^T\mathbf{M} = \mathbf{M}\mathbf{M}^T$, and the spectral theorem asserts that each normal $\mathbf{M}$ has a set of n pairwise orthonormal eigenvectors of $\mathbf{M}$. By contrast, if $\mathbf{M}$ is *non-normal*, $\mathbf{M}^T\mathbf{M} \neq \mathbf{M}\mathbf{M}^T$, no such basis exists. Since symmetric matrices are always normal, it is a necessary, but not a sufficient condition that matrix (4.2.2a) represents directed interactions to be considered non-normal.

If λ is an eigenvalue of $\mathbf{M}$, the *resolvent matrix* $\mathbf{M} - \lambda\mathbf{I}$ is not invertible since there exists an eigenvector $\vec{v}$ with $(\mathbf{M} - \lambda\mathbf{I})\vec{v} = 0$. An alternative definition of the spectrum $\sigma(\mathbf{M})$ is thus the set of points $\lambda \in \mathbb{C}$ where the resolvent matrix does not exit. But the question "Does $(\mathbf{M} - \lambda\mathbf{I})^{-1}$ exist?" is binary and may change from "yes" to "no" by just a tiny ϵ-perturbation of λ. In the presence of noise, a better question to ask is whether $\left|\left|(\mathbf{M} - \lambda\mathbf{I})^{-1}\right|\right|$ is large with respect to some matrix norm $||\cdot||$. This leads to the definition of the ϵ-*pseudospectrum*, defined as the set of points where $\left|\left|(\mathbf{M} - \lambda\mathbf{I})^{-1}\right|\right|$ is large (larger then ϵ^{-1}), or formally, $\sigma_\epsilon(\mathbf{M}) \equiv \left\{\lambda \in \mathbb{C} \ : \ \left|\left|(\mathbf{M} - \lambda\mathbf{I})^{-1}\right|\right| > \epsilon^{-1}\right\}$. The ϵ-pseudospectrum is the open subset of the complex plane bounded by the ϵ^{-1} level-curve of the norm of the resolvent. Intuitively, one can then assume that the ϵ-pseudospectrum is closely confined around the eigenvalues of $\mathbf{M}$. For normal matrices, this assumption is correct. However, for non-normal matrices it is not, and $\left|\left|(\mathbf{M} - \lambda\mathbf{I})^{-1}\right|\right|$ may be large even when λ is far away from the spectrum (Figure 4.1 (a) and (b)).

Consider the proportional growth equation $\mathrm{d}\vec{s}/\,\mathrm{d}t = \mathbf{M}\vec{s}$ with explicit solution $\vec{s}(t) = e^{t\mathbf{M}}\vec{s}(0)$. It is well-known that the asymptotic behavior for $t \to \infty$ is governed by the largest real-part of all eigenvalues of $\mathbf{M}$. For the short-term behavior, $t \downarrow 0$, it can be shown that $\frac{\mathrm{d}}{\mathrm{d}t}\left|\left|e^{t\mathbf{M}}\right|\right|\Big|_{t=0} = \omega(\mathbf{M}) \equiv \sup \sigma\left(\frac{1}{2}\left(\mathbf{M} + \mathbf{M}^T\right)\right)$ where $\omega(\mathbf{M})$ is called the *numerical abscissa* of $\mathbf{M}$ (Figure 4.1(c)). Our main interest are, however, intermediate values of t. To describe such transient behavior, one has to consider ϵ-*spectral abscissa* of a matrix $\mathbf{M}$ defined by $\alpha_\epsilon(\mathbf{M}) = \sup \mathrm{Re}\left(\sigma_\epsilon(\mathbf{M})\right)$, i.e. the supremum of the real part of the ϵ-pseudo-spectrum. An important special case is the *spectral abscissa* $\alpha(\mathbf{M}) \equiv \alpha_{\epsilon=0}(\mathbf{M})$, defined as the largest real-part of all eigenvalues of $\mathbf{M}$.

We now consider the case where $\alpha(\mathbf{M}) < 0$, i.e. where the long-term behavior is asymptotically stable (Figure 4.1(c)), but $\alpha_\epsilon(\mathbf{M}) > 0$ for some $\epsilon > 0$. In that case, the pseudospectra

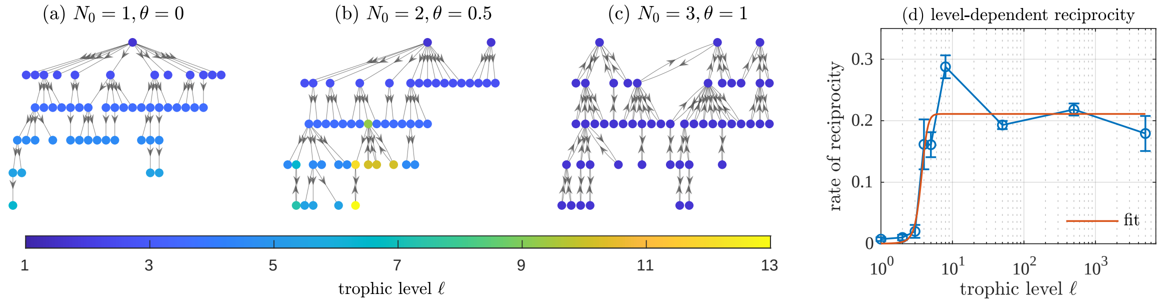

Figure 4.2: (a-c) Three examples of directed networks with different levels of non-normality, different number of top nodes, and different rates of reciprocity θ. The colors indicate the hierarchical (in general non-integer) level ℓ of the nodes. A value of $\theta = 0$ means no edge can be reciprocated, whereas a value of $\theta = 1$ means every edge is reciprocated (since level-dependence is ignored). Network (c) is then normal, since it is symmetric. (d) Empirical analysis of the Blackberry subreddit network. The rate of reciprocity is not constant, but a function of the hierarchical level ℓ. The higher the level, the higher the rate of reciprocity, up to some level of saturation. The red line shows the sigmoid function that best fits the data.

of $\mathbf{M}$ protrude significantly into the right-half plane of $\mathbb{C}$, such that the real-parts of the pseudo-spectrum remain positive (Figure 4.1 (e)). For any such non-normal matrix $\mathbf{M}$ the *Kreiss constant*

$$\mathcal{K}(\mathbf{M}) \equiv \sup_{\epsilon>0} \frac{\alpha_\epsilon(\mathbf{M})}{\epsilon} \tag{4.3.1a}$$

is well-defined, and it can be shown that, for intermediate times (Figure 4.1(c)), there is transient growth according to

$$\sup_{t\geqslant 0} \left|\left|e^{t\mathbf{M}}\right|\right| \geqslant \mathcal{K}(\mathbf{M}). \tag{4.3.1b}$$

Given an interaction matrix $\mathbf{M}$ as in (4.2.2b), we can calculate the Kreiss constant (4.8.26) to obtain lower bounds for the transient growth of net magnetization. An example of such transient growth is shown in the inset Figure 4.1(e). As we shall see, these transients are responsible for socio-economic bubbles. By contrast, if the ϵ-spectrum is confined to the negative half-plane, no transient growth is observed (Figure 4.1(d)).

4.4 Parametrization of Non-Normal Matrices with Level-Dependent Reciprocal Connections

A system such as eqs. (4.8.21) can be interpreted as a dynamical process on a complex network with non-normal $\mathbf{M}$ representing its adjacency matrix. Such *non-normal networks* have been observed in a wide variety of biological and soclo-economic networks [94, 95], and their role in the transmission of noise has been studied [108]. Recall that asymmetry of $\mathbf{M}$ is a necessary, but not a sufficient condition for non-normality. For instance, consider a simple cyclical, directed network of three nodes $\{A, B, C\}$ where $A \to B, B \to C$ and $C \to A$. An adjacency matrix with such cyclical symmetry still gives rise to a normal adjacency matrix. The condition $\mathbf{M}\mathbf{M}^T \neq \mathbf{M}^T\mathbf{M}$ is instead satisfied when the network is directed and hierarchical, which are both intrinsic properties of socio-economic systems [109, 110]. Recently, new methods to generate non-normal networks have been proposed by taking into consideration asymmetrical reciprocity [94] and hierarchy [95] that are typical of non-normal systems. Drawing from these insights, we implement here an algorithm that allows us to control the rate of non-normality along with the number of top nodes, that can be interpreted as thought leaders. In contrast to previous approaches, our rate of reciprocity explicitly depends on the hierarchical level which is a realistic addition as reflected in our empirical analysis below.

The non-normal network with a total of N nodes is initialized with N_0 so-called *top nodes*. These top nodes account for the largely independent N_0 backbones of the communication network common to typical hierarchically, non-normal networks [95]. The remaining $N - N_0$ nodes are added to the existing network sequentially, one node at a time. Each newly added node receives m in-edges, i.e. channels of communication through which it can be influenced. The source of each such edge is selected with probability proportional to the existing nodes' out-degree. As is well-known, this type of preferential attachment creates a skewed degree distribution whereby the network is dominated by a few central nodes [111,112]. Once the m source nodes are determined, each of the m newly formed directed edges may be reciprocated with some independent probability θ. The case $\theta = 1$ recovers a symmetric, i.e. normal, network, whereas levels $\theta \ll 1$ give rise to strongly non-normal systems [94] . Examples of such networks are shown in Figure 4.2 (a-c).

Based on our empirical analysis and to reflect the fact that nodes that are higher up in the hierarchy are harder to be influenced, we assume further that this probability θ is modulated by the hierarchical level ℓ of each node. The lower the node in the hierarchy (the larger ℓ), the more likely the node is reciprocated. Loosely speaking, the hierarchical level ℓ of any node i is defined as the shortest path from a top-node to node i. More precisely, the level ℓ is defined as the trophic hierarchical level [113, 114]. We have analyzed the Reddit discussion forum for the Blackberry meme stock (see Section 4.6 below). An edge is drawn from user i to user j is j replies to a comment of user i. In Figure 4.2(d), we show that the rate of reciprocity is not constant, but an increasing function of ℓ. In other words, the more popular a user's comments, the less likely that user is to reciprocate (comment on) any given edge. For social communication networks, this observation is natural and rationalized as the approximately constant, finite capacity of any given individual to respond to comments. From hereon, we thus model the rate of reciprocity as a sigmoid-function $\frac{\theta}{1+e^{-a(\ell-b)}}$. Parameter θ now has the interpretation of the asymptotic level of reciprocity at high levels ℓ. A value of $\theta = 1$ implies that most, albeit not all edges are reciprocated. Details are found in SI Appendix 3.

Our algorithm to generate non-normal adjacency matrix $\mathbf{A}$ has six parameters: N, N_0, m, θ, a and b. The four parameters m, N, a and b play a subordinate role in the qualitative interpretation of our results. For the remainder of this paper, we thus fix $N = 1000$, $m = 2$, $a = 2.552$, and $b = 3.668$ unless mentioned otherwise. The parameters θ and N_0, on the other hand, have qualitatively important implications on the behavior of our model. The parameter θ characterizes the hierarchical nature of the system. The smaller θ, the more directed the network, and the less the top nodes may be influenced. The parameter N_0 denotes the number of top nodes. If θ is small, then N_0 may be interpreted as the number of (largely) independent, leading opinions in the system. The price dynamics from eqs. (4.8.21) is not directly governed by $\mathbf{A}$, but rather by the related matrix $\mathbf{M}$. The two parameters κ and $p^{\pm}$ allow us to control the characteristics of $\mathbf{M}$ for given $\mathbf{A}$. In the remainder of this article, we fix $p^{\pm} = 0.05$ and we tune κ. This leaves us with a three-parameter model, N_0, θ and κ. Importantly for what follows, the Kreiss constant $\mathcal{K}$ is strictly decreasing in θ and increasing in κ, irrespective of N_0, as long as $N_0 \ll N$. Throughout this article, we constrain the parameter such that $\alpha_0(\mathbf{M}) < 0$, i.e. the asymptotic system dynamics is stable.

4.5 Transient Bubbles Induced by Non-Normal Interactions

Building on the three previous sections, we now run agent based simulations with non-normal adjacency matrix $\mathbf{M}$. While it has been well-established that Ising-like agent-based models with non-zero net opinion (non-zero magnetization) are responsible for the formation of bubbles [82,85,115], we investigate here the regime where the net magnetization (net opinion) fluctuates around zero (sub-critical phase). In the following analysis, we therefore set the coupling strength κ to a sub-critical value. Figure 4.3(a,b) confirms that m_t fluctuates around zero approximately symmetrically, as expected from the fact that the imitation strength κ has been chosen so that the underlying Ising model is subcritical. Furthermore, for fixed parameters (N_0, κ), we compare two types of social networks: $\theta = 0$ and $\theta = 1$. The former corresponds to a case of small reciprocity, and hence large non-normality of $\mathbf{A}$ and hence $\mathbf{M}$. The later corresponds to an almost symmetric - and hence normal - matrix $\mathbf{A}$, which coincides with a much less non-normal matrix $\mathbf{M}$ (see also SI Appendix 4). Comparing Figure 4.3 (a) and (b), we see that the strongly non-normal case (low θ) corresponds to much more pronounced long-lived deviations of the magnetization from its zero average, as is expected from transient dynamics (Figure 4.1(e)).

In the right most column of Figure 4.3, we additionally show the associated price-time

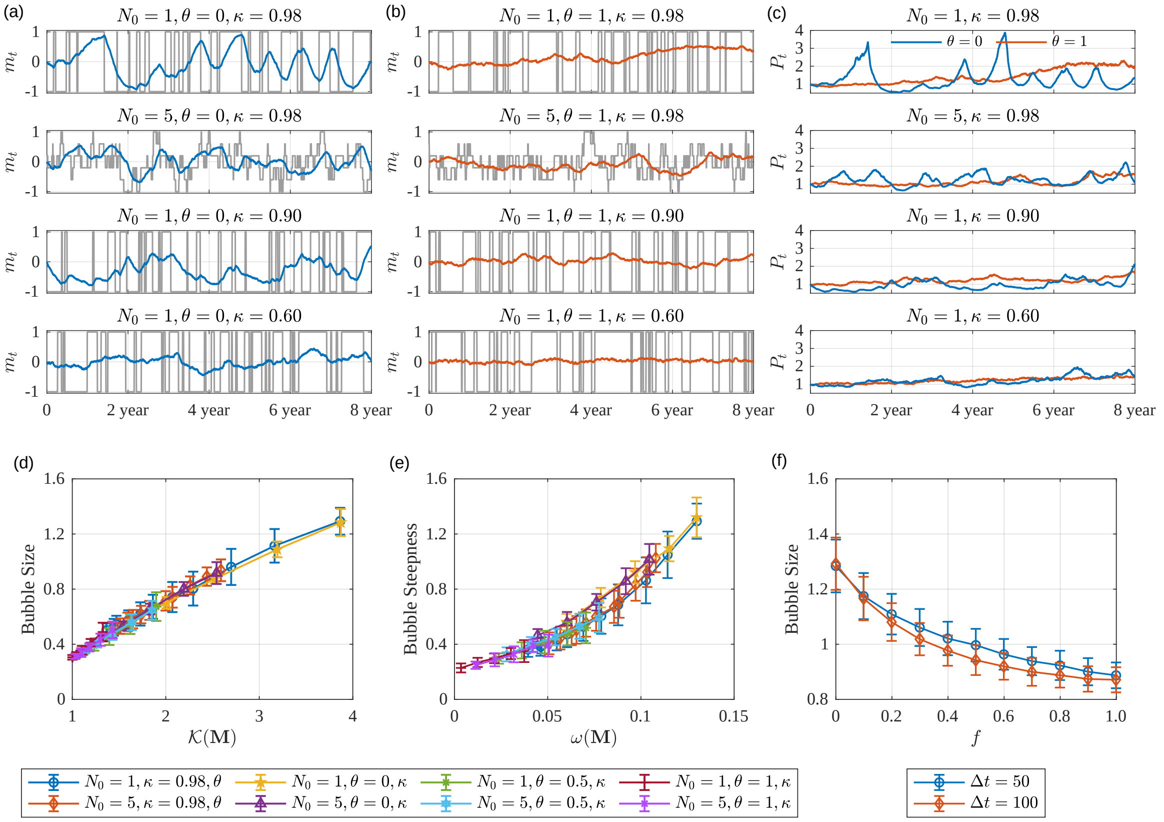

Figure 4.3: (a) Magnetization (4.2.3) for ABM simulation with non-normal interaction matrix $\mathbf{M}$ for different parameter constellations. The net magnetization of the N_0 opinion leaders is shown in grey lines in the background. (b) Same as in (a), but for a symmetrized interaction matrix $\frac{1}{2}\left(\mathbf{M} + \mathbf{M}^T\right)$. The parameter κ is chosen below a sub-critical value, hence the net magnetization is, on average, 0. In contrast to (a), the transients are less pronounced. (c) Price trajectory generated by dynamics with magnetization from (a) and (b), respectively. Only the non-normal matrices induce bubbles. (d) Bubble size as a function of Kreiss constant for different parameter constellations. (e) Bubble steepness as a function of numerical abscissa for different parameter constellations. (f) Bubble size as a function of fraction of nodes receptive to antagonistic opinion.

series P_t obtained as the Walras' equilibrium between fundamentalists and noise traders (SI Appendix 1). It is striking to observe the drastic differences in the price dynamics between highly non-reciprocal (non-normal) interactions compared to reciprocated (normal) ones. In the former, very strong price peaks are preceded by periods of strong price growth, following by fast large asymmetric drawdowns. This qualifies the existence of large amplitude bubbles as a clear diagnostic of this type of non-normal networks. In contrast, for normal networks (and weakly non-normal $\mathbf{M}$ matrices), the price dynamics appears compatible with a standard geometric Brownian motion at least at the qualitative level. As we now show, both of these phenomena are explained as a function of the transients induced by non-normality.

A hallmark of a financial bubble is the existence of unsustainable super-exponential price growth [85, 116, 117]. Within our ABM, it can be shown [96] that the price grows, to first approximation, exponentially as a function of the net magnetization, $P_t = Ce^{cm_t}$, where the scaling coefficient $c > 0$ is a function of the model parameters. We recall that m_t is defined as the average state across all trader states $\vec{s}(t)$. The states $\vec{s}(t)$ are themselves governed by equation (4.2.2b) involving the non-normal interaction matrix $\mathbf{M}$, such that $\vec{s}(t) \sim e^{\mathbf{M}t}\vec{s}(0)$. In a globally stable regime all eigenvalues of $\mathbf{M}$ associated with the stable equilibrium $\vec{s}(t) = 0$ are negative (SI Appendix 4). The standard expectation is thus that m_t remains small and thus P_t should not exhibit abnormal fluctuations. But this is forgetting the transients induced by the non-normality of $\mathbf{M}$. Indeed, as discussed in Section 4.3 (cf. in particular Figure 4.1(c,e)), the asymptotically stable fixed-point $\vec{s} = 0$ is punctuated by repelling dynamics over finite time scales. Furthermore, inequality (4.8.27) provides us with a lower bound of the size of these transients, which are mainly a function of the Kreiss constant $\mathcal{K}(\mathbf{M})$. Combining $P_t = Ce^{cm_t}$ with transient approximately exponential growth of m_t, we thus predict the occurrence of finite lived bubbles qualified as transient super-exponential growth of price. The above considerations lead us to hypothesize that the size of bubbles in the price realizations of our agent-based model are directly proportional to the Kreiss constant $\mathcal{K}(\mathbf{M})$. Moreover, the dependence of the size of these bubbles on parameters (N_0, θ, κ) should only appears through the dependence of bubble sizes on $\mathcal{K}(\mathbf{M})$. To test this hypothesis, we measure the size of the bubbles as the difference in price between the beginning and the end of a regime of super-exponential growth (see SI Appendix 5 for details). For different parameter combinations of (N_0, θ, κ), we generate 100 price simulations according to the following procedure. We first generate a matrix $\mathbf{M}$ (see Section 4.4), and we subsequently simulate a time-series with 25,000 time-steps, corresponding to 100 years (considering 250 trading days per calendar year). On each time series, we measure the size of all bubbles. These sizes are subsequently averaged across all N simulations, with the standard deviation serving as error bars. Figure 4.3(d) demonstrates the existence of a remarkable collapse of all curves when the average bubble sizes are plotted as a function of the Kreiss constant $\mathcal{K}(\mathbf{M})$. A large Kreiss constant is associated with large bubble sizes, even for different network non-normality and social coupling κ. The key insight is that different network parametrization indeed all collapse onto this scaling between Kreiss constant and bubble size.

The theory of transients does not only make a prediction about the size of the transients, but also about their steepness. As visualized in Figure 4.1(c), we expect the steepness of the transients, and therefore of the magnetization and then price, to be increasing in the numerical abscissa $\omega(\mathbf{M})$. Figure 4.3(e) confirms that a large numerical abscissa $\omega(\mathbf{M})$ is indeed associated with large steepness of bubbles.

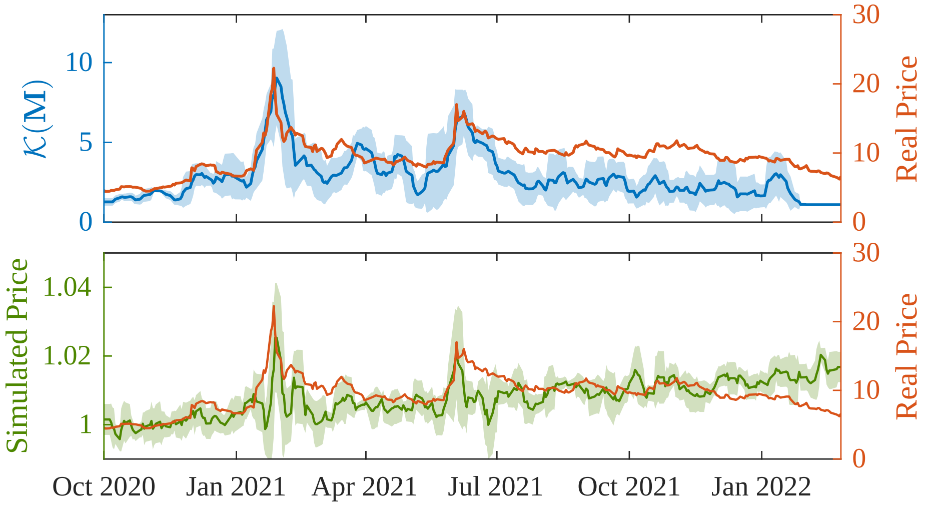

Figure 4.4: (top) Co-evolution of Reddit network Kreiss constant and Blackberry meme stock price. (bottom) ABM price time-series resulting from the simulated (net-zero) magnetization m_t, with distinct peak around the same time as the real price trajectory.

Finally, we test the effect of an influential contrary opinion on the size of the bubble. We first simulate a price dynamics with a single top node, $N_0 = 1$, as well as $N = 1000$, $m = 2$, $N_0 = 1$ and $\theta = 0$. Upon formation of a bubble (continuous price increase for 50 time-steps), we interfere into the system by holding a contrary opinion (opposite state of N_0) for Δt time-steps. That contrary node is connected to a fraction f of all nodes with the exception of the top node. Figure 4.3(f) shows the subsequent decrease of the bubble. The larger f, the larger the effect. This result is encouraging, suggesting that an external controller can counter-balance sub-critical bubbles by rendering the system less non-normal. On the other hand, this approach requires one to have large scale influence. A more scalable approach, that we leave for future research, would be a minimal influence of a few key nodes in order to achieve overall noise-cancellation, as has been recently shown in communication networks [108], and which could improve on the more standard market intervention involving large balance sheet build-up of major financial agents such as a central bank [118].

4.6 Non-Normal Communication in Meme Stock Trading

So far, our assessment of bubbles has relied on agent-based simulations, where we can control the experimental conditions. The difficulty with empirical data is that, in general, one cannot observe the matrix $\mathbf{A}$ that governs trader interactions. On social trading platforms, such as eToro, interactions can be measured precisely [112], but the trading volume relative to the entire market is small, such that influence on the price is, however, negligible. This is not the case for so-called meme stocks which have enjoyed recent popularity. Driven primarily by retail traders, meme stock trading activity has been shown to be largely influenced by Reddit discussion forums [119–121]. Reddit is organized into *subreddits* on which specific topics are being discussed. Users interact by submitting new posts and adding comments to existing posts or comments. Here, we analyze the posts and comments related to four popular meme stocks (Blackberry, Nokia, GameStop and AMC) under the famous subreddit *r/wallstreetbets* (also known as WallStreetBets or WSB) that has become notable for its colorful and profane jargon, aggressive trading strategies, and for playing a major role in the GameStop short squeeze in early 2021. For each of the four stock, at time t, we draw a directed edge from user J to user K if K has commented or replied a stock-related text by user J in the time interval $[t - \Delta t, t]$. In other words, K has been influenced by J's action in the past Δt days. With this procedure, for each meme stock, we extract a dynamically evolving influence network $\mathbf{A}(t)$, of which we can measure the Kreiss constant $\mathcal{K}(t)$. The evolution of the Kreiss constant, along with the trading price, is shown in the top plot of Figure 4.4 for the Blackberry stock (see SI Appendix 6 for similar plots for the other three meme stocks). The two most prominent price peaks around January 2021 and June 2021 coincide with the two largest peaks of the Kreiss constant trajectory. This gives force to our proposition that increased non-normality (quantified by large values of the Kreiss constant) favours the occurrence of transient explosive price behavior (bubbles) associated with the transient growth of perturbations before their relaxation. In other words, we interpret the presence of financial bubbles in these meme stocks as reflecting at least partially the asymmetric hierarchical structure of the reddit discussion forum that induced a polarized bullish opinion among retail traders, which then in-turn pushed the price up. And indeed, the mostly mentioned words in Jan 2021 among the reddit submissions and comments related to BlackBerry are *rocket*, *bb*, *gme*, *shares*, and *buy*. In particular, among the 58,793 mentions of *rocket* and 12,132 mentions of *buy* in 2021, 69% of *rocket* and 43% of *buy* were in January. More generally, as is shown in SI Appendix 6, we do find a positive correlation between the Kreiss constant and price bubbles across meme stocks. This suggests that the non-normal structure of the Reddit meme stock discussion forum is an important driver of the observed price instabilities.

A strong asymmetric hierarchical structure of the reddit discussion forum quantified by a large value of the Kreiss constant provides a powerful catalysis for the emergence of transient price bubbles. But of course, as for general dynamics following non-normal operators, not all perturbations go through a non-monotonous transient amplification. The realized trajectory of the transient very much depends on the projection of the perturbations onto the pseudo-eigenvectors [76,78,122]. Not all large Kreiss constant values should thus lead to a bubble, as the market dynamics is more complex and cannot just be reduced to one source of influence.

The mapping between large Kreiss constant and bubbles becomes rigorous when considered in terms of ensembles of price trajectories. To show this, we insert the empirical Blackberry discussion forum influence network $\mathbf{A}(t)$ as input to our ABM. We simulate the resulting price time-series 100 times and keep track of the average price as well as its average standard deviation (see SI Appendix 6 for details). In the bottom plot of Figure 4.4, one can observe indeed that price spikes coincide - in their ensemble average - with peaks of the Kreiss constant, supporting our hypothesis that the non-normality in the Reddit discussion forum contributes to explain the observed price bubbles.

These simulations exemplify the possibility to diagnose regimes of financial instabilities by measuring the evolution of the Kreiss constant of the underlying network of social interactions between traders. Periods in which the Kreiss constant is large should be interpreted as prone to bubble regimes and large price volatility. The mapping of the detection of (financial) instabilities to the measurement of the Kreiss constant improved conceptually and operationally on the previous approaches attempting to anticipate critical phase transitions [89, 112, 123–125], which do not incorporate the ubiquitous non-normality of complex system dynamics.

4.7 Conclusions

Until now, financial and socio-economic bubbles have been thought of as being associated with special regimes where self-reinforcing interactions strengthen transiently towards a critical point and lead to some form of collective exuberance. This has been formalized by models in physics, ecology and mathematics assuming the presence of an underlying phase transition, criticality, bifurcation or catastrophe. The conceptual breakthrough of the present work is to demonstrate formally, via agent-based model simulations and empirically, that such transient phases of exuberance are generic in real social systems ubiquitously characterized by non-normal properties of asymmetric hierarchical interactions. An important corollary is that financial bubbles should be expected as intrinsic, rather than abnormal monsters appearing in very special conditions. Due to the broad applicability of models involving hierarchical, Ising-like interactions, our framework is expected to explain crowd-forming patterns and collective structures in general hierarchical complex networks, ranging from from biological to artificial intelligent computer systems.

4.8 Appendix

4.8.1 Agent Based Price Simulation

In this section, we explain in detail the agent-based price simulations that we employ. Following refs. [96, 97], it consists of fundamentalists ($\mathcal{F}$) and noise traders ($\mathcal{N}$) who trade a risky and a risk-free asset. The risky asset is a dividend paying stock. The risk-free asset pays a constant return in each time-step and represents a bank account or risk-free government bond. Fundamentalists are rational risk-averse investors who invest by maximizing their expected utility under a constant relative risk aversion (CRRA) utility function in each time-step. The noise traders invest based on social imitation. Compared to previous models where all noise traders interacted with each other, we consider here noise traders that are positioned on a non-normal network. The noise traders are only influenced by their in-neighbours.

4.8.1.1 Assets

The investment universe of the traders consists of two assets, which are equivalent to the set-up in ref. westphal2020market. The risk-free asset pays a constant return r_f in each time-step. It represents a bank account or bond with perfectly elastic supply. The risky asset represents an index fund or stock. It pays a dividend d_t in each time-step and its price P_t is defined endogenously by demand and supply. The dividend process follows a discrete stochastic growth process defined as

$$d_t = d_{t-1}\left(1 + r_t^d\right) \ . \tag{4.8.1}$$

The growth rate $r_t^d = r_d + \sigma_d u_t$ is a Gaussian process with mean value $r_d > 0$, variance σ_d^2, and stochastic increments $u_t \overset{iid}{\sim} \mathcal{N}(0,1)$. The excess return $r_{excess,t}$ of the risky asset with price P_t describes the compensation for holding the risky asset instead of the risk-free asset between $t-1$ and t. It is the sum of the capital return $r_t = \frac{P_t}{P_{t-1}} - 1$ and the return from the dividend d_t minus the risk-free rate r_f:

$$r_{excess,t} = r_t + \frac{d_{t-1} \cdot (1 + r_t^d)}{P_{t-1}} - r_f. \tag{4.8.2}$$

4.8.1.2 Fundamentalists

Fundamentalists are risk-averse investors, endowed with a constant relative risk aversion (CRRA) utility function

$$U(W) = \begin{cases} \log(W) & \text{for } \gamma = 1 \\ \frac{W^{1-\gamma}}{1-\gamma} & \text{for } \gamma \neq 1. \end{cases} \tag{4.8.3}$$

They allocate their wealth among the two assets in order to maximise their expected utility in each time-step.

Each fundamentalist i decides on a fraction $x_t^{\mathcal{F}^i}$ of his wealth $W_t^{\mathcal{F}^i}$ to invest into the risky asset based on the maximization problem

$$\max_{x_t^{\mathcal{F}^i}} \mathbb{E}_{t-1}\left[U(W_t^{\mathcal{F}^i})\right]. \tag{4.8.4}$$

In that sense, the fundamentalists are myopic, as they only consider expected returns one period ahead. The resulting optimal risky fraction $x_t^{\mathcal{F}}$ is identical for each fundamentalist, because they are endowed with the same utility function and have access to the same information. Consequently, the investment of the individual fundamentalists can be considered at the aggregate level as a representative trader investing the risky fraction $x_t^{\mathcal{F}}$ of his wealth $W_t^{\mathcal{F}} = \sum_i W_t^{\mathcal{F}^i}$.

The fundamentalist's wealth at time t can be expressed iteratively as a function of the invested risky fraction, the return on the risky asset, the dividend payment, and the risk-free rate

$$W_t^{\mathcal{F}} = W_{t-1}^{\mathcal{F}} \cdot \left(x_t^{\mathcal{F}} \cdot r_{excess,t} + 1 + r_f\right) = W_{t-1}^{\mathcal{F}} \cdot \left(x_t^{\mathcal{F}} \cdot \left(1 + r_t + \frac{d_t}{P_{t-1}}\right) + (1 - x_t^{\mathcal{F}}) \cdot (1 + r_f)\right). \tag{4.8.5}$$

As derived by [96], the resulting risky fraction $x_t^{\mathcal{F}}$ solving the expected utility maximisation, given in (4.8.4) for the CRRA utility in (4.8.3) is in first order approximation and assuming $d_t \ll P_t$

$$x_{t-1}^{\mathcal{F}} = \frac{1}{\gamma} \frac{E_{t-1}[r_{excess,t}]}{Var_{t-1}[r_{excess,t}]} = \frac{E_{r_t} + \frac{d_{t-1}}{P_{t-1}}(1 + r_d) - r_f}{\gamma(\sigma^2 + \frac{d_{t-1}^2 \cdot \sigma_r^2}{P_{t-1}^2})} \approx \frac{E_{r_t} + \frac{d_{t-1}}{P_{t-1}}(1 + r_d) - r_f}{\gamma \sigma^2} \tag{4.8.6}$$

where E_{r_t} is the expected return of the risky asset and σ^2 is the expected variance. Using the wealth evolution given in (4.8.5) and denoting the number of shares invested in the risky asset by $n_t^{\mathcal{F}} := \frac{x_t^{\mathcal{F}} W_t^{\mathcal{F}}}{P_t}$, the excess demand of the fundamentalist, which is the net money value of the risky asset that the investor wants to buy or sell, is

$$\begin{aligned}\Delta D_{t-1\to t} :=& n_t^{\mathcal{F}} P_t - n_{t-1}^{\mathcal{F}} P_t \\ =& W_{t-1}^{\mathcal{F}} \left(x_t^{\mathcal{F}} \left[1 + r_f + x_{t-1}^{\mathcal{F}} \left(r_t + \frac{d_t}{P_{t-1}} - r_f \right) \right] - x_{t-1}^{\mathcal{F}} \frac{P_t}{P_{t-1}} \right).\end{aligned} \tag{4.8.7}$$

4.8.1.3 Noise Traders

The noise traders' investment strategy is based on social influence. The traders are nodes in a non-normal network and connected through directed edges. Each noise trader is only influenced by its in-neighbours. The set-up is a modification of the noise trader class presented in kaizoji2015super,westphal2020market. The noise traders are described by an Ising-like structure, in which each individual noise trader is either invested in the risky asset or in the risk-free asset. Each of them switches his position to the other asset with a transition probability that depends on the opinion of their in-neighbours. [1]

The opinion of noise traders is implemented as an Ising-like model, in which each node i is considered to be in one of the two possible states $+1$ (the noise trader holds the risky asset), and -1 (the noise trader holds the risk-free asset) denoted as $s^i = \pm 1$. At $t = 0$, one half of randomly selected nodes are initialized in state $+1$, the other half are in state -1. When $t > 0$, the probability of node i to switch its state is given by

$$\pi(s_{t+1}^i = -s_t^i) = \frac{p^{\pm}}{2}(1 - \kappa \frac{1}{k_i^{in}} s_t^i \sum_j a_{ij} s_t^j) \tag{4.8.8}$$

where $p^{\pm}$ controls the average holding time of each asset and the social coupling strength κ determines the noise traders' susceptibility to social imitation. At each time step t, the collective opinion ("magnetization" in the Ising language) m_t is defined as

$$m_t = \frac{1}{N} \sum_{i=1}^{N} s_t^i \quad \in [-1, 1], \tag{4.8.9}$$

Aggregating the independent investment decisions over all noise traders amounts to an equivalent representative noise trader who decides on the fraction $x_t^{\mathcal{N}}$ of his wealth to invest in the risky asset, which is given by

$$x_t^{\mathcal{N}} = \frac{1}{2} + \frac{1}{2N} \sum_{i=1}^{N} s_t^i = \frac{1}{2} + \frac{1}{2} m_t \qquad \in [0, 1]. \tag{4.8.10}$$

[1] Following ref. [96], we do not consider a link between noise traders and the representative fundamental trader.

The noise traders' aggregate wealth evolves equivalently to (4.8.5) as a function of the wealth at the previous time-step, the invested risky fraction, the return on the risky asset, the dividend payment, and the risk-free rate

$$W_t^{\mathcal{N}} = W_{t-1}^{\mathcal{N}} \cdot \left(x_t^{\mathcal{N}} \cdot r_{excess,t} + 1 + r_f\right) = W_{t-1}^{\mathcal{N}} \cdot \left(x_t^{\mathcal{N}} \cdot \left(1 + r_t + \frac{d_t}{P_{t-1}}\right) + (1 - x_t^{\mathcal{N}}) \cdot (1 + r_f)\right). \tag{4.8.11}$$

The resulting aggregated excess demand of the noise traders for the risky asset is described by

$$\Delta D_{t-1\rightarrow t}^{\mathcal{N}} = W_{t-1}^{\mathcal{N}} \left(x_t^{\mathcal{N}} \left[x_{t-1}^{\mathcal{N}} \left(r_t + \frac{d_t}{P_{t-1}} - r_f \right) + r_f + 1 \right] - x_{t-1}^{\mathcal{N}} \frac{P_t}{P_{t-1}} \right). \tag{4.8.12}$$

4.8.1.4 Equilibrium Market Price

Following Walras' theory of general equilibrium [126], the market clearing condition requires an equilibrium between total supply and total demand at each time step. Each trader formulates their excess demand for the next time step and the price is calculated as the equilibrium in which supply equals demand. This is formulated as:

$$\Delta D_{t-1\rightarrow t}^{\mathcal{F}} + \Delta D_{t-1\rightarrow t}^{\mathcal{N}} = 0, \tag{4.8.13}$$

where $\Delta D_{t-1\rightarrow t}^{\mathcal{F}}$ and $\Delta D_{t-1\rightarrow t}^{\mathcal{N}}$ are aggregated excess demands from the fundamentalists and the noise traders respectively for the risky asset. Using the fundamentalists excess demand (4.8.7) with the risky fraction (4.8.6) and the noise traders excess demand (4.8.12), the market clearing condition (4.8.13) can be reformulated as a function of the price P_t. This results in the following equation determining the price

$$a_t P_t^2 + b_t P_t + c_t = 0, \tag{4.8.14}$$

where a_t, b_t and c_t are given by:

$$a_t = \frac{1}{P_{t-1}} [W_{t-1}^{\mathcal{N}} x_{t-1}^{\mathcal{N}} (x_t^{\mathcal{N}} - 1) + W_{t-1}^{\mathcal{F}} x_{t-1}^{\mathcal{F}} (\frac{E_{r_t} - r_f}{\gamma \sigma^2} - 1)] \tag{4.8.15}$$

$$b_t = \frac{W_{t-1}^{\mathcal{F}}}{\gamma \sigma^2} \{x_{t-1}^{\mathcal{F}} \frac{d_t(1 + r_d)}{P_{t-1}} + (E_{r_t} - r_f)[x_{t-1}^{\mathcal{F}} (\frac{d_t}{P_{t-1}} - r_f) + r_f]\} + W_{t-1}^{\mathcal{N}} x_t^{\mathcal{N}} [x_{t-1}^{\mathcal{N}} (\frac{d_t}{P_{t-1}} - 1 - r_f) + r_f] \tag{4.8.16}$$

$$c_t = W_{t-1}^{\mathcal{F}} \frac{d_t(1 + r_d)}{\gamma \sigma^2} [x_{t-1}^{\mathcal{F}} (\frac{d_t}{P_{t-1}} - r_f) + r_f] \tag{4.8.17}$$

Then, at time step t, the unique positive solution to (4.8.14) yields the trade price.

4.8.1.5 State Switching Dynamics

The N noise traders are organized on a network with adjacency matrix $\mathbf{A} = a_{ij}$, where $a_{ij} = 1$ if there exists a directed edge from node j to node i, and $a_{ij} = 0$ otherwise. The in-degree of node i is the number of directed edges pointing to node i, which is given by

$$k_i^{in} = \sum_{j=1}^{N} a_{ij}. \tag{4.8.18}$$

The transition probability ρ that trader i flips its state s_t^i at time t depends on the opinion of their in-neighbours, according to In the Ising model, if node i switches its state from time step t to time step $t+1$, i.e. $s_{t+1}^i = -s_t^i$, the change of the value of node i's state is $-2s_t^i$. Given the probability (4.8.8) of node i to switch its state, the average rate of change of the spin starting in the state ^i_t is given by

$$\triangle s_t^i = -2s_t^i \pi(s_{t+1}^i = -s_t^i) = -p^{\pm} \left(s_t^i - \kappa \frac{1}{k_i^{in}} \sum_j a_{ij} s_t^j \right) . \tag{4.8.19}$$

We introduce the n-dimensional state vector $\vec{s}(t) = (s_t^1, s_t^2, ..., s_t^N)$. We then rewrite (4.8.19) in vectorial form as

$$\triangle \vec{s}(t) = \vec{s}(t+1) - \vec{s}(t) = p^{\pm}(\kappa \mathbf{\Lambda} \mathbf{A} - \mathbf{I})\vec{s}(t) \tag{4.8.20}$$

where $\mathbf{\Lambda}$ is an $N \times N$ diagonal matrix with $1 \big/ k_i^{in}$ on the i-th diagonal entry and $\mathbf{I}$ is the identity matrix. Introducing the matrix

$$\mathbf{M} \equiv p^{\pm}(\kappa \mathbf{\Lambda} \mathbf{A} - \mathbf{I}) \ , \tag{4.8.21a}$$

equation (4.8.20) exhibits the simple structure

$$\triangle \vec{s}(t) = \mathbf{M} \vec{s}(t) \tag{4.8.21b}$$

of a general linear stability analysis. This system remains stable as long as the largest eigenvalue of $\mathbf{M}$ remains negative. As is well known, by continuously tuning $\mathbf{M}$, systems whose linear stability is controlled by (4.8.21b) can undergo a bifurcation (or phase transition) from a stable fixed point with zero average change of spin to a state where all spin change state to align to each other (a state described by higher-order terms beyond the linear expansion

$\mathbf{M}\vec{s}(t)$). The existence of such states has been related to the emergence of financial bubbles (crashes), diagnosed by the existence of transient super-exponential growth (loss). In the remainder of this article, we will analyze the sub-critial regime of $\mathbf{M}$, but with $\mathbf{M}$ being non-normal. We will show that this non-normal structure gives rise to transient dynamics that induce bubbles and crashes much like above criticality in normal symmetric networks.

sub

4.8.2 Parameter values

The parameters of the agent-based models are summarized in Table 4.1.

Parameter name	Explanation	Value
Assets		
r_f	Risk free interest rate	0.00004
d_0	Initial dividend	0.00016
r_d	Average growth rate of the dividend	0.00016
σ_d	Standard deviation of the dividend growth rate	0.000016
P_0	Initial price of the risky	1
Noise Traders		
$x_0^{\mathcal{N}}$	Initial fraction of the risky asset held by the noise traders	0.5
$W_0^{\mathcal{N}}$	Initial wealth of the noise traders	10^9
Fundamentalists		
$x_0^{\mathcal{F}}$	Initial fraction of the risky asset held by the noise traders	0.3
$W_0^{\mathcal{F}}$	Initial wealth of the fundamentalists	10^9
E_{r_t}	Expected return of the risky asset	0.00016
σ_r	Expected standard deviation of the risky asset price	0.02

Table 4.1: Parameter values for the agent-based-model of a financial market with traders on a network.

4.8.3 Theory of Non-Normal Matrices

In this section, we follow a classic textbook [76] to summarize the most important, basic concepts about non-normal matrices.

4.8.3.1 Definition of Non-Normal Matrices

Consider a linear operator $\mathbf{M}$. From hereon, we assume that $\mathbf{M}$ is a finite-dimensional, real $N \times N$ matrix. [2] We call a vector v an eigenvector of $\mathbf{M}$ when $\mathbf{M}v = \lambda v$ for some corresponding eigenvalue $\lambda \in \mathbb{C}$. The set of all eigenvalues of $\mathbf{M}$ is called the *spectrum* $\sigma(\mathbf{M})$ of $\mathbf{M}$.

By definition, a matrix is called *normal* if $\mathbf{M}^T\mathbf{M} = \mathbf{M}\mathbf{M}^T$. The *spectral theorem* asserts that $\mathbf{M}$ *is normal if and only if* $\mathbf{M}$ *is diagonalizable by a unitary matrix, if and only if, there exists a set of N eigenvectors of* $\mathbf{M}$ *that form an orthonormal basis for* $\mathbb{C}^N$*.* By contrast, if $\mathbf{M}$ is *non-normal*, i.e. $\mathbf{M}^T\mathbf{M} \neq \mathbf{M}\mathbf{M}^T$, then no such basis exist. As we shall see below, it is exactly this non-orthogonality that is responsible for the transient dynamics surmised to induce bubbles. Note that symmetric matrices ($\mathbf{M} = \mathbf{M}^T$) are always normal. Therefore, all undirected Ising-like interactions are normal by definition. In other words, in order for $\mathbf{M}$ in (4.8.21b) to be non-normal, it is a necessary, but not a sufficient condition that the matrix $\mathbf{M}$ represents directed interactions.

4.8.3.2 Pseudospectra

Assume λ is an eigenvalue of the (invertible) matrix $\mathbf{M}$ with eigenvector v and consider the *resolvent matrix* $\mathbf{M} - \lambda\mathbf{I}$. The matrix $\mathbf{M} - \lambda\mathbf{I}$ does not have full rank, since there exists a vector v which is mapped to zero, $(\mathbf{M} - \lambda\mathbf{I})v = 0$. Because $(\mathbf{M} - \lambda\mathbf{I})$ does not have full rank, it is not invertible and $(\mathbf{M} - \lambda\mathbf{I})^{-1}$ does not exist. Therefore, one can also define the spectrum $\sigma(\mathbf{M})$ as the set of points $\lambda \in \mathbb{C}$ where the resolvent matrix $(\mathbf{M} - \lambda\mathbf{I})^{-1}$ does not exit.

The answer to the question "does $(\mathbf{M} - \lambda\mathbf{I})^{-1}$ exist" is binary and may change from "yes" to "no" by just a tiny ϵ-perturbation of λ. In the presence of noise, a better question to ask is therefore: "is $\left|\left|(\mathbf{M} - \lambda\mathbf{I})^{-1}\right|\right|$ large? Here $||\cdot||$ is some matrix norm. This leads to the definition of the ϵ-*pseudospectrum*, defined as the set of points where $\left|\left|(\mathbf{M} - \lambda\mathbf{I})^{-1}\right|\right|$ is large (larger then ϵ^{-1}). In mathematical terms,

$$\sigma_\epsilon(\mathbf{M}) \equiv \left\{\lambda \in \mathbb{C} \ : \ \left|\left|(\mathbf{M} - \lambda\mathbf{I})^{-1}\right|\right| > \epsilon^{-1}\right\}. \tag{4.8.22}$$

The ϵ-pseudospectrum is the open subset of the complex plane bounded by the ϵ^{-1} level-curve of the norm of the resolvent. Intuitively, one can then assume that the ϵ-pseudospectrum is closely confined around the eigenvalues of $\mathbf{M}$. For normal matrices, this assumption is

[2] Since we consider only real matrices, $\mathbf{M} \in \mathbb{R}^{N\times N}$, it holds that $\mathbf{M}^* = \mathbf{M}^T$ where $\mathbf{M}^*$ is the Hermitian conjugate of $\mathbf{M}$. Therefore, we will always write $\mathbf{M}^T$ and assume $\mathbf{M}$ is real even if we could write $\mathbf{M}^*$ and assume $\mathbf{M}$ is complex for more generality. See ref. [76] for a generalization s to complex matrices or to infinite dimensional vector spaces.

correct. However, for non-normal matrices it is not, and $\left|\left|(\mathbf{M}-\lambda\mathbf{I})^{-1}\right|\right|$ may be large even when λ is far away from the spectrum. The concept of the ϵ-spectrum provides an appealing geometric interpretation of non-normality. One can get a good start in predicting their behavior if, in additional to the calculation of eigenvalues, one plots a few contour lines of the ϵ-pseudospectrum. And as we shall see next, pseudo spectra are useful to describe transient growth phenomena.

Before we move the next section, we need to define one more quantity: The *ϵ-spectral abscissa* of a matrix $\mathbf{M}$ is defined as

$$\alpha_\epsilon(\mathbf{M}) = \sup \mathrm{Re}\left(\sigma_\epsilon(\mathbf{M})\right), \tag{4.8.23}$$

i.e. $\alpha_\epsilon(\mathbf{M})$ is the supremum of the real part of the ϵ-pseudo-spectrum. Here, we have to use the supremum rather than the maximum since (4.8.22) is in general an open set. However, the spectrum $\sigma(\mathbf{M})$ is a closed set, given by the set of eigenvalues. Therefore, an important special case is the *spectral abscissa* $\alpha(\mathbf{M}) = \alpha_0(\mathbf{M})$, defined as the largest real-part of all eigenvalues of $\mathbf{M}$.

We are interested in cases in which the transient behavior of this system differs from the behavior at large times, for reasons of non-normality. If eigenvalues fail to capture the transients, can pseudospectra do better? The answer is yes: Though pseudospectra rarely give an exact answer, they detect and quantify transients that eigenvalues miss.

4.8.3.3 Magnitude of Transient Excursions

We assume a matrix $\mathbf{M} \in \mathbb{R}^{N\times N}$ and are concerned with the growth and decay of solutions $\vec{s}(t)$ to the time-dependent equation $\mathrm{d}\vec{s}/\,\mathrm{d}t = \mathbf{M}\vec{s}$ with explicit solution $\vec{s}(t) = e^{t\mathbf{M}}\vec{s}(0)$. Specifically, we want to know something about the size of $\left|\left|e^{t\mathbf{M}}\right|\right|$. Figure 1(c) in the main paper shows three types of regimes. The asymptotic behavior for $t \to \infty$ is well-known, $\lim_{t\to\infty} t^{-1}\log\left|\left|e^{t\mathbf{M}}\right|\right| = \alpha(\mathbf{M})$ which is to say that the long-term behavior is governed by the largest real-part of all eigenvalues of $\mathbf{M}$. The typical stability criterion is therefore $\alpha(\mathbf{M}) < 0$.

The short-term behavior as $t \downarrow 0$ is less well-known. For that limit, it can be shown that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left|\left|e^{t\mathbf{M}}\right|\right|\Big|_{t=0} = \omega(\mathbf{M}) \equiv \sup\,\sigma\left(\frac{1}{2}\left(\mathbf{M}+\mathbf{M}^T\right)\right) \tag{4.8.24}$$

where $\omega(\mathbf{M})$ is called the *numerical abscissa* of $\mathbf{M}$.

Our main interest is not $t \downarrow 0$ or $t \to \infty$ but intermediate values of t. A useful lower bound for practical purposes is the inequality

$$\sup_{t\geqslant 0} \left|\left|e^{t\mathbf{M}}\right|\right| \geqslant \frac{\alpha_\epsilon(\mathbf{M})}{\epsilon} \quad \forall\epsilon > 0. \tag{4.8.25}$$

Particularly interesting are cases where $\sigma(\mathbf{M}) < 0$, i.e. where the long-term behavior is asymptotically stable, but $\alpha_\epsilon(\mathbf{M}) > \epsilon$ for some $\epsilon > 0$. In that case, the pseudospectra of $\mathbf{M}$ protrude significantly into the right-half plane of $\mathbb{C}$ (i.e. the positive real-part plane). Despite the asymptotic behavior being stable, it follows from (4.8.25) that there must be transient growth. A visualization of this concept is shown in Figure 1(d,e) of the main paper. A useful constant in this respect is the *Kreiss constant* which is defined by

$$\mathcal{K}(\mathbf{M}) \equiv \sup_{\epsilon>0} \frac{\alpha_\epsilon(\mathbf{M})}{\epsilon} \tag{4.8.26}$$

such that (4.8.25) implies

$$\sup_{t\geqslant 0} \left|\left|e^{t\mathbf{M}}\right|\right| \geqslant \mathcal{K}(\mathbf{M}). \tag{4.8.27}$$

Inequality (4.8.27) is central for our application. Given an interaction matrix $\mathbf{M}$ as in (4.8.21b), we can calculate the Kreiss constant (4.8.26) to obtain lower bounds for the transient growth of net magnetization. Similarly, albeit somewhat less relevant for our prediction of bubbles is an upper bound. It can be shown that $\left|\left|e^{t\mathbf{M}}\right|\right| \leqslant eN\mathcal{K}(\mathbf{M})\ \forall t$ where e is Euler's number and N the matrix dimensionality.

4.8.3.4 Geometric Interpretation of Transients

The above inequalities provide a quantitative description of the transient growth. But there is also a geometric interpretation. A non-normal matrix $\mathbf{M}$ cannot be diagonalized, i.e. one cannot find an orthogonal set of basis vectors. It may then happen that some eigenvectors have small angles between them. In light of the spectral theorem mentioned above, transients occur because the transformation that takes a vector $\vec{s}$ to the eigenbasis of $\mathbf{M}$ is not unitary if the eigenvectors of $\mathbf{M}$ are not orthogonal, and thus does not preserve the norm of $\vec{s}$ [78]. We refer to FIG. 1 in ref. [122] and FIG. 3 in ref. [78] for illuminating visualizations.

4.8.3.5 Henrici's Departure From Normality

Denote by $\mathbf{A}$ the adjacency matrix that represents the network resulting from the above algorithm. It has been shown [94] that the stronger the inequalities (quantified by taking small θ values), the stronger the non-normality of the network, as measured by (the normalized version of) *Henrici's departure from normality*

$$d_F(\mathbf{A}) = \sqrt{||\mathbf{A}||_F^2 - \sum_{\lambda\in\sigma(\mathbf{A})} |\lambda|^2} \Big/ ||\mathbf{A}||_F^2 \, . \tag{4.8.28}$$

Henrici's index is based on the observation that the Frobenius norm of a normal matrix is given by $||\mathbf{A}||_F^2 = \mathrm{tr}\left(\mathbf{A}^T\mathbf{A}\right) = \sum_{\lambda\in\sigma(\mathbf{A})} |\lambda|^2$. The measure (4.8.28) then attains its minimum at zero once the matrix is normal and increases the more the matrix deviates from normality. For example, the values of the adjacency matrices $\mathbf{A}$ depicted in Figure 2 (a,b,c) in the main paper are equal to 1, 0.9007 and 0.8057, respectively.

4.8.3.6 Hierarchies, Trophic Coherence and their Relationship with Non-Normality

Here, we elaborate more on the relationship between the hierarchical structure of a network and its non-normality. One can tune the "level of hierarchy" of a network by picking up the *trophic coherence* measure q previously developed in the context of predator-prey webs [113, 114]. Trophic coherence q has been called a measure of how similar a graph is to a hierarchy, and it is given by

$$q = \sqrt{\frac{1}{L}\sum_{ij} a_{ij}(\ell_i - \ell_j)^2 - 1} \tag{4.8.29}$$

where L is the total number of links and ℓ_i the trophic level of node i. For nodes with zero in-degree, the trophic level is 1. For nodes with an in-degree larger than 0, the trophic level is obtained by solving the following linear system of equations,

$$\ell_i = 1 + \frac{1}{k_i^{in}}\sum_j a_{ij}\ell_j, \tag{4.8.30}$$

where a_{ij} is a coefficient in the adjacency matrix and k_i^{in} is the in-degree of node i. The case $q = 0$ is a complete military hierarchy, where no subordinate has any influence on any superior. The case of large q is the opposite, where everybody can influence everybody else, at least indirectly. Directed graphs that have high trophic coherence are tree like, and can be drawn with all edges pointing in the same direction. Directed graphs with low trophic coherence do not have edges pointing in one clear direction, and appear more random. In ref. [114] a network generating mechanism is introduced that generates a network with given trophic coherence value q and fixed mean degree $\langle k\rangle$. The ensemble over different realizations of such networks with fixed q the *coherence ensemble.*

Trophic coherence and non-normality are related. Highly trophic networks are non normal, as is formalized in the following theorem [109]: *The expected deviation from normality* $d_F(\mathbf{M})$ *as defined in* (4.8.28)*, for directed graphs (without loops) drawn from the coherence ensemble tend to* 1 *with increasing trophic coherence, that is* $\lim_{q\downarrow 0} d_F = 1$*. Furthermore,*

$$d_F > \sqrt{1 - \frac{1}{\langle k \rangle}} \tag{4.8.31}$$

where $\langle k \rangle$ *is the mean degree.*

4.8.4 Growing Non-Normal, Scale-Free Networks

Here, we explain how to grow a non-normal network that consists of a total of N nodes. It is initialized at time $t = N_0$ with N_0 so-called *top nodes.* We denote the set of top nodes by $\mathcal{N}_0 \equiv \{1, \ldots, N_0\}$. At initialization, none of these N_0 nodes has any in-ward or out-ward directed edge. The remaining $N - N_0$ nodes are added to the existing network sequentially, one node at a time. The first node that is not a top node, i.e. node $N_0 + 1$, is added at time $t = N_0 + 1$. The next node is added at time $t = N_0 + 2$, and so forth, until the last node is added to the network at time $t = N$. We denote by $k_i^{\text{in/out}}$ the in/out-degree of node $i \in \{1, \ldots, N\}$ (time-dependence omitted for notational brevity). The generic in-degree of any given node $i > N_0$ is fixed to $m > 0$. Node $i > N_0$ is added at time $t = N_0 + i$ as follows:

1. Add node i to the network with m in-ward directed edges. More precisely, we have

$$k_i^{\text{in}} = \min\{m, N_0 + i - 1\} \tag{4.8.32}$$

   to account for the case where there are less than m nodes in the network at time $t = i$.

2. Select each of the k_i^{in} source-nodes with probability proportional to the out-degree of the existing nodes. More precisely, we denote a directed edge from j to i by $e_{j\to i}$. Then, the probability $\rho_{j\to i}$ that node j is source node of edge $e_{j\to i}$ is given by

$$\rho_{j\to i} = \frac{k_j^{\text{out}} + 1}{\sum_{\ell=1}^{i-1} k_\ell^{\text{out}} + 1} \tag{4.8.33}$$

   where there '+1'-term serves as regularization, so that even disconnected nodes can serve as source with non-zero probability.

3. We assign to node i its (hierarchical) level $\ell = \ell(i)$. In the case where no edge may be reciprocated, ℓ has the intuitive interpretation of the path length (+1) from the top node to node i. More formally, it is given by (4.8.30), which is a function of the

network as a whole and hence changes over time. In each iteration step, we thus have to recalibrate ℓ for each node in the system by solving (4.8.30).

4. Any of the new edges $e_{j\to i}$ may be reciprocated from i back to j with probability $\rho_{i\to j}$. To obtain a non-normal network, this reciprocation could happen with fixed probability $\theta \ll 1$, as implemented in ref. [94]. However, our empirical analysis (cf. Figure 2(d) in the main paper) suggests that in social communication networks θ is not constant. Instead, there is a level-dependence, such that higher levels are morel likely to reciprocate. In other words, $\rho_{j\to i}$ is increasing in $\ell(j)$, up to some level of saturation θ. We parametrize this observation with a sigmoid function

$$\rho_{i\to j} = \frac{\theta}{1+e^{-a(\ell(j)-b)}} - \underbrace{\frac{\theta}{1+e^{-a(1-b)}}}_{\equiv\gamma}. \tag{4.8.34}$$

In the remainder of this article, we fix $\theta = 0.2110, a = 2.552$ and $b = 3.668$ as determined empirically on meme-stock reddit data in Section 4.8.7 below. The offset $\gamma \approx 0.0002$ has been added so that nodes in the high-test level of the hierarchy, $\ell = 1$ are never reciprocated. This offset is merely for convenience, so that we can keep the number of top nodes ('opinion leaders') fixed. For large values of ℓ, we converge to the constant rate of reciprocity $\theta - \gamma \approx \theta$ as in [94].

4.8.5 Parametric Dependence of Non-Normality

Our algorithm to generate non-normal networks has six parameters: N, N_0, m, θ, a and b. The four parameters m, N, a and b play a subordinate role in the qualitative interpretation of our results. We fix them to $N = 1000, m = 2, a = 2.552, b = 3.668$ unless specified otherwise. The parameters θ and N_0, on the other hand, have qualitatively important implications on the behavior of our model. The parameter θ characterizes the hierarchical nature of the system. The smaller θ, the more directed the network. The parameter N_0 denotes the number of top nodes. Here, we analyze in more detail how these parameters effect matrix non-normality.

4.8.5.1 Characteristics of the Social-Influence Matrix M

The price dynamics is not directly governed by $\mathbf{A}$, but rather by the related matrix $\mathbf{M}$ defined by expression (4.8.21). The two parameters κ and $p^{\pm}$ allow us to control the characteristics of $\mathbf{M}$ for given $\mathbf{A}$. In the remainder of this article, we fix $p^{\pm} = 0.05$ and focus on parameter

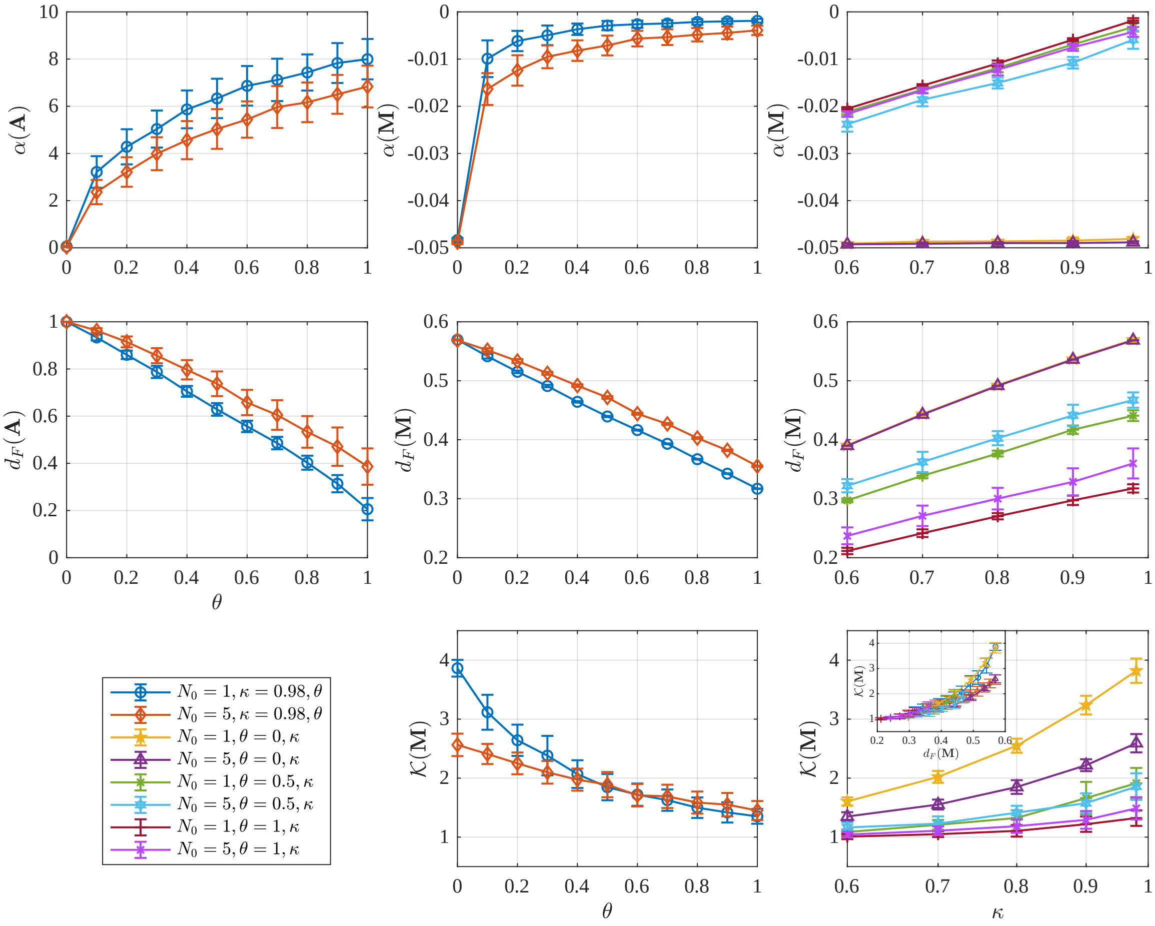

Figure 4.5: Properties of $\mathbf{A}$ and $\mathbf{M}$ for different parameters. Left panels: Properties of the adjacency matrix $\mathbf{A}$, characterizing the network of interactions. Middle and right panels: Properties of the interaction matrix $\mathbf{M}$ derived from $\mathbf{A}$ via equation (4.8.21a). Top left panel: $\alpha(\mathbf{A})$, the largest real part of the eigenvalues of $\mathbf{A}$. Middle left panel: Henrici's departure from normality $d_F(\mathbf{A})$ as a function of backward link probability θ. Top middle and right panels: $\alpha(\mathbf{M})$, the largest real part of the eigenvalues of $\mathbf{M}$. Note the negative values, indicating that the dynamics around the $\Delta\mathbf{s} = 0$ fixed-point is stable. Middle panel: As the rate of reciprocation increases, $\mathbf{M}$ becomes less non-normal. Middle right panel: The larger the coupling κ, the less normal the matrix $\mathbf{M}$. Bottom panels: The behavior of the Kreiss constant as a function of θ and κ. Bottom right inset panel: Monotonic relationship between the level of non-normality, d_F and the Kreiss constant $\mathcal{K}$. Error bars are obtained as standard errors across 20 different realizations of the matrix $\mathbf{A}$.

κ. [3]

Figure 4.5 presents some properties of $\mathbf{A}$ and $\mathbf{M}$ as a function of backward link probability

[3] The value of $p^{\pm}$ has been calibrated such that the annualized volatility is on par with typical values for public equity. We have checked that changing this parameter does not qualitatively effect our results.

θ and imitation strength κ. For all parameter combinations, the largest real part $\alpha(\mathbf{M})$ of the eigenvalues of $\mathbf{M}$ remains negative, ensuring that the dynamics around the $\Delta\mathbf{s} = 0$ fixed-point is stable [4]

As anticipated, the matrix $\mathbf{A}$ becomes less and less non-normal as θ increases (middle, left plot). For $\mathbf{A}$ to become fully normal ($d_F(\mathbf{A}) = 1, \mathcal{K} = 0$), we would need to make sure that every edge is reciprocated. This is achieved by removing the level-dependence (4.8.34), which is straight forward to implement (see also ref. [94]). However, $\mathbf{A}$ becoming (approximately) normal does not imply that $\mathbf{M}$ becomes (approximately) normal (middle plot). Recall from (4.8.21a) that $\mathbf{M} \propto \mathbf{\Lambda A}$ where $\mathbf{\Lambda}$ is a diagonal matrix with $1/k_i^{\text{in}}$ on the i-th diagonal entry. The symmetry, and hence normality, of $\mathbf{A}$ does not imply the symmetry of $\mathbf{M}$. To see this, assume $\mathbf{A}$ is symmetric. It then holds that $(\mathbf{A\Lambda})^T = \mathbf{\Lambda}^T\mathbf{A}^T = \mathbf{\Lambda A}$. In general, a diagonal matrix $\mathbf{\Lambda}$ does not commute with $\mathbf{A}$, i.e. $\mathbf{\Lambda A} \neq, \mathbf{A\Lambda}$. Therefore, in general, $(\mathbf{A\Lambda})^T \neq \mathbf{A\Lambda}$ and a symmetric $\mathbf{A}$ does not imply a symmetric $\mathbf{M}$. However, as can be seen in the middle plot of Figure 4.5, as $\mathbf{A}$ becomes less non-normal, so does $\mathbf{M}$, which manifests in $d_F(\mathbf{M})$ decaying to low values as $\theta \to 1$. The larger the rate of reciprocity θ, the lower the non-normality of $\mathbf{M}$.

Since we have established that $\alpha(\mathbf{M}) < 0$, the Kreiss constant (4.8.26) is well-defined. Hence, we may examine the structural dependence of the Kreiss constant as a function of the model parameters N_0, p and κ. Figure 4.5 (bottom panels) shows the Kreiss constant $\mathcal{K}(\mathbf{M})$ as a function of θ and κ. Generally, $\mathcal{K}$ is a decreasing function of the level of reciprocity θ. This is to be expected and in line with the behavior of d_F as θ increases. The less non-normal the matrix, the smaller $\mathcal{K}$ is.

4.8.6 Financial Bubbles

4.8.6.1 Measurement of Financial Bubbles

In this section, we explain in more detail how the size of a bubble is measured. We apply the method used in ref. [97] to identify the bubbles in the price time series. The first step it to identify peaks and valley in the price path. A peak at time scale Δt in the price time series occurs at time step t_i if

$$P_{t_i} \geq P_{t_j} \quad \forall t_j \in [t_i - \Delta t, t_i + \Delta t] , \tag{4.8.35}$$

[4] Theoretically, this follows from the fact that $\Lambda\mathbf{A}$ is a Markov transition matrix (the sum across each row is equal to 1). A well known property of such transition matrices is that the largest absolute value of all its eigenvalues are less or equal to 1. It follows that $\kappa\Lambda\mathbf{A} - \mathrm{I}$ has negative eigenvalues so long as $\kappa < 1$, as is the case in this paper.

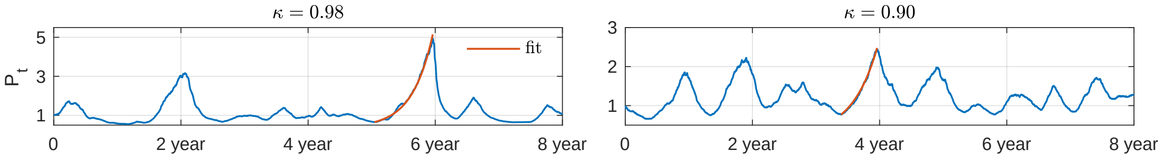

Figure 4.6: The evolution of the price P_t with $\kappa = 0.98$ and $\kappa = 0.90$. The interaction network is non- normal, generated by our attachment algorithm with $N = 1000$, $m = 2$, $N_0 = 1$ and $\theta = 0$. Several price-peaks are detected, but only one of those (indicated by the red line) in each of the two plots corresponds to a bubble with super-exponential growth $(\alpha > 1)$.

where P_t is the price at time step t. This peak detection is an a posteriori measure, which is much simpler than the the notoriously difficult task of real-time price peak prediction.

Once the peaks are determined, a valley is defined as the time at which the price is minimal between two consecutive peaks. Potential candidates for bubbles are then the examined as the price-time series between any valley and its consecutive peak. For a given candidate, denote by t_v the time of a valley and by t_p the time of the subsequent peak. Following [96], we fit the price time series P_t

$$P_t = P_{t_v} \exp\left[\beta \; x_v^{\mathcal{N}} \; \left(\alpha^{t-t_v} - 1\right)\right] \tag{4.8.36}$$

where t runs from t_v up to t_p and $x_v^{\mathcal{N}}$ denotes the risky fraction at time t_v. The parameters α and β are determined with a least-squares method. A value of $\alpha > 1$ indicates super-exponential growth, and hence a bubble.

For values where $\alpha > 1$, we define the height of the bubble as $log(P_{t_p}) - log(P_{t_v})$ and its steepness as $\frac{log(P_{t_p}) - log(P_{t_v})}{t_p - t_v}$.

Given the nature of our price simulations in units of trading days, we select $\Delta t = 1$ year, such that the minimal distance between any two peaks is one year. Therefore, a bubble can develop over the time-scale of months. Figure 4.6 shows two examples of such fits.

Although the hall-mark characteristic of an unsustainable bubble is super-exponential growth [96], we can also relax the $\alpha > 1$ constraint and consider general price movements between subsequent peaks. This is what we have done in Figure 4.7(b) below to obtain more data-points. In real price data, there is a lot more noise than in the simulation, thereby making it difficult to always clearly distinguish between the $\alpha > 1$ and $\alpha < 1$ regime. However, when enforcing the $\alpha > 1$ constraint, the same trends are observed, albeit with much less data.

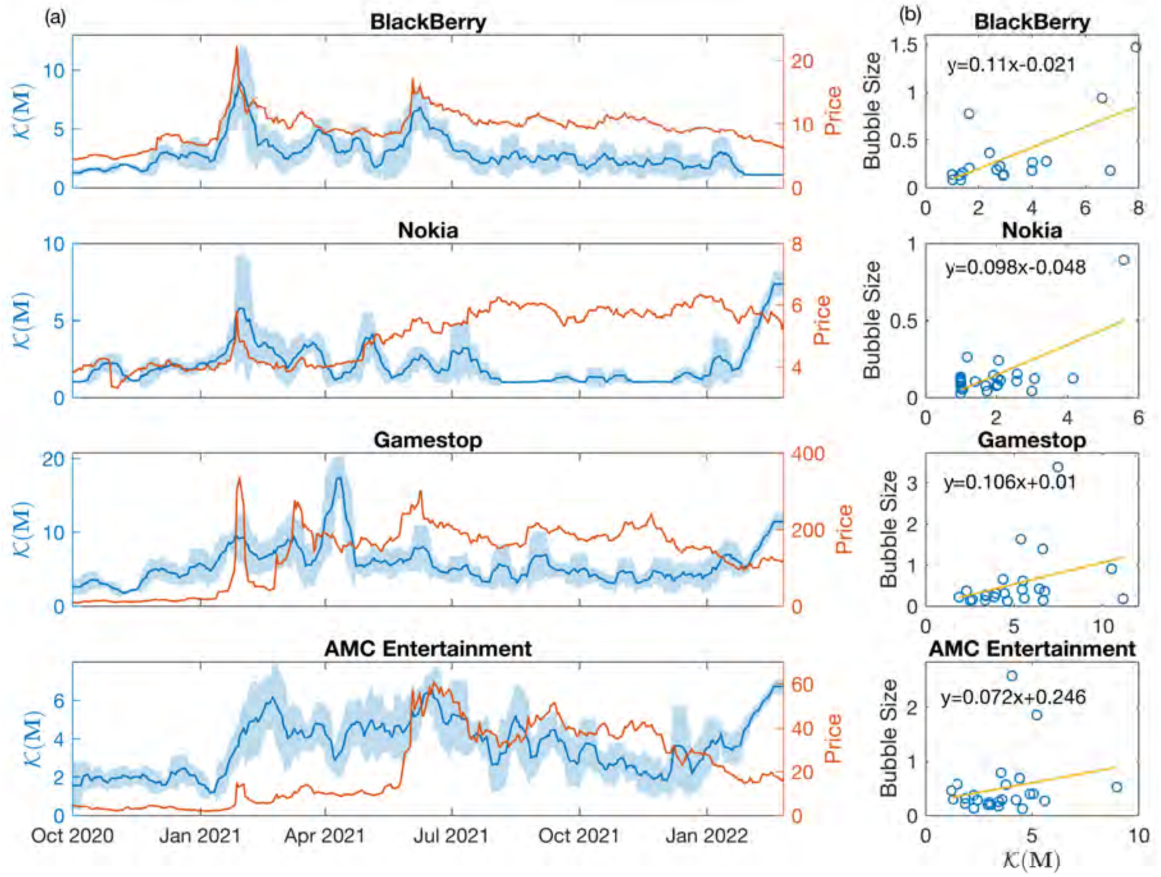

Figure 4.7: (a) Left y-axes: evolution Kreiss-constant 4.8.26 of meme-stock discussions on reddit. Right y-axes: Meme stock prices. (b) Correlation between bubble size and Kreiss constant.

4.8.6.2 Scaling Laws

In Figure 4.8 we show the dependence of bubble size and bubbles steepness as a function of our model parameters, along with the subsequent collapse when considered a function of $\mathcal{K}$ and ω, respectively.

4.8.6.3 Parameter Sensitivity

In Figure 4.9 we show that the scaling laws holds for a wide variety of parameter constellations.

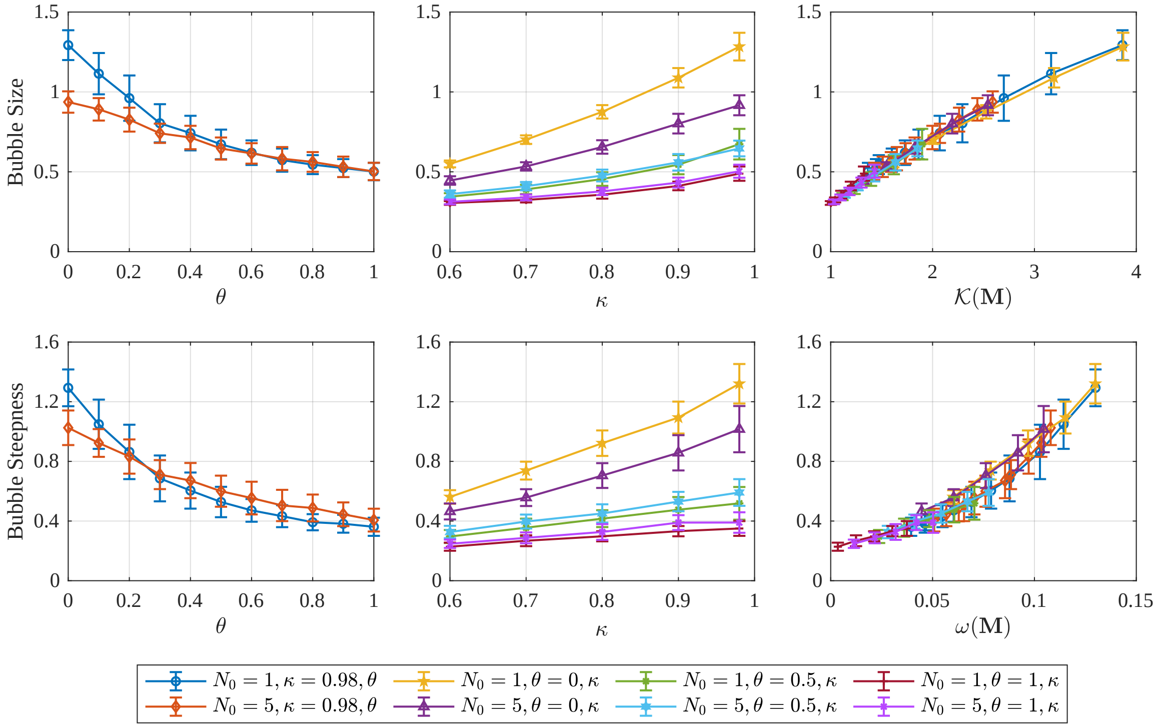

Figure 4.8: Dependence of bubble size and bubbles steepness as a function of different model parameters, and subsequent collapse when considered a function of $\mathcal{K}$ and ω, respectively.

4.8.7 Empirical Analysis of Meme Stocks

4.8.7.1 Meme Stocks and Reddit Data

A meme stock is a stock that gains popularity among retail investors through social media. The popularity of meme stocks is generally based on internet memes shared among traders, on platforms such as Reddit [119–121].

Reddit is an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, images, and videos, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "communities" or "subreddits", which cover topics such as news, politics, religion, science, movies, video games, music, books, sports etc. Under each subreddit, registered members can post their submissions, which are like topics they want to discuss with others, and others can comment on the submissions. A submission is at the highest level, and can contain thousands of comments. Each comment can contain replies.

In our paper, we collected data of all submissions, comments and replies related to four popular meme stocks from Oct 1st, 2020 to Feb 25th, 2022. The data is extracted from the famous subreddit *r/wallstreetbets* (also known as WallStreetBets or WSB) that has become

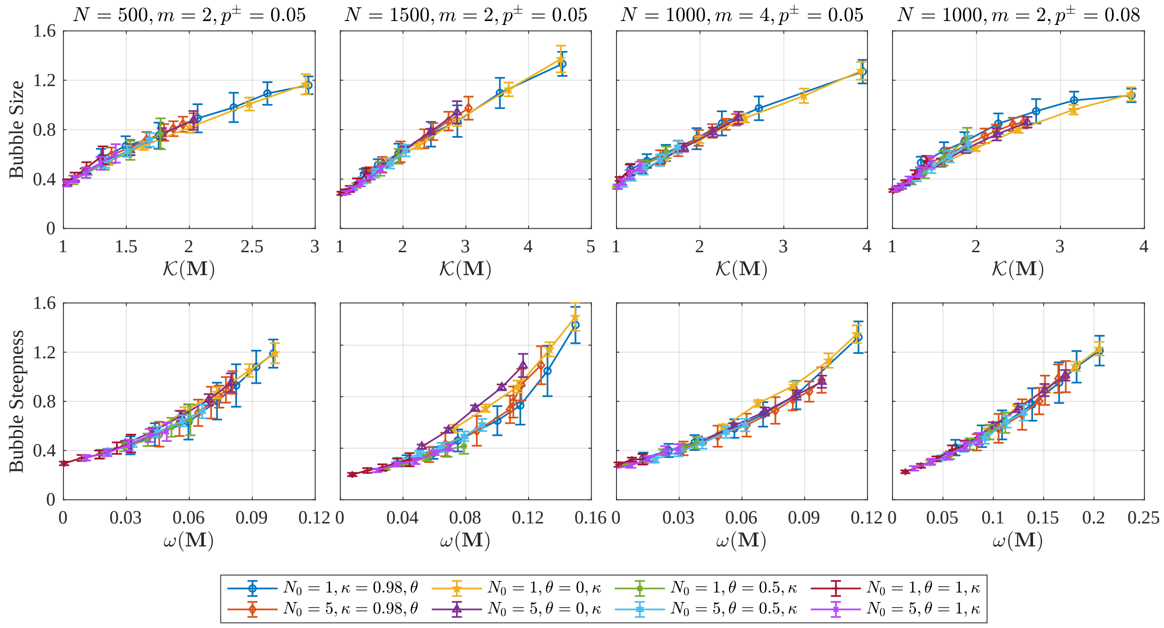

Figure 4.9: Top panel: average bubble size as a function of $\mathcal{K}(\mathbf{M})$ for different combination of $(N, m, p^{\pm})$. Bottom panel: average bubble steepness as a function of $\omega(\mathbf{M})$ for different combination of $(N, m, p^{\pm})$. Each data point is obtained by averaging over 100 simulations each lasting $T_{total} = 25,000$ time-steps.

	BB	NOK	GME	AMC
keyword searches	bb, blackberry	nok, nokia	gme, gamestop	amc
number of submissions	11,076	6,858	36,304	20,510
number of comments and replies	222,733	67,476	1,064,624	337,933

Table 4.2: Overview of Reddit Meme stock discussion topics.

notable for its colorful and profane jargon, aggressive trading strategies, and for playing a major role in the GameStop short squeeze in early 2021 [119]. The four meme stocks include BlackBerry Limited (BB), Nokia Oyj (NOK), GameStop (GME) and AMC Entertainment Holdings Inc (AMC). Apart from GameStop, many other heavily shorted securities saw a huge increase of their price volatilities in early 2021, and were considered to be driven by retail investors on social platforms [127–130]. We have limited our attention to the above mentioned four meme stocks since they had been most actively discussed on reddit.

We used the open-source python package PRAW[5] and Pushshift[6] to collect data from the subreddit r/wallstreetbets. The searching key words are the name or symbol of the four stocks. Within the subreddit r/wallstreetbets, we collect all submissions whose context

[5] available at https://praw.readthedocs.io/en/stable/

[6] available at https://github.com/pushshift/api

include one of the keywords, and collect all comments and replies under this submission. Table 4.2 summarizes the keywords and the number of submissions, comments and replies.

Based on this crawled data, we construct a dynamic network $\mathbf{A}(t)$ of users for each stock. For each of the four stocks, at time t, we draw a directed edge from user i to user j if user j has commented a submission by user i or user j has replied to a comment/repliy by user i in the time interval $[t - \Delta t, t]$. In other words, j has been influenced by i'a action in the past Δt days. With this procedure, for each meme stock, we extract a dynamically evolving influence network $\mathbf{A}(t)$, of which we can measure the Kreiss constant $\mathcal{K}$ and other related quantities. Similarly, we can measure the number of bubbles and their size for the associated daily stock price. The 1-minute price data of these four stocks for the time from Oct 1st, 2020 to Feb 25th, 2022 is obtained from Refinitiv Eikon.

In Figure 4.10, we show the monthly evolution of word frequency of five frequently mentioned words within the meme-stock discussions on Reddit. In Jan-2021, the mostly frequently mentioned word is *rocket* in all four meme stocks. *Rocket* is a symbol to express the expectation of price rocketing at Reddit. Thus we can see that this word appears mainly in January and June when there were two huge price bubbles.

4.8.7.2 Network Properties of Reddit Discussion Forums

In Section 4.8.4 we have discussed how to grow non-normal networks with preferential-attachment and level-dependent rates of reciprocity. Here, we provide empirical evidence for these assumptions by analyzing the Blackberry stock from Oct 1st, 2020 to Jul 31st, 2021.

We start by analyzing basic network properties for the network across all time. Figure 4.11(a) confirms the heavy-tailed degree-distribution of both in- and out- degree as is ubiquitous in socio-economic networks [131]. In Figure 4.11(b), we see that the hierarchical level ℓ spans multiple orders of magnitude. Here, we have calculated the level ℓ of each node based on its in-degree via (4.8.30).

To confirm the presence of preferential attachment, we measure for each user the total number of previously received comments, as well as the additional number of received comments in each subsequent week. We assign the total number of received comments into bins and calculate the average number of received replies in the following week. Figure 4.11(c) shows the average trend on a double-logarithmic scale, together with a linear fit (grey dashed lined) obtained via least-squares regression. The slope of that least-squares fit is equal to 0.82, which is short of a slope equal to one expected from pure preferential attachment. Instead, there seems to be a sub-linear preferential attachment, which has been well-studied both theoretically and empirically [132]. However, for simplicity and because our results are not expected to change qualitatively, we grow our simulated networks based on pure preferential attachment via (4.8.33).

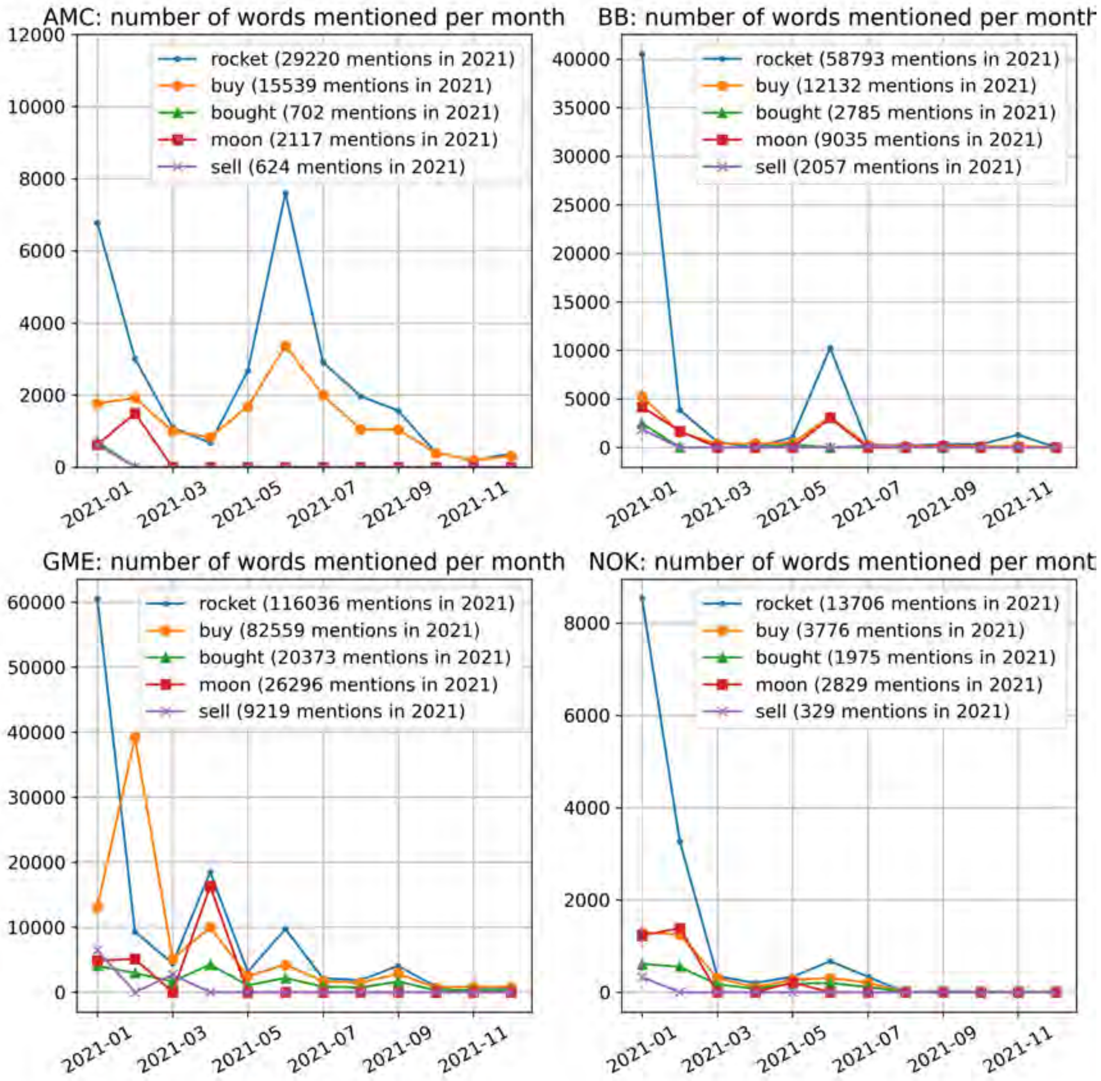

Figure 4.10: The number of words mentioned per month within the meme-stock discussions on Reddit for stock AMC (upper-left), BB (upper-right), GME (lower-left) and NOK (lower-right). Word frequency of five words (*rocket* in blue line, *buy* in orange line, *bought* in green line, *moon* in red line, and *sell* in purple line) are plotted, and the total number of mentions in 2021 is displayed in the legend.

Next, we group the nodes into 9 bins, according to their level ℓ. The bin boundaries are as follows: $\left[0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 10.5, 10^2, 10^3, 10^5\right]$. For each such bin, we calculate the average rate of at which users respond to reddit posts simply as the normalized count of replies per user, averaged across all users. The result is reported in Figure 2(d) of the main paper and shows a sigmoid-shape as in (4.8.34). With least-squares optimization, we determine the parameters a, b and θ as 2.552, 3.668 and 0.2110, respectively.

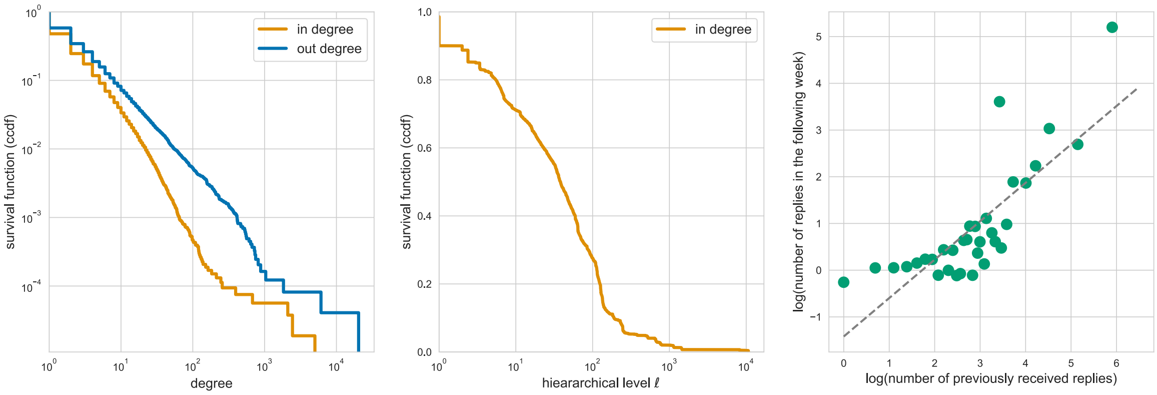

Figure 4.11: Analysis of Blackberry Reddit meme stock discussion forum. (a) Survival function of in- and out-degrees. (b) Survival function of hierarchical levels. (c) Number of previously received comments vs. number of future comments. The grey dashed line indicates a linear fit with slope 0.82.

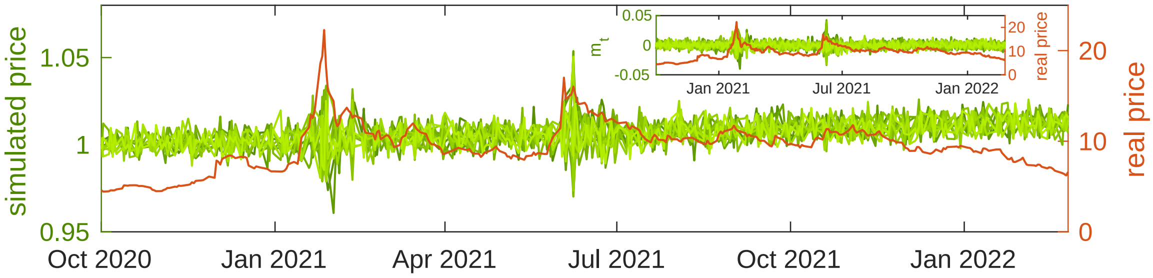

Figure 4.12: Same as Figure 4 of the main paper, but for multiple realizations of the net magnetization $m(t)$.

4.8.7.3 ABM Simulation with Reddit Network

To paint a more causal picture, we use the Blackberry discussion network $\mathbf{A}(t)$ as input to our agent-based model (see also Section 5 in the main paper). Rather than simulating once, we simulate 20 times due to the inherent stochasticity in the spin flip (4.8.8). Figure 4.12 shows the price evolution of each realizations. We notice two very distinct transients around each of the two distinct bubbles in January 2021 and June 2021, respectively. However, the simulated price is not necessarily increasing, but also decreasing, i.e. we observe both positive and negative bubbles ("crashes"). This is a direct consequence of the inversion symmetry in (4.8.8). The negative transients could be made positive by means of an external, positive news source (symmetry breaking in the Ising jargon), and justifies our selection of a positive realization in the main paper.

Chapter 5

Rank the spreading influence of nodes using dynamic Markov process

Ranking the spreading influence of nodes is of great importance in practice and research. The key to ranking a node's spreading ability is to evaluate the fraction of susceptible nodes been infected by the target node during the outbreak, i.e., the outbreak size. In this paper, we present a dynamic Markov process (DMP) method by integrating the Markov chain and the spreading process to evaluate the outbreak size of the initial spreader. Following the idea of the Markov process, this method solves the problem of nonlinear coupling by adjusting the state transition matrix and evaluating the probability of the susceptible node being infected by its infected neighbours. We have employed the susceptible-infected-recovered (SIR) and susceptible-infected-susceptible (SIS) models to test this method on real-world static and temporal networks. Our results indicate that the DMP method could evaluate the nodes' outbreak sizes more accurately than previous methods for both single and multi-spreaders. Besides, it can also be employed to rank the influence of nodes accurately during the spreading process.

Based on Jian-Hong Lin, Zhao Yang, Jian-Guo Liu, Bo-Lun Chen, and Claudio J. Tessone, "Rank the spreading influence of nodes using dynamic Markov process." 2022.
(submitted to *Journal of The Royal Society Interface*)

5.1 Introduction

Complex networks are widely used to represent interactions between people, technology, and various entities. Among all the studies within the area of network theory, understanding the dynamics of spreading processes is of particular interest. Although the spreading dynamics on networks are not a new phenomenon, studies in this field lead to better understandings of many important social and natural processes [133], such as the spreading of infectious diseases [134], the propagation of computer virus [135], the cascading process [136], traffic congestion [137], the centralization in Bitcoin system [7, 8], and so on. One important approach in studying the spreading dynamics is to estimate and rank nodes' spreading abilities. Through this approach, one might first locate influential nodes of complex networks and later on control the outbreak of epidemics [27, 28], target the opinion leaders in social networks [25], quantify the scientific impact [29, 30], and accelerate the adoption of innovation [26], etc.

Classical centrality measures have been developed to identify the spreading influence of nodes. The degree centrality [138] is probably the most straightforward one. Nodes with larger degree centrality are considered to have better spreading abilities than the other nodes within a graph. The betweenness centrality [139], which calculates the number of shortest paths cross through a certain node, represents the controllability of information flow over the networks. The closeness centrality [140] measures the inverse of the mean geodesic distance from a certain node to all other nodes. The more central a node is, the closer it is to all the other nodes. The eigenvector centrality [36] assigns relative scores to all nodes in the network based on the concepts that connections to influential nodes, i.e., high-scoring nodes, would be more important than that to low-scoring nodes. The k-shell decomposition method [141] assigns nodes to different shells and considers those located within the core of the network are the most efficient spreaders. Furthermore, a lot of methods for identifying the node spreading influence have been developed from different perspectives [142–148]

The classical centrality measures are based on the network topological structure solely. However, recent studies have shown that the nodes' spreading influence is determined not only by the network structure but also the parameters of the dynamical processes. Therefore, various structural based centralities cannot properly identify nodes' influences since the rankings remain the same under different dynamical parameters. Šikić *et al.* [31] argued that for a given susceptible-infected-recovered (SIR) model [149], the rank of nodes' influence largely depends on the spreading rate and recovering rate. Klemn *et al.* [32] suggested that the eigenvector centrality could only identify the nodes' spreading influence accurately when the spreading rate is close to the inverse of the largest eigenvalue of the network [149]. Considering the susceptible-infected-susceptible (SIS) model [150], Ide *et al.* [151] have proposed a numerical framework that uses the importance of the centrality type to determine how the vulnerable nodes change along the diffusion phases. Besides, Liu *et al.* [33] have described the infectious probabilities of nodes by a matrix differential function and have developed the dynamics-sensitive (DS) centrality to predict the outbreak size for ranking nodes' spreading influence.

The centrality measures proposed in [31–33] are all linear methods based on discrete Markov process. However, it is important to note that the spreading process in SIR and

SIS model is usually non-linear couple process. Therefore, without taking the non-linear couple process into consideration, all the nodes' influence would be over estimated. For instance, if a susceptible node has $n^{'}$ infected nodes, the probability of this node to be infected is $1-(1-\beta)^{n^{'}}$ instead of $n^{'}\beta$ approximated by the linear methods, where β is the spreading rate in the SIR or SIS model. In this paper, we present a dynamic Markov process (DMP) to evaluate the outbreak size of the nodes at given time steps. This method can be directly applied in ranking nodes' spreading influence. It overcomes the problem of nonlinear coupling by calculating the susceptible node to be infected by its neighbours sequentially and adjusting the state transition matrix during the spreading process. Our simulation results on susceptible-infected-recovered (SIR) model show that the DMP method has comparable accuracy to the linear methods [31–33] for both single spreader and multi-spreaders [152]. Furthermore, we have employed the susceptible-infected-recovered (SIR) and susceptible-infected-susceptible (SIS) models to test the DMP method on real static and dynamic networks [149, 150, 153, 154]. The simulation results show that the DMP method can rank nodes' spreading influence accurately.

5.2 Methods

5.2.1 Centrality Measures.

A network $G=(V,E)$ with $n=|V|$ nodes and $e=|E|$ links could be described by an adjacency matrix $\mathbf{A}=\{a_{ij}\}$ where $a_{ij}=1$ if node i is connected to node j, and $a_{ij}=0$ otherwise. For directed network, if only node i is pointing to node j, then $a_{ij}=1$ and $a_{ji}=0$.

The degree of node i is defined as the number of its neighbors, namely

$$k_i=\sum_{j=1}^{n}a_{ij}, \tag{5.2.1}$$

where a_{ij} is the element of matrix $\mathbf{A}$.

The main idea of eigenvector centrality is that a node's importance is not only determined by itself, but also by its neighbours' importance [36]. Accordingly, eigenvector centrality of node i, i.e. v_i, is defined as

$$v_i=\frac{1}{\lambda}\sum_{j=1}^{n}a_{ij}v_j, \tag{5.2.2}$$

where λ is a constant. Obviously, Eq. 5.2.2 can be written in a compact form as

$$\mathbf{Av}=\lambda\mathbf{v}, \tag{5.2.3}$$

where $\mathbf{v} = (v_1, v_2, \cdots, v_n)^T$. That is to say, $\mathbf{v}$ is the eigenvector of the adjacency matrix $\mathbf{A}$ and λ is the corresponding eigenvalue. According to Perron-Frobenius Theorem [155], the elements of the leading eigenvector are positive. Since the influences of nodes should be positive, $\mathbf{v}$ must be the leading eigenvector corresponding to the largest eigenvalue of $\mathbf{A}$, therefore we have $\mathbf{v} = \mathbf{q}_1$.

5.2.2 Dynamic Markov Process Method

In order to calculate the probabilities of nodes to be infected, one needs to solve the problem of nonlinear couple during the spreading process. Consider the union of a finite number of event $A_1^{'}, \dots, A_{n'}^{\prime}$, the probability of the event $\cup_{i=1}^{n'} A_i^{\prime}$ could be written as [156]:

$$Pr\cup_{i=1}^{n'} A_i^{\prime} = 1-\prod_{i=1}^{n^{'}}(1-Pr(A_i^{'})) = Pr(A_1^{'})+(1-Pr(A_1^{'}))Pr(A_2^{'})+\cdots+\prod_{i=1}^{n^{'}-1}(1-Pr(A_i^{'}))Pr(A_{n^{'}}^{'}). \tag{5.2.4}$$

The Eq. 5.2.4 means that the probability of the event $\cup_{i=1}^{n'} A_i^{\prime}$ equals to the probability of the event $A_1^{'}$ plus the probability of the event $A_2^{'}$ while event $A_1^{'}$ doesn't happen plus the probability of the event $A_3^{'}$ while both event $A_1^{'}$ and $A_2^{'}$ do not happen, so on and so forth. Based on the above-described probability theorem, for a susceptible node with $n^{'}$ infected neighbors, the probability of being infected during the spreading process can be represented in the non-linear format as $1-(1-\beta)^{n^{'}}$, or as $\beta + (1-\beta)\beta + (1-\beta)^2\beta + \cdots + (1-\beta)^{n^{'}-1}\beta$, where the first term β represents the probability that this node has been infected by its first infected neighbor, the second term $(1-\beta)\beta$ represents the probability that it has not been infected by its first infected neighbor but has been infected by its second infected neighbor, and the third one $(1-\beta)^2\beta$ represents the probability that the node has not been infected by its first infected neighbor nor its second infected neighbor but has been infected by its third infected neighbor, etc.

By combining the above-described process with the standard SIR model where an infected node would infect its susceptible neighbors with a spreading rate β and recover immediately, we propose a dynamics Markov process method as follows: Define $\mathbf{x}(t)$ $(t \geq 0)$ as an $n \times 1$ vector whose components are approximated as the probabilities of nodes to be infected at time step t. Especially, if node i is the initially infected node, then $x_i(0) = 1$ and $x_{j \neq i}(0) = 0$. In the dynamics Markov process, the initial Markov state transition matrix $\mathbf{M} = \mathbf{A}^T$, where $\mathbf{A}^T$ is the transpose of $\mathbf{A}$. If $m_{ij} = 0$, node j could not be infected by node i anymore. Otherwise m_{ij} is the probability of node i to be susceptible. When $t = 0$, if node i is the initial infected node, it could not be susceptible anymore. Therefore we have $m_{ij} = 0$, where $j = 1, 2, \dots n$. When $t \geq 1$, we denote $\mathbf{C}(t)$ as an $n \times n$ matrix, where $c_{ji}(t)$ is the probability of node j to be infected by node i at time step t.

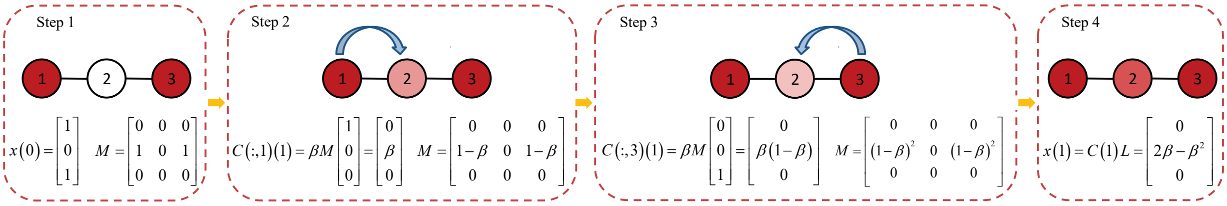
Step 1
1
2
3
$x(0)=\begin{bmatrix}1\\0\\1\end{bmatrix}$ $M=\begin{bmatrix}0&0&0\\1&0&1\\0&0&0\end{bmatrix}$
Step 2
1
2
3
$C(:,1)(1)=\beta M\begin{bmatrix}1\\0\\0\end{bmatrix}=\begin{bmatrix}0\\\beta\\0\end{bmatrix}$ $M=\begin{bmatrix}0&0&0\\1-\beta&0&1-\beta\\0&0&0\end{bmatrix}$
Step 3
1
2
3
$C(:,3)(1)=\beta M\begin{bmatrix}0\\0\\1\end{bmatrix}=\begin{bmatrix}0\\\beta(1-\beta)\\0\end{bmatrix}$ $M=\begin{bmatrix}0&0&0\\(1-\beta)^2&0&(1-\beta)^2\\0&0&0\end{bmatrix}$
Step 4
1
2
3
$x(1)=C(1)L=\begin{bmatrix}0\\2\beta-\beta^2\\0\end{bmatrix}$

Figure 5.1: (Color online) An example network with 3 nodes, where node 1 and node 3 is the initial infected nodes. The probability of node 2 $x_2(1)$ to be infected at time step 1 is $2\beta - \beta^2$ generated by the DMP method.

The updating rules are described as below. We first calculate the influence of node 1 to all its susceptible neighbors. If $x_1(t-1) > 0$, node 1 would infect its susceptible neighbors with a probability β at time step t. The probability of node j to be infected by node 1 is then

$$c_{j1}(t) = \beta m_{j1} x_1(t-1), \tag{5.2.5}$$

where $j = 1, 2, \ldots n$. After that, for all the nodes j, if $m_{jl} > 0$, where $l = 1, 2, \ldots n$, we calculate the probability of node j to be susceptible, i.e. the probability of node j that has not been infected by node 1. It equals to $m_{jl} - c_{j1}(t)$. After that, for all the nodes j, we update the element of the state transition matrix as follows:

$$m_{jl} := m_{jl} - c_{j1}(t), \tag{5.2.6}$$

where $l = 1, 2, \ldots n$. Once the state transition matrix has been updated, we continue to calculate the impact of node $2, 3, 4, ...n$ sequentially in the same way. In the end, the probabilities of node i been infected at time step t is

$$x_i(t) = \sum_{j=1}^{n} c_{ij}, \tag{5.2.7}$$

The spreading influence of the target node within a certain time T^* is $\sum_{t=1}^{T^*} \sum_{j=1}^{n} x_j(t)$. During the spreading process, the DMP method solves the problem of nonlinear coupling by calculating the probability of the node to be infected by its infected neighbours sequentially via adjusting the state transition matrix $\mathbf{M}$. As shown in Fig. 5.1, node 1 and node 3 are the initial spreaders. According to the DMP method, we firstly calculate the probability of node 2 to be infected by node 1, which is β. Then we update the state transition matrix $\mathbf{M}$. After that, we calculate the probability of node 2 to be infected by node 3, which is $\beta(1-\beta)$. And in the end, we get the probability of node 2 to be infected by its both infected neighbours, which equals to $\beta + \beta(1-\beta) = 1 - (1-\beta)^2$. We note that this is the exact probability of the node 2 to be infected.

The DMP method could be extended to the SIS model with $\gamma = 1$, where γ is the probability of the infected nodes enter the susceptible state (see the details in the Data Analysis section). In SIS model, the difference is that at each time step t, the transition matrix $\mathbf{M}$ could be updated by $m_{ij} = a_{ji}(1 - x_i(t-1))$. For the temporal network, the network could be described by $\mathbf{A}(t)$ at each time step t. Thus, at time t the transition matrix $\mathbf{M}$ could be updated by $m_{ij} = a_{ji}(t)(1 - \sum_{r=0}^{t-1} x_i(r))$.

5.3 Data Analysis

5.3.1 Data description

We have tested the performance of DMP method in estimating the nodes' spreading influence according to the SIR and SIS models on four real networks. The first network is "C. elegans", a directed network representing the neural network of Caenorhabditis elegans [63]. The data is available at https://snap.stanford.edu/data/C-elegans-frontal.html. The second network is a scientific collaboration network, "Erdös", where nodes are scientists and edges represent the co-authorships. The data can be freely downloaded from the web site http://wwwp.oakland.edu/enp/thedata/. The third one is an email communication network of University Rovira i Virgili (URV) of Spain, involving faculty members, researchers, technicians, managers, administrators, and graduate students [157]. The data can be found at http://konect.cc/networks/arenas-email/. The last network is a directed network based on the ODLIS dictionary network. This a hypertext reference resource for library and information science professionals, university students and faculty, and users of all types of libraries. The node represents web site of Odlis and the edge represents the network are the connection between two web sites. This data is available at http://networkdata.ics.uci.edu/netdata/html/ODLIS.html. Basic statistical properties of these four networks are presented in Table 5.1.

Besides, we have also analyzed four real-world dynamic networks in order to evaluate the effectiveness of the DMP method. The first temporal network is *Contacts in a workplace* (CW) network. This data includes contacts between individuals measured in an office building in France, from June 24 to July 3, 2013 [158]. The second one is the *Primary school* (PS) temporal network, where nodes are the children and teachers, and edges represent the contacts between them [159, 160]. The CW and PS network could be downloaded at http://www.sociopatterns.org/datasets/. The *email-Eu-core-temporal-Dept1* (EM01) and *email-Eu-core-temporal-Dept2* (EM02) [161] temporal network are generated by using email data from a large European research institution, where edges present email between members of the research institution. These two datasets are available at http://snap.stanford.edu/data/index.html. In Table 5.2, we provide the detailed statistical properties of the above temporal networks.

Network	n	e	$\langle k \rangle$	$1/\lambda_1$
C. elegans	297	2345	7.896	0.109
Erdös	454	1313	5.784	0.079
Email	1133	5451	9.622	0.048
Odlis	2900	18241	6.290	0.077

Table 5.1: Basic statistical features of C. elegans, Erdös, Email, and Odlis networks, including the number of nodes n, the number of the edges e, the average degree $\langle k \rangle$ or $\langle k_{out} \rangle$ (for directed networks) and the reciprocal of the largest eigenvalue $1/\lambda_1$.

Network	n	e
CW	92	1492
PS	242	21295
EM01	309	11106
EM02	162	7758

Table 5.2: Basic statistical features of CW, PS, EM01, and EM02 temporal networks, including the number of nodes n and the number of the edges e respectively.

5.3.2 The SIR and SIS Model

We apply the susceptible-infected-recovered (SIR) model [149] and the susceptible-infected-susceptible (SIS) [150] model to simulate the spreading process and record the nodes' spreading influence at each time step. In the SIR model, there are three kinds of individuals: (i) susceptible individuals that could be infected, (ii) infected individuals which are able to infect their susceptible neighbors, and (iii) recovered individuals that will never be infected again. At each time step, every infected node will contact its neighbors and each of its susceptible neighbors will be infected with a probability β. Then the infected nodes enter the recovered state with a probability μ. While in SIS model, there are only two kinds of individuals, i.e. the susceptible individuals and the infected ones. The infected nodes would infect its susceptible neighbors with the probability β and enter the susceptible state with a probability γ. For single-node spreading, only one seed node is infected at the beginning, and all the other nodes are susceptible. While for multiple-nodes spreading, a set of nodes are infected and the rests are initially susceptible. At each time step t, the number of nodes that switch from the susceptible state to the infected state represents the node's spreading

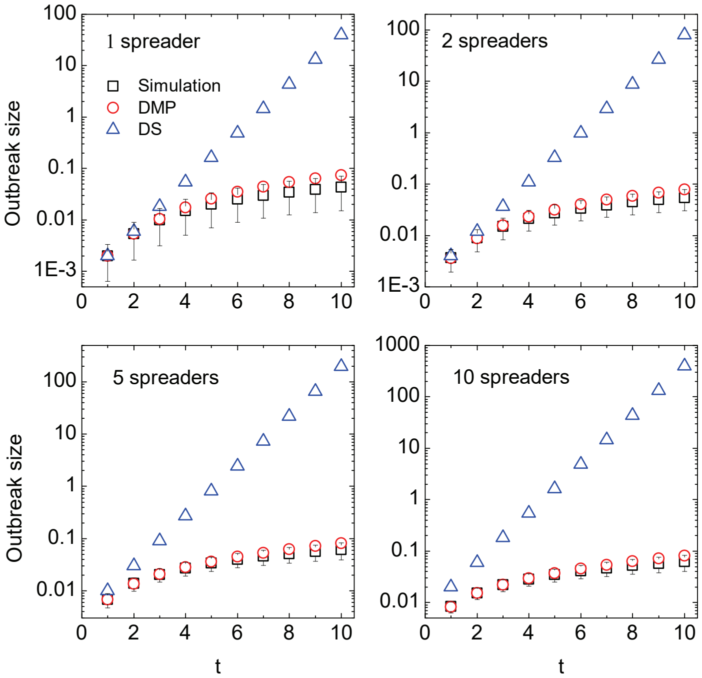

Figure 5.2: The performance of the DMP method and the DS centrality for evaluating the outbreak generated by both single spreader and multiple spreaders on regular network of $N = 1000$ and $\langle k \rangle = 20$ during the SIR spreading process with spreading rate $\beta = 0.1$. The symmetric bars indicate the fluctuations around the average value computed on 10^5 realizations of the stochastic process.

influence. For simplicity, we set $\mu = 1$ in SIR model and $\gamma = 1$ in SIS model. In this paper, all the analysis are based on the discrete-time dynamics.

5.3.3 Kendall's Tau

In this paper, we use the Kendall's tau to measure the correlation between the nodes' spreading influence and centrality measures (e.g., degree, eigenvector centrality and DMP method). For each node i, we denote y_i as its spreading influence and z_i as the target centrality measure, the accuracy of the target centrality in evaluating nodes' spreading influences can be quantified by the Kendall's Tau [162], as

$$\tau = \frac{2}{\sqrt{(n(n-1)/2 - n_1)(n(n-1)/2 - n_2)}} \sum_{i<j} \mathrm{sgn}[(y_i - y_j)(z_i - z_j)], \tag{5.3.1}$$

where $n_1 = \sum_i v_i(v_i - 1)/2$, v_i is the number of the i^{th} group of ties for the first quantity and $n_2 = \sum_j u_j(u_j - 1)/2$, u_j is the number of the j^{th} group of ties for the second quantity

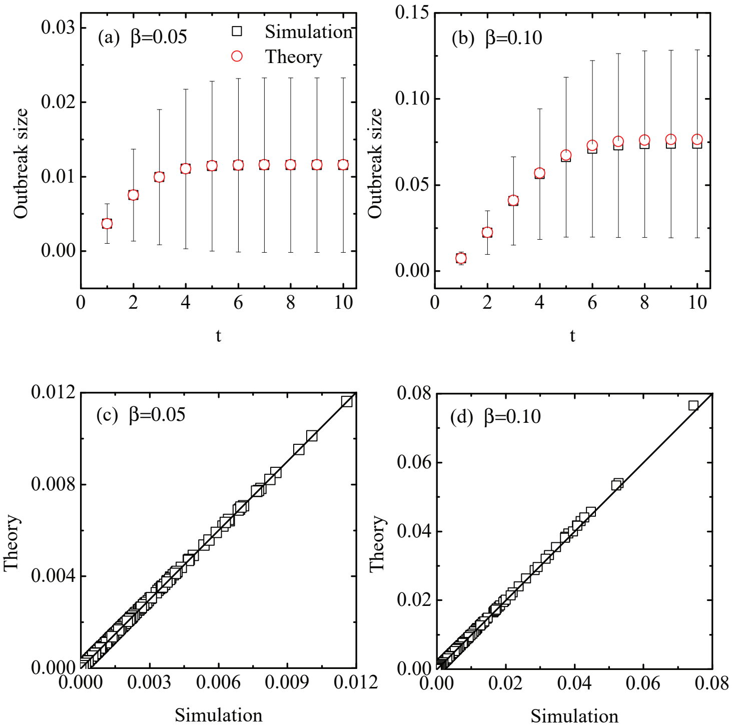

Figure 5.3: (Color online) The accuracy of the DMP method for evaluating nodes' spreading influence in a model network without loop of $N = 500$ and $\langle k \rangle = 8$ during the SIR spreading process. The subplot (a) and (b) show the outbreak size of node 1 in model network generated by DMP method at each time steps when the spreading rate β is 0.05 and 0.1 respectively. The subplot (c) and (d) show the outbreak size of all nodes in model network generated by the DMP method when spreading rate β is 0.05 and 0.1 respectively. The symmetric bars indicate the fluctuations around the average value computed on 10^5 realizations of the stochastic process.

and $\mathrm{sgn}(y)$ is a piecewise function: when $y > 0$, $\mathrm{sgn}(y) = +1$; $y < 0$, $\mathrm{sgn}(y) = -1$; when $y = 0$, $\mathrm{sgn}(y) = 0$. τ measures the correlation between two ranking lists, whose value is between $[-1, 1]$ and a larger τ corresponds to a better performance.

5.3.4 Numerical Result

Figure 5.2 shows the comparison between the DMP method and the DS method for evaluating the nodes' spreading influence on four empirical networks. The results suggest that the DMP method could evaluate the outbreak size for both single and multi spreaders more accurately than the DS method. One could easily observe that the theoretical results generated by DMP method are over estimated. The main reason of this overestimation is that the nodes

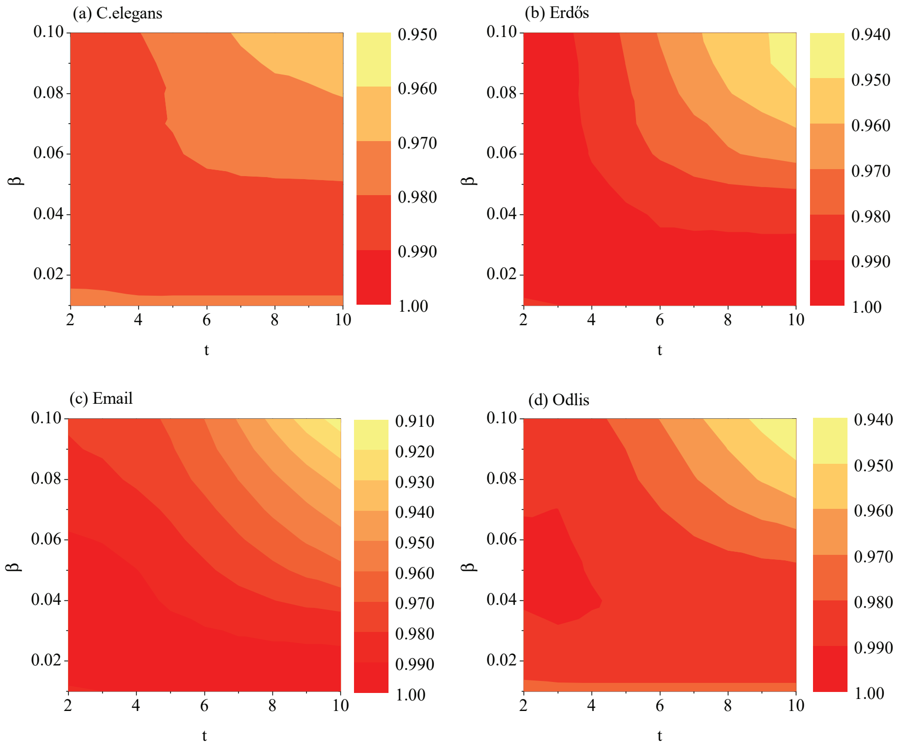

Figure 5.4: (Color online) The accuracy of the DMP method in evaluating nodes' spreading influences according to the standard SIR model in the four real networks, quantified by the Kendall's Tau. The spreading rate β varies from 0.01 to 0.10, Each data point is obtained by averaging over 10^5 independent runs.

infected by the initial node would infect themselves when time steps $t \geq 3$. One could then expect that a network without any loop would diminish this bias. As shown in Fig.5.3, in a network without loop, the DMP method could evaluate the nodes' the spreading scope accurately compared with the simulation result on network.

We also test the performance of the DMP method in ranking nodes' spreading influence during the spreading process on SIR model with different spreading rates β. The spreading influence of an arbitrary node i is quantified by the number of infected nodes and recovered nodes at t, where the spreading process starts with only node i being initially infected. Here the Kendall's tau τ is used to evaluate the correlation between the nodes' spreading influence and the centrality measures (DMP method, degree and eigenvector centrality), where τ is in the range $[-1, 1]$, and a larger value of τ indicates a better performance. As shown in Fig. 5.4, in all the cases, the values of τ of the DMP method is always between 0.915 and 1.0, which suggests that the ranking lists generated by the DMP method is almost the same as the ones generated by the simulation result.

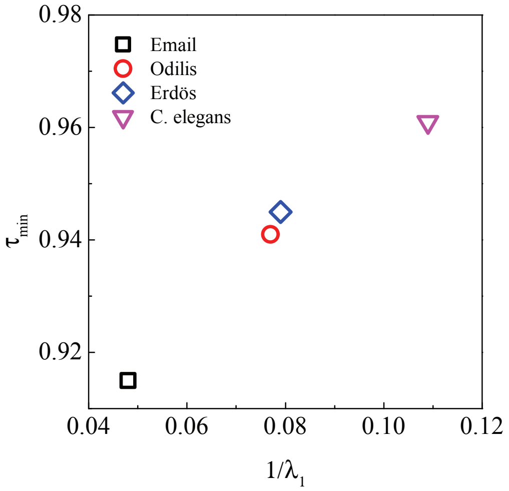

Figure 5.5: (Color online) The correlation between the minimum Kendall's tau τ_{min} and the inverse of the largest eigenvalue of the network $1/\lambda_1$.

Furthermore, one can find that the accuracy of the DMP method for ranking nodes' spreading influence is affected by the network structure. As shown in Fig. 5.4, the descent speed of τ in the C. elegans network are smaller than the ones in the Email network. To get deeper insight of how the network structure affect the performance of the DMP method, we analyze the correlation between the minimum Kendall's tau of the networks in Fig. 5.4 and the inverse of the largest eigenvalue of the network $1/\lambda_1$. The results are shown in Fig. 5.5. With increasing value of $1/\lambda_1$, the minimum Kendall's tau increases. For instance, in the C. elegans network, the reciprocal of the largest eigenvalue $1/\lambda_1$ is 0.109, which is significantly larger than that of the Email network (0.048). The fact that the DMP method performs particularly well in C. elegans network for ranking nodes' spreading influence indicates that the largest eigenvalue of the network is the main factor affecting the accuracy of the DMP method. A larger value of $1/\lambda_1$ would lead to a better performance of the DMP method.

In Fig.5.4, we have shown the spreading influence of nodes during the whole process. In the current experiment, we rank nodes' spreading influence in a special situation: the running time of the simulation is long enough such that there is not any infected node in the network. For each node, since we do not the exact convergent time step in the simulation, here we set a fix time step T^* as 5 in the DMP method. The results are shown in Fig. 5.6. The Kendall's tau τ of the DMP method is between 0.893 to 0.995, which indicates that the ranking lists generated by the DMP method and the SIR spreading process are almost the same. Compared with the degree and eigenvector centrality, the DMP method could locate the influential spreaders more accurately.

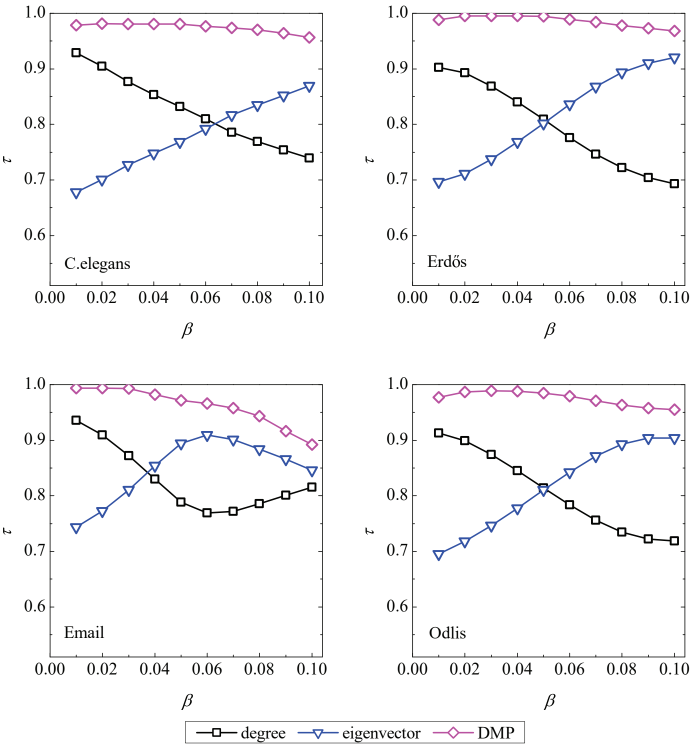

Figure 5.6: (Color online) The comparison among the DMP method, degree and eigenvector in evaluating nodes' spreading influences according to the standard SIR model with enough time steps until there is not any infected in the networks, quantified by the Kendall's Tau. The spreading rate β varies from 0.01 to 0.10, Each data point is obtained by averaging over 10^5 independent runs.

The DMP method can also be used to evaluate the outbreak size in SIS model. For simplicity, we set $\gamma = 1$ in SIS model. In this case, nodes in the network could be infected several times for high spreading rates β and long time steps t. We set the final time step t^* to 30. The results are shown in Fig. 5.7. The results are very similar to the ones in SIR model. The Kendall's tau τ of the DMP method is between 0.865 and 1.0. This indicates that the DMP method could also rank the nodes' spreading influence accurately in SIS model.

We have extended the DMP method to temporal networks. The results are shown in Fig. 5.8. The Kendall's tau τ is between 0.923 to 0.992, which indicates that the ranking lists generated by the DMP method and the real SIR model on temporal are highly identical to each other. Therefore, the DMP method could be used to detect the influential nodes in temporal network accurately.

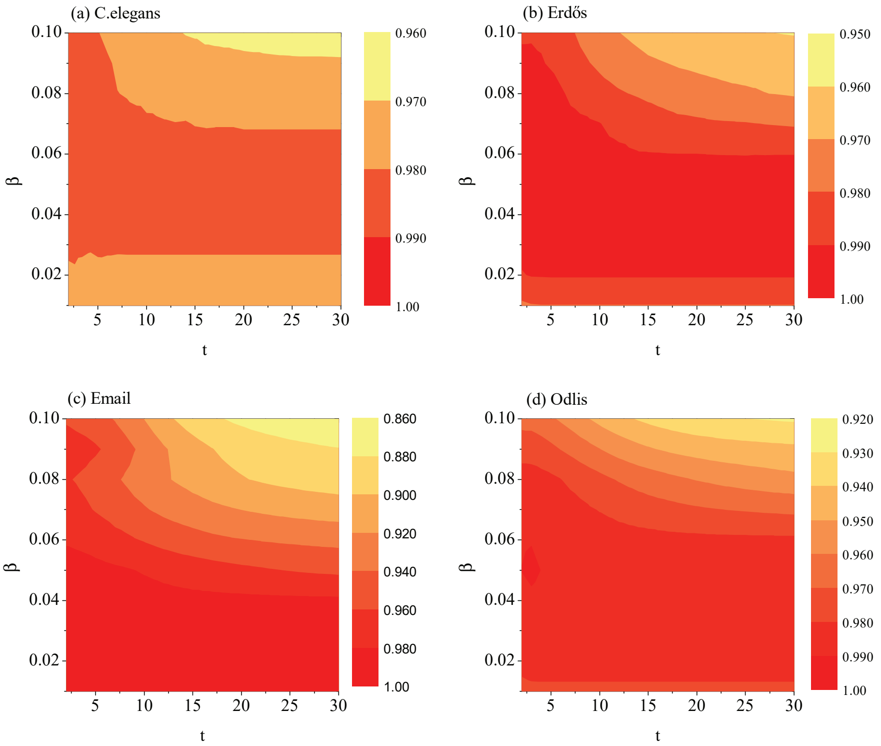

Figure 5.7: (Color online) The accuracy of the DMP method in evaluating nodes' spreading influences according to the standard SIS model in the four real networks, quantified by the Kendall's Tau. The spreading rate β varies from 0.01 to 0.10, Each data point is obtained by averaging over 10^5 independent runs.

5.4 Discussions

The essential question of ranking the node spreading influence is how to estimate the outbreak size of the initial spreader [163, 164]. To answer this question, one needs to fix the nonlinear coupling issue during the spreading process. In this paper we present a new method to evaluate the spreading scope from the perspective of Markov chain process, namely dynamics Markov process (DMP). This method solves the problem of nonlinear coupling by adjusting the state transition matrix, in which the elements of the matrix are the probabilities of nodes in susceptible state. The simulation results show that the DMP method could estimate the nodes' spreading scope at each time steps accurately in directed network without loop. Furthermore, according to the empirical results on four real networks, for both the SIR and SIS model, the ranking list generated by the DMP method is very close to the the ones of the simulation results, especially when the spreading rate and time step is small.

The DMP method could also be used to evaluate the nodes' spreading scope generated

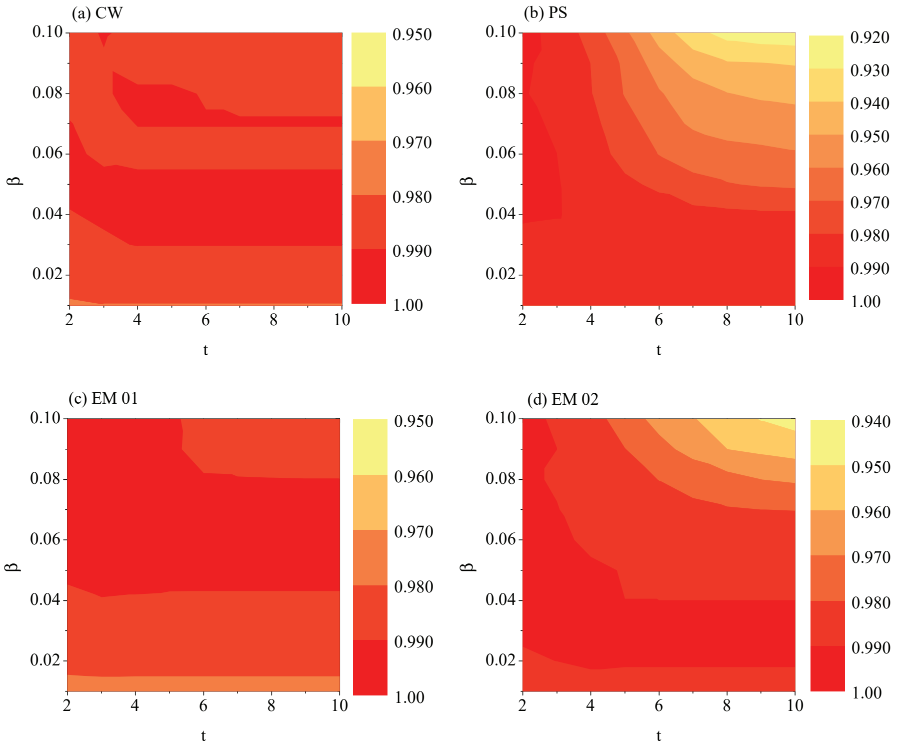

Figure 5.8: (Color online) The accuracy of the DMP method in evaluating nodes' spreading influences according to the standard SIR model in the four real temporal networks, quantified by the Kendall's Tau. The spreading rate β varies from 0.01 to 0.10, Each data point is obtained by averaging over 10^5 independent runs.

by multi-spreaders. Our simulation results indicate that when there exist multiple spreaders, the DMP method significantly outperforms the DS [33] centrality with increasing values of spreaders and time steps. The key to identifying multiple influential spreaders is to solve the overlap problem [165], which is the non-linear couple problem during the spreading process. Given the fact that the DMP method is a non-linear method, it will be able to identify multiple influential spreaders by using the greedy approach [166–168]. Moreover, the DMP method is also suitable for detecting the influential nodes in dynamic networks.

Comparing to the other methods in evaluating the outbreak size of the spreading dynamics, e.g., the Message-Passing Techniques [166] and the Percolation [169], the DMP method evaluates the spreading scope from the perspective of Markov process, and provides a general framework for ranking node spreading influence. Therefore. it can be extended and applied in modeling many other important dynamics such as Ising model [116], Boolean dynamics [170], voter model [171], synchronization [172], and so on.

Chapter 6

Nestedness maximization in complex networks through the fitness-complexity algorithm

Nestedness refers to the structural property of complex networks that the neighborhood of a given node is a subset of the neighborhoods of better-connected nodes. Following the seminal work by Patterson and Atmar (1986), ecologists have been long interested in revealing the configuration of maximal nestedness of spatial and interaction matrices of ecological communities. In ecology, the BINMATNEST genetic algorithm can be considered as the state-of-the-art approach for this task. On the other hand, the fitness-complexity ranking algorithm has been recently introduced in the economic complexity literature with the original goal to rank countries and products in World Trade export networks. Here, by bringing together quantitative methods from ecology and economic complexity, we show that the fitness-complexity algorithm is highly effective in the nestedness maximization task. More specifically, it generates matrices that are more nested than the optimal ones by BINMATNEST for 61.27% of the analyzed mutualistic networks. Our findings on ecological and World Trade data suggest that beyond its applications in economic complexity, the fitness-complexity algorithm has the potential to become a standard tool in nestedness analysis.

Based on Jian-Hong Lin, Claudio J. Tessone, and Manuel Sebastian Mariani. ”Nestedness maximization in complex networks through the fitness-complexity algorithm.” *Entropy* 20.10 (2018): 768.

6.1 Introduction

Network representations of complex interacting systems provide simple and powerful frameworks to characterize the topology of interactions and understand its impact on the emergence of collective phenomena [34, 173]. Some topological properties are found in a wide variety of real networks, which has led scholars to investigate possible interaction mecha-

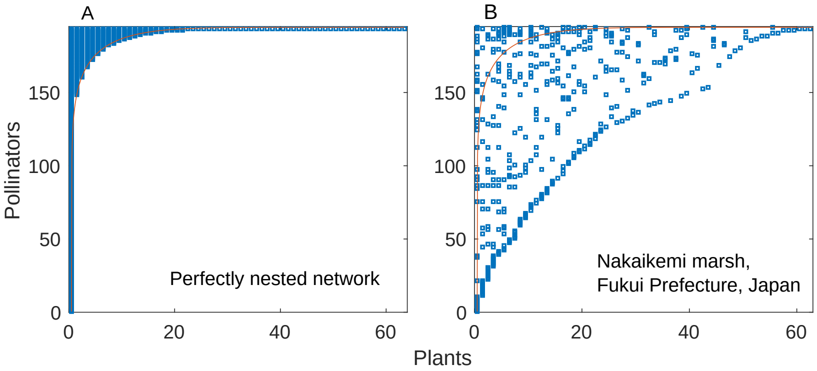

Figure 6.1: An illustration of the interaction matrix of a perfectly nested network as compared to the interaction matrix of a non-nested network (Nakaikemi marsh pollination network) composed of the same number of nodes and links. In a perfectly nested network (left panel), one can define a line (marked in red) that perfectly partitions the matrix into a filled region (i.e., the region above the line) and an empty region (i.e., the region below the line). The same feature does not hold for a non-nested network (right panel).

nisms behind their emergence. An example is the heavy-tailed distribution of the number of links per node (degree): its ubiquity has motivated the study of various network growth mechanisms that can generate networks with that property [173]. First conceived [174] and measured [37, 40] in biogeographic studies, *nestedness* [175] is one of such pervasive properties. In a perfectly nested bipartite network, the interaction partners of a given node are also partners of more generalist nodes. This property results in a ”triangular” shape of the network's interaction matrix (i.e., the binary matrix whose elements denote the presence or absence of a link, see Fig. 6.1).

While perfectly nested networks are unambiguously defined, they are also rarely found in real systems. However, many real networks exhibit a high degree of nestedness. The degree of nestedness of a bipartite network has not been uniquely defined in the literature [175]. In the widely-adopted definition by Atmar and Patterson [40], which is the one we consider here, a network is highly nested if the rows and columns of its interaction matrix can be ordered in such a way that one can find a line that separates almost perfectly the filled and empty regions of the matrix. It is essential to notice that this definition involves a reordering of the interaction matrix's rows and columns; alternative definitions of nestedness [176, 177] (not considered here) do not involve any matrix reordering.

Based on various metrics and definitions, nestedness has indeed been found in systems as diverse as spatial patterns of species distribution [37, 175], mutualistic plant-animal networks [178], manufacturer-contractor networks [179, 180], country-product export networks [181, 182], spatial patterns of firm distribution [181, 183], among others. The ubiquity of the pattern has naturally led scholars to investigate how nestedness relates to other network properties [184–186], which mechanisms can possibly explain its emergence in ecological [187–189] and socio-economic [179, 190, 190] networks, and its implications for the stability and feasibility of ecological systems [191, 192].

One of the most popular algorithms to quantify the degree of nestedness of a given network is the *Nestedness Temperature Calculator* [40]. Introduced by Atmar and Patterson in 1993 [40], the algorithm first determines a line of perfect nestedness by defining a perfectly nested interaction matrix with the same number of links as the original matrix. Then, it seeks to find the ranking of rows and columns that minimizes the average distance ("temperature" [40]) of observed "unexpected" matrix elements from the line of perfect nestedness – the unexpected matrix elements are those that are different from the corresponding ones in a perfectly nested matrix with the same number of links as the original matrix. Lower temperatures correspond to more nested topologies.

While the original Nestedness Temperature Calculator (NTC) by Atmar and Patterson [40] has been widely used in ecology [175], it exhibits some shortcomings that have been later overcome by the BINMATNEST algorithm [38]. BINMATNEST minimizes nestedness temperature through a genetic algorithm that confers higher chance to reproduce upon lower-temperature orderings [38]. The optimal matrices by BINMATNEST exhibit substantially lower temperature than those ranked by the NTC [38], which is why BINMATNEST can be considered as the state-of-the-art approach for nestedness temperature minimization in ecology.

Here, we explore an alternative approach to nestedness temperature minimization inspired by the recent Economic Complexity literature [41, 193]. Originally introduced to rank countries and products in the country-product export network [41], the fitness-complexity algorithm ranks the countries and products in such a way that the resulting incidence matrix exhibits a (typically imperfect) "triangular" shape [41, 193–195]. In World Trade, this suggests that the most competitive countries tend to diversify their export baskets, whereas the most sophisticated products can be only fabricated by the most competitive countries [41, 193]. The country score produced by the algorithm, referred to as country fitness, is positively correlated with country GDP per capita [41, 193]. Importantly, deviations from the linear-regressed trend are highly informative about the future economic development of the country [196, 197], resulting in GDP predictions often more accurate than those by the International Monetary Fund [198, 199].

The fact that matrices sorted according to the fitness-complexity algorithm exhibit a neater "triangular" shape than those sorted by degree [194] suggests that the algorithm might be competitive with algorithms typically adopted in ecology for nestedness temperature minimization [200]. The main goal of this article is to extensively compare the fitness-complexity algorithm and BINMATNEST according to their ability to minimize nestedness temperature. To this end, we analyze 142 mutualistic networks from http://www.web-of-life.es/ and 14 years of World Trade country-product networks from https://atlas.media.mit.edu/en/resources/data/. We compare the nestedness temperature of the matrices as ranked by BINMATNEST with those of the same matrices as ranked by the fitness-complexity algorithm.

We find that the fitness-complexity algorithm generates sorted matrices that exhibit a lower temperature than the optimal matrices by BINMATNEST for the 61.27% of the analyzed ecological networks. The only matrices where BINMATNEST outperforms substantially the fitness-complexity algorithm are low-size and high-density ones. The FCA is marginally outperformed by BINMATNEST for World Trade networks which exhibit higher density than mutualistic networks of similar size. Our findings suggest that while originally introduced as a ranking algorithm in economic production networks, the fitness-complexity algorithm has the potential to become a standard tool for nestedness detection in complex networks.

6.2 Materials and Methods

This paper focuses on binary bipartite networks. We label row-nodes (countries/pollinators) and column-nodes (products/plants) through Latin ($i \in \{1, \dots, N\}$) and Greek ($\alpha \in \{1, \dots, M\}$) letters, respectively. The total number of row-nodes and column-nodes is denoted as N and M, respectively, whereas the total number of links is denoted as L. The $N \times M$ network's *incidence matrix* [34] is denoted as B: its element $B_{i\alpha}$ is equal to one (”filled” element) if link (i, α) is observed, zero (”empty” element) otherwise. We refer to the incidence matrix of mutualistic networks as *interaction matrix* [178]. The density Φ of the network is defined as $\Phi = L/(M\,N)$.

6.2.1 Nestedness temperature minimization (NTM) problem

Nestedness temperature is determined through three steps: determination of the line of perfect nestedness, node ranking, and temperature calculation. We provide below the details of the three steps, and state the NTM problem.

First, to compute the nestedness temperature of a given matrix, one needs to determine its *line of perfect nestedness*. In this work, we use the definition provided by Rodríguez-Gironés and Santamaría [38] which overcomes some of the shortcomings of the original geometrical construction by Atmar and Patterson [40]. By rescaling the row and columns labels in such a way that they range from 0 to 1, the line of perfect nestedness is determined through the following shape function [38]

$$f(x;p) = \frac{0.5}{N} + \frac{N-1}{N}\left(1 - \left(1 - \frac{M\,x - 0.5}{M-1}\right)^p\right)^{1/p}. \tag{6.2.1}$$

This function depends on a single parameter, p, which is determined by imposing that the area above the curve in the interval $(0,1)$ equals the fill of the matrix Φ.

Second, matrix temperature depends on the order of rows and columns. The *nestedness temperature minimization (NTM) problem* (or, equivalently, the *nestedness maximization problem*) consists in determining the ranking of rows and columns that produces a ranked matrix of minimal temperature T (defined below). The output of this step is, therefore, a pair of rankings, one for rows and one for columns. Equivalently, we can say that the output of the ranking is a *ranked matrix*. Due to the large number of possible permutations of rows and columns, a combinatorial search is infeasible [38], which has motivated ecologists to search for fast ranking methods [38, 40, 201]. The main goal of this paper is to compare two alternative ranking algorithms, the one adopted by BINMATNEST (details in Section 6.2.2) and the fitness-complexity algorithm (details in Section 6.2.3).

Third, for a given network and a given ranking of its row-nodes and column-nodes, one calculates nestedness temperature T as follows. The unexpected elements of the ranked matrix are the the empty elements above and the filled elements below the line of perfect nestedness (as determined through Eq. (6.2.1)). We denote by $\mathcal{U}$ the set of unexpected elements. For each unexpected element (i,α), one draws a straight line of slope -1 in the interaction matrix (after having normalized to one the column and row labels, as described above). On this line, one compute the distance $d_{i\alpha}$ of unexpected element (i,α) from the line of perfect nestedness, and the distance $D_{i\alpha}$ between the intersection points of this line with the x-axis and y-axis (see Fig. 1 in [38] for an illustration). The total unexpectedness U of the ranked matrix is given by [38, 40]

$$U = \frac{1}{N\,M}\sum_{(i,\alpha)\in\mathcal{U}}\left(\frac{d_{i\alpha}}{D_{i\alpha}}\right)^2. \tag{6.2.2}$$

Matrix temperature is defined as $T = 100\,U/U_{max}$, where $U_{max} = 0.04145$ [38, 40]. A perfectly nested matrix has zero temperature (”perfect order” [40]), whereas random, noisy matrices have large temperature.

We stress that the key point in our analysis is that the calculation of nestedness temperature T requires a ranked matrix as input: different rankings of rows and columns lead to different matrix temperatures. This allows us to compare different ranking algorithms with respect to the nestedness temperature they produce. We expect the rankings by effective algorithms for NTM to produce ranked matrices that exhibit lower temperature than the ranked matrices by other algorithms.

6.2.2 Genetic algorithm approach: BINMATNEST (BIN)

The BINMATNEST algorithm [38] adopts a genetic-algorithm approach [202] to the NTM problem. As the computational steps of the ranking algorithm are detailed in [38], we only discuss here the main ideas behind the algorithm. The goal is to find a "solution" to the NTM problem, i.e., the minimal-temperature ranking of the nodes. The algorithm starts with a set of candidate solutions ("chromosomes" in the genetic-algorithm language [202]) – among these solutions, the rankings by degree and by the Nestedness Temperature Calculator by Atmar and Patterson [40]. In each generation, the algorithm considers a well-performing solution, and it generates an "offspring" solution o by probabilistically combining elements of the well-performing solution w with elements of a randomly-selected "partner" solution p.

More specifically, let us consider the ranking of the row-nodes. Given a well-performing solution $w = \{w_1, \dots, w_N\}$ and a partner solution $p = \{p_1, \dots, p_N\}$, the each element of the offspring solution is given by the corresponding element of w with probability $1/2$; otherwise, it is determined by the following steps:

- We randomly select an integer k between 1 and N.
- We set $o_i = w_i$ for $i \in \{1, \dots, k\}$.
- We set $o_i = p_i$ for $i \in \{k+1, \dots, N\}$, if and only if $p_i \notin \{w_1, \dots, w_k\}$.
- If $p_i \in \{w_1, \dots, w_k\}$, we assign one of the ranking positions that have not yet appeared in o to 0_i.

One applies the same steps to the ranking of the column-nodes. Besides, after these steps are performed, the offspring solution can undergo a mutation with a given probability (set to 0.1 in [38]). If the mutation happens, in the case of row-nodes, one extracts uniformly at random two integers $k_1, k_2 \in \{1, \dots, N\}$ ($k_1 < k_2$), and cyclically permutes the elements $\{o_{k_1}, \dots, o_{k_2}\}$. The process described above is iterated for a given number of generations, and the minimal-temperature solution is eventually selected to determine the network nestedness temperature.

The output of the BINMATNEST algorithm is therefore a ranking of the rows and columns that minimizes nestedness temperature T. Importantly, the optimal rankings by BINMATNEST lead to temperature values that are substantially lower than those determined by the widely-used Nestedness Temperature Calculator [40] – see Figs. 4-5 in [38], for example. Based on those results, BINMATNEST can be considered as the state-of-the-art

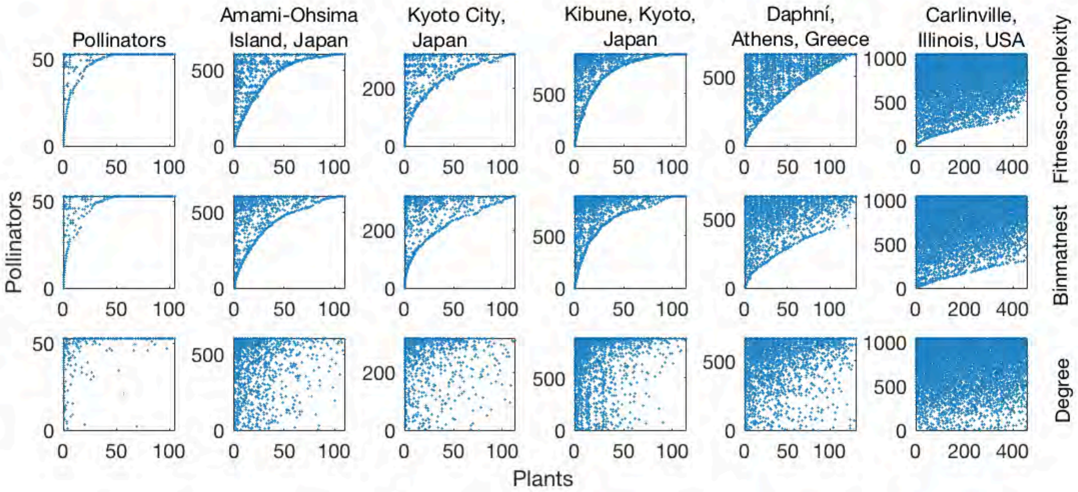

Figure 6.2: Six empirical mutualistic matrices of different density packed according to three different methods: fitness-complexity algorithm (top row), BINMATNEST (intermediate row), and degree (bottom row). The matrices ranked by fitness-complexity and BINMATNEST are significantly more nested than those ranked by degree.

approach for NTM in ecological networks. In this paper, we implement the BINMATNEST algorithm by using the function `nestedrank` from the R package `bipartite` with argument `method="binmatnest"`. This function gives as output the ranking of row-nodes and column-nodes by the BINMATNEST algorithm.

6.2.3 Non-linear iterative algorithms: Fitness-Complexity algorithm (FCA)

Originally introduced to rank countries and products in the bipartite country-product export network [41], the fitness-complexity algorithm has been applied to diverse systems including ecological mutualistic networks [200], knowledge production networks [203], food production networks [204]. In its formulation for countries and products [41], the algorithm aims to find a vector of "fitness" scores $F = \{F_i\}$ for countries and "complexity" scores $Q = \{Q_\alpha\}$ for products, respectively. The algorithm starts from a uniform initial condition [41]

$$\begin{aligned} F_i^{(0)} &= 1, \\ Q_\alpha^{(0)} &= 1, \end{aligned} \tag{6.2.3}$$

and it subsequently refines the fitness and complexity scores according to the following non-linear iterative equations:

$$\begin{aligned} \tilde{F}_i^{(n)} &= \sum_\alpha B_{i\alpha} Q_\alpha^{(n-1)}, \\ \tilde{Q}_\alpha^{(n)} &= \frac{1}{\sum_i B_{i\alpha}/F_i^{(n-1)}}. \end{aligned} \tag{6.2.4}$$

After each iterative step, the scores are normalized by their mean:

$$\begin{aligned} F_i^{(n)} &= \tilde{F}_i^{(n)} / \overline{\tilde{F}_i^{(n)}}, \\ Q_\alpha^{(n)} &= \tilde{Q}_\alpha^{(n)} / \overline{\tilde{Q}_\alpha^{(n)}}. \end{aligned} \tag{6.2.5}$$

Differently from widely-used spectral ranking algorithms (see [197] for a review), the second line of Eq. (6.2.4) is markedly non-linear. Such non-linearity is motivated by economic-complexity considerations. Empirical evidence indicates indeed that competitive countries tend to diversify their export baskets, which makes it reasonable to quantify the score of a given country as the sum over the scores of its exported products. At the same time, the fact that a product is exported by many countries (in particular, developing countries) suggests that the product might require few productive capabilities to be made and it is unlikely to be a sophisticated one. This motivates the non-linear dependence of product score $\tilde{Q}_\alpha^{(n)}$ on country score $F_i^{(n-1)}$: $\tilde{Q}_\alpha^{(n)}$ is heavily penalized if α is exported by a low-fitness country.

Do the iterations above converge to a unique fixed point? Scholars have found that while the answer is positive, the scores of several nodes can potentially converge to a zero value, which reduces the discriminative power of the ranking based on the fixed point of the map [205]. Besides, this convergence to zero tends to be relatively slow, and it strongly depends on the density and shape of the incidence matrix [195,205]. To prevent this potential issue, we adopt a convergence criterion based on ranking: we stop the iterations at step n^* if and only if the ranking of countries and products at step n^* is almost exactly the same as the ranking at step $n^* + \Delta n$, i.e., if few ranking variations occurred in the subsequent Δn steps. In practice, the stopping iteration n^* is defined as the smallest iteration such that both Spearman's correlation coefficients $\rho(F^{(n^*)}, F^{(n^*+\Delta n)})$ and $\rho(Q^{(n^*)}, Q^{(n^*+\Delta n)})$ are larger than $1 - 10^{-3}$. Unless otherwise stated, the results presented in this manuscript refer to $\Delta n = 10$ – the criterion allows us to stop the algorithm after a finite number of iteration for all the analyzed networks. We find that results for $\Delta n = 20$ and $\Delta n = 30$ are in qualitative agreement with those obtained with $\Delta n = 10$; the same holds for results obtained by running a fixed number $n^* = 100$ of iterations of the FCA – details are provided in the Results section.

While we formulated the algorithm for the country-product network, the algorithm can

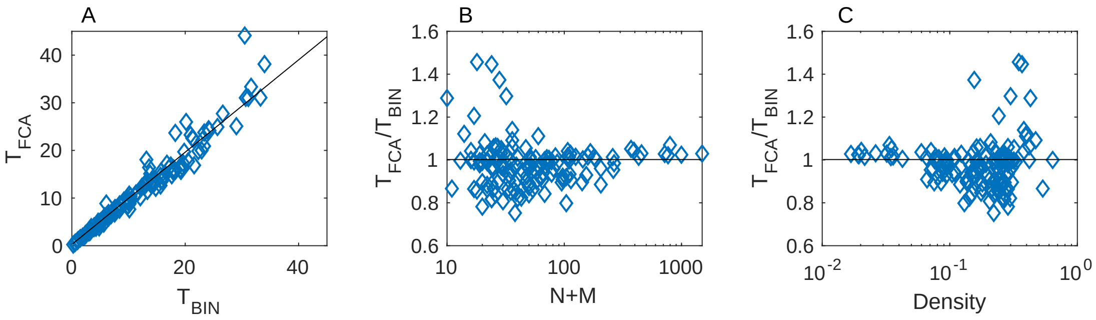

Figure 6.3: Results on mutualistic networks: a comparison of the nestedness temperature T_{FC} of the matrices ranked by the FCA with the nestedness temperature T_{BIN} of the optimal matrices found by the BINMATNEST genetic algorithm. The two temperatures are positively correlated (panel A), yet the temperature measured by the fitness-complexity algorithm is lower than that by BINMATNEST for the majority of analyzed networks. The only networks where BINMATNEST produces a substantially lower temperature ($T_{FCA}/T_{BIN} > 1$) are characterized by small size $N + M$ (panel B) and high density Φ (panel C).

be applied to any bipartite network by replacing ”countries” with the system’s row-nodes (e.g., animals in mutualistic networks [200]) and ”products” with the system’s column-nodes (e.g., plants). In this paper, we apply it not only to the country-product network, but also to mutualistic networks: the fitness score of animal and plant species represents their importance and vulnerability, respectively [200].

6.3 Results

6.3.1 Mutualistic networks

We analyzed the 142 pollination networks provided by The Web of Life (www.web-of-life.es) project. The species are plants (rows) and pollinators (columns) and the type of interaction is Pollination. The main goal of our paper is to compare the FCA and the BINMATNEST algorithm with respect to their performance in the NTM problem. Fig. 6.2 shows that qualitatively, the matrices produced by the fitness-complexity algorithm are substantially more nested than those produced by ranking the nodes by degree, and their nestedness might be comparable or even larger than that of the matrices ranked by BINMATNEST.

The reason why the FCA produces highly nested structures is that the score of a plant/product is mostly determined by the least-fit pollinator/country[1]: a plant/product that is pollinated/produced by a generalist pollinator/country – i.e., many pollinators/countries can pollinate/produce it – is heavily penalized and achieves a low complexity score Q; whereas a plant/product that is only pollinated/produced by specialist pollinator/country – i.e., few pollinators/countries can pollinate/produce it – attains a high complexity score. Hence, when sorting plants/products and pollinators/countries by the FCA, the plants/products are essentially ranked by the degree of generalization of their least-fit pollinators/exporters, which naturally results in a nested structure.

We now proceed in a more quantitative fashion by comparing, for all the analyzed empirical networks, the temperature values produced by the FCA with those by BINMATNEST. To do this, for the rankings determined by both methods, we determine the corresponding matrix temperature T according to Eq. (6.2.2). We find that while the temperature values achieved by the two methods are positively correlated (Fig. 6.3A), the temperature T_{FCA} by the FCA is lower than the temperature T_{BIN} by BINMATNEST for 61.27% of the networks. This result is stable with respect to variations in the convergence criterion adopted for the FCA[2].

The only matrices where the FCA is substantially outperformed by BINMATNEST are characterized by small size (Fig. 6.3B) and high density (Fig. 6.3C), yet these two properties seem necessary but not sufficient for BINMATNEST to outperform the FCA. Interestingly, among matrices that are found to be "colder" by the FCA, the lowest T_{FCA}/T_{BIN} ratio ($T_{FCA}/T_{BIN} = 0.75$) was observed in the M_PL_060_13 network ($N = 31, M = 7, L = 48$); in this dataset, $T_{BIN} = 10.15$ whereas $T_{FCA} = 7.64$. By contrast, among matrices that are found to be "colder" by BINMATNEST, the highest T_{FCA}/T_{BIN} ratio ($T_{FCA}/T_{BIN} = 1.46$) was observed in the M_PL_042 network ($N = 6, M = 12, L = 18$).

To deepen our understanding of the relation between the rankings by the FCA and BINMATNEST, we study their correlation and how such correlation depends on network properties. The Spearman's correlation coefficient [206] between the rankings by the two methods is positive and relatively high for both plants and pollinators (Fig. 6.4). Yet, as we have seen in Fig. 6.3, discrepancies between the two rankings point to a better ability of the FCA to "pack" the matrix in such a way that it displays a nested structure. The networks where we observe the largest discrepancies between the rankings by BINMATNEST

[1] Such dependence can be even sharpened by replacing $1/F^{(n)}$ with $(1/F^{(n)})^{\gamma}$ (with $\gamma > 0$) in the dependence of the complexity score on fitness score (second line of Eq. (6.2.4)) [194, 205], or by defining the complexity of a product directly as the minimum fitness of its interaction partners [195]. However, we do not explore these possibilities here.

[2] This result was obtained with $\Delta n = 10$. The fraction of datasets where $T_{FCA} < T_{BIN}$ is equal to 61.97% and 61.97% for $\Delta n = 20$ and $\Delta n = 30$, respectively. Besides, the same fraction was equal to 62.68% when using a fixed number $n^* = 100$ of iterations for all the networks. We conclude that the fraction of datasets where $T_{FCA} < T_{BIN}$ is not substantially affected by the adopted convergence criterion for the FCA.

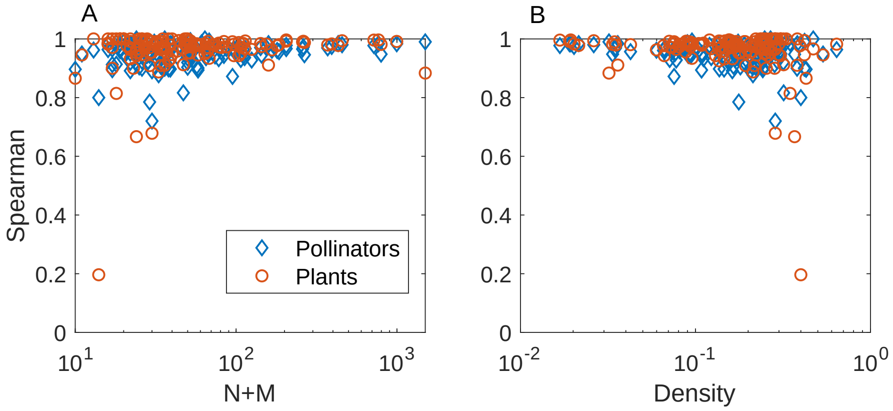

Figure 6.4: Results on mutualistic networks: Spearman's rank correlation coefficient ρ between the rankings by BINMATNEST and the fitness-complexity algorithm, for the rankings of pollinators (rhombuses) and plants (circles). Panels A and B represent ρ as a function of size $N + M$ and density Φ, respectively. The two methods produce highly-correlated rankings: the networks where we observe the lowest values of correlation are the small (panel A) and high-density ones (panel B).

and the FCA are the small and high-density ones – for example, the minimal observed correlation for the rankings of pollinators is $\rho = 0.20$, observed for one of the smallest networks [M_PL_069_02 which has $N = 4, M = 10, L = 16$]. All the other Spearman's coefficient values are above 0.67.

6.3.2 Country-product networks

We analyzed 14 years of World Trade data obtained from https://atlas.media.mit.edu/en/resources/data/. The raw data include information on which country exported which products to which countries, and the volume (measured in US dollars) of each trade relation. For each country-product pair (i, α), we denote by $w_{i\alpha}$ the volume of product α exported by country i. In line with the Economic Complexity literature [41, 193, 207], we construct a binary country-product network by only keeping the links between those country-product pairs such that $R_{i\alpha} \geq 1$, where $R_{i\alpha} := w_{i\alpha}/w_{i\alpha}$ is referred to as *revealed comparative advantage* [41], $w_{i\alpha} = w_i\, w_\alpha/W$ denotes the expected weight based on the total export volume $w_i := \sum_\beta w_{i\beta}$ of country i, the total export volume $w_\alpha := \sum_j w_{j\alpha}$ of product α, and the total export volume $W = \sum_{j\beta} w_{j\beta}$ in the system. In other words, a given country i is connected to a given product α in the bipartite country-product network if and only if the export volume $w_{i\alpha}$ exceeds the expected export volume. Based on this assumption, we construct 14 binary networks corresponding to the 2001-2014 period.

Fig. 6.5 compares the temperature by the FCA and BINMATNEST in the size-density

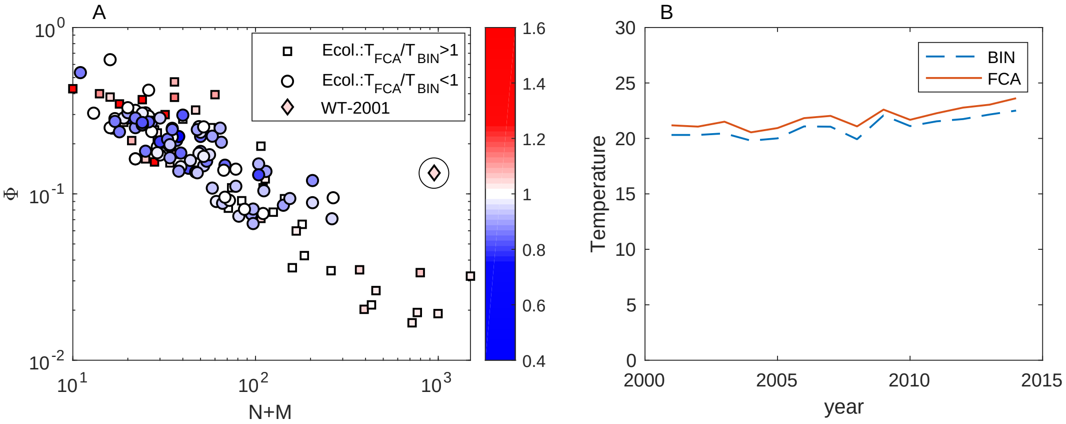

Figure 6.5: Results on mutualistic and World Trade networks. In panel A, each dot represents a network in the size-density plane; the dots' shape and color depend on the T_{FCA}/T_{BIN} ratio, in such a way that mutualistic networks with a ratio larger or smaller than one are represented by red squares or blue circles, respectively. This illustration confirms that the mutualistic networks where T_{FCA} is substantially larger than T_{BIN} are characterized by small size and high density. The World Trade network from 2001 (represented by the circled rhombus) exhibits relatively high density compared to mutualistic networks of comparable size; World Trade networks from other years (2002-2014) exhibit a similar size and density as the one from 2001, and they are not shown here. Panel B shows that the temperature T_{BIN} by BINMATNEST is marginally smaller than the temperature T_{FCA} by the FCA for all the analyzed years of World Trade, and the temperature values do not exhibit wide fluctuations over time.

plane, for all the analyzed mutualistic networks and the World Trade networks. The figure reveals that compared to the mutualistic networks analyzed above, the obtained country-product networks turn out to have a similar size as the largest mutualistic networks, but substantially larger density (see Fig. 6.5A). For all the analyzed World Trade networks, the temperature by BINMATNEST is marginally smaller than the one by the FCA, and both temperatures are stable over the years (see Fig. 6.5B): the average of T_{FCA}/T_{BIN} over the 14 analyzed years is equal to 1.04.

6.4 Discussion

We showed that the fitness-complexity ranking algorithm [41] is a highly effective method to ”pack” the incidence matrix of a given bipartite network in order to maximize its nestedness. In particular, an extensive comparison with BINMATNEST – the state-of-the-art nestedness maximization method in ecology – revealed that the FCA produces ranked matrices with temperature values substantially lower than those of the optimal matrices by BINMATNEST for the majority of analyzed datasets. Small-size and high-density ecological matrices are those where the rankings by the two methods differ the most, and where BINMATNEST has a chance to produce matrices of significantly smaller temperature than those ranked by the fitness-complexity algorithm.

Importantly, the Nestedness Temperature Minimization problem is not only a theoretical one, but it has also implications for the important problem of forecasting of the secondary effects of species' extinctions [200]. More specifically, recent works [194, 200] have pointed out that the rankings of active and passive species (countries and products, in World Trade analysis [194]) that result in the most packed matrices are also those that best reproduce the rankings of the nodes according to their structural importance and vulnerability (as determined by numerical simulations of ranking-based targeted attacks to the network). Maximizing nestedness is therefore highly informative on the structural importance of active species and vulnerability of passive species.

Finally, recent literature has reinterpreted nestedness as a mesoscopic property instead of a macroscopic one [186, 208, 209]. This means that nestedness can be interpreted not as a hierarchical organization of interactions between all pairs of nodes (as in Fig. 6.1), but as a property of subcomponents of the network. While our results show that the fitness-complexity algorithm can be used as a nestedness detection tool, whether it can be exploited (and arguably, generalized) to detect network compartments that exhibit an internal nested topology remains an intriguing open question.

Chapter 7

Conclusions

This thesis contributes to a better understanding of the structure and dynamics of financial, socio-economic, and ecological systems. Chapters 2 and 3 show that the Bitcoin ecosystem is evolving towards an increasingly centralised system from the binary and weighted versions. This is illustrated by the inequality of the bitcoin distribution in the BLN, where only about 10% of the nodes hold 80% of the bitcoins. And the averaged Gini coefficient of the binary and weighted centrality measures steadily increases throughout the entire BLN history. Furthermore, removing influential nodes leads to the collapse of the BLN into many components. In future research, we will first explore the mechanisms that induce the centralisation of the BLN. To this end, we will analyse the evolution of users' centrality in the BLN. On the other hand, we will explore the mechanism underlying the evolution of the BLN structure. Then, how the removal of important nodes and links impacts the failure rates of transitions is still an open question. To address this issue, we will propose a model for payment flow dynamic simulation on BLN and analyse the influence of important nodes and links. Finally, it is interesting to study the influence maximization in BLN, i.e. to locate a subset of important nodes which have a large influence on the structure and transition of the BLN.

The main contributions of Chapter 4 include three aspects. First, we develop an algorithm to generate non-normal networks based on the empirical evidence that nodes with a higher hierarchy are harder to be influenced. Second, we present a more realistic model for the transient explosive growth using an Ising-like model on non-normal networks. We show that non-normality leads to financial bubbles and crashes generically at subcriticality. This conceptually and operationally improves previous models aimed at anticipating critical phase transitions, which do not consider the non-normality of complex systems. Third, we reveal that in financial systems bubble size is directly controlled by the Kreiss constant of the non-normal matrix and bubble steepness is determined by the numerical abscissa of the non-normal matrix. In future works, our model can be extended to temporal networks and weighted networks. Furthermore, we will also analyse how non-normality affects other dynamical models in complex systems.

In Chapter 5, we present a dynamic Markov process to evaluate the outbreak size of nodes at a given time step, which can be directly applied in ranking the spreading influence of nodes. This method overcomes the problem of nonlinear coupling by calculating the probability of susceptible nodes being infected by their neighbours sequentially and adjusting the state transition matrix during the spreading process. Simulation results on SIR and SIS models show that the dynamic Markov process can rank the spreading influence of nodes accurately. In the future, we will extend our method to solve the influence maximization problem on complex networks, where influence maximization is an NP-hard problem to identify a subset of nodes to maximize the spread of influence.

In Chapter 6, we apply the fitness-complexity algorithm to maximize the nestedness of matrices. We reveal that matrices reordered by the fitness–complexity and BINMATNEST - the state-of-the-art nestedness maximization method in ecology - are substantially better "packed" than those reordered by degree centrality. Furthermore, the fitness-complexity algorithm can reorder matrices with substantially lower temperature values than those reordered by BINMATNEST. As a part of future work, we will apply the fitness complexity algorithm to detect network compartments that exhibit an internally nested topology, and extend the fitness-complexity algorithm to the nestedness of temporal and multilayer networks.

Bibliography

[1] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White. Economic networks: The new challenges. *Science*, 325(5939):422–425, 2009.

[2] D. Sornette and P. Cauwels. Financial bubbles: mechanisms and diagnostics. *Review of Behavioral Economics*, 2(3):279–305, 2015.

[3] N. Luhmann. *Social systems.* Stanford University Press, 1995.

[4] C. S. Holling. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, pages 1–23, 1973.

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[6] I. Gertsbakh and Y. Shpungin. *Network reliability and resilience.* Springer Science & Business Media, 2011.

[7] J. H. Lin, K. Primicerio, T. Squartini, C. Decker, and C. J. Tessone. Lightning network: a second path towards centralisation of the bitcoin economy. *New Journal of Physics*, 22(8):083022, 2020.

[8] J. H. Lin, E. Marchese, C. J. Tessone, and T. Squartini. The weighted bitcoin lightning network. *arXiv preprint arXiv:2111.13494*, 2021.

[9] D. Sornette, S. C. Lera, J. H. Lin, and K. Wu. Non-normal interactions create socio-economic bubbles. *Swiss Finance Institute Research Paper*, (22-43), 2022.

[10] J. H. Lin, Z. Yang, J. G. Liu, B. L. Chen, and C. J. Tessone. *preparing*, 2022.

[11] J. H. Lin, C. J. Tessone, and M. S. Mariani. Nestedness maximization in complex networks through the fitness-complexity algorithm. *Entropy*, 20(10):768, 2018.

[12] J. Poon and T. Dryja. The bitcoin lightning network: Scalable off-chain instant payments, 2016.

[13] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260, 2008.

[14] S. Wheatley, D. Sornette, T. Huber, M. Reppen, and R. N. Gantner. Are bitcoin bubbles predictable? combining a generalized metcalfe's law and the log-periodic power law singularity model. *Royal Society Open Science*, 6(6):180538, 2019.

[15] H. Vranken. Sustainability of bitcoin and blockchains. *Current Opinion in Environmental Sustainability*, 28:1–9, 2017.

[16] A. Urquhart. The inefficiency of bitcoin. *Economics Letters*, 148:80–82, 2016.

[17] M. A. Javarone and C. S. Wright. From bitcoin to bitcoin cash: a network analysis. In *Proceedings of the 1st Workshop on Cryptocurrencies and Blockchains for Distributed Systems*, pages 77–81, 2018.

[18] D. Kondor, M. Pósfai, I. Csabai, and G. Vattay. Do the rich get richer? an empirical analysis of the bitcoin transaction network. *PLoS ONE*, 9(2):e86197, 2014.

[19] A. Chauhan, O. P. Malviya, M. Verma, and T. S. Mor. Blockchain and scalability. In *2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 122–128. IEEE, 2018.

[20] I. A. Seres, L. Gulyás, D. A. Nagy, and P. Burcsi. Topological analysis of bitcoin's lightning network. In *Mathematical Research for Blockchain Economy*, pages 1–12. Springer, 2020.

[21] S. Martinazzi and A. Flori. The evolving topology of the lightning network: Centralization, efficiency, robustness, synchronization, and anonymity. *PLoS ONE*, 15(1):e0225966, 2020.

[22] S. Valverde and R. V. Solé. Self-organization versus hierarchy in open-source social networks. *Physical Review E*, 76(4):046118, 2007.

[23] M. Gupte, P. Shankar, J. Li, S. Muthukrishnan, and L. Iftode. Finding hierarchy in directed online social networks. In *Proceedings of the 20th international conference on World wide web*, pages 557–566, 2011.

[24] B. Corominas-Murtra, J. Goñi, R. V. Solé, and C. Rodríguez-Caso. On the origins of hierarchy in complex networks. *Proceedings of the National Academy of Sciences*, 110(33):13316–13321, 2013.

[25] F. Bodendorf and C. Kaiser. Detecting opinion leaders and trends in online social networks. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, pages 65–68. ACM, 2009.

[26] S. Pei and H. A. Makse. Spreading dynamics in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(12):P12002, 2013.

[27] R. Pastor-Satorras and A. Vespignani. Immunization of complex networks. *Physical Review E*, 65(3):036104, 2002.

[28] R. Cohen, S. Havlin, and D. Ben-Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24):247901, 2003.

[29] R. Sinatra, D. Wang, P. Deville, C. Song, and A. L. Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

[30] D. Wang, C. Song, and A. L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[31] M. Šikić, A. Lančić, N. Antulov-Fantulin, and H. Štefančić. Epidemic centrality?is there an underestimated epidemic impact of network peripheral nodes? *The European Physical Journal B*, 86(10):440, 2013.

[32] K. Klemm, M. Serrano, V. M. Eguíluz, and M. S. Miguel. A measure of individual role in collective dynamics. *Scientific Reports*, 2(1):1–8, 2012.

[33] J. G. Liu, J. H. Lin, Q. Guo, and T. Zhou. Locating influential nodes via dynamics-sensitive centrality. *Scientific Reports*, 6:21380, 2016.

[34] M. Newman. *Networks*. Oxford University Press, 2018.

[35] F. A. Rodrigues. Network centrality: an introduction. In *A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems*, pages 177–196. Springer, 2019.

[36] P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201, 2001.

[37] B. D. Patterson and W. Atmar. Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biological Journal of the Linnean Society*, 28(1-2):65–82, 1986.

[38] M. A. Rodríguez-Gironés and L. Santamaría. A new algorithm to calculate the nestedness temperature of presence–absence matrices. *Journal of Biogeography*, 33(5):924–935, 2006.

[39] M. S. Mariani, Z. M. Ren, J. Bascompte, and C. J. Tessone. Nestedness in complex networks: observation, emergence, and implications. *Physics Reports*, 813:1–90, 2019.

[40] W. Atmar and B. D. Patterson. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia*, 96(3):373–382, 1993.

[41] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero. A new metrics for countries' fitness and products' complexity. *Scientific Reports*, 2:723, 2012.

[42] S. Bartolucci, F. Caccioli, and P. Vivo. A percolation model for the emergence of the bitcoin lightning network. *Scientific Reports*, 10(1):1–14, 2020.

[43] E. Rohrer, J. Malliaris, and F. Tschorsch. Discharged payment channels: Quantifying the lightning network's resilience to topology-based attacks. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 347–356. IEEE, 2019.

[44] M. E. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[45] M. E. Newman, A. L. Barabási, and D. J. Watts. *The structure and dynamics of networks.* Princeton University Press, 2006.

[46] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[47] S. P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005.

[48] M. E. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.

[49] R. Pfitzner, I. Scholtes, A. Garas, C. J. Tessone, and F. Schweitzer. Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks. *Physical Review Letters*, 110(19):198701, 2013.

[50] P. Bonacich. Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555–564, 2007.

[51] J. Morgan. The anatomy of income distribution. *The Review of Economics and Statistics*, pages 270–283, 1962.

[52] P. Crucitti, V. Latora, and S. Porta. Centrality measures in spatial networks of urban streets. *Physical Review E*, 73(3):036125, 2006.

[53] J. Park and M. E. Newman. The statistical mechanics of networks. *Physical Review E*, 70:066117, 2004.

[54] T. Squartini and D. Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 13(8):083001, 2011.

[55] V. J. L. de Jeude, G. Caldarelli, and T. Squartini. Detecting core-periphery structures by surprise. *Europhysics Letters*, 125:68001, 2019.

[56] M. E. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[57] S. Lee and H. Kim. On the robustness of lightning network in bitcoin. *Pervasive and Mobile Computing*, 61:101108, 2020.

[58] Y. Guo, J. Tong, and C. Feng. A measurement study of bitcoin lightning network. In *2019 IEEE International Conference on Blockchain (Blockchain)*, pages 202–211. IEEE, 2019.

[59] A. Mizrahi and A. Zohar. Congestion attacks in payment channel networks. In *International Conference on Financial Cryptography and Data Security*, pages 170–188. Springer, 2021.

[60] M. Conoscenti, A. Vetrò, and J. C. De Martin. Hubs, rebalancing and service providers in the lightning network. *IEEE Access*, 7:132828–132840, 2019.

[61] C. Decker. Lightning network research; topology datasets. `https://github.com/lnresearch/topology`. Accessed: 2020-10-01.

[62] B. S. Srinivasan and L. Lee. Quantifying decentralization. *https://news.earn.com/quantifying-decentralization-e39db233c28e*, 2017.

[63] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[64] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001.

[65] L. A. N. Amara, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. In *The structure and dynamics of networks*, pages 207–210. Princeton University Press, 2011.

[66] E. Marchese, G. Caldarelli, and T. Squartini. Detecting mesoscale structures by surprise. *Communications Physics*, 5(1):1–16, 2022.

[67] F. Parisi, T. Squartini, and D. Garlaschelli. A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks. *New Journal of Physics*, 22:053053, 2020.

[68] N. Vallarano, M. Bruno, E. Marchese, G. Trapani, F. Saracco, T. Squartini, G. Cimini, and M. Zanon. Fast and scalable likelihood maximization for exponential random graph models. *arXiv preprint arXiv:2101.12625*, 2021.

[69] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[70] Y. A. Kuznetsov. *Elements of applied bifurcation theory*, volume 112. Springer, 1998.

[71] D. Sornette. Critical phenomena in natural sciences (chaos, fractals, self-organization and disorder: Concepts and tools). *Springer Series in Synergetics, Heidelberg*, 2004.

[72] M. Scheffer. Critical transitions in nature and society. *Princeton University Press*, 2009.

[73] J. C. Rocha, G. Peterson, Ö. Bodin, and S. Levin. Cascading regime shifts within and across scales. *Science*, 362(6421):1379–1383, 2018.

[74] T. M. Bury, R.I. Sujith, I. Pavithran, M. Scheffer, T. M. Lenton, M. Anand, and C. T. Bauch. Deep learning for early warning signals of tipping points. *Proceedings of the National Academy of Sciences*, 118(39), 2021.

[75] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, and T. A. Driscoll. Hydrodynamic stability without eigenvalues. *Science*, 261(5121):578–584, 1993.

[76] M. Embree and L. N. Trefethen. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators.* Princeton University Press Princeton, 2005.

[77] B. K. Murphy and K. D. Miller. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.

[78] T. Biancalani, F. Jafarpour, and N. Goldenfeld. Giant amplification of noise in fluctuation-induced pattern formation. *Physical Review Letters*, 118(1):018101, 2017.

[79] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Positive feedback investment strategies and destabilizing rational speculation. *the Journal of Finance*, 45(2):379–395, 1990.

[80] J. Y. Campbell, A. W. Lo, A. C. MacKinlay, and R. F. Whitelaw. The econometrics of financial markets. *Macroeconomic Dynamics*, 2(4):559–562, 1998.

[81] M. P. Scholl, A. Calinescu, and J. D. Farmer. How market ecology explains market malfunction. *Proceedings of the National Academy of Sciences*, 118(26), 2021.

[82] T. Lux. Herd behaviour, bubbles and crashes. *The Economic Journal*, 105(431):881–896, 1995.

[83] R. Cont and J. P. Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(2):170–196, 2000.

[84] J. P. Bouchaud and M. Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge University Press, 2003.

[85] D. Sornette. *Why stock markets crash: critical events in complex financial systems*, volume 49. Princeton University Press, 2nd printing, 2017.

[86] R. N. Mantegna and H. E. Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge University Press, 1999.

[87] S. Galam. Sociophysics: A review of Galam models. *International Journal of Modern Physics C*, 19(03):409–440, 2008.

[88] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010.

[89] M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. Van de Koppel, I. A. Van de Leemput, S. A. Levin, E. H. Van Nes, and M. Pascual. Anticipating critical transitions. *Science*, 338(6105):344–348, 2012.

[90] S. Battiston, J. D. Farmer, A. Flache, D. Garlaschelli, A. G. Haldane, H. Heesterbeek, C. Hommes, C. Jaeger, R. May, and M. Scheffer. Complexity theory and financial regulation. *Science*, 351(6275):818–819, 2016.

[91] V. H. Sridhar, L. Li, D. Gorbonos, M. Nagy, B. R. Schell, T. Sorochkin, N. S. Gov, and I. D. Couzin. The geometry of decision-making in individuals and collectives. *Proceedings of the National Academy of Sciences*, 118(50), 2021.

[92] A. D. Sánchez, J. M. López, and M. A. Rodriguez. Nonequilibrium phase transitions in directed small-world networks. *Physical Review Letters*, 88(4):048701, 2002.

[93] F. W. S. Lima and J. A. Plascak. Kinetic models of discrete opinion dynamics on directed barabási–albert networks. *Entropy*, 21(10):942, 2019.

[94] M. Asllani, R. Lambiotte, and T. Carletti. Structure and dynamical behavior of non-normal networks. *Science Advances*, 4(12):eaau9403, 2018.

[95] J. D. O'Brien, K. A. Oliveira, J. P. Gleeson, and M. Asllani. Hierarchical route to the emergence of leader nodes in real-world networks. *Physical Review Research*, 3(2):023117, 2021.

[96] T. Kaizoji, M. Leiss, A. Saichev, and D. Sornette. Super-exponential endogenous bubbles in an equilibrium model of fundamentalist and chartist traders. *Journal of Economic Behavior & Organization*, 112:289–310, 2015.

[97] R. Westphal and D. Sornette. Market impact and performance of arbitrageurs of financial bubbles in an agent-based model. *Journal of Economic Behavior & Organization*, 171:1–23, 2020.

[98] H. P. Minsky. Monetary systems and accelerator models. *The American Economic Review*, 47(6):860–883, 1957.

[99] S. Keen. Emergent macroeconomics: deriving Minsky's financial instability hypothesis directly from macroeconomic definitions. *Review of Political Economy*, 32(3):342–370, 2020.

[100] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591, 2009.

[101] M. Gisler, D. Sornette, and R. Woodard. Innovation as a social bubble: The example of the human genome project. *Research Policy*, 40:1412–1425, 2011.

[102] E. Samanidou, E. Zschischang, D. Stauffer, and T. Lux. Agent-based models of financial markets. *Reports on Progress in Physics*, 70(3):409, 2007.

[103] A. G. Haldane and A. E. Turrell. Drawing on different disciplines: macroeconomic agent-based models. *Journal of Evolutionary Economics*, 29(1):39–66, 2019.

[104] G. Dosi and A. Roventini. More is different... and complex! the case for agent-based macroeconomics. *Journal of Evolutionary Economics*, 29(1):1–37, 2019.

[105] T. Ott, P. Masset, T. S. Gouvea, and A. Kepecs. Apparent sunk cost effect in rational agents. *Science Advances*, 8(6):eabi7004, 2022.

[106] M. G. Neubert and H. Caswell. Alternatives to resilience for measuring the responses of ecological systems to perturbations. *Ecology*, 78(3):653–665, 1997.

[107] S. Nicoletti, D. Fanelli, N. Zagli, M. Asllani, G. Battistelli, T. Carletti, L. Chisci, G. Innocenti, and R. Livi. Resilience for stochastic systems interacting via a quasi-degenerate network. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(8):083123, 2019.

[108] G. Baggio, V. Rutten, G. Hennequin, and S. Zampieri. Efficient communication over complex dynamical networks: The role of matrix non-normality. *Science Advances*, 6(22):eaba2282, 2020.

[109] S. Johnson. Digraphs are different: Why directionality matters in complex systems. *Journal of Physics: Complexity*, 1(1):015003, 2020.

[110] M. Kawakatsu, P. S. Chodrow, N. Eikmeier, and D. B. Larremore. Emergence of hierarchy in networked endorsement dynamics. *Proceedings of the National Academy of Sciences*, 118(16), 2021.

[111] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[112] S. C. Lera, A. Pentland, and D. Sornette. Prediction and prevention of disproportionally dominant agents in complex networks. *Proceedings of the National Academy of Sciences*, 117(44):27090–27095, 2020.

[113] S. Johnson and N. S. Jones. Looplessness in networks is linked to trophic coherence. *Proceedings of the National Academy of Sciences*, 114(22):5618–5623, 2017.

[114] C. Pilgrim, W. Guo, and S. Johnson. Organisational social influence on directed hierarchical graphs, from tyranny to anarchy. *Scientific Reports*, 4388(10), 2020.

[115] Z. Q. Jiang, W. X. Zhou, D. Sornette, R. Woodard, K. Bastiaensen, and P. Cauwels. Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 chinese stock market bubbles. *Journal of Economic Behavior & Organization*, 74(3):149–162, 2010.

[116] D. Sornette. Physics and financial economics (1776–2014): puzzles, Ising and agent-based models. *Reports on Progress in Physics*, 77(6):062001, 2014.

[117] A. Hüsler, D. Sornette, and C. H. Hommes. Super-exponential bubbles in lab experiments: evidence for anchoring over-optimistic expectations on price. *Journal of Economic Behavior & Organization*, 92:304–316, 2013.

[118] J. Miao and P. Wang. Asset bubbles and credit constraints. *American Economic Review*, 108(9):2590–2628, 2018.

[119] Š. Lyócsa, E. Baumöhl, and T. Vỳrost. YOLO trading: Riding with the herd during the GameStop episode. *Finance Research Letters*, page 102359, 2021.

[120] L. Lucchini, L. M. Aiello, L. Alessandretti, G. D. F. Morales, M. Starnini, and A. Baronchelli. From Reddit to Wall Street: the role of committed minorities in financial collective action. *Royal Society Open Science*, 9(211488), 2022.

[121] A. Betzer and J. P. Harries. How online discussion board activity affects stock trading: the case of gamestop. *Financial Markets and Portfolio Management*, pages 1–30, 2022.

[122] Z. G. Nicolaou, T. Nishikawa, S. B. Nicholson, J. R. Green, and A. E. Motter. Non-normality and non-monotonic dynamics in complex reaction networks. *Physical Review Research*, 2(4):043059, 2020.

[123] M. Scheffer, J. Bascompte, W. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara. Anticipating critical transitions. *Nature*, 461:53–59, 2009.

[124] I. A. Van de Leemput, M. Wichers, A. O. J. Cramer, D. Borsboom, F. Tuerlinckx, P. Kuppens, Egbert H. van Nes, W. Viechtbauer, E. J. Giltay, S. H. Aggen, et al. Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, 111(1):87–92, 2014.

[125] J. Ma, Y. Xu, Y. Li, R. Tian, and J. Kurths. Predicting noise-induced critical transitions in bistable systems. *Chaos*, 29(8):081102, 2019.

[126] L. Walras. *Elements of pure economics.* Routledge, 1954.

[127] J. Jackson. How the 'SaveAMC' campaign caused the movie theater company's stocks to soar. Newsweek, 2021. Retrieved January 28, 2021.

[128] M. Egkolfopoulou. Stock Investors Are Hunting for the Next GameStop on Reddit and Twitter. Bloomberg News. Retrieved January 27, 2021.

[129] M. Fitzgerald. Bed Bath & Beyond, AMC rally with GameStop as little investors squeeze hedge funds in more stocks. CNBC. Retrieved January 27, 2021.

[130] T. Kilgore. Nokia's stock soars toward a record gain on record volume, for no apparent reason. MarketWatch. Retrieved January 27, 2021.

[131] L. Muchnik, S. Pei, L. C. Parra, S. D. S. Reis, J. S. Andrade J., S. Havlin, and H. A. Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific Reports*, 3(1):1–8, 2013.

[132] P. L. Krapivsky and S. Redner. A statistical physics perspective on web growth. *Computer Networks*, 39(3):261–276, 2002.

[133] T. Zhou, Z. Fu, and B. Wang. Epidemic dynamics on complex networks. *Progress in Natural Science*, 16(5):452–457, 2006.

[134] M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals.* Princeton University Press, 2008.

[135] J. O. Kephart, G. B. Sorkin, D. M. Chess, and S. R. White. Fighting computer viruses. *Scientific American*, 277(5):88–93, 1997.

[136] A. E. Motter. Cascade control and defense in complex networks. *Physical Review Letters*, 93(9):098701, 2004.

[137] D. Li, B. Fu, Y. Wang, G. Lu, Y. Berezin, H. E. Stanley, and S. Havlin. Percolation transition in dynamical traffic network with evolving critical bottlenecks. *Proceedings of the National Academy of Sciences*, 112(3):669–672, 2015.

[138] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[139] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[140] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[141] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.

[142] J. G. Liu, Z. M. Ren, and Q. Guo. Ranking the spreading influence in complex networks. *Physica A: Statistical Mechanics and its Applications*, 392(18):4154–4159, 2013.

[143] Z. M. Ren, A. Zeng, D. B. Chen, H. Liao, and J. G. Liu. Iterative resource allocation for ranking spreaders in complex networks. *EPL (Europhysics Letters)*, 106(4):48005, 2014.

[144] J. H. Lin, Q. Guo, W. Z. Dong, L. Y. Tang, and J. G. Liu. Identifying the node spreading influence with largest k-core values. *Physics Letters A*, 378(45):3279–3284, 2014.

[145] L. Lü, T. Zhou, Q. M. Zhang, and H. E. Stanley. The h-index of a network node and its relation to degree and coreness. *Nature Communications*, 7:10168, 2016.

[146] A. J. Alvarez-Socorro, G. C. Herrera-Almarza, and L. A. González-Díaz. Eigencentrality based on dissimilarity measures reveals central nodes in complex networks. *Scientific Reports*, 5:17095, 2015.

[147] G. Lawyer. Understanding the influence of all nodes in a network. *Scientific Reports*, 5:8665, 2015.

[148] F. D. Malliaros, M. E. G. Rossi, and M. Vazirgiannis. Locating influential nodes in complex networks. *Scientific Reports*, 6, 2016.

[149] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[150] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200, 2001.

[151] K. Ide, R. Zamami, and A. Namatame. Diffusion centrality in interconnected networks. *Procedia Computer Science*, 24:227–238, 2013.

[152] B. L. Chen, W. X. Jiang, Y. X. Chen, L. Chen, R. J. Wang, S. Han, J. H. Lin, and Y. C. Zhang. Influence blocking maximization on networks: Models, methods and applications. *Physics Reports*, 976:1–54, 2022.

[153] I. Scholtes, N. Wider, R. Pfitzner, and C. J. Tessone. Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nature Communications*, 5:5024, 2014.

[154] R. Pfitzner, I. Scholtes, A. Garas, C. J. Tessone, and F. Schweitzer. Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks. *Physical Review Letters*, 110(19):198701, 2013.

[155] R. A. Hom and C. R Johnson. Matrix analysis. *Cambridge University Express*, 1985.

[156] M. H. DeGroot and M. J. Schervish. Probability and statistics, 2012.

[157] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68(6):065103, 2003.

[158] M. Génois, C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3(3):326–347, 2015.

[159] V. Gemmetto, A. Barrat, and Cattuto C. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infectious Diseases*, 14(1):695, 2014.

[160] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J. F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176, 2011.

[161] A. Paranjape, R. Benson, A., and J. Leskovec. Motifs in temporal networks. *In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610, 2017.

[162] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[163] D. B. Chen, H. L. Sun, Q. Tang, S. Z. Tian, and M. Xie. Identifying influential spreaders in complex networks by propagation probability dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(3):033120, 2019.

[164] D. Bucur and P. Holme. Beyond ranking nodes: Predicting epidemic outbreak sizes by network centralities. *PLOS Computational Biology*, 16(7):e1008052, 2020.

[165] X. Wang, X. Zhang, D. Yi, and C. Zhao. Identifying influential spreaders in complex networks through local effective spreading paths. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(5):053402, 2017.

[166] F. Altarelli, A. Braunstein, L Dall'Asta, J. R. Wakeling, and R. Zecchina. Containing epidemic outbreaks by message-passing techniques. *Physical Review X*, 4(2):021024, 2014.

[167] F. Morone and H. A. Makse. Influence maximization in complex networks through optimal percolation. *Nature*, 524(7563):65–68, 2015.

[168] L. Guo, J. H. Lin, Q. Guo, and J. G. Liu. Identifying multiple influential spreaders in term of the distance-based coloring. *Physics Letters A*, 380(7):837–842, 2016.

[169] Y. Hu, S. Ji, Y. Jin, L. Feng, H. E. Stanley, and S. Havlin. Local structure can identify and quantify influential global spreaders in large scale social networks. *Proceedings of the National Academy of Sciences*, 115(29):7468–7472, 2018.

[170] S. A. Kaufmann. The origins of order, 1993.

[171] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591, 2009.

[172] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou. Synchronization in complex networks. *Physics Reports*, 469(3):93–153, 2008.

[173] A. L. Barabási and M. Pósfai. *Network science*. Cambridge University Press, 2016.

[174] P. J. Darlington. *Zoogeography*. John Wiley: New York, 1957.

[175] W. Ulrich, M. Almeida-Neto, and N. J. Gotelli. A consumer's guide to nestedness analysis. *Oikos*, 118(1):3–17, 2009.

[176] M. Almeida-Neto, P. Guimaraes, P. R. Guimaraes Jr, R. D. Loyola, and W. Ulrich. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, 117(8):1227–1239, 2008.

[177] P. P. Staniczenko, J. C. Kopp, and S. Allesina. The ghost of nestedness in ecological networks. *Nature Communications*, 4:1391, 2013.

[178] J. Bascompte, P. Jordano, C. J. Melián, and J. M. Olesen. The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences*, 100(16):9383–9387, 2003.

[179] S. Saavedra, F. Reed-Tsochas, and B. Uzzi. A simple model of bipartite cooperation for ecological and organizational networks. *Nature*, 457(7228):463, 2009.

[180] S. Saavedra, D. B. Stouffer, B. Uzzi, and J. Bascompte. Strong contributors to network persistence are the most vulnerable to extinction. *Nature*, 478(7368):233, 2011.

[181] S. Bustos, C. Gomez, R. Hausmann, and C. A. Hidalgo. The dynamics of nestedness predicts the evolution of industrial ecosystems. *PLoS ONE*, 7(11):e49393, 2012.

[182] F. Saracco, R. Di Clemente, A. Gabrielli, and T. Squartini. Detecting early signs of the 2007–2008 crisis in the world trade. *Scientific Reports*, 6, 2016.

[183] A. Garas, C. Rozenblat, and Schweitzer F. Economic specialization and the nested bipartite network of city-firm relations. *Multiplex and Multilevel Networks*, pages 74–83, 2019.

[184] S. Johnson, V. Domínguez-García, and M. A. Muñoz. Factors determining nestedness in complex networks. *PLoS ONE*, 8(9):e74025, 2013.

[185] S. H. Lee. Network nestedness as generalized core-periphery structures. *Physical Review E*, 93(2):022306, 2016.

[186] A. Solé-Ribalta, C. J. Tessone, M. S. Mariani, and J. Borge-Holthoefer. Revealing in-block nestedness: Detection and benchmarking. *Physical Review E*, 97(6):062302, 2018.

[187] S. Suweis, F. Simini, J. R. Banavar, and A. Maritan. Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature*, 500(7463):449–452, 2013.

[188] S. Valverde, J. Piñero, B. Corominas-Murtra, J. Montoya, L. Joppa, and R. Solé. The architecture of mutualistic networks as an evolutionary spandrel. *Nature Ecology & Evolution*, 2(1):94, 2018.

[189] D. S. Maynard, C. A. Serván, and S. Allesina. Network spandrels reflect ecological assembly. *Ecology Letters*, 21(3):324–334, 2018.

[190] M. D. König and C. J. Tessone. Network evolution based on centrality. *Physical Review E*, 84(5):056108, 2011.

[191] S. Allesina and S. Tang. Stability criteria for complex ecosystems. *Nature*, 483:205–208, 2012.

[192] R. P. Rohr, S. Saavedra, and J. Bascompte. On the structural stability of mutualistic systems. *Science*, 345(6195):1253497, 2014.

[193] M. Cristelli, A. Gabrielli, A. Tacchella, G. Caldarelli, and L. Pietronero. Measuring the intangibles: A metrics for the economic complexity of countries and products. *PLoS ONE*, 8(8):e70726, 2013.

[194] M. S. Mariani, A. Vidmer, M. Medo, and Y. C. Zhang. Measuring economic complexity of countries and products: which metric to use? *The European Physical Journal B*, 88(11):293, 2015.

[195] R. J. Wu, G. Y. Shi, Y. C. Zhang, and M. S. Mariani. The mathematics of non-linear metrics for nested networks. *Physica A: Statistical Mechanics and its Applications*, 460:254–269, 2016.

[196] M. Cristelli, A. Tacchella, and L. Pietronero. The heterogeneous dynamics of economic complexity. *PLoS ONE*, 10(2):e0117174, 2015.

[197] H. Liao, M. S. Mariani, M. Medo, Y. C. Zhang, and M. Y. Zhou. Ranking in evolving complex networks. *Physics Reports*, 689:1–54, 2017.

[198] M. C. A. Cristelli, A. Tacchella, M. Z. Cader, K. I. Roster, and L. Pietronero. On the predictability of growth. *World Bank Policy Research Working Paper*, (8117), 2017.

[199] A. Tacchella, D. Mazzilli, and L. Pietronero. A dynamical systems approach to gross domestic product forecasting. *Nature Physics*, 14(8):861–865, 2018.

[200] V. Domínguez-García and M. A. Muñoz. Ranking species in mutualistic networks. *Scientific Reports*, 5:8182, 2015.

[201] P. R. Guimarães and P. Guimaraes. Improving the analyses of nestedness for large sets of matrices. *Environmental Modelling & Software*, 21(10):1512–1513, 2006.

[202] D. Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4(2):65–85, 1994.

[203] G. Cimini, A. Gabrielli, and F. S. Labini. The scientific competitiveness of nations. *PLoS ONE*, 9(12):e113470, 2014.

[204] C. Tu, J. Carr, and S. Suweis. A data driven network approach to rank countries production diversity and food specialization. *PLoS ONE*, 11(11):e0165941, 2016.

[205] E. Pugliese, A. Zaccaria, and L. Pietronero. On the convergence of the fitness-complexity algorithm. *The European Physical Journal Special Topics*, 225(10):1893–1911, 2016.

[206] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[207] C. A. Hidalgo and R. Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, 2009.

[208] A. Grimm and C. J. Tessone. Detecting nestedness in graphs. In *International Workshop on Complex Networks and their Applications*, pages 171–182. Springer, 2016.

[209] A. Grimm and C. J. Tessone. Analysing the sensitivity of nestedness detection methods. *Applied Network Science*, 2(1):37, 2017.