Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
COPPE UFRJ

TRANSFER ENTROPY NETWORKS TO RANK INFLUENCE OVER TIME WITH APPLICATION TO THE BRAZILIAN STOCK MARKET

José de Paula Neves Neto

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Daniel Ratton Figueiredo

Rio de Janeiro
Agosto de 2023

TRANSFER ENTROPY NETWORKS TO RANK INFLUENCE OVER TIME WITH APPLICATION TO THE BRAZILIAN STOCK MARKET

José de Paula Neves Neto

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientador: Daniel Ratton Figueiredo

Aprovada por: Prof. Daniel Ratton Figueiredo
Prof. Ana Paula Couto da Silva
Prof. Antonio Augusto de Aragão Rocha
Prof. Franklin de Lima Marquezino
Prof. Gerson Zaverucha

RIO DE JANEIRO, RJ – BRASIL
AGOSTO DE 2023

Neves Neto, José de Paula

Transfer Entropy Networks to Rank Influence over Time with Application to the Brazilian Stock Market/José de Paula Neves Neto. – Rio de Janeiro: UFRJ/COPPE, 2023.

XIV, 69 p.: il.; 29, 7cm.

Orientador: Daniel Ratton Figueiredo

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2023.

Referências Bibliográficas: p. 65 – 69.

1. transfer entropy. 2. information transfer. 3. network science. 4. network centrality. 5. stock market. 6. dynamic networks. I. Ratton Figueiredo, Daniel. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

To my father Zezé (in memoriam) and my mother Mari (in memoriam), who inspired me to walk this path, and to my more than beloved children Mariana, Lucas and Thiago, who were the light of the lighthouse that I always followed.

Thanks

I would like to thank many people who were important in many aspects of my life during the walking on my doctoral study, mainly because I underwent some unfortunate events in my private life, which often made me discredit myself.

I was the first from a poor family to reach the University, the first to obtain a master's degree and until the moment the only one to reach a doctorate. This is thanks to my father and mother, both now deceased, who, despite all the great difficulties, gave me the best instruments to get me here: shelter, food, health, education and love. Thanks Dad. Thank you mom.

My children, although adults, are always my babies, and will always be. They bring me sense to life. I love them and almost everything I do, I do it for them. I waited a very long time to begin my doctoral study to let them grow and I am proud of them, of what they became to be.

As nowadays I work at Cuiabá since 1997, my mother's share family made my life sweeter as I came back to Rio de Janeiro, my home city. I love each one of them.

During my course I have discovered many good friends to whom I have joined in the dance activity at UFRJ, that became my exhaust valve and a pleasure. I thank them for the companionship and learning. There are many names here, but I would like to mention specialy Beatriz Machado, Larissa Fernandes, Joyce Carias, Jonatha Gomes. And beyond University walls, thanks to Lívia Veras, Márcia Gravina and Ariadne do Couto.

Same way I would like to thank two dance families: Jaime Arôxa Tijuca, through Lídio Freitas and Monique Marculano, and Artance Danças, through Fernando Schellenberg and Nayara Melo. I made many friends in those houses.

I would like to thank two very special women, friends: Márcia Moura Silva (*in memoriam*) and Jandira Ferreira Novais Marmello. And although they had their own issues in life, even if they were far away, they took their time and mind to support me and that helped me a lot. My love and my affection.

And without quite expecting it, on that journey, in the environment of my doctoral study, I found who would become a wonderful companion, of peace, affection, quietness and requited love: Joyce Carias (yes, the same one above). Thinking of her gave me and gives me a lot of strength to keep on going.

My colleagues in the LAND lab and COPPE in general took an important role in my learning. I found knowledge and inspiration on them. They always showed me there is no limit to learn. Remarkably I would like to thank Giulio Iacobelli, Joanna Manjarres, David Brilhante, Fábio David and Jefferson Simões who were always closer to me, not only for academic reasons. And here, a very special thank to beloved Carolina Vieira (our Carol)!

Special thanks to my advisor, Daniel Ratton Figueiredo, who is one of the best I could have, technically and personally. He was the best human being I could have crossed paths with in at this point of my life and he always lifted me up. In fact, he is an inspiration and I thank him very much.

I thank my academic institution, Universidade Federal do Mato Grosso, for its support during my leave for the doctoral study, supporting me with my salaries. And, in particular, I would like to my former student, my current colleague and now my boss Allan Gonçalves de Oliveira, for his support in many situations.

And finally, thanks God. I trust Him.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

TRANSFER ENTROPY NETWORKS TO RANK INFLUENCE OVER TIME WITH APPLICATION TO THE BRAZILIAN STOCK MARKET

José de Paula Neves Neto

August/2023

Advisor: Daniel Ratton Figueiredo

Department: Systems Engineering and Computer Science

Transfer entropy is a relatively recent technique to measure the information transfer or influence between two random events represented by their time series. However, in many scenarios a set of events can mutually influence one another, both directly and indirectly. This thesis proposes a methodology to characterize influence in such scenarios by building transfer entropy networks (weighted, directed) and using classic network centrality metrics to determine top-ranked influential and influenced nodes (events). The methodology is applied to the Brazilian stock market exchange using 32 years of historical data (daily stock prices) to build a single network as well as a sequence of networks over time (from overlapping time windows). Among the many findings, the static analysis indicates that influence is not moderately correlated with classic financial indicators (e.g., stock volume), and that some stocks are both influential and influenced. Moreover, top-ranked influential and influenced stocks are both very dynamic, in contrast to stocks that are top-ranked according to classic financial indicators (e.g., stock volume). Last, a simple analytical model to predict the mutual information between two events as well as the transfer entropy is also proposed.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

José de Paula Neves Neto

Agosto/2023

Orientador: Daniel Ratton Figueiredo

Programa: Engenharia de Sistemas e Computação

A entropia de transferência é uma técnica relativamente recente para medir a transferência de informações ou influência entre dois eventos aleatórios representados por suas séries temporais. No entanto, em muitos cenários, um conjunto de eventos podem se influenciar mutuamente, tanto direta quanto indiretamente. Esta tese propõe uma metodologia para caracterizar a influência em tais cenários através da construção de redes de entropia de transferência (arestas ponderadas, direcionadas) e usando métricas clássicas de centralidade de rede para determinar os nós (eventos) influentes e influenciados mais bem classificados. A metodologia é aplicada ao mercado de ações brasileiro usando 32 anos de dados históricos (cotações diárias das ações) para construir uma única rede, bem como uma sequência de redes ao longo do tempo (a partir de janelas de tempo com sobreposição). Entre as muitas descobertas, a análise estática indica que a influência não está nem moderadamente correlacionada com indicadores financeiros clássicos (por exemplo, volume monetário) e que algumas ações são influentes e influenciadas ao mesmo tempo. Além disso, as ações influentes e influenciadas mais bem classificadas são muito dinâmicas, em contraste com as ações mais bem classificadas de acordo com indicadores financeiros clássicos (por exemplo, volume monetário). Por fim, também é proposto nesa tese um modelo analítico simples para prever a informação mútua entre dois eventos, bem como a entropia de transferência.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Influence is a fundamental concept in nature and society as all objects both exert and suffer some kind of influence from one another. Human intuition says that climate influences crop or the unemployment rate influences salary, for example. Although it is an intuitive concept, a fundamental question arises already at the outset: what is influence? Influence can be vaguely defined as the change in the dynamics of an object in response to the change of dynamics in another object. This work will address this central question and characterize very influential and very influenced objects.

There are many situations in which the action of an object clearly affects the behavior of another. This kind of relationship between objects is what we will refer to as *influence* or *dependence*. Influence can also be seen as *information*, in particular when one object influences another, one can say that the first object has passed information to the second one. Thus, influence can be intuitively interpreted as *information transfer*. An important question is determining if the dynamics of two objects are related in some kind of *dependence* or *influence*. For example, does heart beat rate influence blood pressure? Does the population size of an insect influence a given animal disease? Does the price of a share in the stock market influence the price of another share? Does Facebook friendship influence WhatsApp contacts? In any case, we are only able to answer these questions if we can measure and compare the dynamics of such events over time.

In order to measure influence, we consider objects that are associated with some kind of time series over the same time period, so that we can "compare" the time series of two objects in order to measure the amount of influence that one object has over the other. In order to measure influence or *information transfer*, and not simply correlation between two time series, an important notion is that of *present* and *future*. Influence comes after the fact. Information passed in the present affects the future. This will also induce a direction in the relationship, as the present of an object may influence the future of another, but not necessarily vice-versa.

Transfer entropy [1, 2] is a relatively recent metric proposed to measure the influence of one object over another from their corresponding time series. Thus, it embodies all of the above requirements and it has been applied to a myriad of scenarios, and in particular to financial markets (see discussion in Chapter 2. The concept of transfer entropy is the cornerstone of this thesis as well as its application to the Brazilian stock market exchange. However, different from most prior works, the main objective of this thesis is to characterize and identify top-ranked stocks according to the influence they exert or suffer in the stock market.

Thus, the main focus is characterizing the influence between shares in the stock market exchange. Is there influence between pairs of shares? Can such pairwise influence be used to determine the most influential and the most influenced shares in the stock market? To achieve this goal, we consider time series to represent the price evolution of shares in the Brazilian stock market. In fact, the entire 32 years of historical data with daily stock prices negotiated in the Brazilian stock market exchange is considered in this work. Using a daily price change reference, we build for each stock a time series of symbols representing the direction of price change (up, down, no change, no negotiated). We then measure the information transfer between a single pair of symbols time series using the measure for transfer entropy. Finally, using the information transfer among all pairs of shares, we build a directed weighted network where nodes are shares and edges have weights with transfer entropy representing information transfer. Thus, nodes of this network can be ranked using classic network centrality metrics to assess their importance, as discussed in Chapter 3.

The previous analysis comprises a very long period of time (32 years) and the vast majority of shares (nodes) are removed from the analysis due to low liquidity in the period (i.e., not traded on most days). Moreover, the time varying dynamics of the stock market intuitively will lead to a time varying influence among the stocks. In order to characterize the influence over time, the proposed methodology is applied to a sequence of overlapping time windows that lead to a sequence of networks over time, as discussed in Chapter 4.

The empirical computation of transfer entropy from time series can require significant computational effort depending on the number of symbols and length of the time series. Thus, models that can predict the transfer entropy from simple parameters extracted from the time series can be very useful. Chapter 5 proposes a simple model to predict the mutual information between two random variables. The model has two parameters and a simple analytical expression. This model can also be used to predict the transfer entropy in a simplified model. This model can be of independent interest, beyond the problem of characterizing influence.

1.1 Objective

So, this work has two main objectives:

1. Define a metric to gather influence, giving it a number;

2. Apply that metodology over long historical time series.

In order to do that we will use concepts as transfer entropy, network centrality measures, time series and node ranking.

1.2 Main contributions

The main contributions of this work are summarized as follows:

1. A methodology to characterize influence (influential and influenced) from empirical time series. This methodology uses the notion of transfer entropy to build a weighted directed network that is then used to rank nodes according to the influence they exert or suffer in the network. This work was published in SBC WPerformance 2018 and received the *Best Paper Award* [3].

2. An application of the proposed methodology to the Brazilian stock market using historical data comprising of 32 years. The original stock price time series was converted into a price variation time series and filtered for liquidity before applying the proposed methodology. Results indicate several findings concerning influence in the Brazilian stock market and its correlation with financial indicators. Part of this work was published in SBC WPerformance 2018 and received the *Best Paper Award* [3].

3. An application of the proposed methodology over time to the Brazilian stock market comprising of a sequence of 65 networks (time windows). The characterization of influence over time indicates that top influential and influenced nodes are highly dynamic, among several other findings. This work was published in Physica A in 2023 [4].

4. A simple model to predict the mutual information for two random variables. The model has two parameters and yields an analytical expression that can be easily computed. The model is also related to a simplified model for predicting the transfer entropy between three random variables. This work has yet to be published.

1.3 Organization

The remainder of this dissertation is organized as follows:

- Chapter 2 presents important concepts and related works concerning information transfer, network centrality (ranking), and stock market analysis.
- Chapter 3 presents the proposed methodology for characterizing influence using transfer entropy networks. The methodology is applied to the Brazilian stock market to determine top influential and influenced stocks of that market.
- Chapter 4 applies the proposed methodology to characterize influence over time in the Brazilian stock market. Using a sliding window methodology, 65 different entropy networks are constructed and analyzed to characterize influential and influenced stocks over time.
- Chapter 5 presents a simple and parsimonious model to predict the mutual information between two random variables, as well as its adaptation to predict transfer entropy.
- Chapter 6 presents a brief conclusion and outlines potential future work.

Chapter 2

Background and Related Work

The idea of measuring time series dependency is old, and there are many ways to measure dependency and information. Possibly the most widely accepted notion is that of *entropy*, introduced by Shannon [5].

2.1 Shannon Entropy

Before one aims measuring dependente between two time series, it is imperative to gauge each time series by itself.

One way to do that is to use the notion of *entropy* introduced by Shannon [2, 5]. In his work, Shannon grounded the basis of *information theory* and the way to store and transmit information through a noisy channel without loss of information.

His work is in the basis of modeling and building communication systems, as he has proved mathematical limit of how well a receiver is able to identify what data was generated by transmitter through a channel, with applications, for example in the Ethernet (IEEE 802.3), the Wi-Fi (IEEE 802.11) or cell phone systems (FDMA, GSM, TDMA, CDMA, UMTS, LTE, NR *etc*).

2.1.1 Definition

Given a random variable I, with probability distribution $p(i)$, the Shannon entropy H_I is defined as

$$H_I = -\sum_{i \in D(I)} p(i) \log_2 p(i) \tag{2.1}$$

where the sum extends over all possible values i for I, as $D(I)$ is the image of random variable I. The higher the value of H_I the more information it carries out. In particular, the maximum value of H_I is achieved for a uniform random variable.

2.2 Mutual Information

When dealing with two time series simultaneously a useful measure for their mutual dependence is the *mutual information* [6, 7].

We can see it as the amount of information that can be obtained about one random variable by observing the other random variable.

2.2.1 Definition

The mutual information M_{IJ} of two random variables I and J, with distribution $p(i)$ and $p(j)$ respectively, and joint-probability $p(i,j)$ is defined as

$$M_{IJ} = \sum_{\substack{i \in D(I) \\ j \in D(J)}} p(i,j) \log_2 \frac{p(i,j)}{p(i)p(j)} \tag{2.2}$$

Mutual information does not fit our goals because its result does not consider the direction of the relationship between the two elements.

2.3 Transfer entropy

To account for these directed relationship between two objects we use the concept of transfer entropy.

Transfer entropy was introduced by Schreiber in [1]. In his article, Schreiber approached the subject of measuring information transfer between two systems represented by their time series. He was interested in understanding how individual values produced by a system could guide information production so that another system were sensitized in a probable way.

Schreiber's work is based on the concept of *entropy* of a random variable and *mutual information* [5]. The novelty of Schreiber's proposal with respect to mutual information is the introduction of a time delay in one of the series.

Bossomaier *et al.* [2] make a compilation of transfer entropy and some fundamental concepts about this subject.

2.3.1 Definition

Unlike other methods for measuring coherence or correlation (relationship in general) between two time series, transfer entropy has a directional methodology, through introduction of a time lag between time series, which compares an event (a value) in a time series to an event shifted in the other time series.

In [1] Schreiber defines transfer entropy $T_{J \to I}$ as

$$T_{J\to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log \frac{p(i_{n+1} \mid i_n^{(k)}, j_n^{(l)})}{p(i_{n+1} \mid i_n^{(k)})} \tag{2.3}$$

Where:

- I and J are numerical series with same length N;
- i_{n+1} is an element of I, for $n \in \{1, 2, \ldots, N-1\}$;
- $i_n^{(k)}$ and $j_n^{(l)}$ are segments, respectively, from I (with lenght k) and J (with lenght l), backwards starting in n. Then, $i_n^{(k)} = i_{n-k+1}, \ldots, i_{n-1}, i_n$ and $j_n^{(k)} = j_{n-l+1}, \ldots, j_{n-1}, j_n$;
- $p(\cdot)$ is the joint-probability density function.

For computational reasons, we only consider transfer entropy where J and I are discrete random variables, assuming a finite number of symbols, and we consider the case in which the segments $i_n^{(k)}$ and $j_n^{(l)}$ have both length 1 ($k = l = 1$). Under these assumptions and using the definition of conditional probability Equation (2.3) can be rewritten as

$$\mathrm{T}_{J\to I} = \sum p(i_{n+1}, i_n, j_n) \log \frac{p(i_{n+1}, i_n, j_n) \cdot p(i_n)}{p(i_{n+1}, i_n) \cdot p(i_n, j_n)} \tag{2.4}$$

where summation runs over all possible values for i_{n+1}, i_n, j_n.

2.3.2 Computing Transfer Entropy

Consider two time series I and J, both with length N. In order to apply Equation 2.3 to empirical data the (joint and conditional) probabilities of $p(i_+, i, j)$, $p(i_+ \mid i, j)$ and $p(i_+ \mid i)$, where $i_+, i \in I$ and $j \in J$, must be properly estimated. This is done using the time series to compute the relative frequency of the observed values.

To ease calculation, those probabilities are transformed into their versions without conditionals, yielding only the (joint) empirical probabilities $p(i)$, $p(i_+, i)$, $p(i, j)$ and $p(i_+, i, j)$. Thus, the number of occurrences for each instance (single value, pair of values and triple values) is counted across the two time series to determine the corresponding relative frequencies, which are used as their probabilities.

Finally, the transfer entropy from J to I is determined by the summation described below, running over all values contained in I and J series.

$$T_{J\to I} = \sum_{\substack{i_{n+1},\, i_n \in D(I) \\ j_n \in D(J)}} p(i_{n+1}, i_n, j_n) \log_2 \frac{p(i_{n+1}, i_n, j_n) \cdot p(i_n)}{p(i_{n+1}, i_n) \cdot p(i_n, j_n)} \tag{2.5}$$

where $n \in 1, 2, \ldots, N-1$ and i_k, j_k are the k-th values into I and J series, respectively.

As discussed ahead in Section 3.1.2, values for i_{n+1}, i_n and j_n are discrete and can assume only four different symbols, instead of a larger number of values or even a continuous scale. Note that the space of possible combinations of values among these three variables grows exponentially in the number of symbols, and can thus hamper estimation of their respective probabilities and, consequently, lead to a noisy computation of the transfer entropy.

2.3.3 Applications of transfer entropy

Since its introduction, transfer entropy has proved useful in several applications.

Market networks can be defined as the network of equities created by enterprises and their relationships within the financial market. Market networks take advantage of network analysis tools and metrics which are leveraged to analyze the market, understand relationship between enterprises, support forecasting and portfolio strategies [8, 9].

There are various works on market network analysis [10, 11], market network measurements [12], market network correlations [13–16], stock market network forecasting [17] and financial modeling [18]. For example, Tabak et al. [10] considers a market network where nodes are shares in a stock market and undirected edge weights denote a distance based on cross-correlation between the prices time series. In particular, correlation values of $+1$, 0 and -1 were mapped into distances of 0, 1.4142 and 2, respectively, which are then used to compute the Minimal Spanning Tree (MST) of the network.

Approaches based on transfer entropy have recently increased and are being addressed in different scenarios, including financial markets [1, 2, 19]. In particular, there are recent studies on information transfer between stock markets (of different countries) using transfer entropy methods [19–23], including the use of variations like *Effective Transfer Entropy*, *Phase Transfer Entropy*, *Rényi's Transfer Entropy* and *Effective Phase Transfer Entropy*, that are discussed in the following. However, our focus is information transfer between shares of the same stock market.

Kwon & Yang [20] used transfer entropy to assess strength and direction of information flow between global stock indexes, in a total of twenty-five ones. They made use of most important stock market over four continents, and, aggregating them into three major sets, measured how fluctuation in one set of markets could be reflected on the other ones. Via transfer entropy they were capable of elucidating that there is a correlation between some of those markets and it was not obtained by chance.

Marschinski & Kantz [21] also have a study wrapping information flow between stock markets using a slightly modified version of transfer entropy, called *effective transfer entropy* (ETE) as the measuring tool due to limited availability amount of data. Their study object were US-American Dow Jones Industrial Average (DJ) and German DAX Xetra (DAX) with a time scale of one minute between them.

Yang, Shang & Lin [19] analyze financial time series proposing a novel Effective Phase Transfer Entropy (EPTE), using signal phase shifts.

Korbel, Jiang & Zheng [22] analyze information flows among shares over the main five stock market, using transfer entropy, as complex networks communities. In special, they try to find out information flow among economic sectors, represented by sets of shares.

Jiayi He and Pengjian Shang [23] use transfer entropy to analyze stock market indexes influence and to compare the results with other three methods: effective transfer entropy (ETE), Rényi transfer entropy (RTE) and propose a effective Rényi transfer entropy (ERTE).

2.4 Network Centrality

There are real world systems which, for their complexity in quantity of elements and their relationship, have an intrinsic difficulty to be analyzed.

A common need in a study into this subject is to rank those elements by some kind of metric or characteristic, so we can measure importance of each element compared to the others.

There are many available models and metrics for ranking those elements. In many cases those systems are modeled as graphs, where elements are represented by nodes (vertices) and their relationships by links (edges) between each pair of vertices.

In particular, there are situations where those connection lines have a particular characteristic of directionality. In other words, given a pair of nodes A and B, a link from A to B is not necessarily equal to the one from B to A. So here we have a directed graph.

2.4.1 Node Ranking in Networks

There are many available models and metrics for ranking nodes in directed weighted networks. Directed edges give us to distinguish two types of nodes in the network: *authorities* and *hubs*. Authorities are nodes which act as sinks, receiving most of the links or the most valuable links, while hubs are nodes with many outgoing edges links or the ones with higher values.

In order to rank network nodes two techniques that can be applied are Pagerank [24, 25] and HITS [26, 27]. Pagerank ranks web pages by their importance, considering the importance of the pages that point to it. HITS has a similar recursive structure, but considering that nodes are both hubs and authorities, and uses this categorization in defining importance. In particular, an important authority is a node that is pointed to by important hubs. Similarly, a hub is important if it points to important authorities.

Since the graph that represents the market network is complete (all edges are present), because we seek relationships between all stocks, such algorithms are only meaningful in their extended version that uses edge weights: *weighted Pagerank* [28] and *weighted hyperlink-induced topic search* [29, 30].

2.4.2 Pagerank

Pagerank is an algorithm that was originally produced to sort web pages by their importance, counting **how many** links from other important pages point to a specific page. The more a page gets indications from other important pages, more it is important. This definition is clearly recursive and that algorithm is proven to be convergent to the extent that it is iteratively executed. The algorithm assigns a unique number to every node (web page). The higher the number the more important the page is relatively to all other pages. For this work a slightly different version of Pagerank must be used, as we need to deal with links weight instead of links amount. In the end, we have a unique number that means how important a node is regardless of whether it is the source or sink for the link, that is, it does not differentiate important sources (hubs) and sinks (authorities) from each other.

2.4.3 HITS

The HITS algorithm, on the other hand, uses a similar approach, but assigns two numbers to each node, one for how important it acts as a hub (origin of links) and a second number for how important it acts as an authority (destination of links). Again, it is necessary that a modified version of the algorithm is used to compute links weight instead of links amount.

2.5 Stock Market Analysis

Stock market is an environment where share prices can change due external factors like politics, economics, wars, speechs of authorities, publication of economic data, some company balance sheets, crop data, relevant weather events and many other variables, even rumors. All those external data are out of control of shareholders and stock brokers, but may impact share prices and seriously affect their stock positions.

As well, we can think stock market exchange internal factors can affect share prices, for example, the rise or fall in the value of shares of a given company, a given sector or a given stock index may suggest or influence operators to sell or buy certain shares, which would influence their prices and, consequently, their upward and downward movements.

A common picture is that of a stock exchange trader, in front of a monitor, watching the rise and fall of stock prices, waiting for the right moment to place his orders.

And then there are those traders who perform so-called fundamental analysis, who base their decisions on trying to look at the big picture and predict the direction in which the market (which is nothing more than people) will go.

But all this market analysis with the purpose of supporting stock buying and selling decisions requires economic study, stock market research and a good deal of expertise.

But a possible trend analysis of market share price movements could, in theory, be carried out by basically looking at the values and movements of shares within the market itself, believing that the movement of possibly a single share or a set of shares can cause another share or another set of shares to be price-modified up or down in a ripple effect or, in other words, with the former influencing the latter.

That is what this work tries to do, a endogenous analysis of share prices movements. But, instead looking at the raw share prices, we will focus on the upward and downward movements, in a historic base, using data publicly available at [B]^3 website [31].

Chapter 3

Characterizing Influence using Transfer Entropy Networks

Shares traded in a stock market intuitively influence one another, as the price variation in a share may trigger (or be correlated with) a subsequent price variation in another share. Intuitively, an investor or stockbroker observes price movements of shares today to make decision about buying or selling shares tomorrow.

Our focus is the *influence* between shares in a stock market exchange. By considering a pairwise directed measure of influence, we construct a directed weighted network and leverage its structure to determine the most *influential* and the most *influenced* shares in the stock market.

To achieve this goal, we consider the time series to represent the share price move direction (not the share price value) and measure the information transfer between every two ordered pair of shares. We construct a directed weighted network where nodes are shares and edges weights represent transfer entropy values. Thus, we can analyze this network using classic centrality measures to assess the importance of nodes (shares).

Note that the edge weight $w_{i,j}$ denotes how much share i influences share j. Intuitively the sum of incoming edges to node i characterizes how influenced is this share, while the sum of outgoing edges of node i characterizes its influence power. Ranking shares based on its influence power or its subjection to influence yields a metric of *importance* for shares.

The ranking of shares can be done using algorithms like Pagerank or HITS, originally developed for ranking web pages [24–26], which leverage the weights and structure of the network. Moreover, nodes can also be ranked by their incoming and outgoing total (sum) edge weights, here called *Node Weight*.

We apply this methodology to shares traded in the [B]^3, the most important stock market exchange in Brazil. Using publicly available data (time series of share daily prices over a 32 year period), we measure pairwise transfer entropy and construct the network. We rank nodes according to three different metrics yielding for each metric a ranking for influential and influenced shares. We assess the coherency among the top ranked shares across the different rankings to better characterize influenced and influential shares. We also investigate the relationship between influence (according to the ranking of different metrics) and traded volume of shares, and results indicate there is virtually no correlation.

The novelty of this work is to consider the relationship between present and future price movements of different shares. We are not aware of other works that leverage influence among present and future share movements to rank influential and influenced shares.

3.1 Methodology

This section discusses the time series dataset that was constructed as well as the proposed methodology to rank influential and influenced stocks.

3.1.1 Time series

Stock price is the most basic information concerning a stock and most analysis use the stock price over time. Instead, the focus of this work is on price movement: the direction of change in the stock price within a trading day. In particular, our focus is to measure the influence that a price movement of a given stock *today* has on the price movement of another stock *tomorrow*.

Several studies make comparisons between stock prices which are synchronized in time, that is, prices of two stocks A and B are compared in the same time interval $t_i, \ldots, t_f$, where t_i and t_f are initial and final time instants, respectively. Such studies investigate the binding between the prices of two stocks, trying to measure and understand when they might move together (e.g., correlated) over time.

In contrast, our focus is on stock price movements in subsequent days. This is motivated by trading decisions (to sell or buy a stock) that often occur after the observation and analysis of prior but recent stock price movements. Intuitively, there is a time delay between observing (stock price movements) and taking action (trading a stock). Thus, the proposed methodology follows along this intuition.

Thus, daily couple of stock prices are converted into symbols that are the basic elements of the time series, namely price movements. One advantage of using price movements is to have just three symbols (e.g., *up*, *same*, *down*) for each time series, which allows for a more accurate estimation of the empirical distribution of a pair of time series (joint and marginal, as required by the calculation of transfer entropy). Another advantage is being independent of price magnitude, as different stocks have

t	0	1	2	3	4	5	6	7	8	9	10
OP	8.69	0.28	9.18	5.66	2.77		1.51		8.85	8.88	4.98
CP	0.28	9.18	5.66	2.77	7.18		8.85		8.88	4.98	3.12
MD	$\downarrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$	$\varnothing$	$\uparrow$	$\varnothing$	$\approx$	$\downarrow$	$\downarrow$

t = time (day); OP = opening price; CP = closing price; M = movement direction

Figure 3.1: Mapping stock prices series into stock movements (biases) series.

trading prices with different orders of magnitude.

3.1.2 Movement symbols

A set of symbols indicating price movements are used to investigate influence in the stock prices movements. Using the daily *opening* and *closing* prices for stocks we map a price series ("continuous" float type) into a symbol series (discrete "integer" type) as shown in Figure 3.1. If a stock price goes up from stock market exchange opening time to its close time, we have some "$\uparrow$" symbol for that stock in that day; if it goes down, we have some "$\downarrow$" symbol; if it remains *almost the same*, we have a "$\approx$" symbol; and there are cases where we have no opening neither/or closing price for the stock, because it is not traded in that day, and in this case we have a "$-$" symbol.

The term "*almost the same*" means that the magnitude of the difference between opening and closing prices are withing a factor of $\varepsilon > 0$. The idea is to avoid false *up* or *down* movements due to very small price variations in the day. While ε can be chosen arbitrarily, this work considers the factor $\varepsilon = 4.5 \cdot 10^{-5} = 0.0045\%$ because it represents a daily percentage that corresponds to a 1% (one percent) monthly variation divided by 22 (twenty two) business days in a month.

Therefore each stock has a time series whose elements are one of four possible symbols ($\uparrow, \downarrow, \approx, \varnothing$) and its length is the number of business days in time period.

3.1.3 Pairwise transfer entropy

The time series of stock movements is used calculate the transfer entropy between each ordered pair of stocks. A lag of one day in the time series is used. Thus, the transfer entropy will measure the information transferred from the price movement today of a given stock to the price movement tomorrow of another stock. The necessary empirical probability distributions (joint and marginal) required in Equation 2.3 are first computed. Using these empirical distributions, the transfer entropy is computed.

Note that transfer entropy is a directed measure: given two stock time series A and B, the lag of one day of A with respect to B will yield a different joint empirical distribution from corresponding empirical distribution for the lag of one day of B with respect to A.

3.1.4 Network and ranking

All ordered pairs of stocks are considered in constructing the network of stocks. Thus, we have a complete directed graph (all possible edges are present) where nodes are stocks and directed edges have weights that correspond to the transfer entropy between the ordered pair.

In order to rank nodes and capture the set of influential and influenced nodes in the network, Pagerank [24, 25], HITS [26, 27] and node weights (incoming and outgoing) are used.

3.1.5 Similarity coefficients

In order to assess results from several rankings, we need a tool to compare them to each other.

Jaccard similarity

The *Jaccard similarity coefficient* is a statistic used to determine the similarity and diversity of two sets and is defined as the size of the intersection divided by the size of the union of the two sets. For two sets A and B, the Jaccard similarity is given by

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|},$$

where $J(A,B) = 1$ if the sets are identical and $J(A,B) = 0$ if they have no elements in common. The Jaccard similarity is often used to determine the similarity between the top-k elements of two rankings.

Spearman's correlation coefficient

The *Spearman's rank correlation coefficient* (ρ) measures the correlation between two rankings. It does not use the values associated with the data points but instead their relative ordering. The coefficient is calculated over the entire ranking and it is $+1$ when the two rankings are identical and -1 when one of them is a reverse ordering of the other. The value 0 indicates their is no correlation between the two rankings (i.e., one of them is random permutation of the other).

3.2 Dataset Details

[B]^3, which stands for *Brasil, Bolsa, Balcão*, the main brazilian stock market exchange, is the target case of this work. There are more than 1500 companies which negotiate their bonds in [B]^3, as seen on data records for 2019, the last year of study. And such information, including bond prices, is publicly available [32].

Companies can have a large number of different bonds negotiated in the stock exchange and are also responsible for various actions concerning them, such as creation, discontinuation, name changes, joining or splitting bonds. However, such activities impose limitations in the data analysis since those records are not always easily available, being hard to track name changes, or bonds that have joined or split. This poses some difficulty in tracking a particular bond over long time periods.

3.2.1 Dealt stocks and records

[B]^3 offers an electronic data set containing daily transactions for each stock since January 1986. Our initial evaluation, with data until 2019, revealed 183,127 different bonds negotiated in the stock market within that period, where each bond within [B]^3 is identified by a unique code. However, many of them have a short life time (e.g., one month) or a very low *liquidity* (fraction of days that it is traded). In terms of an analysis over a large time period, stocks with one of these two characteristics are not interesting. For instance, a particular type of bond, the *stock option* has not been considered in this work, even if it corresponds to more than 91% of all papers traded since it is also not an enterprise stock.

Therefore, the following two filters were applied to the original data discarding stocks that did not pass this criteria:

1. Stock code is dealt in at least in 50% of the business days within the period considered (i.e., liquidity of at least 0.5).
2. Stock code is not a *stock option*.

When considering the entire 32-years period, this filter reduces the amount of stock codes from tens of thousands to a few hundred. Moreover, since the liquidity criteria is for a particular period, it is possible that stock does meet this criteria for a certain period (e.g., one year) but does for another period (e.g., a different year). The relative high requirement on the minimum liquidity (traded in at least half of the business days) is to ensure that the stock is traded often enough such that information flow (transfer entropy) from this stock to others and from others to this stock can be computed more reliably.

3.2.2 Daily basis time series

In order to obtain biases (direction of variation in share price) series from price series of each share, we have chosen a daily basis of its closing price relation to its opening price. Chosing a daily basis instead of a, for instante, a weekly basis was done having in mind to have longer time series, with more date, and obtain more accurate transfer entropy values, as this computation is too sensitive to series length. Furthermore, we use the maximum $[\mathsf{B}]^3$ gives us as source data.

Beyond that, it looks like there isto no reason, tha is strong enough, to choose a time scale other than daily over the latter. Therefore, the choice, for the purposes of the study, is reasonably arbitrary and, if necessary, such a study can be replicated using the same methodology on another time scale.

3.3 Results on Overall Network

This section presents empirical results obtained by applying the proposed methodology to data from $[\mathsf{B}]^3$ considering a single transfer entropy network comprising the entire 32-years period, ranging from January 1986 to December 2019. Raw data for this period consists of 8,125,691 records (each record is a daily transaction of a bond), wrapping 183,127 unique stock codes, on 8,385 working days, carried out by 4,730 companies. However, recall that some stocks are not considered due to their type (stock option) or very low liquidity over the period. After removing them, only a total of 98 liquid stocks remain, a fraction of $0,000535148$. This reflects the fact that very few stocks are actively traded over the entire 32-years period.

As a remark, results throughout this section are presented using the empirical complementary cumulative distribution function (CCDF) which captures the percentage of objects (edges or nodes) that are above a certain value, and are suited for heavy tailed distributions (in log-log scale). A point (x, y) in this plot indicates the fraction y of objects that have a value greater than x. The plots also indicate with grid lines the 50% (median value, in red), 10% (green) and 5% (yellow) of the distribution.

3.3.1 Edges weight: transfer entropy values between stocks

We start by presenting results on the edge weights among all 98 stocks. Note that the network is a complete graph with directed edges and, thus has a total of 9,506 edges. Recall that edge weights represent transfer entropy values between respective nodes, according to the direction of information transfer.

Figure 3.2 shows the complementary cumulative distribution function (CCDF) of edge weights. The additional colored grid lines indicate 50% (red), 10% (green)

Table 3.1: Top-20 transfer entropy values for 32-years period, from stock j to stock i.

#	i (to)	j (from)	T_{ji} (transfer entropy)
1	CSNA3T	USIM5T	0.0381664944121
2	FNAM11	GGBR3F	0.0343417549928
3	FNAM11	BRKM5F	0.0332163988323
4	FNAM11	BRKM5	0.0326865832172
5	RSID3	GGBR3F	0.0316434026287
6	CSNA3T	GGBR4T	0.0302360042251
7	ALPA4	GGBR3F	0.0301210219964
8	FNAM11	CCRO3	0.0297901606699
9	FNAM11	SAPR4	0.0295252827106
10	ITSA3	SLED4	0.0295212046787
11	TELB3F	CRUZ3	0.0292981139248
12	POMO4	BRKM5F	0.0292606302150
13	RSID3	SLED4	0.0291596116402
14	FNAM11	IDNT3	0.0291281610330
15	ITSA3	RSID3	0.0288580360821
16	FNAM11	KLBN4F	0.0287563399963
17	FNAM11	POMO4F	0.0287128955369
18	PETR4T	CSNA3T	0.0286861454868
19	FNAM11	SLED4	0.0285840734402
20	POMO4	BRKM5	0.0283279060196

and 5% (yellow) of 9,506 total points of the distribution. Note that a small fraction of edges have a very large value for transfer entropy. The shape of the curve exposes a heavy tail feature where most edges have very small transfer entropy values indicating they transfer little or no information, while a small fraction of edges transfer much more information. Calculating loglikelihood ratio among some candidate distributions indicates, however, it does not follow a power law distribution nor present a power law tail. The 20 largest values (Table 3.1) are represented by the green dots in Figure 3.2. Note that 50% of the values have a transfer entropy less than 1.984×10^{-3}, whereas 5% have a value larger than 1.662×10^{-2}, with an average of 4.141×10^{-3}. Negative transfer entropy values are possible and indicate that no information is transferred.

Note that some stocks in Table 3.1 appear multiple times as the source (j column) of information transfer among the largest edge weights, while other stocks appear multiple times as the target (i column), suggesting that there are bonds with greater presence in each side of information transfer. For example, SLED4 (3x), GGBR3F (3x), BRKM5 (2x) and BRKM5F (2x) appear more than once as source of information transfer among the top 20 (number of appearances are in parenthesis) of *most influential* stocks. If we do not consider the stocks but the corporation (companies), aggregating their bonds, BRKM (4x), GGBR (4x) and SLED (3x) occur several times as influential source of information transfer. On the other side, CSNA3T (2x), RSID3 (2x), FNAM11 (9x), ITSA3 (2x), POMO4 (2x) appear frequently as target of information transfer

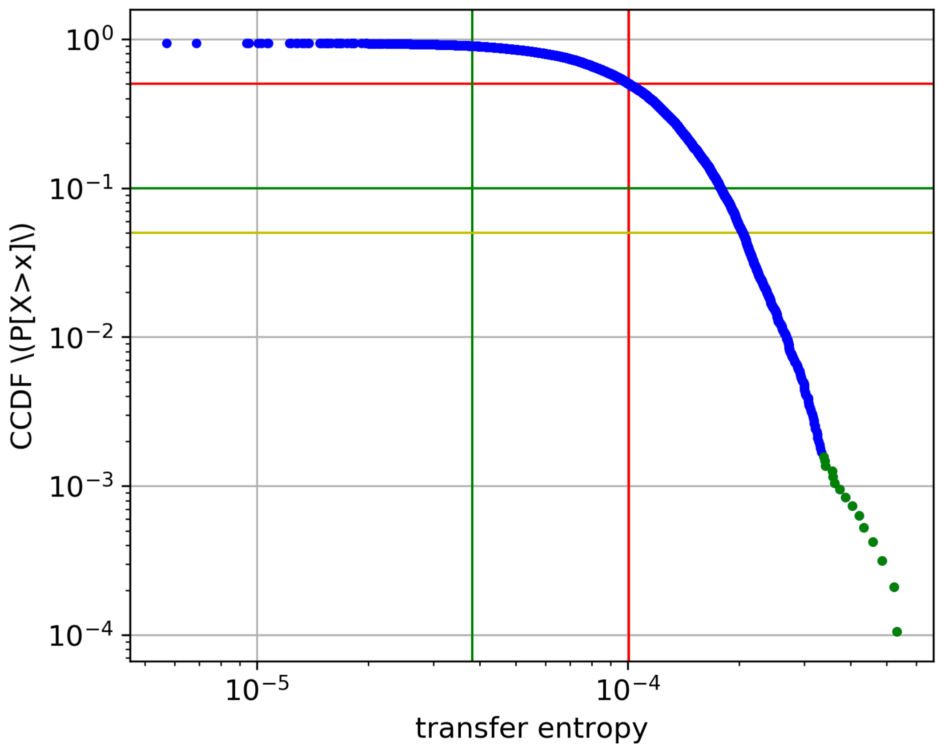

Figure 3.2: Complementary cumulative distribution function of weight (transfer entropy) values for all edges for the 32-years period, in log-log scale.

among the top 20. Contrary to the situation with influential stocks, different bonds do not represent the same corporation, so the number of bonds is the same as the number of corporations.

Figure 3.3 shows all transfer entropy values at once, in matrix form, as a unicolor map, with color intensity corresponding to transfer entropy values in linear scale. Vertical axis are stocks that act as information transmitters whilst horizontal axis stocks act as information receivers. For example, the 5 top edges values (for pair of stocks) mentioned above are those in Table 3.2, while the 5 bottom values are those ones in Table 3.3. In general, a lighter line in such a map suggests that the corresponding stock is not an important source of information transfer, whereas a darker line indicate that the stock transfers significant information to other stocks. The top 5 most influential stocks appears in Table 3.4 and 5 least influential in Table 3.6. Analogously, a lighter column suggests that the stock does not receive much information from other stocks, being therefore non susceptible to the action of others, and a darker column suggests a stock receives information from many other stocks. Tht top 5 most influenced stocks are shown in Table 3.5 and the 5 least ones in Table 3.7.

Table 3.2: Overall top 5 transfer entropy values, also illustrated in colormap of Figure 3.3.

line	column	from	to	transfer entropy value
93	33	USIM5T	CSNA3T	0.0381664944121
47	45	GGBR3F	FNAM11	0.0343417549928
16	45	BRKM5F	FNAM11	0.0332163988323
15	45	BRKM5	FNAM11	0.0326865832172
47	78	GGBR3F	RSID3	0.0316434026287

Table 3.3: Overall bottom 5 transfer entropy values, also illustrated in colormap of Figure 3.3.

line	column	from	to	transfer entropy value
30	46	CRUZ3	FNOR11	-0.00908997612219
97	46	VALE5F	FNOR11	-0.00952680861834
30	61	CRUZ3	ITSA4T	-0.01118301432840
30	45	CRUZ3	FNAM11	-0.01120112405660
30	25	CRUZ3	CMIG4T	-0.01350899110990

Table 3.4: 5 most influential stocks, also illustrated in Figure 3.3.

line	stock	transfer entropy summation
45	FESA4	1.0788611224182590
75	PETR4F	1.0174284791173107
0	ALPA4	1.0167043055403462
33	CSNA3	0.9603095956577570
74	PETR4	0.9412726806220999

Table 3.5: 5 most influenced stocks, also illustrated in Figure 3.3.

column	stock	transfer entropy summation
47	FNAM11	1.0085356414726480
54	GOAU4	0.9074556562171353
90	UNIP6	0.8916304188223408
17	BRKM5	0.8896364765005400
82	SBSP3	0.8727675593858700

Table 3.6: 5 least influential stocks, also illustrated in Figure 3.3.

line	stock	transfer entropy summation
83	SBSP3F	0.019319701110951003
39	ELET6	−0.003204582188684090
18	BRKM5F	−0.048637158118732540
10	BBDC4F	−0.050704479162321590
46	FJTA4	−0.136756721479942400

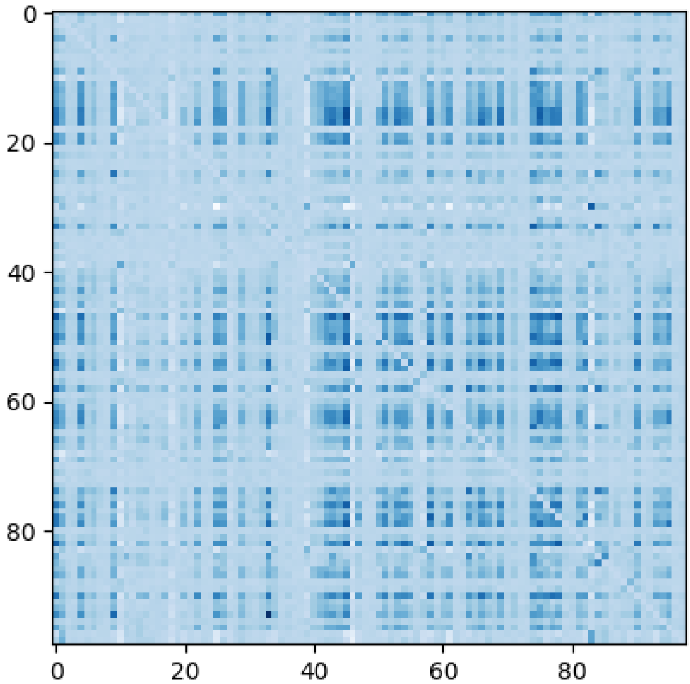

Figure 3.3: Colormap of weight (transfer entropy) values for all edges for the 32-years period, among 98 nodes, in linear scale.

Table 3.7: 5 least influenced stocks, also illustrated in Figure 3.3.

column	stock	transfer entropy summation
18	`BRKM5F`	0.100734284889238110
96	`VALE3`	0.082605480412294110
30	`CPLE6`	0.078982563300848980
68	`LAME4F`	0.070089128376402700
46	`FJTA4`	0.056565966447729994

3.3.2 Node weight

Node weight is what we call the summation of all incoming or outgoing edges on that node. Incoming weight is the sum of incoming edges, whilst outgoing weight is the sum of outgoing edges of a specific node.

The outgoing and incoming weights of nodes allow us to understand their importance in transferring (both sending and receiving) information in the stock network. Thus, each stock could now be ranked both in terms of playing as *influenced* or *influential* of information transfer in the graph.

Stocks with higher incoming weight are those more susceptible to price variations of other stocks, whereas stocks with lower incoming weight are relatively proof of other stocks movements. Both kinds of stocks are interesting, depending on the goals and stock market expectations. Stocks with higher outgoing weight are the most influential ones, whereas stocks with lower outgoing weight are relatively invisible to the other stocks.

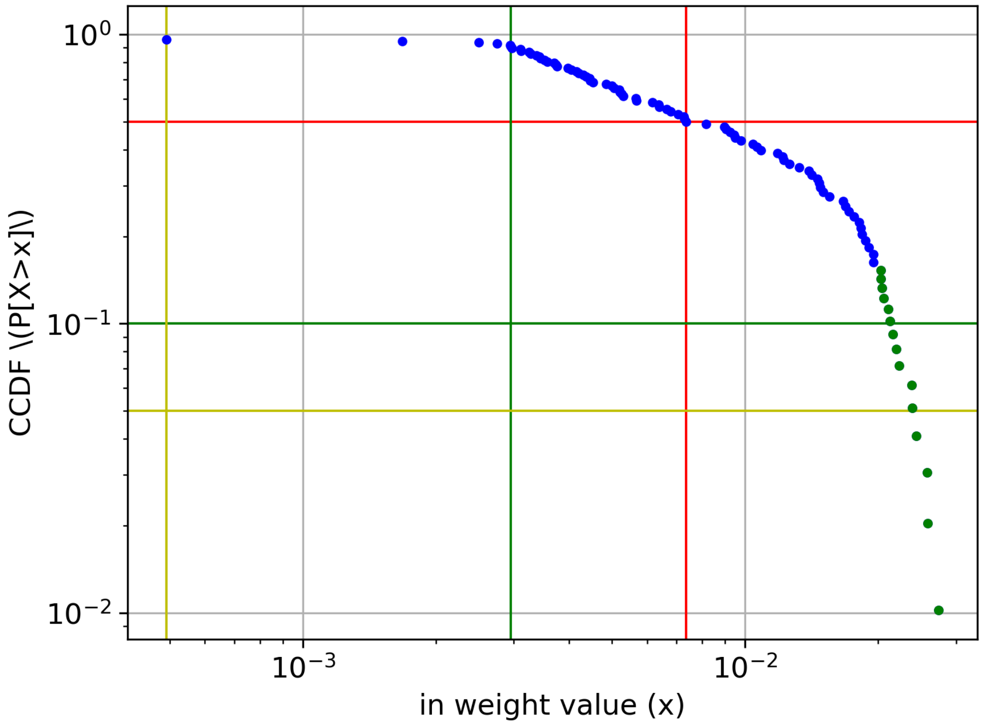

Figure 3.4: Complementary cumulative distribution function of node incoming weights (influenced by others), in log scale.

Figure 3.4 shows the complementary cumulative distribution functions for incoming and Figure 3.5 for outgoing node weights, in logarithmic scale. In both plots, their shapes show that most stocks have a very small weight while a small fraction of the stocks have weights much larger than the average, indicating a heavy tail distribution. However, such distributions fail to pass an statistical test for power law distributions. In any case, the heavy tail nature of the distributions indicates that only a small fraction of the stocks are actually influential and influenced.

Notably, the incoming weight of top 20 stocks among 98 ones represent 43.6% of total incoming weight (top 5 ones can be seen in Table 3.5), whilst top 20 ones with outgoing weight represent 41.1% of total outgoing weight (top 5 ones can be seen in Table 3.4).

Interestingly, there is an overlap of 40% (8 stocks) between the top 20 largest incoming and outgoing weight values, as shown in Table 3.8, whereas Spearman ρ coefficient between the entire incoming weight ranking and outgoing weight ranking is 51.13%. This shows that stocks can play both roles of being influential (to a set

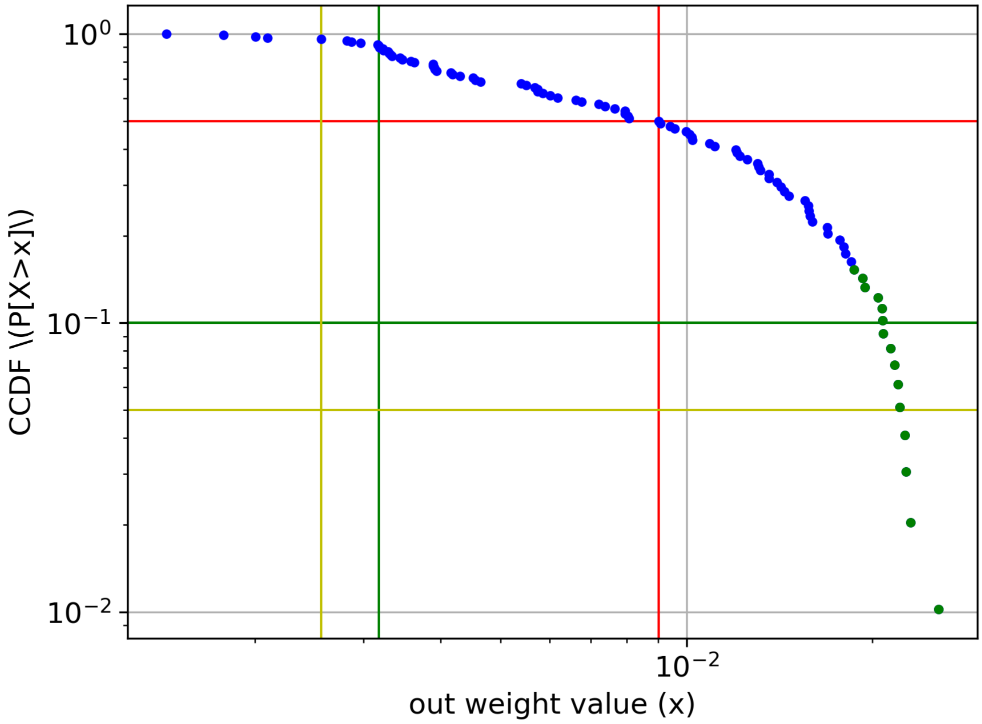

Figure 3.5: Complementary cumulative distribution function of node outgoing weights (that influence others), in log scale.

of stocks) and influenced (by another set of stocks), as the two metrics (Jaccard and Spearman) indicate a moderate degree of correlation (similarity).

3.3.3 Pagerank

The weighted version of Pagerank [24, 25] was used to rank the stocks. Note that the weighted Pagerank algorithm computes how important a stock is with respect to being the target of information transfer; in other words, how much the stock market influences that stock. Larger Pagerank values indicates stocks that are more influenced.

On the other hand, in order to capture how influential are the stocks using Pagerank, we reverse the direction of the edges in the graph. With this transformation, a node with high summation of “incoming” edges is a great source of information transfer. The weighted version of Pagerank is applied to this network (with reversed edges) to identify the most influential nodes.

Figure 3.6 shows the CCDF of Pagerank values for the original network, whereas Figure 3.7 shows the CCDF of Pagerank values for the reverse edge network. Again, very few nodes have Pagerank values much larger than the average, being very much influenced by others (Figure 3.6) or very much influential on others (Figure 3.7).

Table 3.8: Stocks both in top 20 incoming and top 20 outgoing weight rankings.

stock code	incoming ranking #	outgoing ranking #
GUAR3	14	2
USIM3	15	3
GOAU3	12	8
ITSA3	8	10
CSNA3T	4	15
PETR4T	5	18
RSID3	6	19
ALPA4	3	20

Table 3.9: Stocks both in top 20 Pagerank and top 20 reverse edge Pagerank rankings.

stock code	straight ranking #	reverse ranking #
GUAR3	13*	2
USIM3	15	4*
GOAU3	12	8
ITSA3	7*	9*
CSNA3T	4	16*
PETR4T	6*	19*
RSID3	5*	18*
ALPA4	2*	20

Same way as edge weights and node weights, such distributions fail to pass an statistical test for power law distributions. These charts can be compared to those in Figures 3.4 and 3.5 which shows incoming outgoing node weight, respectively. Interestingly, note that incoming weight and Pagerank look very similar as well as outgoing weight resembles reverse edge Pagerank.

The Jaccard similarity of top 20 stocks in the Pagerank rankings for the two networks (original and reversed edges) is 40% (8 stocks), as shown in Table 3.9. The Spearman similarity is $\rho = 50.67\%$ for the entire Pagerank rankings of the two networks. Again, there is significant overlap between the top influenced and influential stocks, as well as significant similarity in their rankings.

Remarkably, comparing Table 3.8 (for node weight) and Table 3.9 (for Pagerank), note that 8 stocks are exactly the same and are either in the same ranking position or at most 1 position different (different positions are indicated by an asterisk sign).

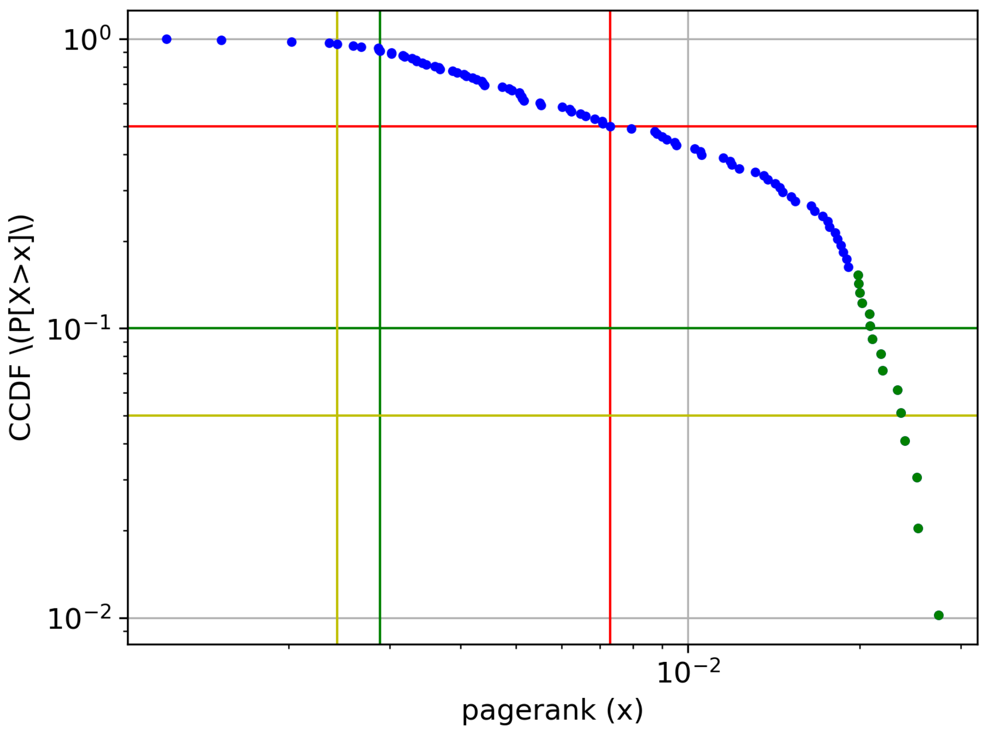

Figure 3.6: Complementary cumulative distribution function for Pagerank values of stocks (influenced), in log-log scale.

3.3.4 Authorities and hubs

We also use HITS algorithm [27] to rank stocks in the graph. HITS algorithm assigns two numbers to every node in the graph, namely a *hub factor* and an *authority factor*. A *hub* is related to the *source* and an *authority* to the *target* of information transfer. Figures 3.8 and 3.9 show the complementary cumulative distribution function for authority and hub values.

As Pagerank, HITS graphics show that there are some shares whose values are very far from the average. We can compare graphics on Figure 3.8 with those ones on Figure 3.6 and Figure 3.4, as well it is possible to compare Figure 3.9 with those ones on Figure 3.7 and Figure 3.5. Shape of curves representing incoming edges (incoming weight, Pagerank and HITS authority) look very similar to each other. We can say the same about those representing outcoming edges (outgoing weight, *reversed edges* Pagerank and HITS hub).

The Jaccard similarity of top 20 stocks between the two HITS rankings (authority and hub) is 35% (7 stocks), as shown in Table 3.10. The Spearman similarity is $\rho = 52.19\%$ for the entire HITS rankings of the two networks. Again, there is significant overlap between the top influenced and influential stocks, as well as significant similarity in their rankings.

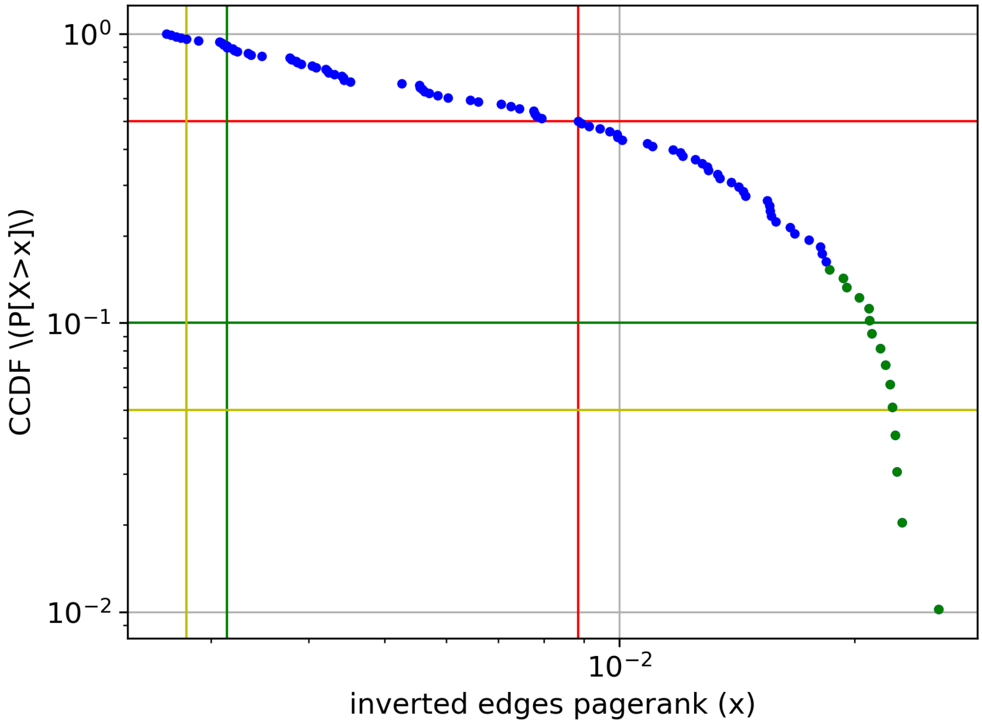

Figure 3.7: Complementary cumulative distribution function for reverse edge Pagerank values of stocks (influential), in log-log scale.

Again, comparing Table 3.8 (node weight) and Table 3.9 (Pagerank) with Table 3.10, notice that seven stocks are exactly the same and are either in the same ranking position or at most one position different.

Indeed, when those rankings are put side by side as in Table 3.11, we have a consolidated sight of stocks that appear in all generated rankings.

Table 3.10: Stocks both in top 20 authority and top 20 hub rankings.

stock code	authority ranking #	hub ranking #
GUAR3	13	3
USIM3	14	7
GOAU3	11	9
ITSA3	7	12
CSNA3T	3	17
PETR4T	6	19
RSID3	5	18

Table 3.11: Common stocks in top 20 rankings.

stock code	incoming weight ranking#	outgoing weight ranking#	straight pagerank ranking#	reverse pagerank ranking#	authority hits ranking#	hub hits ranking#
GUAR3	14	2	13	2	13	3
USIM3	15	3	15	4	14	7
GOAU3	12	8	12	8	11	9
ITSA3	8	10	7	9	7	12
CSNA3T	4	15	4	16	3	17
PETR4T	5	18	6	19	6	19
RSID3	6	19	5	18	5	18
ALPA4	3	20	2	20	—	—

Table 3.12: Jaccard similarity and Spearman's ρ among rankings for influential and influenced stocks for the same metric (node weight, Pagerank and HITS).

Metric	**weight**	**Pagerank**	**HITS**
Jaccard	0.40	0.40	0.35
Spearman's ρ	0.5113	0.5067	0.5219

3.3.5 Self-considered metrics

In what follows we consider the similarity between the rankings for influential and influenced stocks for each of the metrics. In particular, the incoming node weight ranking is compared to outgoing node weight ranking, Pagerank ranking is compared to the *reversed edges* Pagerank ranking and authority ranking is compared to hub ranking.

Table 3.12 shows the Jaccard similarity for the top 20 for each metric as well as the Spearman ρ similarity coefficient. In both cases and for the three metrics (node weight, Pagerank and HITS) there is some level of similarity between the influential and influenced rankings (Jaccard ≈ 0.4 and $\rho \approx 0.51$). This moderate level of similarity indicates that some stocks can be influential and be influenced at a relatively similar ranking positions, an observation that was not expected.

3.3.6 Comparison among ranking methods

If we cut off complementary cumulative distribution functions curves in Figure 3.4 (node weight), Figure 3.6 (Pagerank) and Figure 3.8 (HITS) in three segments according to their "inflection points", one can realize they look almost the same.

Interestingly, as seen in Table 3.13, there is a large intersection between top-20 incoming weight ranking and top-20 Pagerank ranking, with 19 over 20 bonds appearing in both rankings, resulting in a Jaccard coefficient of 0.95 for the incoming metric. Moreover, the full ranking of these two metrics exhibit very strong corre-

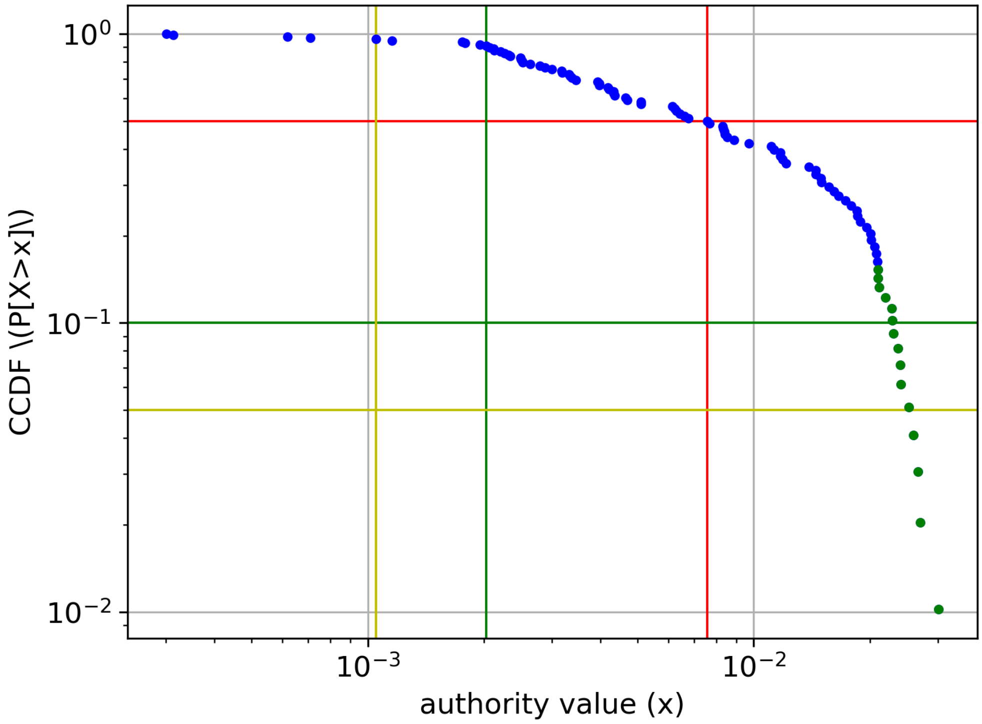

Figure 3.8: Complementary cumulative distribution function for authorities values of stocks (influenced), in log-log scale.

Table 3.13: Jaccard similarity and Spearman's ρ among rankings for influenced among metrics (node weight, Pagerank and HITS).

	incoming weight *vs.* Pagerank	**Pagerank *vs.* HITS authority**	**HITS authority *vs.*incoming weight**
Jaccard	0.95	0.95	0.90
Spearman's ρ	0.9951	0.9918	0.9968

lation, with a Spearman coefficient of $\rho = 0.9951$. Similar results were obtained when comparing the other two pairs of rankings: Pagerank *vs.* HITS authority and incoming weight *vs.* HITS authority, as can be seen in Table 3.13.

Furthermore, as seen at Table 3.14, all stocks appearing in top 20 reverse edge Pagerank ranking are also present in the top 20 outgoing weight ranking. While full the rankings are not identical, only a few exchanges occur either going up or down a single position in the ranking, 3 with 4, 9 with 10, 15 with 16 and 18 with 19. Besides that, the full ranking of these two metrics are very consistent, and have a Spearman coefficient of $\rho = 0.9914$. Thus, there is a very strong coherence between node weights and Pagerank for both influential and influenced rankings of stocks. Same occurs between reverse Pagerank *vs.* HITS hub as well between HITS hub *vs.*

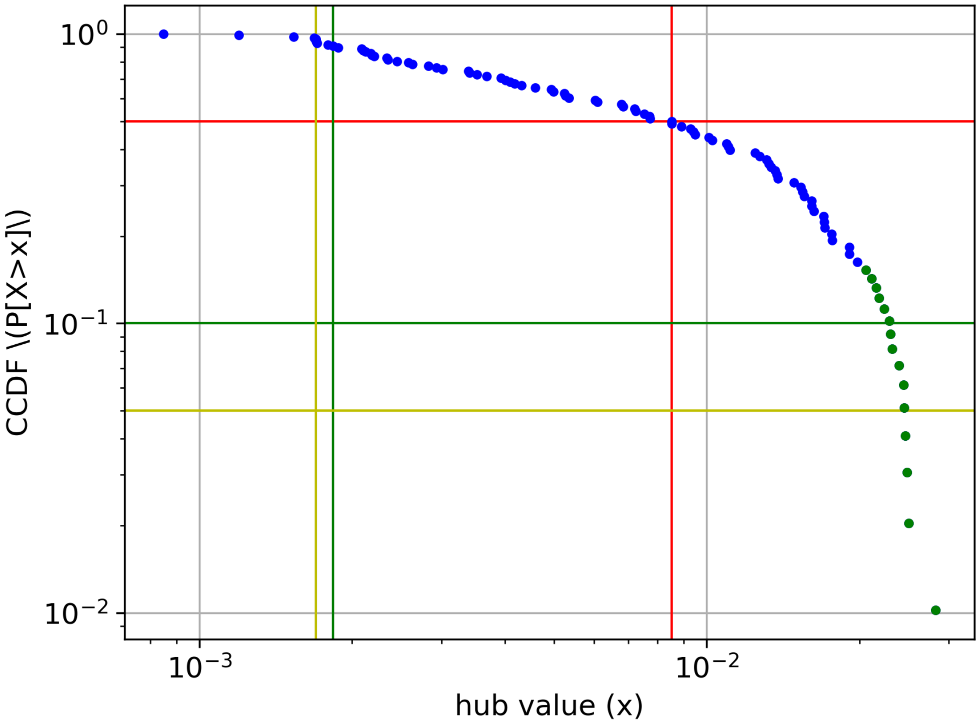

Figure 3.9: Complementary cumulative distribution function for hub values of stocks (influential), in log-log scale.

Table 3.14: Jaccard similarity and Spearman's ρ among rankings for influential among metrics (node weight, Pagerank and HITS).

	outgoing weight *vs.* Pagerank	**reverse Pagerank *vs.* HITS hub**	**HITS hub *vs.* outgoing weight**
Jaccard	1.00	0.95	0.95
Spearman's ρ	0.9914	0.9881	0.9951

outgoing weight.

3.3.7 Transaction volume

There are some financial indicators that reveal financial power of stocks under the umbrella term *transaction volume*. We will consider three of them with slightly differences: **(a)** *financial volume*, meaning total amount of money negotiated in financial transactions of a stock; **(b)** amount of *deals*, meaning number of trades carried out with the stock; and **(c)** total amount of stock dealt, meaning total number of stocks traded with this code. Besides these ones, another important indicator is **(d)** *liquidity* of a stock, meaning percentage of business days that stock is negotiated in the stock market exchange during a time period.

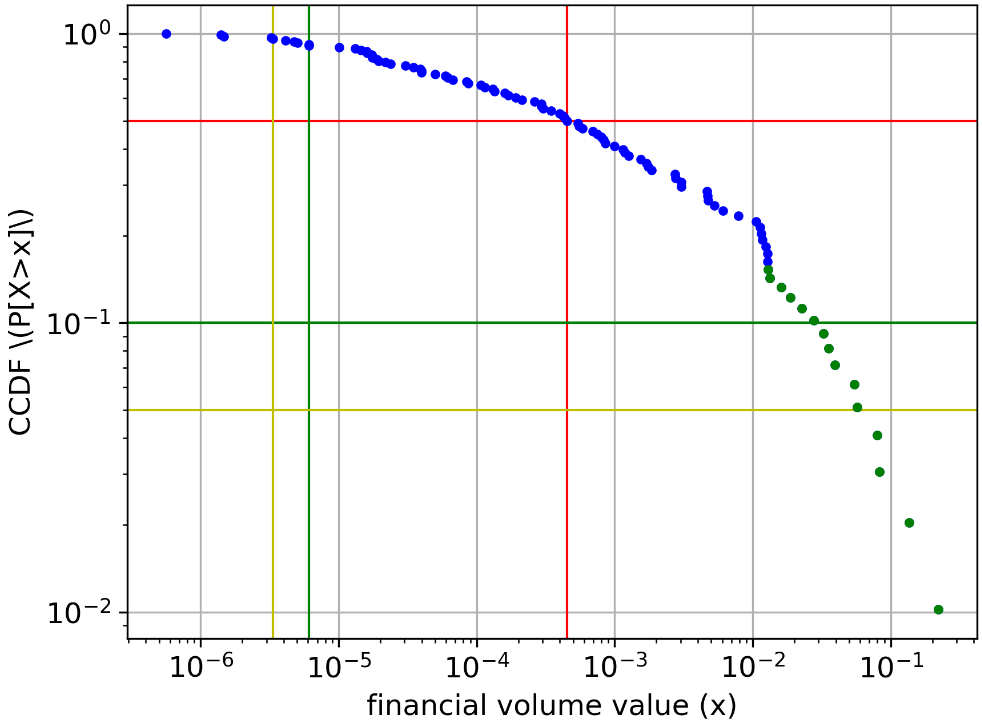

Figure 3.10: Complementary cumulative distribution function for financial volume of stocks, in log scale.

Figure 3.10 shows the CCDF for financial volume values for all stocks in the entire 32 years. Its shape follows a power law distribution with an estimated exponent $\alpha = 1.32$, and shows the striking difference among the stocks with the larger values, as shown in Table 3.15(a). The α exponent was estimated using the method proposed by Clauset et. al [33] implemented in Python package `powerlaw`.

Figure 3.11 shows the CCDF for liquidity indicating a curve that resembles two lines: the leaning portion of the curve shows that, in general, liquidity values are separated by orders of magnitude, wheres vertical portion of the curve brings together a group of about 20 equally liquid stocks with same liquidity. The top 20 are shown in Table 3.15(b).

As a remark about liquidity values, there is a discontinuity on the names of the stocks mainly at March 16, 1998, when a three letter code for the stocks names moved to a four letter code. So, value of 0.643 means, actually, 1.0 for period since March 16, 1998, because that bond appeared in all 5392 dates (business days) after March 16, 1998. Unfortunately, this discontinuity cannot be easily tracked for all stocks and was ignored in this study (treated as different stocks).

The Jaccard coefficient using the the top 20 stocks of each ranking, and Spearman coefficient ρ using the entire ranking were also computed among all transaction

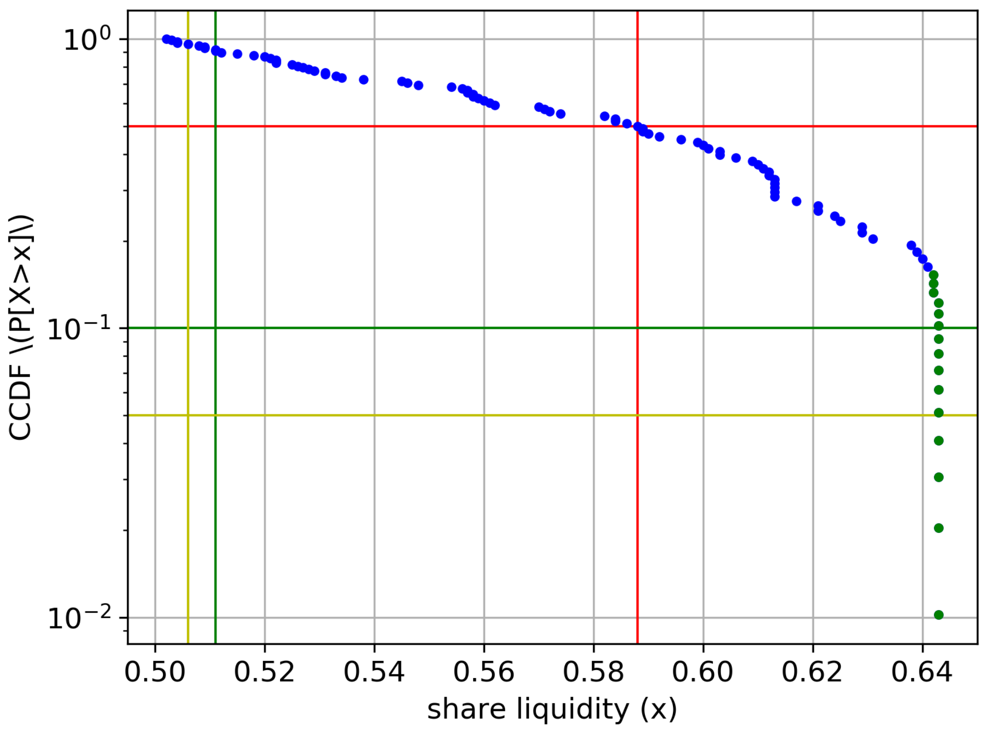

Figure 3.11: Complementary cumulative distribution function, for liquidity of stocks, in log scale.

metrics: liquidity, financial volume, bonds dealt and deals amount, shown in Table 3.16. The values indicate that for most pair of metrics there is a good agreement in both the top 20 and the entire ranking, such as financial volume (F) and deals (D). This indicates that there is no preferred financial indicator from the set analyzed as to determine influence between stocks.

3.3.8 Weight and Pagerank *vs.* transaction volume

Financial volume is an important indicator in stock market analysis. Thus, an important consideration is the relationship between the financial volume and their influence with respect to information transfer. Indeed, stocks with have very large transaction volumes (compared to the average) could be influential in driving the price variations in the stock market.

Recall that Figure 3.10 shows the distribution of the total traded financial volume of the stocks. Note that the two top stocks, `PETR4` and `VALE5`, are one order of magnitude larger than 98% of the smaller stocks. The top 7 stocks with most traded financial volume are shown in Table 3.17, along their position in according to the different rankings (recall that the network has 98 stocks).

Table 3.17 shows that financial volume and influence are not correlated. A

Table 3.15: In order, first 20 positions for (a) financial volume (b) liquidity.

#	(a) stock	(a) money negotiated	(b) stock	(b) liquidity
1	PETR4	258944682957090	BBDC3	0.643
2	VALE5	159003384346700	BBDC4	0.643
3	VALE3	96291893620200	CMIG4	0.643
4	BBDC4	93095339913060	CPLE6	0.643
5	BBAS3	66420109179700	CSNA3	0.643
6	PETR3	63473593815020	ELET3	0.643
7	ITSA4	45900906054600	ELET6	0.643
8	GGBR4	41376134847600	GOAU4	0.643
9	USIM5	38006196242200	ITSA4	0.643
10	CSNA3	32298868621000	PETR3	0.643
11	CMIG4	26426116845400	PETR4	0.643
12	CCRO3	21873335852700	SBSP3	0.643
13	PCAR4	18705702050600	BBAS3	0.642
14	LAME4	15564124239600	CMIG3	0.642
15	ELET3	15135315589000	UNIP6	0.642
16	GOAU4	15006607161759	TELB4	0.641
17	BRAP4	14982982833500	VALE3	0.640
18	ELET6	14552182848500	LAME4	0.639
19	BRKM5	13698972693790	INEP4	0.638
20	EMBR3	13459153406000	CPLE3	0.631

Table 3.16: The 20-size Jaccard j and Spearman ρ among all transaction indicators: financial volume (F), amount of bonds dealt (B), amount of deals done (D) and liquidity (L).

	L *vs.*F	L *vs.*D	L *vs.*B	F *vs.*D	D *vs.*B	B *vs.*F
j	0.6	0.65	0.6	0.95	0.45	0.5
ρ	0.397387	0.555200	0.452963	0.802944	0.474220	0.694815

Table 3.17: Position of 7 largest financial volume bonds, in order from left to right, in the 4 rankings of influence metrics, considering 98 bonds.

Ranking	PETR4	VALE5	VALE3	BBDC4	BBAS3	PETR3	ITSA4
incoming weight	80	81	33	70	66	68	53
outgoing weight	88	95	60	90	79	84	76
Pagerank	81	82	33	71	67	70	53
reverse Pagerank	91	85	60	92	82	86	81
HITS authority	77	79	34	68	67	69	55
HITS hub	87	95	63	86	79	81	76

Table 3.18: Spearman's ρ between financial volume ranking and influence rankings.

Metric	**volume *vs.* weight**	**volume *vs.* Pagerank**	**volume *vs.* HITS**
incoming	-0.074409	-0.098547	-0.065815
outgoing	-0.197317	-0.197993	-0.195073

highly traded stock is neither very influential nor very influenced. In general, those 7 high valuable stocks are in the half bottom of all 98-size influence rankings, around position 65 for incoming edges and 82 for outgoing edges, considering the arithmetic mean of the values.

Using the similarity between the full ranking, the Spearman ρ coefficient between financial volume ranking and each one of the rankings metrics is shown in Table 3.18. Indeed, the near zero values for the coefficient, ranging from -0.074409 to -0.197993, indicates that there is no correlation between the rankings of traded volume and its influence in terms of information transfer. The negative values (near zero) in the Spearman coefficient suggests that there is a weak inversion between rankings.

Chapter 4

Characterizing Influence Over Time

The previous chapter considered a single transfer entropy network constructed from the dataset in order to characterize influence. In particular, the transfer entropy of the edges were computed using the entire time series of the corresponding nodes. Under this simplified model lies the assumption that transfer entropy is homogeneous and consistent over time. Alternatively, this simplified model offers a general or average description of the system (in terms of influence) over the entire time period.

However, transfer entropy (and other pairwise interaction processes) are most often not homogeneous or consistent over time. Indeed, they are subject to other external factors and can vary significantly over time. Consequently, the characterization of influence will also vary over time. This chapter focuses on understanding influence over time by considering not a single network but a series of transfer entropy networks, as discussed below. Indeed, it will be shown that top level influence (the most influential or the most influences) is highly dynamic over time.

4.1 Networks over time

Recall that pairwise transfer entropy is computed using the corresponding time series over some period of time. Thus, given a time period, the transfer entropy can be computed for all pairs and a transfer entropy network can be constructed. This idea will be used to build a sequence of networks. Finally, each transfer entropy network can be analyzed independently, characterizing its influential and influenced nodes. The influence characterized in each network will provide a characterization of the influence over time.

One important consideration is determining the periods of time that will be used to build the transfer entropy network. In order to allow for a smoother and more flexible analysis, this work adopts a sliding window approach. The window determines a period of time over the entire time series that is used to build the the transfer entropy network. There are two important parameters for the time window: size and shift. The *size* determines the duration of the period of time of a single window. The *shift* determines the amount of time that changes between two consecutive windows, and thus the overlap between them. Window size must not be too short so that there is not enough data to yield relevant information (estimation of transfer entropy). Likewise, it must not be too long so that different dynamics are not merged into a single network.

Figure 4.1 illustrates how two consecutive time windows overlap. The window size is 24 months (4 periods of 6 months) and the shift is 6 months. This gives an overlap of 18 months (3 periods of 6 months) between two consecutive windows. The first window has an additional 6 months before the common period while the second has 6 months after that common period. Therefore, two consecutive time windows have 3/4, 75%, of the same raw data, those coming from the daily trading records of [B][3] shares. This overlap is important because the analysis of each window should have some continuity with the analysis of its adjacent windows.

Applicable values for window sizes were 12, 24 and 36 months and shift values were 6, 12, and 24 months. While experiments were performed using a combination of these values, in the following results window size is defined as 24 months and window shift is 6 months.

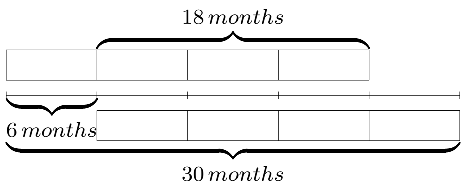

Figure 4.1: The overlap between two consecutive time windows.

Note that the entire time period of the dataset consists of 384 months (32 years, from 1986-01 to 2019-12) yielding a total of 65 time windows, as ilustraded in Figure 4.2. For each time window, a transfer entropy network is generated. Only the time series within that time window is used to determine the stocks that have enough liquidity and then calculate the transfer entropy between those pairs, constructing the transfer entropy network. Note that the number of nodes can vary among the networks, as this corresponds to number of stocks with enough liquidity within that time window (soon to be discussed).

4.2 Results over time

The previous chapter used three metrics to assess strength of ties among nodes and determine influence: node weight, Pagerank and HITS. In that follows, due to similarity of the results, HITS metric was set aside and all measurements are done

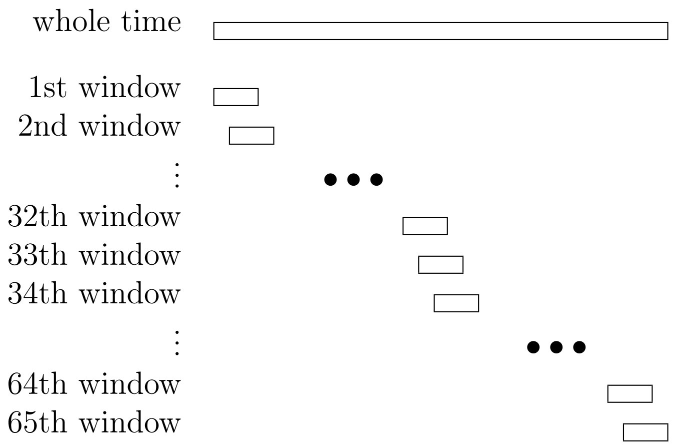

Figure 4.2: Operation of 65 sliding time windows, where whole time consists of 384 months and each time windows consists of 24 months, separated by shifts of 6 months.

with node weight and Pagerank. This allows a better focus on other important aspects concerning dynamics.

4.2.1 Number of stocks by time window

We start by considering the number of stocks in each time window. Recall that for each time windows, stocks are removed if they do not meet the minimum liquidity or represent a *stock option*. Figure 4.3 shows the number of stocks in each time window, indicating that this number varies significantly and has a growing trend. The minimum value of 135 is reached in the window #23; a local peak of 762 happens in window #43; and it ends with a maximum value of 910 in window #65.

4.2.2 Edge weight (transfer entropy) over time

Recall that time window is analyzed independently, so for each time window a network is constructed using the data corresponding to that time window. Figure 4.4 shows the CCDF for the edge weights of 4 time windows (from a total of 65) separated by 5 year intervals. The figure shows that the average and the tail of the distributions grows significantly over time.

This growth of information transfer in the network is also illustrated in Figure 4.5 which shows the average edge weight over time. Clearly, information transfer between stocks have becomes more intense (higher average) and more concentrated (heavier tail) over time.

Figure 4.6 shows a *heatmap* (roughly speaking, a sequence of linearized *colormaps*) where each column represents a time window and each line corresponds to a directed edge. The color indicates the relative magnitude of the transfer entropy of the edge in that time window with respect to the total information transfer.

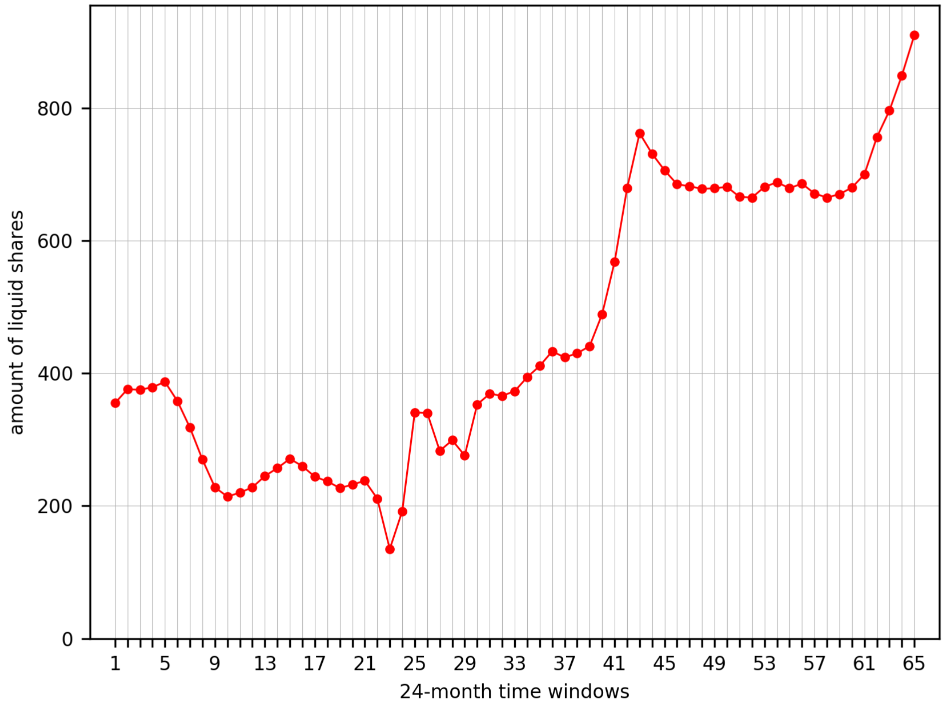

Figure 4.3: Number of liquid stocks (network size) in each time window.

Darker columns indicate higher average values and more concentrated values. The figure shows that a few specific edges over time have a significant increase in their information transfer.

Note the abrupt change in the line pattern in the heatmap in Figure 4.6 around window 25. This was caused by the change of stock names. All stock names changed on March 16, 1998 and since windows size is 384 months, all windows including this key date will be affected (i.e., the same stock can appear twice). The first window affected is #23, beginning at July 1996 and ending at June 1998 whilst the last window affected is #26, beginning at January 1998 and ending at December 1999. Subsequent windows only have the new stock names.

Edges can be ranked by the amount of information they transfer. Thus, for each time window, consider the edge with the largest edge weight. Does this edge remain a top ranked edge in other time windows? The answer is no in general, and the result is shown in Figure 4.7. The plot shows all edges that are the top ranked in a time window and track their relative position in other time windows (among the top 20). It can be seen as a line chart, where each line represents one edge, that indicates which position a edge reaches in the ranking of the windows. Although, as it is not possible to show all edges due great amount of lines, and would not be readable, only upper part of the chart with 20 positions are shown, and only those

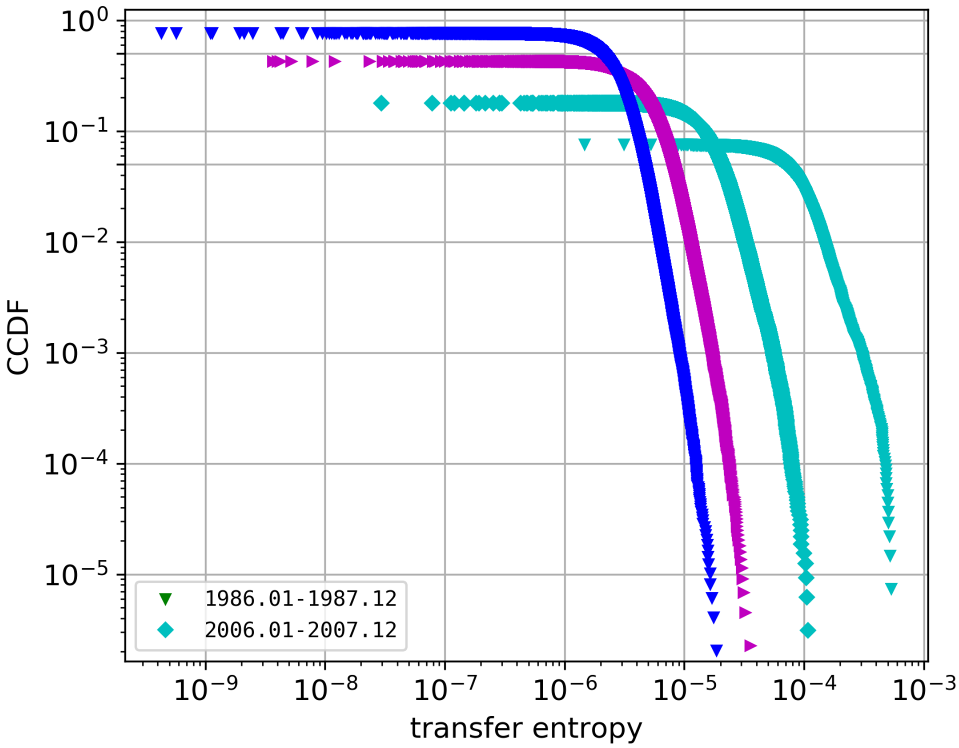

Figure 4.4: CCDF of edge weights for four different time windows separated by 5 years, in log-scale

edges which reach number 1 position in at least one window. Note that very few edges that are top 1 in a time window appear as top 20 in a different time window. Thus, there is significant change in the ranking of the information transfer among top ranked edges.

More specifically, almost all 65 #1 positions are occupied by different edges, with the exception of two: edge `CPFL4`→`BBDC3F` in windows #27–28 and edge `BESP3`→`BESP4` in windows #29–30. Moreover, edge `TELB4F`→`TELB3F` is #1 in window #24 and #2 in window #25, and `BESP3`→`BESP4`, just seen before, was also #4 in window #28. In total, there were 73 appearances of these #1 edges in the top 20 positions along 65 time windows, 12,4% of a total of 591 occurrences. Thus, some edges do maintain their significance in transferring information over time.

4.2.3 Node weight over time

As previously discussed, the node outgoing and incoming weight reflects its role as being influential and influenced, respectively. Thus, for each time windows the node outgoing and incoming weight is computed and used to rank the nodes in that time window. Do top ranked nodes in a given time window are also top ranked in other time windows? Figure 4.8 answers this question, as it is a line chart, where each

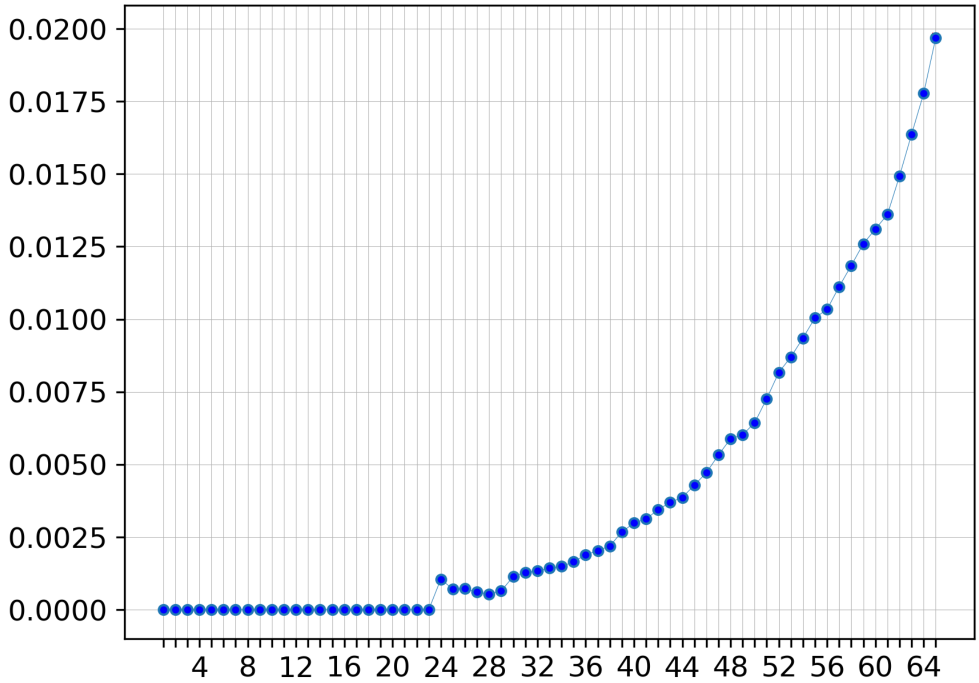

Figure 4.5: Average edge weight (information transfer) in each time window.

line represents one share, that indicates which position a share reachs in the rank of the windows. Although, as it is not possible to show all shares due great amount of lines, and would not be readable, only upper part of the chart with 20 positions are shown, and only those shares which reach number 1 position in at least one window. So, a total of 57 different stocks appear 170 times in this chart showing while the top position varies significantly, some stocks remain as the most influenced over time.

More precisely, 6 stocks appear as top ranked more than once, but not more then three times and not always consecutively: `FES_4`, in windows #20–#22, `LAB_2`, in windows #2–#4, `LCAM3F`, in windows #56 and #63, `OGXP3T`, in windows #46 and #47, `SMAL11F`, in windows #60–#61, and `VALE3T`, in windows #43 and #50.

Figure 4.9 shows the CCDF of incoming node weight for seven different time windows in different years. Clearly, the average and the tail of the distribution are increasing over time, indicating that nodes are becoming more influenced over time. This increase in incoming node weight is related to the increase in information transfer among network edges, as previously discussed. This increase will also manifest itself on outgoing node weight, as discussed below.

Recall that outgoing node weight reflects how influential a stock is in the market, since its outgoing weight represents the total amount of information it transfers to other nodes. Do influential stocks remain influential over time? Figure 4.10 shows

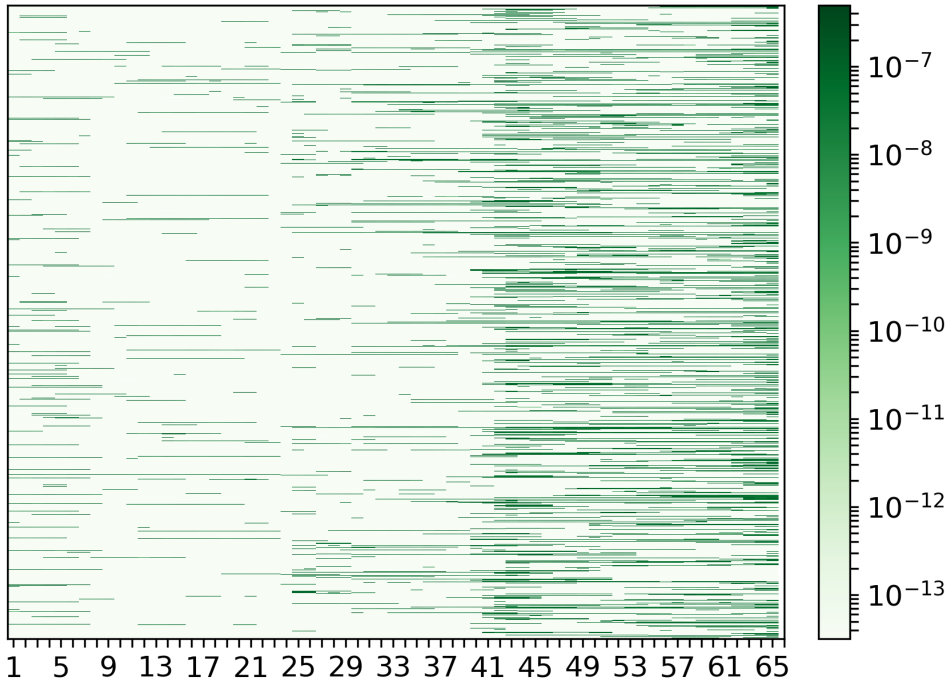

Figure 4.6: Relative information transfer of edges across different time windows: Heatmap of all edges in each time window.

the ranking of nodes that appeared as top ranked (number 1) in at least one time window. While no stock appears more than once as top ranked, together they appear 126 times in this chart. Clearly, there is significant change in the top ranked influential nodes (by outgoing weight), being even more dynamic then top ranked influenced nodes (by incoming weight).

Figure 4.11 shows the CCDF of outgoing node weight for seven different time windows in different years. As for incoming node weight, the average and the tail of the distribution are increasing over time, indicating that nodes are becoming more influential over time.

4.2.4 Pagerank over time

The weighted version of Pagerank is used to determine the ranking of nodes on the network corresponding to each time window. Recall that this ranking reflects nodes that are more influenced by others (larger incoming weight). Again, consider the stability of nodes in the top of this ranking.

Figure 4.12 shows the ranking of nodes that appear as top ranked (number 1) in at least one time window, as it is a line chart, where each line represents one share, that indicates which position a share reachs in the rank of the windows. Although,

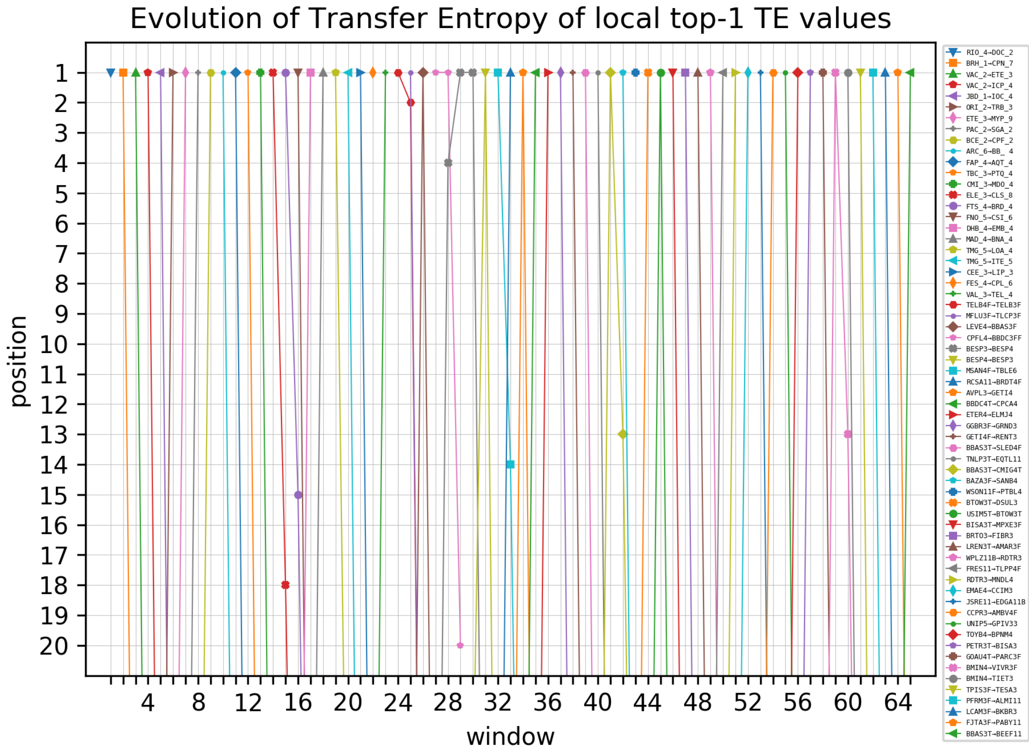

Figure 4.7: Ranking of edges that are the highest ranked edge in at least one time window.

as it is not possible to show all shares due great amount of lines, and would not be readable, only upper part of the chart with 20 positions are shown, and only those shares which reach number 1 position in at least one window. A total of of 56 different stocks appear a total of 173 times in this chart, again showing that there is significant change among the top ranked influenced nodes. Interestingly, note the similarity of this dynamics with that of the ranking based on incoming node weights, shown in Figure 4.8. This indicates that ranking based on incoming node weight is consistent with Pagerank.

Recall that by reversing the edges of the network and applying Pagerank on this modified network will yield a ranking that is based on the original outgoing edges (that have been reversed) and thus will capture how influential are the nodes. This procedure can be applied to the network of each time window in order to determine the influential nodes over time. Is there stability among the top ranked influential nodes?

Figure 4.13 shows the ranking for node that were top ranked (number 1) in at least one time window. A total of 65 different stocks appear a total of 117 times in this chart, again indicating significant movement among top ranked influential stocks. Again, this dynamics is similar to the one found in the top rankings based

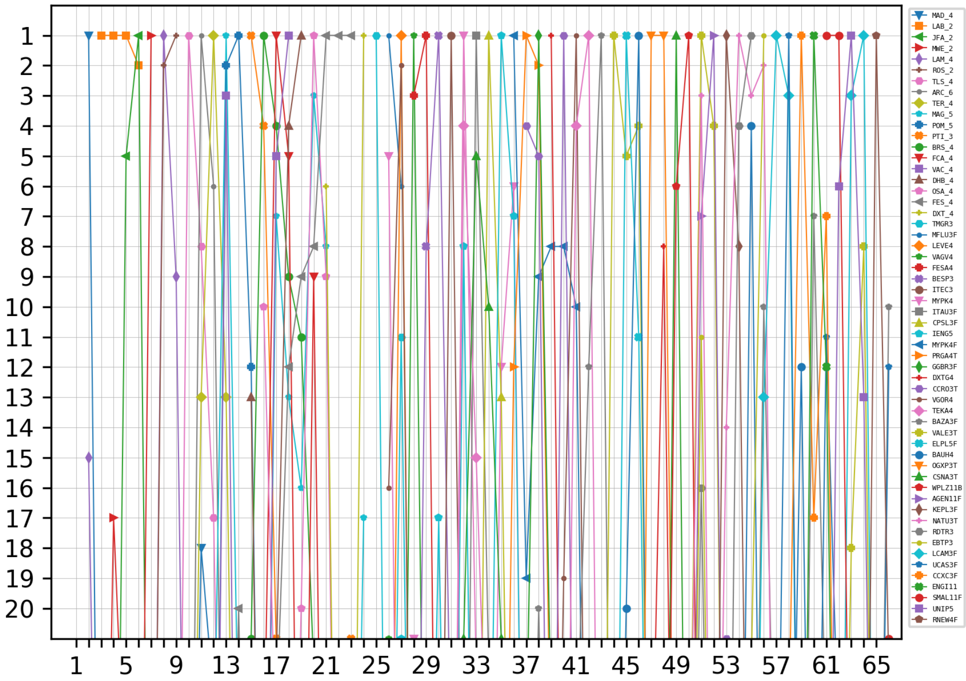

Figure 4.8: Ranking according to incoming edge weight for nodes that are highest ranked in at least one time window.

on outgoing edge weights, shown in Figure 4.10.

The different rankings for top ranked stocks have a similar characteristic: besides being very dynamic with many different stocks taking the top ranked position, top ranked stocks rarely appear in the top 20 positions of other time windows. In order to quantify this finding, consider a stock that its first appearance in the top 20 is the top ranked position, as well as a stock that its last appearance in the top 20 is the top ranked position. Table 4.1 shows the number of stocks that have these characteristics for the four rankings considered. Note that incoming node weight and Pagerank show very similar statistics, as well as outgoing node weight and reverse Pagerank. But more interesting, when considering influential nodes (outgoing node weight and reverse Pagerank), more than 50% of the time (34/64 or 35/65) the last recorded position of a stock is the top ranked position. In other words, a stock that reaches the number 1 position of the ranking disappears from the top 20 positions in subsequent time windows in 50% of cases. It seems that top ranked influential stocks significantly lose their influence afterwards.

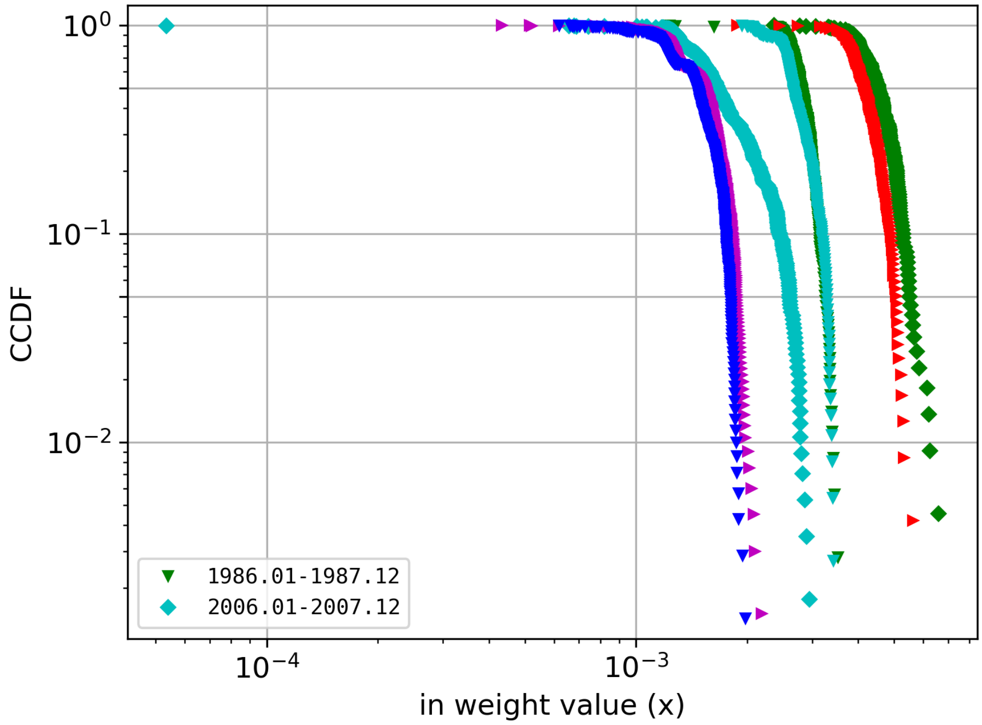

Figure 4.9: CCDF of incoming node weights for seven different time windows in different years, in log-scale.

Table 4.1: Number of times that a top ranked (number 1) stock is its first or last appearance among the top 20 positions in other time windows, for different ranking metrics.

	incoming weight	ordinary Pagerank	outgoing weight	reverse Pagerank
first time	11	11	28	29
last time	11	10	34	35

4.2.5 Transaction volume over time

This section considers financial indicators over time, using each time window to compute the different financial indicators. This allows understanding of these financial indicators over time, as well as the stability of highly ranked stocks.

Figures 4.14 and 4.15 show the ranking for stocks that were top ranked (appeared as number 1) in at least one time windows with respect to financial volume and amount of stocks dealt, respectively. Interestingly, there is significant stability with respect to top ranked stocks in terms of financial volume. More precisely, only 7 different stocks appear as top ranked and in total appear 137 times in this chart. Note that almost all stocks that reach the top ranked position remain in the top rank for years (2 windows corresponds to one year).

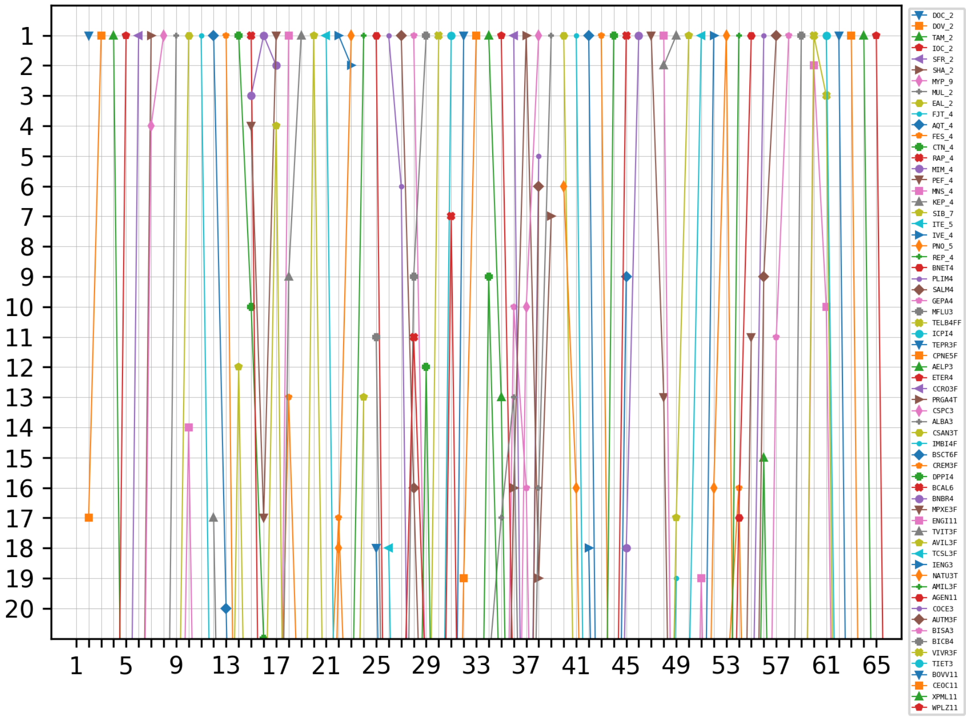

Figure 4.10: Ranking according to outgoing edge weight for nodes that are highest ranked in at least one time window.

The top ranked stocks in terms of amount of stocks dealt is also quite stable, as shown in Figure 4.15. In particular, only 16 different stocks appear in the top ranked position and in total they appear 213 times in this chart. Interestingly, `FNAM11` has been the top ranked stock since window 50 (for over 7 years) while all other have disappeared from the top 20.

The stability of top ranked stocks with respect to these two financial indicators is in sharp contrast to the lack of stability of top ranked influential or influenced stocks. Clearly, top ranked influence in the stock market is much more dynamic than top ranked financial indicators.

4.2.6 Node weight and Pagerank *vs.* transaction volume

An important consideration is the relationship between the rankings produced by financial indicators and the rankings of influence proposed in this work. This relationship is investigated across the different time windows using the top 20 ranked stocks (measured by the Jaccard similarity) as well as the entire ranking (measured by the Spearman coefficient). Since each time window provides a similarity metric between the rankings, we report on the average and variance across all 65 time windows.

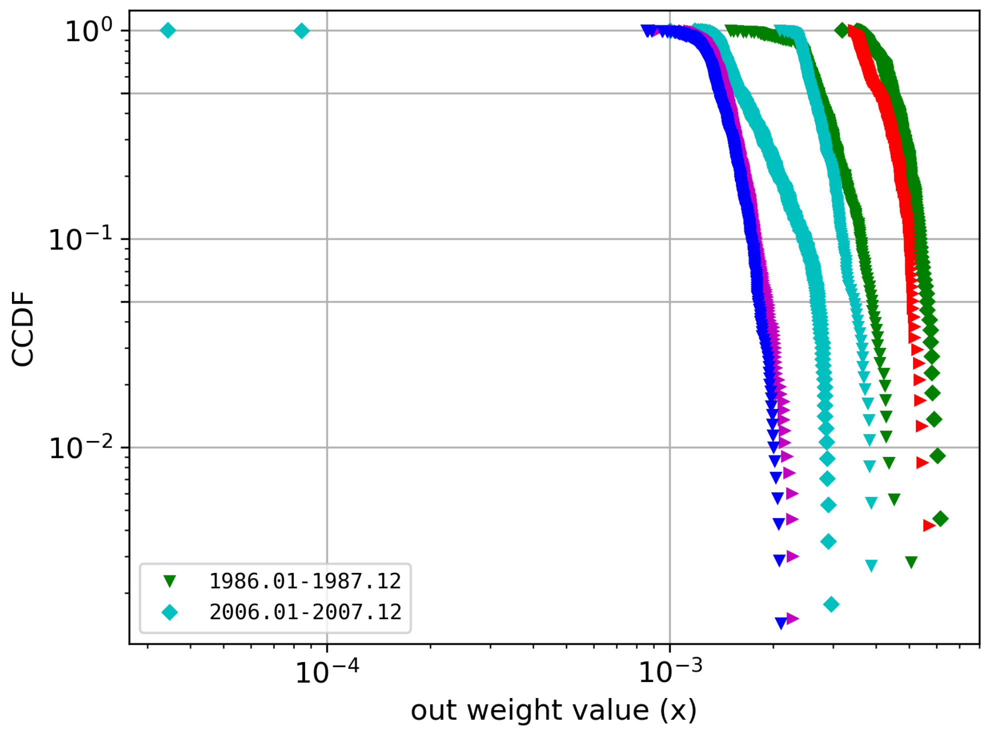

Figure 4.11: CCDF of outgoing node weights for seven different time windows in different years, in log-scale.

Table 4.2 shows the result of comparing the different rankings of influence with financial volume for the four ranking metrics. Clearly, when considering the top 20 of each ranking, there is no similarity as the average Jaccard coefficient is near zero when comparing with all four influence rankings. When considering the entire rankings, results using Spearman coefficient shows a weak negative value in three rankings (the exception is Pagerank), indicating that such rankings are mostly inverted. Thus, financial volume alone also does not capture the overall ranking of influential and influenced stocks.

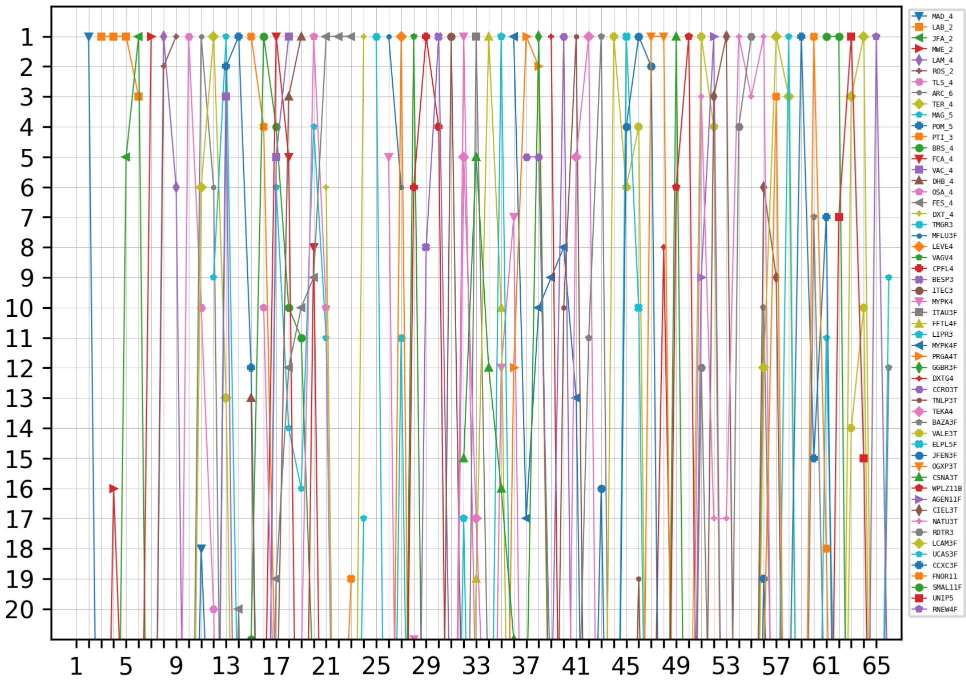

Figure 4.12: Ranking according to Pagerank for nodes that are highest ranked in at least one time window (influenced nodes).

Table 4.2: Comparing weight and Pagerank rankings with financial volume ranking, with mean (μ) and variance (σ^2) of all 65 windows, using Jaccard and Spearman coefficients.

Jaccard

	incoming metric		**outgoing metric**	
	weight	Pagerank	weight	Pagerank
μ	0.009231	0.009231	0.012308	0.013077
σ^2	0.000992	0.000992	0.001002	0.001021

Spearman

	incoming metric		**outgoing metric**	
	weight	Pagerank	weight	Pagerank
μ	-0.375695	0.381413	-0.434775	-0.436110
σ^2	0.032814	0.032198	0.015576	0.014370

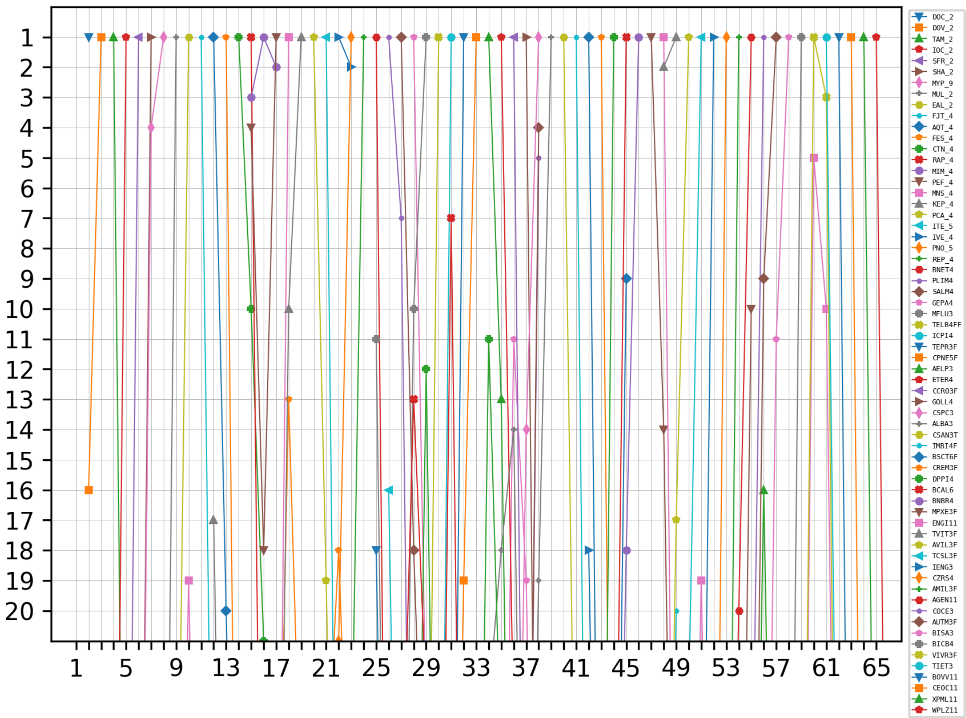

Figure 4.13: Ranking according to Pagerank on the reverse network for nodes that are highest ranked in at least one time window (influential nodes).

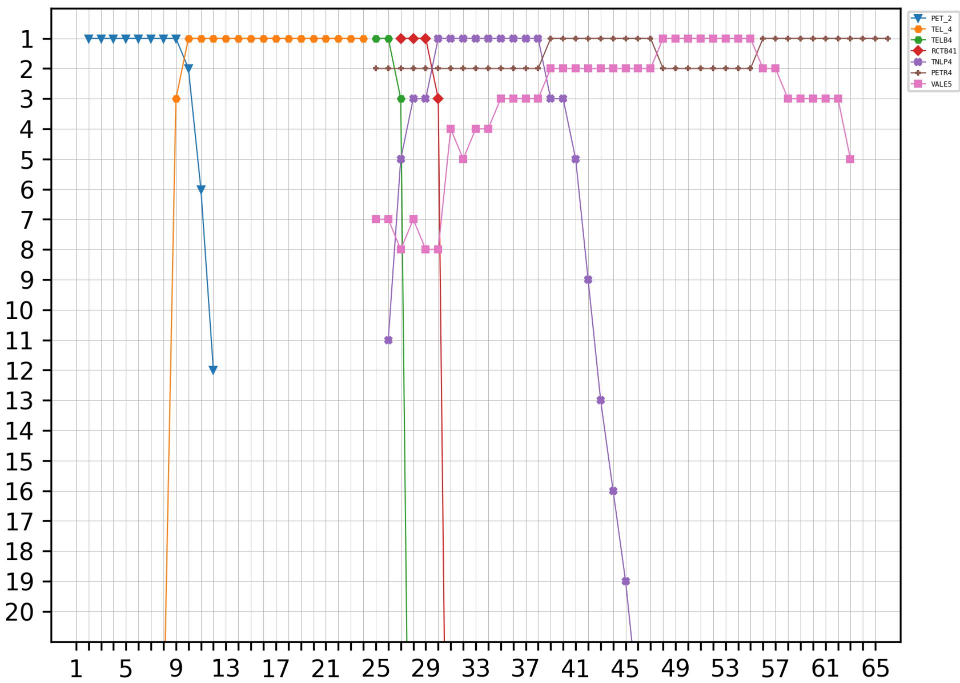

Figure 4.14: Evolution of financial volume (*aka* financial).

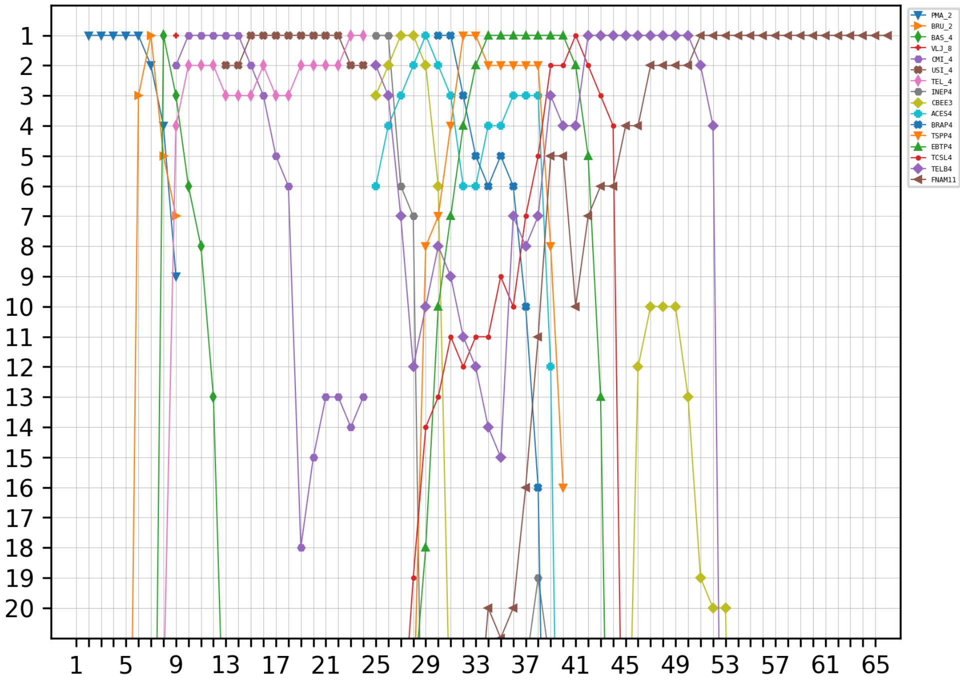

Figure 4.15: Evolution of amount of stocks dealt (*aka* bonds).

Chapter 5

Model for Predicting Mutual Information

Since the beginning of the Information Age, the amount of data generated and processes by various sectors of society has increased exponentially. This advance first fueled by the *Big Data* movement has more recently lead to the emergence of the new and promising field of *Data Science* [34].

Much of the data collected can be naturally modeled by time series. For example, the daily patterns of different stocks within the stock market, or the hourly patterns of different *memes* in online social networks. The dynamics of such time series are often related to one another, and sometimes in quite unexpected scenarios. For example, the daily number of searches for the word "influenza" in web search engines has strong correlation to the number of people infected by influenza [35, 36].

Given the intuitive dependencies that arise between time series dynamics, a fundamental problem is to characterize and measure their relationship. Being so fundamental, it is not surprising that there are various approaches and metrics to measure the dependency between time series. One such approach has originated from information theory and is well-grounded on probability theory. Within this approach, metrics such as joint entropy, relative entropy (also known as Kullback-Leibler distance in statistics), and mutual information have been proposed and are widely used [37].

Mutual information intuitively measures the amount of information that the dynamics of a time series contains of another. It measures how much the knowledge of a time series reduces the uncertainty of another. Mutual information is widely applied to problems in various fields such as machine learning [38, 39], computer vision [40, 41], finance [7, 42], and bioinformatics [6, 43].

While the mutual information between two time series is a straightforward calculation, the computational cost is quadratic on the number of symbols present in the time series. While this is feasible for many scenarios, it can become prohibitive when the number of symbols is extremely large, such as in computer images or DNA sequences. In such scenarios, a model can be used to predict the mutual information between the two time series based on a small number of parameters and having reduced computational cost.

This work proposes a simple and parsimonious model for predicting the mutual information between two time series. The model is based on two independent and identically distributed sequences, and has two key parameters: number of symbols, and probability of copying an element in the other sequence. More importantly, the model requires a constant amount of time to provide a prediction, independently of the number of symbols. We derive an analytical expression for the mutual information of the model, and show numerical results under synthetic data. Last, we show a relationship between the proposed prediction model and transfer entropy.

In this chapter, we present a background on mutual information and some related work; the proposed model to predict mutual information; a numerical evaluation of the proposed model; and an adaptation of the proposed model to transfer entropy.

5.1 Background and Related Work

The idea of measuring the relationship or dependence between time series or random objects is quite fundamental since it finds applications in various domains of knowledge. While there are various approaches, one that is widely adopted emerged from information theory and is grounded on probability theory. This approach started with the concept of *entropy*, introduced by Shannon around 70 years ago to measure information [5]. Given a random variable X and image v_X and probability distribution $p_X(x)$, the Shannon entropy H_X is defined as

$$H_X = -\sum_{x \in v_X} p_X(x) \log_2 p_X(x) \tag{5.1}$$

Intuitively, the entropy measures the amount of information that a random variable carries. The smallest value of zero occurs when there is no randomness ($p_X(x) = 1$, for some $x \in v_X$) while the largest value of $\log_2 |v_X|$ when the randomness is the largest possible ($p_X(x) = 1/|v_X|$, for all $x \in v_X$, the uniform distribution).

While the notion of entropy applies to a single random variable, the general idea can be expanded to accommodate two random variables. In this case, the goal is to measure the relationship or the influence between the two variables. One particular extension is *mutual information* that intuitively captures the amount of information that one random variable has about the other. In particular, it measures the deviation of the joint distribution from the product of the marginals. It is defined as follows:

$$M_{XY} = \sum_{\substack{x \in v_X \\ y \in v_Y}} p_{XY}(x,y) \log_2 \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \ , \tag{5.2}$$

where, $p_{XY}(x,y)$ denotes the joint-probability distribution, and $p_X(x)$, $p_Y(y)$ the X and Y marginals, respectively (in the sequel we shall omit the base of log). Note that if X and Y are independent, then the mutual information is zero. Moreover, note that M_{XY} is symmetric under the exchange of X and Y.

5.1.1 Mutual information on time series

While the concept of mutual information is defined in the context of random variables and their probability distributions, the concept is widely used in practice. Moreover, it is also used to assess the mutual information between time series. While there are different approaches to measure the mutual information between two time series, one simple approach leverages the empirical probability distributions observed in the two time series.

Specifically, let $I = (x_1, \ldots, x_n)$ and $J = (y_1, \ldots, y_n)$ denote two time series of length n, with $x_i \in v_I$ and $y_i \in v_J$ for every $1 \leq i \leq n$. Note that v_I and v_J denote the set of symbols that can appear in time series I and J, respectively. The mutual information between I and J is defined as

$$\widehat{M}_{IJ} = \sum_{\substack{x \in v_I \\ y \in v_J}} \widehat{p}_{IJ}(x,y) \log_2 \frac{\widehat{p}_{IJ}(x,y)}{\widehat{p}_I(x)\widehat{p}_J(y)} \ , \tag{5.3}$$

where, $\widehat{p}_{IJ}$, $\widehat{p}_I$ and $\widehat{p}_J$ denote the empirical distributions, respectively. In particular, for all $x \in v_I$ and $y \in v_J$, $\widehat{p}_{IJ}(x,y) = 1/n \sum_{i=1}^{n} \mathbb{I}(x_i = x \text{ and } y_i = y)$, $\widehat{p}_I(x) = 1/n \sum_{i=1}^{n} \mathbb{I}(x_i = x)$ and similarly for $\widehat{p}_J$, where $\mathbb{I}(\cdot)$ denotes the indicator function.

Note that if the sequence of pairs (x_i, y_i) are independently and identically distributed according to random variables (X, Y), then $\widehat{p}_{IJ}$, $\widehat{p}_I$ and $\widehat{p}_J$ will converge to p_{XY}, p_X, and p_Y, respectively, as n goes to infinity (by the law of large numbers). Therefore, $\widehat{M}_{IJ}$ converges to M_{XY} and we recover the definition provided in Equation (5.2).

5.1.2 Computational aspects of mutual information

The computation for mutual information as defined in Equation (5.2) requires unrolling a double sum, one over the set of values for X and another over the set of values for Y. Hence, the computational complexity is $\Theta(|v_X||v_Y|)$, where $|v_X|$ and $|v_Y|$ denote the number of symbols that can be assumed by X and Y, respectively. Thus, this computation has quadratic complexity in the number of symbols, assuming $|v_X|$ and $|v_Y|$ differ by only a constant.

Note that computing the mutual information for time series data has the same complexity, given that that relative frequencies are available, and is given by $\Theta(|v_I||v_J|)$. However, the computation of the relative frequencies require iterating through the time series, and thus have computational complexity $\Theta(n)$, where n is the length of the time series. Thus, the final computational complexity is $\Theta(|v_I||v_J| + n)$ which is dominated by the first term in scenarios where there are many symbols.

5.2 Prediction model for mutual information

This section presents two related models that can predict the mutual information between two time series. The model assumes that values appearing in both time series come from the same symbol set and their dependency is encoded as follows: given an entry of one series, the corresponding entry of the other series is either a copy or it may be re-sampled, according to a given probability. The re-sample procedure can be performed in two different ways, giving rise to two model variations. In the first model re-sample occurs using the full set of available symbols (repetition are thus allowed), whereas, in the second model re-sampled is restricted to the set of symbols that excludes the symbol assumed by the corresponding entry of the other series (repetition are not allowed).

The model has two key parameters, s and p; $s \geq 1$ is a positive integer denoting the number different symbols and $p \in [0, 1]$ is the re-sampling probability.

With repetitions: Let X be a random variable with uniform distribution on $\{1, 2, \ldots, s\}$, and let Y be another random variable defined as follows:

$$Y = \begin{cases} X & \text{with probability } 1 - p \ , \\ U \sim Unif\{1, \ldots, s\} & \text{with probability } p \ . \end{cases} \tag{5.4}$$

In words, with probability $1-p$ (independent of X) the random variable Y is a copy of X, whereas with probability p, Y is a uniform random variable on $\{1, 2, \ldots, s\}$ independent from X. Note that, in this scenario, the variable U is allowed to take the value of X, thus repetition are allowed even when the value of Y is not copied.

As it turns out, the parameter p can be thought of as a level of dependence between X and Y. In particular, if $p = 1$, X e Y are independent, whereas when $p = 0$, $Y = X$.

Proposition 1. *Let $s \geq 1$ and $p \in [0, 1]$. Let X be a uniform random variable on $\{1, \ldots, s\}$ and Y defined as in Equation* (5.4)*. Then, the mutual information* *between X and Y is given by:*

$$M_{XY}(s,p) = \frac{1}{s}(p+s(1-p))\log(p+s(1-p)) + \frac{s-1}{s}p\log p\ . \tag{5.5}$$

Note that $M_{XY}(s,1)=0$, in accordance with the fact that X and Y are independent when $p=1$. On the other hand, $M_{XY}(s,0) = \lim_{p\to 0} M_{XY}(s,p) = \log s$, which corresponds to the Shannon entropy of a uniform random variable on $\{1,\dots,s\}$ (see, Section 5.1)[1].

Proof. Let us denote $v_Y = v_X = \{1,\dots,s\}$. By definition of mutual information (see, Equation (5.2)) we have that

$$M_{XY} = \sum_{\substack{x\in v_X\\ y\in v_Y}} p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\ .$$

Using that X is uniform on v_X and that Y is given by Equation (5.4), we obtain: $p_{XY}(x,y) = \frac{p}{s^2} + \frac{1-p}{s}\mathbb{I}(x=y)$, $p_X(x) = \frac{1}{s}$, and $p_Y(y) = \left(1-p+\frac{p}{s}\right)\frac{1}{s} + \frac{p}{s}\frac{s-1}{s} = \frac{1}{s}$; i.e., the marginal of Y is also uniform on $\{1,\dots,s\}$. Thus,

$$\begin{aligned}
M_{XY} &= \sum_{x=1}^{s}\sum_{y=1}^{s}\left(\frac{p}{s^2}+\frac{1-p}{s}\mathbb{I}(x=y)\right)\log\left(\frac{\frac{p}{s^2}+\frac{1-p}{s}\mathbb{I}(x=y)}{s^{-2}}\right)\\
&= \sum_{x=1}^{s}\sum_{\substack{y=1\\ j\neq x}}^{s}\frac{p}{s^2}\log(p) + \sum_{x=1}^{s}\left(\frac{p+s(1-p)}{s^2}\right)\log(p+s(1-p))\\
&= \frac{s-1}{s}p\log p + \frac{p+s(1-p)}{s}\log(p+s(1-p))\ .
\end{aligned}$$

□

As mentioned above, intuitively the re-sampling probability p may be thought of as a knob tuning the degree of dependency between X and Y. As it turns out, the mutual information is indeed monotonically decreasing, showing that higher is the value of p lower is the mutual information (degree of dependency), with the minimal value $M_{XY}=0$ attained for $p=1$. Indeed,

$$\frac{\partial}{\partial p}M_{XY}(s,p) = \frac{(s-1)(\log p - \log(s+p-sp))}{s}\ ,$$

and $\frac{\partial}{\partial p}M_{XY}(s,p) < 0$ for all $p<1$ and $s>1$.

[1] In fact, the mutual information between a random variable X and itself (which is the case when $p=0$) is just its entropy, and thus $M_{XX} = H_X$.

Without repetition: This model is a minor variation of the previous; as before, let X be uniform in $\{1, 2, \ldots, s\}$ and Y be now defined as:

$$Y = \begin{cases} X & \text{with probability } 1-p\ , \\ U \sim Unif\left(\{1, \ldots, s\} \setminus \{X\}\right) & \text{with probability } p\ . \end{cases} \tag{5.6}$$

In words, with probability $1-p$ the random variable Y is a copy of X, whereas with probability p, Y is a uniform random variable on $\{1, 2, \ldots, s\}$ minus the value assumed by the variable X. Note that, in this scenario, the range of the variable U depends on the value of X (repetition is not allowed), and therefore U is not independent of X. Thus, X and Y are dependent random variable, even when $p = 1$. Due to the fact that repetition is forbidden (with probability p), for the model to be well-defined we shall assume $s \geq 2$.

Proposition 2. *Let $s \geq 2$ and $p \in [0, 1]$. Let X be a uniform random variable on $\{1, \ldots, s\}$ and Y defined as in Equation* (5.6)*. Then, the mutual information between X and Y is given by:*

$$M_{XY}(s, p) = (1-p)\log(s(1-p)) + p\log\left(\frac{sp}{s-1}\right)\ . \tag{5.7}$$

Note that $M_{XY}(s, 0) = \lim_{p\to 0} M_{XY}(s, p) = \log s$, thus, also in this case, we have that the mutual information between X and Y converges to the Shannon entropy of a uniform random variable on $\{1, \ldots, s\}$ when $p \to 0$. On the other hand, $M_{XY}(s, 1) = \log\left(\frac{s}{s-1}\right)$. As expected, the mutual information between X and Y is not equal to 0, when $p = 1$, because the two random variables are dependent.

Proof. The proof follows the very same steps of the proof of Proposition 1. Simply observe that in the case without repetition we have: $p_{XY}(x, y) = \frac{p}{s(s-1)}\mathbb{I}(x \neq y) + \frac{1-p}{s}\mathbb{I}(x = y)$, $p_X(x) = \frac{1}{s}$, and $p_Y(y) = \frac{1-p}{s} + \frac{p}{s-1}\frac{s-1}{s} = \frac{1}{s}$; i.e., the marginal of Y is again uniform on $\{1, \ldots, s\}$. Thus,

$$\begin{aligned} M_{XY} &= \sum_{x=1}^{s}\sum_{y=1}^{s}\left(\frac{p}{s(s-1)}\mathbb{I}(x \neq y) + \frac{1-p}{s}\mathbb{I}(x = y)\right)\log\left(\frac{\frac{p}{s(s-1)}\mathbb{I}(x \neq y) + \frac{1-p}{s}\mathbb{I}(x = y)}{s^{-2}}\right) \\ &= \sum_{x=1}^{s}\sum_{\substack{y=1 \\ y\neq x}}^{s}\frac{p}{s(s-1)}\log\left(\frac{sp}{s-1}\right) + \sum_{x=1}^{s}\left(\frac{1-p}{s}\right)\log\left(s(1-p)\right) \\ &= p\log\left(\frac{sp}{s-1}\right) + (1-p)\log(s(1-p))\ . \end{aligned}$$

$\square$

In the model without repetition, the role of the re-sampling probability is not so obvious as before. In particular, even when $p = 1$, $M_{XY}(s, 1) = \log \frac{s}{s-1}$, implying that X e Y are not independent. It is quite natural to ask whether, for any given s, there exists a value of p for which the mutual information is 0. Interestingly, computing the mutual information setting $p = \frac{s-1}{s}$, one obtains $M_{XY}(s, \frac{s}{s-1}) = 0$. Indeed, the value $p = \frac{s-1}{s}$ corresponds to the value for which the variable X and Y are independent. Also, whenever repetition is not allowed, the mutual information is monotonically decreasing in p in the interval $(0, (s-1)/s)$, going from the value $\log s$ to 0 and monotonically increasing on $((s-1)/s, 1)$, rising from 0 to $\log \frac{s}{s-1}$. This can be verified using that

$$\frac{\partial}{\partial p} M_{XY}(s, p) = \log \frac{sp}{s-1} - \log(s - sp) \ .$$

Note that the dependency between X and Y is stronger when $p = 0$ than when $p = 1$.

In Figure 5.1, both Equations (5.5) and (5.7) are plotted for different values of s and p. Note that, the value of the mutual information at $p = 0$ (both in the case with repetition and without) coincides with the Shannon entropy of a uniform variables on $\{1, \ldots, s\}$; for example, when $s = 16 = 2^4$ the values at $p = 0$ is 4. Also notice that the mutual information without repetition (dashed lines) attains the minimum value (equal to zero) at a value of $p = \frac{s}{s-1}$; thus for example, for $s = 2$ the minimum is attained at $p = 1/2$, and for $s = 4$ at $p = 3/4$. On the contrary, when repetition is allowed (solid lines) the minimum (equal to zero) is always attained at $p = 1$. Another point worth mentioning, is that the mutual information with/without repetition tend to be very close to each other when s grows. For example for $s = 32$, the two curves are already quite close. Intuitively, this phenomenon is due to the fact that when s is big enough, re-sampling with or without repetition are statistically equivalent. Formally, we can prove this phenomenon noticing that, when s tends to infinity, $\frac{s}{s-1} \approx 1$ and $\frac{1}{s}(p + s(1-p)) \approx (1-p)$. Thus, using the latter into Equation (5.5) and Equation (5.7) we obtain that the mutual information with or without repetition tend to each other when s grows.

5.3 Numerical Evaluation

Let us now consider a sequence of i.i.d. random vectors $(Z_i)_{i=1}^N$, where $Z_i = (X_i, Y_i)$. The joint probability of (X_i, Y_i) may be given either by the model with repetition (Equation (5.4)) or without (Equation (5.6)). We can then compute the mutual information between the vectors $I = (X_i)_{i=1}^N$ and $J = (Y_i)_{i=1}^N$ as discussed in Sec-

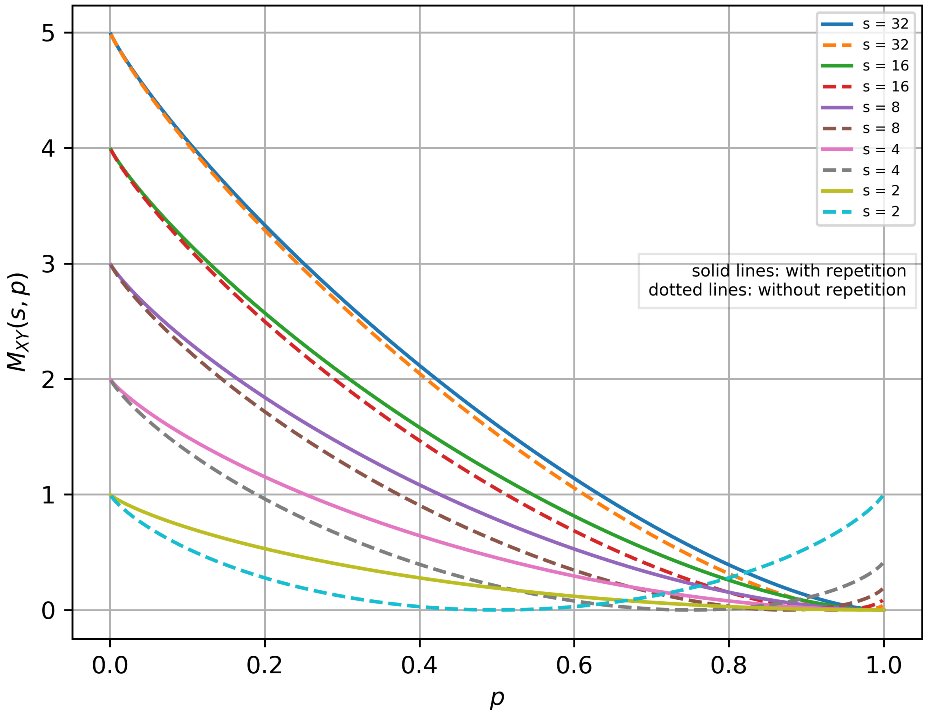

Figure 5.1: Mutual information between X and Y as a function of number of symbols (s) and re-sampling probability (p).

tion 5.1.1. In order to do that we use the observation to compute $\widehat{p}_{IJ}$, $\widehat{p}_I$ and $\widehat{p}_J$ (see Equation (5.3)). Clearly, by the Law of Large Numbers, these values converge to the actual joint probability, and corresponding marginals. Therefore, the mutual influence between the two series $I = (X_i)_{i=1}^N$ and $J = (Y_i)_{i=1}^N$ converges, as $N \to \infty$, to the mutual influence of X and Y; i.e., to Equation (5.5) (when considering the model with repetition) and to Equation (5.7) (for the model without repetition).

To experimentally observe the convergence of $\widehat{M}_{IJ}$ to M_{XY}, we have built simulations (with/without repetition) using several values for: number of symbols (s), time series length (N), and number of rounds.

Figure 5.2 and 5.3 depict the empirical mutual information $\widehat{M}_{XY}$ with/without repetition respectively, as a function of p, for $s = 128$ and different values of time series length ($N = 10^5, 10^6, 10^7, 10^8$), with repetition, and for $s = 4$ and different values of time series length ($N = 10, 50, 100, 500$), without repetition. As it turns out, a good fit with the analytical mutual information M_{XY} is obtained for $N = 10^7$ in first case and 500 in second case.

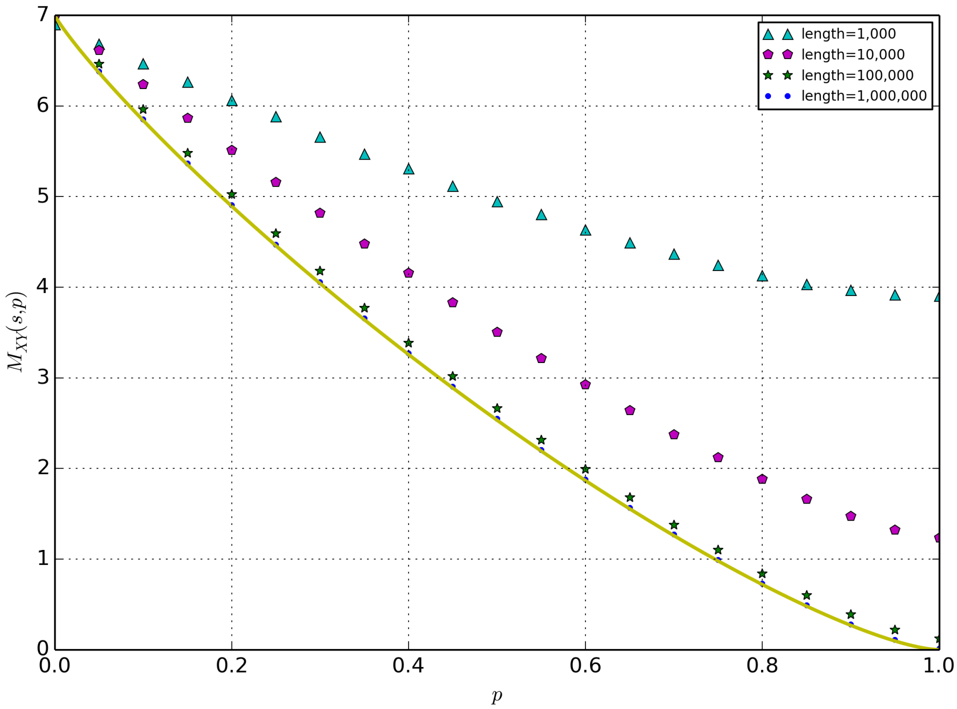

Figure 5.2: Empirical mutual information $\widehat{M}_{XY}$ (with repetition) with $s = 128$ for different time series lengths. The solid line represents the theoretical value for the mutual information M_{XY} (with repetition).

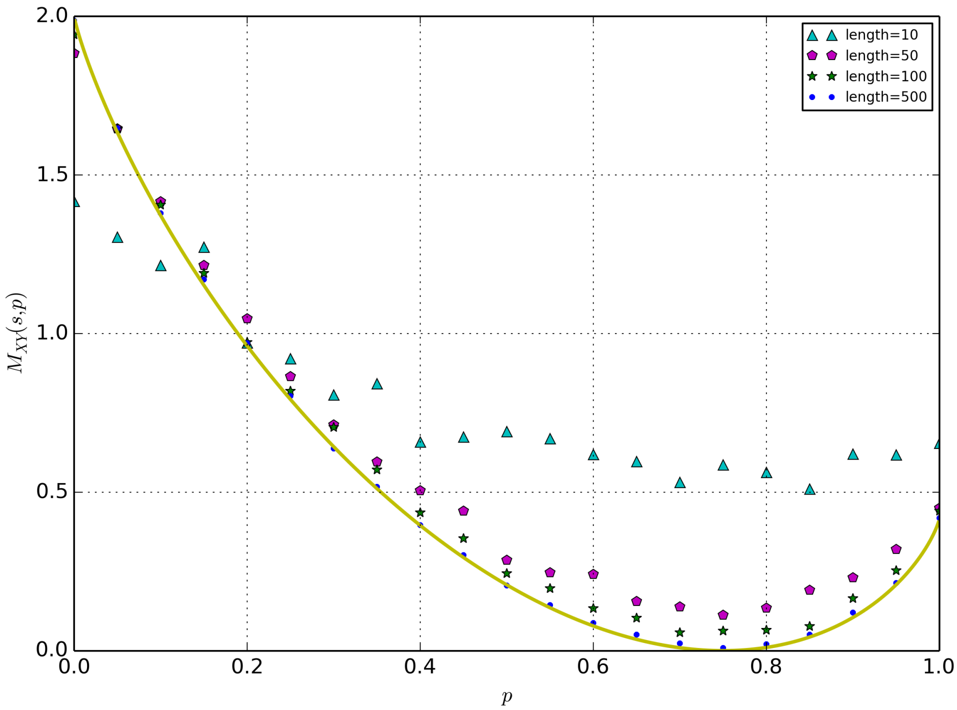

Figure 5.3: Empirical mutual information $\widehat{M}_{XY}$ (without repetition) with $s = 4$ and different time series lengths. The solid line represents the mutual information M_{XY} (without repetition).

5.4 Relationship to Transfer Entropy

A more recently proposed measure of the dependence between two time series is *transfer entropy* [1]. As opposed to mutual information, which uses the static probability distribution of the two time series (joint and marginals), transfer entropy takes the dynamical structure of the time series into account. Specifically, transfer entropy incorporates the dynamical structure by studying the transition probabilities rather than static probabilities (as mutual information).

Intuitively, if the dynamics of a time series $I = (x_1, \ldots, x_n)$ is described by the transition probabilities $p(x_{i+1}|x_i)$ (Markovian dynamics), transfer entropy measures the amount of information transferred by the knowledge of another time series $J = (y_1, \ldots, y_n)$, onto the dynamics of I by comparing $p(x_{i+1}|x_i)$ and $p(x_{i+1}|x_i, y_i)$. Differently from mutual information, transfer entropy is not symmetric, as one time series might reveal information about another, but not vice-versa.

Transfer entropy has been applied in various domains, and in particular in the financial sector where time series are derived from the stock market [19, 20]. A recent book reviews the applications for transfer entropy while also providing some of its fundamental aspects [2].

While originally defined for time series (i.e., empirical data), transfer entropy can also be defined in terms of random variables and their (conditional) probability distributions. Let $X = (X_1, X_2)$ denote a random vector and Y a random variable. The transfer entropy from Y to X is defined as follows:

$$T_{Y \to X} = \sum_{\substack{(x_1,x_2) \in v_X \times v_X \\ y \in v_Y}} p_{XY}(x_1, x_2, y) \log \frac{p_{XY}(x_2|x_1, y)}{p_X(x_2|x_1)} \,, \tag{5.8}$$

where p_{XY} denotes the joint probability distribution of X and Y, p_X the corresponding marginal distribution for X, and v_X and v_Y their images, respectively. Note that $T_{Y \to X}$ measures the influence of the knowledge of Y in predicting X_2, given that X_1 is known. In essence, transfer entropy measures how much the knowledge of Y reveals about the *transition probability* $p_X(x_2|x_1)$. In particular, whenever the transition probability does not depend on Y, i.e., $p_X(x_2|x_1) = p_{XY}(x_2|x_1, y)$, for all x_1, x_2, y, the transfer entropy between Y and $X = (X_1, X_2)$ is equal to zero.

Note that the calculation of Equation (5.8) requires unrolling a triple sum across the possible values taken by $X = (X_1, X_2)$ and Y. Thus, its computational complexity of computing the transfer entropy directly from its definition is $\Theta(|v_X|^2 |v_Y|)$ which is cubic on the number of symbols if $|v_X|$ and $|v_Y|$ differ by only a constant fraction. Therefore, it has a significantly higher complexity than mutual information. Thus, a simple model to predict the transfer entropy could be quite useful, if it can provide an estimation for the transfer entropy in constant time. Could the model proposed in this paper do the job?

In order to answer this question, consider the following adaptation of the proposed model. Lets assume that X_1 and X_2 are i.i.d. with uniform distribution on $\{1, 2, \ldots, s\}$, and let $Y = X_2$ with probability $1 - p$, or be chosen uniformly at random on $\{1, 2, \ldots, s\}$ with probability p, independently from X_2 and X_1. Thus, Y can exactly predict X_2 with probability $1 - p$, and intuitively, as p decreases more information will be transferred from Y to X_2. Moreover, since X_1 and X_2 are independent, X_1 has no information about X_2, and thus, p and s are the only parameters of this model.

The model above induces an extremely simplified scenario for measuring the transfer entropy, as there are no correlation between X_1 and X_2, but the second term X_2 depends on Y. Thus, the transfer entropy from Y to X in this scenario can be easily computed, and the next proposition shows that it is equal to the mutual information between Y and X_2.

Proposition 3. *Let $s \geq 1$ and $p \in [0, 1]$. Let $X = (X_1, X_2)$ be a uniform random variable on $\{1, \ldots, s\} \times \{1, \ldots, s\}$ and Y as defined above. Then, the transfer entropy from Y to X is given by:*

$$T_{Y \to X}(s, p) = M_{X_2, Y}(s, p) \ , \tag{5.9}$$

where $M_{X_2,Y}(s, p)$ is the mutual information between X_2 and Y, as given by Equation (5.5).

Proof. Let $v = \{1, \ldots, s\}$. Note that if the random vector (X_1, X_2) is uniformly distributed on $v \times v$, then X_1 and X_2 are independent and identically distributed, with uniform distribution on v. This implies that $p_X(x_2|x_1) = p_X(x_2)$. Also, it is not difficult to see that X_1 and X_2 are conditionally independent given Y, i.e., $p_{XY}(x_1, x_2|y) = p_{XY}(x_1|y)p_{XY}(x_2|y)$. Since X_1 is independent from Y, i.e., $p_{XY}(x_1|y) = p_X(x_1)$, we obtain that $p_{XY}(x_1, x_2|y) = p_X(x_1)p_{XY}(x_2|y)$. Thus,

$$p_{XY}(x_2|x_1, y) = \frac{p_{XY}(x_2, x_1, y)}{p_{XY}(x_1, y)} = \frac{p_{XY}(x_2, x_1|y)}{p_{XY}(x_1|y)} = \frac{p_{XY}(x_2|y)p_X(x_1)}{p_X(x_1)} = p_{XY}(x_2|y) \ .$$

Substituting in Equation (5.8), it follows that

$$
\begin{aligned}
T_{Y\to X} &= \sum_{\substack{(x_1,x_2)\in v_X\times v_X \\ y\in v_Y}} p_{XY}(x_2,x_1,y)\log\frac{p_{XY}(x_2|x_1,y)}{p_X(x_2|x_1)} \\
&= \sum_{\substack{(x_1,x_2)\in v_X\times v_X \\ y\in v_Y}} p_{XY}(x_2,x_1,y)\log\frac{p_{XY}(x_2|y)}{p_X(x_2)} \\
&= \sum_{\substack{x_2\in v_X \\ y\in v_Y}} p_{XY}(x_2,y)\log\frac{p_{XY}(x_2,y)}{p_X(x_2)p_Y(y)} \\
&= M_{X_2,Y}(s,p)\ ,
\end{aligned}
$$

which is the mutual information between X_2 and Y. $\square$

Thus, the transfer entropy from Y to X of the adaptation of proposed model is identical to the mutual information between X_2 and Y. Another less trivial adaptation could allow for X_2 to depend on X_1, and thus provide a more adequate model for the concept of transfer entropy. However, this further investigation is left for future work.

It is important to clarify that the model proposed in this chapter was not used to calculate transfer entropy. Also mutual information was not used to generate data that was used to build those rankings.

Chapter 6

Conclusion

Transfer entropy has recently emerged as an ubiquitous metric to measure the influence between two dynamics, capturing the notion of information transfer. This work has leveraged this metric to measure influence among shares traded in the Brazilian stock market exchange, [B][3] (former BM&FBOVESPA). In particular, we have considered a 32-years daily record of stock price movement to characterize the influence (or information transfer) between shares in the market. Beyond computing the pairwise transfer entropy among the time series shares, we proposed a network-based approach to identify both shares that are influential and shares that are influenced across the market.

We construct transfer entropy network where shares correspond to nodes and transfer entropy values correspond to weighted directed edges between shares. Classic network centrality metrics such as Pagerank and HITS were used to rank nodes, both in terms of incoming and outgoing edges, revealing influential and influenced shares in the market. We show that a small fraction of the shares have ranking metric values that are significantly larger than average, indicating their prevalence as either influential or influenced shares. This observation is consistent across rankings.

We also find very good consistency among the three rankings (Weight, Pagerank and HITS) concerning the most (top 20) influenced shares in the market, while the agreement concerning the most influential is less pronounced among the three rankings. Moreover, our analysis indicates that the most traded shares (evaluated according to their transaction financial volume) in the market are not the most influential nor the most influenced, indicating a lack of correlation between traded volume and information transfer.

The initial stock market exchange network was constructed using the entire time series of price movements, across a period of 32 years. In this situation a share, identified by its name, is assessed over the whole period. A few 98 shares succeeded such condition.

Next in the work, we have characterized influence over time by using same methodology but considering shorter time series, with 2 years long. Thus, it was easier to a share to bypass the tough filter of appearing along 21 years to be assessed. So, it was possible to understanding the persistence of top influential and influenced shares over time.

An important remark is that it was not an objective of this work to establish any demonstration of cause and effect between the obtained rankings and external factors that could lead to these results. Furthermore, that analysis involves knowledge of economics, financial markets trading *etc*. One of the main proposals used in this work was exactly to try not to worry about such external factors to keep focus on the numbers.

Thus, this research work intended to answer the question of whether information transfer, here handled as *influence* and measured using transfer entropy metric, among a set of nodes, here representing shares of a stock market exchange, could be revealed from data to enable ranking of those nodes into the most influenced or most influential ones, over a period of 32 years. Some hypothesis were confirmed and others refuted, as previously discussed and summarized as follows:

1. Regardless of the influence metric used, be it *sum of incoming edge weights*, *pagerank* or *HITS authority*, numerical results and their respective charts exhibit similar shapes for corresponding input data, *i.e.*, the weight sum of incoming edges, the pagerank computation and HITS authority calculation are very similar and, look like they might be interchangeable used to meter influenced nodes. Likewise, *sum of outgoing edge weights*, *reverse pagerank* or *HITS hub* show similar patterns, so one could choose any one of them to unveil influential shares (or nodes).

2. Data unveiled it is possible to realize some shares as hubs and other ones as authorities with a much higher value for influence than the average, be it incoming or outgoing. Shares in the top of rankings capture our attention as candidates to be followed closer. That situation might guide some stock broker to watch some specific shares in order to trigger some sell/buy order depending on his strategy, based on values for influence of shares, by instance.

3. As well, data unveiled it is possible to notice some shares which can't influence other ones and some that are prove to be influenced by other, as their calculated valued for influence are very low. Almost like a contrary sense, shares in the bottom of rankings also capture our attention as candidates to be followed closer. That situation might be used as an strategy of conservative stock brokers to stay safe during a turbulent stock market time, holding positions with these stocks.

4. It is clear from historical data of financial indicators, such as financial volume, corroborating common sense, that a few stocks move much more money than others, often participating in stock exchange price indexes, such as Ibovespa, as in other parts of the world there are indices such as Down Jones (USA), DAX (Germany), FTSE 100 (United Kingdom) and Nikkei 500 (Japan). They have their importance into stock market exchange due the amount of money they move on and, almost as a consequence, their high liquidity, which is useful to make money as soon as needed by their shareholders.

5. From the data, it was proved that financial indicators (volume, bonds and deals) are not correlated to influence indicators (node weight, pagerank, HITS), which sounds to go against common sense, as correlation coefficients (Jacquard and Spearman) between these rankings reveal a low value. In other words, the most important shares from the point of view of financial value are not the ones that appear as the most influenced or the most influential shares.

Finally, in this work we have proposed a simple and parsimonious model for predicting the mutual information value between two time series without the cost of an empirical calculation, grounded on two simple parameters: number of symbols and length of series, besides the with/without repetition model for generating a copied series for the controlled study.

We have then found an analytical formula for that purpose, proving that, for a certain combination of large series lengths, according to number of symbols, it works. Also, we notice that the novel transfer entropy metric, adapted to the proposed model is identical to the mutual information, although at a higher cost, left for further studies.

6.1 Future Work

Some important issues directly related to the main object of study were not addressed here, but could be be addressed in future work, as follows:

1. A single organization deals with several shares in the stock market exchange. Some shares give owners rights to vote at shareholders' meetings or preference in receiving dividends, for example. But share names do not necessarily follow a strict rule. As long as time passes by those names can change, be split, be joined or disappear, and even new ones can emerge. And that can happen individually anytime. But the problem lies in the fact that these changes cannot be easily tracked. This fact means that a share do not carry a same name for the whole period of time or that there is no share name that lasts

for the whole period of time. So, this causes a discontinuity in the time series every time this name change occurs. Soon, this temporal analysis of a stock is impaired and, consequently, the temporal analysis of pairs of stocks is also impaired.

As a suggestion a research could be done to normalize share names over time, giving them a unique id over the whole period of time, regardless of its name over time, and, thus, allowing a more accurate analysis for pairs of time series of shares.

2. This research transform time series of (almost continuous) share prices into time series of discrete symbols representing shares moves. Using four symbols, for example, it is possible to capture upward moves, downward moves, lack of move or a "zero" move. Using more than four discrete symbols it is possible to represent more than one level of upward moves and, idem, of downward moves. If we name, say, 3 levels of upward moves as *up1*, *up2* and *up3*, and 3 same 3 levels of downward moves, say, as *down1*, *down2* and *down3*, a pair, for example, (*up1*,*up2*) is obviously different of (*up2*,*up3*), but they sound like same thing, which is not considered during analysis.

   So, there are some pairs considered completely different, regardless some of them are next to each other, and in some way are similar. So, some research about a model that could capture this behavior should be done.

3. Unfortunately, the simple mathematical model proposed to predict the mutual information and also the transfer entropy between two random variables was not evaluated with many real data sets.

   Research needs to be carried out to verify if the proposed model can predict mutual information and transfer entropy of real time series, under what conditions and in which domains.

References

[1] SCHREIBER, T. "Measuring information transfer", *Physical Review Letters*, v. 85, pp. 461–464, Jul 2000.

[2] BOSSOMAIER, T., BARNETT, L., HARRÉ, M., et al. *An introduction to transfer entropy: information flow in complex systems*. Gewerbestrasse 11, 6330 Cham, Switzerland, Springer Nature, 2016.

[3] DE PAULA NEVES NETO, J., FIGUEIREDO, D. R. "Ranking Influential and Influenced Shares Based on the Transfer Entropy Network". In: *Anais do XVII Workshop em Desempenho de Sistemas Computacionais e de Comunicação*, Porto Alegre, RS, Brasil, 2018. SBC. Disponível em: <https://sol.sbc.org.br/index.php/wperformance/article/view/3324>.

[4] DE PAULA NEVES NETO, J., FIGUEIREDO, D. R. "Ranking influential and influenced stocks over time using transfer entropy networks", *Physica A: Statistical Mechanics and its Applications*, p. 129119, 2023. ISSN: 0378-4371. Disponível em: <https://www.sciencedirect.com/science/article/pii/S037843712300674X>.

[5] SHANNON, C. E. "A mathematical theory of communication", *Bell System Technical Journal*, v. 27, pp. 379–423, Jul 1948.

[6] STEUER, R., KURTHS, J., DAUB, C. O., et al. "The mutual information: detecting and evaluating dependencies between variables", *Bioinformatics*, v. 18, n. suppl_2, pp. S231–S240, 2002.

[7] DIONISIO, A., MENEZES, R., MENDES, D. A. "Mutual information: a measure of dependency for nonlinear time series", *Physica A: Statistical Mechanics and its Applications*, v. 344, n. 1-2, pp. 326–329, 2004.

[8] BAKER, W. E. "Market networks and corporate behavior", *American journal of sociology*, v. 96, n. 3, pp. 589–625, 1990.

[9] KAUÊ DAL'MASO PERON, T., DA FONTOURA COSTA, L., RODRIGUES, F. A. "The structure and resilience of financial market networks", *Chaos:*

An Interdisciplinary Journal of Nonlinear Science, v. 22, n. 1, pp. 013117, 2012.

[10] TABAK, B. M., SERRA, T. R., CAJUEIRO, D. O. “Topological properties of stock market networks: The case of Brazil”, *Physica A: Statistical Mechanics and its Applications*, v. 389, n. 16, pp. 3240–3249, 2010.

[11] LEAHY, S. P., LEVY CARCIENTE, S., STANLEY, H. E., et al. “Structure and dynamics of the Brazilian stock market: A correlation analysis”, 2014.

[12] LEE, G. S., DJAUHARI, M. A. “An overall centrality measure: The case of US stock market”, *International Journal of Electrical & Computer Sciences*, v. 12, n. 6, 2012.

[13] SHIROKIKH, O., PASTUKHOV, G., BOGINSKI, V., et al. “Computational study of the US stock market evolution: a rank correlation-based network model”, *Computational Management Science*, v. 10, n. 2-3, pp. 81–103, 2013.

[14] LYÓCSA, Š., VỲROST, T., BAUMÖHL, E. “Stock market networks: the dynamic conditional correlation approach”, *Physica A: Statistical Mechanics and its Applications*, v. 391, n. 16, pp. 4147–4158, 2012.

[15] NAMAKI, A., SHIRAZI, A., RAEI, R., et al. “Network analysis of a financial market based on genuine correlation and threshold method”, *Physica A: Statistical Mechanics and its Applications*, v. 390, n. 21, pp. 3835–3841, 2011.

[16] BOGINSKI, V., BUTENKO, S., PARDALOS, P. M. “Statistical analysis of financial networks”, *Computational Statistics & Data Analysis*, v. 48, n. 2, pp. 431 – 443, 2005. ISSN: 0167-9473. Disponível em: <`http://www.sciencedirect.com/science/article/pii/S0167947304000258`>.

[17] ATSALAKIS, G. S., VALAVANIS, K. P. “Surveying stock market forecasting techniques - Soft computing methods”, *Expert Systems with Applications*, v. 36, n. 3, pp. 5932 – 5941, 2009. ISSN: 0957-4174. Disponível em: <`http://www.sciencedirect.com/science/article/pii/S0957417408004417`>.

[18] PETELIN, D., ŠINDELÀF, J., PŘIKRYL, J., et al. “Financial modeling using gaussian process models”, *Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems*, v. 2, pp. 672–677, 2011. Disponível em: <`http://ieeexplore.ieee.org/document/6072854/`>.

[19] YANG, P., SHANG, P., LIN, A. “Financial time series analysis based on effective phase transfer entropy”, *Physica A: Statistical Mechanics and its Applications*, v. 468, pp. 398–408, 2017.

[20] KWON, O., YANG, J.-S. “Information flow between stock indices”, *Europhysics Letters*, v. 82, n. 6, pp. 68003–p1–68003–p4, Jun 2008. Disponível em: <http://iopscience.iop.org/article/10.1209/0295-5075/82/68003>.

[21] MARSCHINSKI, R., KANTZ, H. “Analysing the information flow between financial times series”, *The European Physical Journal B*, v. 30, pp. 275–281, 2002.

[22] KORBEL, J., JIANG, X., ZHENG, B. “Transfer entropy between communities in complex networks”, *ArXiv e-prints*, jun. 2017.

[23] HE, J., SHANG, P. “Comparison of transfer entropy methods for financial time series”, *Physica A: Statistical Mechanics and its Applications*, v. 482, pp. 772–785, 2017.

[24] BRIN, S., PAGE, L. “The anatomy of a large-scale hypertextual Web search engine”, *Computer Networks and ISDN Systems*, v. 30, n. 1, pp. 107 – 117, 1998. ISSN: 0169-7552. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016975529800110X>.

[25] PAGE, L., BRIN, S., MOTWANI, R., et al. *The PageRank citation ranking: bringing order to the web.* Technical Report 1999-66, Stanford University, November 1999. Disponível em: <http://ilpubs.stanford.edu:8090/422/>.

[26] KLEINBERG, J. M. “Authoritative sources in a hyperlinked environment”, *Journal of the ACM (JACM)*, v. 46, n. 5, pp. 604–632, 1999.

[27] CHAKRABARTI, S., DOM, B., RAGHAVAN, P., et al. “Automatic resource compilation by analyzing hyperlink structure and associated text”, *Computer Networks and ISDN Systems*, v. 30, n. 1, pp. 65–74, april 1998. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169755298000877>. Proceedings of the Seventh International World Wide Web Conference.

[28] XING, W., GHORBANI, A. “Weighted pagerank algorithm”. In: *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pp. 305–314. IEEE, 2004.

[29] LI, L., SHANG, Y., ZHANG, W. "Improvement of HITS-based algorithms on web documents". In: *Proc. Int. Conf. on World Wide Web*, pp. 527–535, 2002. ISBN: 1-58113-449-5.

[30] ZHANG, X., YU, H., ZHANG, C., et al. "An improved weighted HITS algorithm based on similarity and popularity". In: *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, pp. 477–480, Aug 2007.

[31] "Bolsa de Valores, Mercadorias e Futuros do Estado de São Paulo". 2017. Disponível em: <http://www.b3.com.br/>.

[32] "Séries históricas de acões da B3". 2022. Disponível em: <https://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/historico/mercado-a-vista/series-historicas/>.

[33] CLAUSET, A., SHALIZI, C. R., NEWMAN, M. E. J. "Power-Law Distributions in Empirical Data", *SIAM Review*, v. 51, n. 4, pp. 661–703, 2009.

[34] DONOHO, D. "50 years of data science", *Journal of Computational and Graphical Statistics*, v. 26, n. 4, pp. 745–766, 2017.

[35] DUGAS, A. F., HSIEH, Y.-H., LEVIN, S. R., et al. "Google Flu Trends: correlation with emergency department influenza rates and crowding metrics", *Clinical infectious diseases*, v. 54, n. 4, pp. 463–469, 2012.

[36] LAZER, D., KENNEDY, R., KING, G., et al. "The parable of Google Flu: traps in big data analysis", *Science*, v. 343, n. 6176, pp. 1203–1205, 2014.

[37] COVER, T. M., THOMAS, J. A. *Elements of information theory.* 605 Third Avenue, New York, NY, USA, John Wiley & Sons, 2012.

[38] BENNASAR, M., HICKS, Y., SETCHI, R. "Feature selection using joint mutual information maximisation", *Expert Systems with Applications*, v. 42, n. 22, pp. 8520–8532, 2015.

[39] BATTITI, R. "Using mutual information for selecting features in supervised neural net learning", *IEEE Transactions on neural networks*, v. 5, n. 4, pp. 537–550, 1994.

[40] HIRSCHMULLER, H. "Accurate and efficient stereo processing by semi-global matching and mutual information". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, v. 2, pp. 807–814. IEEE, 2005.

[41] PLUIM, J. P., MAINTZ, J. A., VIERGEVER, M. A. "Mutual-information-based registration of medical images: a survey", *IEEE transactions on medical imaging*, v. 22, n. 8, pp. 986–1004, 2003.

[42] HLAVÁČKOVÁ-SCHINDLER, K., PALUŠ, M., VEJMELKA, M., et al. "Causality detection based on information-theoretic approaches in time series analysis", *Physics Reports*, v. 441, n. 1, pp. 1–46, 2007.

[43] SONG, L., LANGFELDER, P., HORVATH, S. "Comparison of co-expression measures: mutual information, correlation, and model based indices", *BMC bioinformatics*, v. 13, n. 1, pp. 328, 2012.