INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD • INDIA

Research and Publications

Fractional Differencing: (In)stability of Spectral Structure and Risk Measures of Financial Networks

Arnab Chakrabarti
Anindya S. Chakrabarti

W.P. No. 2020-07-01
July 2020

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD – 380015
INDIA

Fractional Differencing: (In)stability of Spectral Structure and Risk Measures of Financial Networks[1]

Arnab Chakrabarti[2]
Anindya S. Chakrabarti[3]

Abstract

Computation of spectral structure and risk measures from networks of multivariate financial time series data has been at the forefront of the statistical finance literature for a long time. A standard mode of analysis is to consider log returns from the equity price data, which is akin to taking first difference ($d = 1$) of the log of the price data. Sometimes authors have considered simple growth rates as well. Either way, the idea is to get rid of the nonstationarity induced by the *unit root* of the data generating process. However, it has also been noted in the literature that often the individual time series might have a root which is more or less than unity in magnitude. Thus first differencing leads to under-differencing in many cases and over differencing in others. In this paper, we study how correcting for the order of differencing leads to altered filtering and risk computation on inferred networks. In summary, our results are: (a) the filtering method with extreme information loss like *minimum spanning tree* as well as filtering with moderate information loss like *triangulated maximally filtered graph* are very susceptible to such d-corrections, (b) the spectral structure of the correlation matrix is quite stable although the d-corrected market mode almost always dominates the uncorrected ($d = 1$) market mode indicating under-estimation in the standard analysis, and (c) the PageRank-based risk measure constructed from Granger-causal networks shows an inverted U-shape evolution in the relationship between d-corrected and uncorrected return data over the period of analysis 1972-2018 for historical data of NASDAQ.

Keywords: Fractional differencing, Eigenspectrum, Financial network, Network filtering, risk and vulnerability

PhySh codes: Collective behavior in networks, Financial networks, Centrality.

JEL codes: C58, G32, C14, C38.

[1]This research was partially supported by institute grant, IIM Ahmedabad. We have benefited from discussions with Andrey Pogudin. All remaining errors are ours.

[2](Corresponding author) Misra Centre for Financial Markets and Economy, Indian Institute of Management Ahmedabad, Gujarat 380015, INDIA. Email: arnab_c@zohomail.in

[3]Economics Area and Misra Centre for Financial Markets and Economy, Indian Institute of Management Ahmedabad, Gujarat 380015, INDIA. Email: anindyac@iima.ac.in

1 Introduction

Financial markets are complex interactive systems, susceptible to disturbances from internal instability as well as external perturbations. As the returns series from the stocks and other financial assets exhibit complex interdependence, a natural way to infer about the underlying market dynamics has been to analyze the observed cross correlations. Stocks belonging to the same or related sector, tend to exhibit correlated price movements e.g. the historical correlations utilized in pair trading [34]. On the other hand, it is likely that information specific to a particular company will affect that company's stock only. Study of the spectral distribution of the empirical correlation matrix has become a standard toolkit to capture the internal structure of the market by finding out the major and minor modes of fluctuations [31]. Another complementary measure of market instability is given by the notion of *systemic risk*, which provides an idea of vulnerability of the market to small localized shocks. A common feature of almost all the analysis done in the literature are dependent on one standard methodology of converting price data into log return by first-differencing the log price series ($r = \Delta log(P)$ where r is return and P is price). However, there is a known fact in the literature that the first-differencing may not be appropriate for all stocks as it might lead to over-differencing and under-differencing [10]. Yet, surprisingly there is a lacuna in the literature on the dependence of eigenspectra and risk measures to this seemingly innocuous modeling choice. In this paper, we attempt to fill that gap and characterize the structures that are robust to the choice of *en masse* first-differencing.

Before describing the results, we first provide a brief summary of the developments in network studies on financial correlation matrices. In one of the first attempts, Plerou et al. [28] analyzed the cross-correlation matrix of stock price changes of 1000 largest U.S. companies for the period 1994-1995, and they showed that the correlation matrix has the universal properties of the Gaussian orthogonal ensemble of random matrices. Following the lead, a large literature developed on analysis to large dimensional correlation matrices, especially on the applications of random matrix theory to filter statistically significant eigenmodes from the spectral structure of the matrices [31]. A natural extension of this stream of work led to the development of clustering studies in the comovement matrices. In an influential work, Mantegna [22] introduced a graph theoretic visualization of the topological relations, obtained from the correlation matrix. This led to a surge of studies in the intersection of network theory and statistical properties of financial time series. Subsequently, the concepts of asset trees and more generally, asset graphs were introduced. The asset tree or the minimum spanning tree is a connected graph without any closed loop such that the sum of pairwise distance is minimized [22, 24]. In general, the asset graphs provide a network view of the stocks with a hierarchy in the strength of correlation values [25]. Ref. [13] compared how strongly correlated clusters of stocks are expressed as branches in the asset tree and as clusters in asset graphs, and found that the eigenvector corresponding to the largest eigenvalue has larger components on the central nodes and a few eigenvectors corresponding to the next largest eigenvalues are more or less localized. Further developments in this domain have been discussed in Ref. [31].

A parallel development took place in terms of quantification of entanglement in the financial markets. Past few decades have witnessed massive expansion, soaring complexity and inter-connectedness of the financial markets, leading to questions of how to regulate such complexity [6]? Changes in correlation and the corresponding spectral structure poten-

tially impacts the system's vulnerability to a financial shock [17]. The induced risk can lead to a major crisis or depression in the world economy, such as the 2008 global financial crisis that spurred the financial institutions to study and measure this risk accurately. It is a challenging task to identify the extent of the risk beforehand. Several assessment methodologies are in place such as the identification of systemically important financial institutions (SIFIs) developed by the International Monetary Fund, the Bank of International Settlement and Financial Stability Board. Earlier this risk was conceptualised in terms of the financial institutions like banks. But according to the current reality, as companies can access funding without going through banks, the relation between market and institution is more relevant and must be taken into account. Ref. [30] advocated for appropriate regulation to contain the risk. Ref. [9] argued that to comprehend the risk fully, the risk measure has to integrate bank failure contagion with financial markets spillover effects and payment and settlement risks. Ref. [1] showed that financial networks are "*robust-yet-fragile*" in the sense that interconnectedness entails increased ability to absorb shock but interconnectedness beyond a certain level make the financial system susceptible to contagion.

As an increase in the risk would be a direct consequence of the complexity and interconnectedness of the network, it is very important to capture the network relationship among the financial institutions. To do so, three concepts emerged- "*too central to fail*", "*too big to fail*" and "*too connected to fail*" (see [35]). Among these the notion of "*too central to fail*", as argued by [35], extracts most information from the network and a risk measure was crated based on PageRank centrality. Notably, [35] built their measures based on the findings of [3] who described that all the available measures of the risk are based on one of the four L's of financial crises: leverage, liquidity, losses, and linkages. They developed several econometric measures based on linkages, to capture connectedness in the financial network using principal component analysis as well as Granger Causality network. These measures have a focus on the network topology of asset returns. On the other hand, considering publicly traded financial institutions and defining a systemic event as simultaneous losses among multiple financial institutions, CoVaR by [2], and Co-Risk by [7], were developed. Since these two measures do not correlate well with the risk measure capturing the "*too central to fail*" mechanism [35] and do not directly related to the comovement matrices that will be the building block of our work, we will not consider them in the following.

In this paper, we show that the construction of the asset graphs as well as estimation of vulnerability (following the "*too central to fail*" paradigm) depends on the construction of return series. The standard practice to create the log returns is by taking first difference ($d = 1$) of the log of price data, in order to remove the non-stationarity induced by unit roots in the data [10]. However, the time series is likely to have a root more or less than unity. Thus first differencing leads to under-differencing in many cases and over differencing in others. This is a known phenomenon in the literature, however there is surprisingly little work in this domain. As De Prado stated in his book [10] published in 2018: *"After Hosking's paper, the literature on this subject has been surprisingly scarce, adding up to eight journal articles written by only nine authors"* (p. 76), where the original paper by Hosking was written in 1981 [14].

With this motivation, we used fractional differencing instead of integer differencing to construct the return of an asset. This is particularly useful for the processes with long term dependence. In the following, we investigate how would the spectral structure and the risk measures change by the choice of difference parameter. In a nutshell, our comparative

analysis indicates that the risk measure and the topology of networks can change quite a bit depending on the order of differencing fo the primitive return series. We have carried out all of our analysis on historical NASDAQ data (ranging from 1972 to 2018). In particular, we have created moving windows of 300 stocks with the lagrest market capitalization every year, where the windows have a length of four years. This choice of window construction is quite standard in the literature [17] and our findings are robust to alternative specifications. On each of the window, we compute the spectral structure, composition of filtered network and risk measures using the return series constructed from first differencing ($d = 1$) and optimally chosen d based on ARFIMA (*autoregressive fractionally integrated moving average*) specification.

To summarize the results, we make a series of observations. First, optimally chosen order of differencing (d parameter) has a wide range of distribution, roughly from 0.5 to 1.4 with some infrequent outliers, with a mode of the histogram close to 1. Thus on an average, the standard calculations are correct. Second, the largest eigenvalues corresponding to optimally chosen d obtained from all the windows, are very similar in magnitude to those obtained from $d = 1$ although the former typically dominates the latter. Also, the bulk of the distribution of the eigenvalues are very similar to each other. But this is where the similarity ends and dissimilarity begins. The third observation is that a big difference appears on the construction of *minimum spanning tree* (or MST in short). We create a similarity measure based on the number of common edges between two MSTs corresponding to each window for $d = 1$ and optimally chosen d for stocks. In the beginning of our data (in the windows corresponding to 1970s and 80s), the match was barely in the order of 30%. however, there is an almost linear trend of catching up as the share of common edges increased, with almost 70% match towards the end of the data (in 2010s). A very similar pattern is seen with less stringent filtering techniques like TMFG *triangulated maximally filtered graph*. Fourth, we compute risk measure capturing vulnerability of stocks from Granger causal matrices constructed from the time series data of returns. Then we study the similarity between the risk measures obtained via $d = 1$ and optimally chosen d. This relationship shows an inverted U-shaped pattern where peak of the similarity is found in the 1990s, which continues till the beginning of 2000. Thus the similarity in the risk-measure seems to be non-monotonic.

The rest of the paper is arranged as follows. In section 2, we provide a short introduction to fractional differencing and ARFIMA model followed by a brief introduction to the spectral structure and construction of risk measures from Granger-causal networks. In section 2.5, the data, used for statistical analysis, is described. Results are presented in section 3. We summarize the findings and conclude with remarks on the implications of the results in section 4.

2 Comovement Structure of Multivariate Financial Data and Fractional Differencing

In this section, we will discuss the methodology we have adopted in this paper. Our main objective is to detect the extent of changes in a network induced by fractional differencing. We are particularly interested to see the impact on the risk measures and the topological

properties of the network. First, we are going to discuss some of the key concepts that will help us to measure this impact.

2.1 Long Memory and ARFIMA

In this subsection, we will briefly introduce the long memory and ARFIMA model. In statistical analysis of time series data, an *AutoRegressive Integrated Moving Average* model (ARIMA) is a well-known way to model, infer and forecast a time series. If X is not a stationary process then the differencing parameter has to be an integer- commonly taken to be one or two. This method is frequently used in practice for short term dependence in the time series. However, some economic and financial time series exhibits long term dependence- commonly called long memory [27]. For such series, the autocorrelation decays hyperbolically- much slowly than exponential decay. [21] showed that fractionally differentiated series can give rise to long memory. Based on this idea the ARIMA can be generalised as ARFIMA- *AutoRegressive Fractionally Integrated Moving Average*. A series X_t is a time series then ARFIMA(p,d,q) model is defined like the following:

$$\Phi(L)(1-L)^d(X_t-\mu) = \Theta(L)u_t \quad \text{where} \quad u_t \sim iid(0,\sigma_u^2), \tag{1}$$

where L is the backward-shift operator, d is the fractional differencing parameter (takes any real value) and $\Phi(L) = 1-\phi_1 L - ... - \phi_p L^p$ and $\Theta(L) = 1+\theta_1 L - ... + \theta_q L^q$. The stochastic process X_t is stationary if $d \in (-0.5, 0.5)$. If $d > 0.5$, then the process will possess infinite variance and so won't be stationary. An ARFIMA(0, d, 0) process boils down to simply

$$(1-L)^d(X_t-\mu) = u_t, \quad \text{where} \quad u_t \sim iid(0,\sigma_u^2). \tag{2}$$

2.2 Spectral Analysis of Equal Time Correlation Matrix

Basic topological properties of a network can be described by the distribution of the spectrum of the adjacency matrix. It is well-known that networks generated by the same random process have same distribution of the eigenvalues. Spectrum also helps to understand some of the key properties of the interacting units of the network. For an example, the extreme (highest) eigenvalue of the adjacency matrix gives us significant insight about the *market mode* or the collective response of the entire market to external information. To be more precise, spectral analysis of the equal time correlation matrix reveals three types of fluctuations: (i) due to the market mode that is common to all stocks, (ii) sectoral contributions (related to specific business sectors) and (iii) idiosyncratic (i.e., specific to individual stocks' dynamics). These are captured by segregating the network spectra into three parts: (i) largest and most extreme eigenvalue (ii) deviating eigenvalues and (iii) bulk of the eigenspectra respectively. The following equation illustrates the decomposition ([28], [31], [17]):

$$C = C^{market} + C^{group} + C^{random} \tag{3}$$

$$= \lambda_1 u_1 u_1^T + \sum_{i=2}^{N_g} \lambda_i u_i u_i^T + \sum_{i=N_g+1}^{N} \lambda_i u_i u_i^T, \tag{4}$$

where C is the $N \times N$ correlation matrix, decomposed into three parts, and N_g denotes number of eigenvalues deviated from the bulk except the top one (the threshold for the bulk is typically determined by Marcenko-Pastur distribution [31]). λ_is and u_is are eigenvalues and eigenvector respectively. Finally, u^T represents transpose of the u vector. In this work, we will investigate how strongly fractional differencing affects extreme eigenvalues and the rest of the distributions.

2.2.1 Sensitivities of Filtering Algorithms: Minimum Spanning Tree

A method to obtain the graph-theoretic visualization of the network, based on the topological relation of the stocks, was proposed in [22]. Using the equal time correlation coefficients of the daily difference of logarithm of closure price of stocks, ref. [22] obtained the taxonomy of a portfolio of stocks traded in a financial market. As correlation coefficient cannot be considered as a metric (correlations can be negative, whereas a distance measure is always nonnegative), the distance between two stocks is defined in the following way-

$$d_{i,j} = \sqrt{2(1 - \rho_{i,j})} \tag{5}$$

These calculated pairwise distances are then used to construct a *minimum spanning tree* (MST) that connects N nodes of the network with $N-1$ edges, such that the total sum of the pairwise distances, i.e. $\sum_{i,j} d_{i,j}$ is minimum. These constructions played an important role in portfolio optimization, see for example [24]. In this work, we examine the impact of fractional differencing on the constructed MST.

2.2.2 Sensitivities of Filtering Algorithms: Triangulated Maximally Filtered Graph

The TMFG method [23] is a filtering technique used to construct a subnetwork from a correlation matrix which captures the most relevant information between the nodes and minimizes spurious associations. The resultant graph is a clique-tree composed of four nodes cliques connected with three node cliques. TMFG imposes a structural constraint to retain $3(N-2)$ edges of the original network as opposed to filters like MST contains only $N-1$ edges, thus retaining far more information than MST. TMFG produces a planar network, i.e. the network can be drawn on a sphere with no edges crossing. We will examine the impact of fractional differencing on the graph constructed by TMFG method.

2.3 Risk Measure from Granger-Causal Networks

To understand the impact on overall risk, we will adopt a risk measure from the Granger-Causal network as described in [35]. In order to study the propagation of the shock, Ref. [3] first used the notion of Granger causality which captures the directionality of the connections between the stocks. For two return time series (corresponding to ith and jth stock), the stock i is said to *Granger-cause* stock j if past values of stock i contain information that along with the past information contained in stock j predicts the present value of stock j better than that with the past information in stock j alone. Statistically, the prediction-performance is captured through linear regression. let R_{it} and R_{jt} be two stationary time

series with zero mean, then the Granger-causality between ith and jth stock are obtained through the output of the following two regression equations:

$$\begin{aligned} R_t^i &= a^i R_{t-1}^i + b^{ij} R_{t-1}^j + e_t^i \\ R_t^j &= a^j R_{t-1}^j + b^{ji} R_{t-1}^i + e_t^j, \end{aligned} \tag{6}$$

where a and b are regression parameters and e^i, e^j are uncorrelated white noise processes. If b^{ij} is statistically significant then we say that the time series of stock j Granger-causes the time series of stock i. So in the graph representation, the two nodes will be connected, but the direction will be from j to i. If b^{ji} is also significant then the direction is both way i.e. there is a feedback relationship. If none of them are significant, then there is no edges between these two nodes.

If the i-th asset Granger causes j-th asset then the (i, j)th element of the adjacency matrix will be 1, otherwise, it is chosen to be 0:

$$A[i, j] = \begin{cases} 1 & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The *degree of Granger causality* is defined by the fraction of nondiagonal elements of $A[i, j]$ taking value 1, i.e. the fraction of statistically significant Granger-causal relationship. It is also called a *macro measure* or *macro connectedness* as opposed to the *micro connectedness* that estimates the connectedness of individual financial institution [16] or listed firms, in general.

Ref. [35] prescribed a measure of centrality using the PageRank algorithm, originally proposed by [26] to rate or order Web pages, from a "too central to fail" perspective and shown the other measures of systemic risk like conditional value at risk (CoVaR) and marginal expected shortfall (MES) are inconsistent with each other as well as with the Rank measure. PageRank is defined as follows:

$$Rank_{it} = \frac{1-\alpha}{N} + \alpha \sum_j E_{ijt} Rank_{jt}, \tag{8}$$

where α is called a damping factor, N is total number of financial institutions and E_{ijt} is the normalized Granger-causal matrix. It is worthwhile to note here that PageRank is essentially a generalization of *eigenvector centrality*, which in turn is defined to measure the importance of the nodes based on their overall connectivity. In the present discussion, we will not pursue the methodological discussion of eigenvector centralities. Interested readers can consult Ref. [3] for an overview of the construction and application of eigenvector-based measures of risk.

2.4 Description of the Analysis

Now we will provide a brief outline of our methodology and implementation.

1. To determine the appropriate differencing parameter, we first fit the log returns of individual stocks by a ARFIMA$(0, d, 0)$ model. We adopt the semiparametric estimation

procedure proposed by Ref. [12]. this method is based on spectral regression. With this we get the estimated differencing parameter d. For each stock then we calculate two log-return series:- one with the usual method (unit differencing) and the second with the estimated optimal differencing parameter. We call the second return series as d-corrected return series. If the log-price of an asset i is defined as X^i, then the usual log return is defined as $R(t) = (1-L)X(t)$ and the d-corrected log return is defined as $R^d(t) = (1-L)^d X(t)$. For empirical implementation, we have utilized *fracdiff* package in R (see https://cran.r-project.org/web/packages/fracdiff/fracdiff.pdf).

2. We need one more correction in the return data to reliably construct the correlation matrix. It has been recognized in the literature that spurious correlations arise from volatility clustering ([15]; see also [11]). Therefore, we consider the Generalized Auto-Regressive Conditional Heteroskedasticity (GARCH)(1,1) model (see [32] for a textbook treatment) to adjust the return series for latent volatility. Conditioning on the filtration $\mathcal{F}(\{\{R^i_\tau\}_{-\infty}^{t-1}\}_{i=1}^{N})$,

$$R_{i,t} = \mu_i + \sigma_{it}\epsilon_{it}, \tag{9}$$

$$\sigma^2_{it} = \omega_i + \alpha_i(R_{i,t-1} - \mu_i)^2 + \beta_i \sigma^2_{i,t-1} \tag{10}$$

where μ, σ, ω and α are coefficients of the equation and N is total number of stocks. To control for heteroscedasticity-induced correlations, we adjust the returns for each stock by dividing by the latent volatility series estimated using a GARCH(1,1) model on the return series as follows:

$$\tilde{R}^i_t = \frac{R_{it}}{\hat{\sigma}_{i,t}} \quad \text{for the } i\text{-th stock at } t\text{-th time point.} \tag{11}$$

The resultant return series would be free of the effects of latent volatility. For empirical implementation, we have utilized *rugarch* package in R (see https://cran.r-project.org/web/packages/rugarch/rugarch.pdf).

3. To summarize, we construct modified returns for both simple return series (with differencing parameters 1) and d-corrected return series. Let us call them $\tilde{R}$ and $\bar{R}$ respectively.

4. Then we calculate the correlation matrix out of $\tilde{R}$ and $\bar{R}$ for each window, and compare the distributions of the eigenvalues of the correlation matrix (λ in Eqn. 4). In particular, the largest eigenvalues of the correlation matrices of $\tilde{R}$ are calculated over each time window to see the dynamic evolution. Since it represents the *market mode* ([28], [17]), the difference between the two dominant eigenmodes would inform us about the differences in the strength of the *market modes*.

5. Next, we calculate the MST (constructed using *ape* package in R; see https://cran.r-project.org/web/packages/ape/ape.pdf) from the distance matrices obtained from equal time correlation matrices constructed (using the transformation following Eqn. 5) from $\tilde{R}$ and $\bar{R}$ for each window. We define a degree of similarity between two MSTs (from $d = 1$ and d-corrected return series) by finding the fraction of edges common to both graphs with respect to the total number of edges. This gives us an idea of the stability of the simplest form of the filtered subnetwork inferred from the comovement structure.

6. Similar to step 5, we calculate the two TMFGs (constructed using *networktoolbox* package in R; see https://cran.r-project.org/web/packages/NetworkToolbox/NetworkToolbox.pdf) corresponding to $\tilde{R}$ and $\bar{R}$ return series. Unlike the MST, TMFG should be calculated from the correlation matrix (or, square of the correlation matrix.) Like step 5, here also we calculate the degree of similarity by calculating the fraction of edges common to both the graphs.

7. Next, we calculate PageRank vectors from the Granger-causality matrices (the $\{i, j\}$-th element of the Granger causal matrix denotes Granger causation running from the j-th return series to the i-th return series, vide Eqn. 6 and 7; evaluated at 5% level of significance), constructed from simple return series and d-corrected return series. Therefore, for each stock we obtain two ranks corresponding to two return series. These ranks represent a risk measure in the form of vulnerability and shock spillover from one stock to the other.

8. For each window, in order to compare two sets of ranks we calculate the correlation coefficients between them. This gives us a measure indicating how much the choice of differencing parameter affects the risk of the network. We utilize three measures of association: Pearson's, Kendall' τ and Spearman's rank correlation coefficients.

   Pearson's correlation is the usual product moment correlation between variables X and Y,

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}, \tag{12}$$

   where $cov(X, Y)$ denotes the covariance between X and Y and $\sigma(.)$ denotes the standard deviation.
   Spearman's correlation is the Pearson's correlation for the rank values of two variables.

$$\rho_S = \frac{\text{cov}(\text{rank}(X), \text{rank}(Y))}{\sigma_{\text{rank}(X)} \sigma_{\text{rank}(Y)}}, \tag{13}$$

   Finally, Kendall's τ is the difference between the probability of concordance and probability of discordance, which can be represented as-

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(x_i - x_j)\text{sign}(y_i - y_j), \tag{14}$$

   where (x_i, y_i) and (x_j, y_j) are two pairs of observations of (X, Y).

2.5 Data

Our dataset consists of time series data of stocks from NASDAQ (one of the largest trading market in the world in terms of market capitalization; located in the US) from 1972 to 2018. The market started operating in 1971 along with operationalization of the NASDAQ index. Since we are interested in studying the evolution of the market structure, we divide the available time period in overlapping, moving windows of 4 years' length (the windows cover 1972-75, 1973-76 and so on). there are 47 such windows. In each window, we choose top 300 stocks in terms of market capitalization (calculated as equity price multiplied by number of

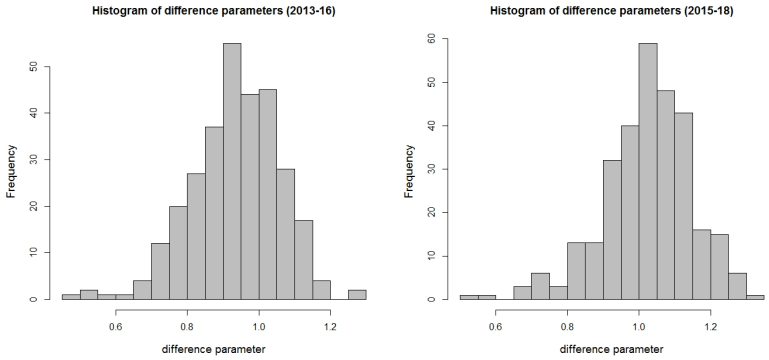


Figure 1: Histogram of GPH (Geweke and Porter-Hudak) estimates of differencing parameters for 300 stocks in time windows 2013-16 and 2015-18. The mode of the distribution is close to one in both cases. However, the distributions exhibit a sizeable range approximately from 0.5 to 1.4. A similar pattern can be found on the other windows as well (not shown here).

outstanding shares). Thus every window in our dataset would consist of 300 stocks (except the first one, which has 124 due to missing data) with daily closing price for four consecutive years (the first closing price recorded on 1st January or the first trading day of the year, and the last data-point being the last recorded closing price on 31st December or the last trading day of the relevant year).

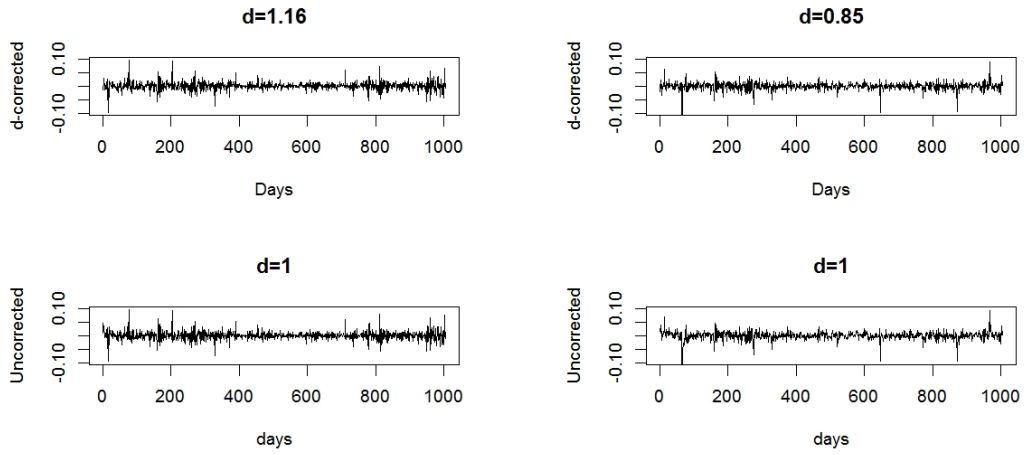
A known problem of historical equity data is that often for some days the is not reported. To make sure that our dataset does not suffer from too many missing data points, we restrict the choice of the stocks within each window, to have at least 95% of the closing price data. therefore, out of around 1000 trading days in four years (market remains closed on weekends and other holidays) each stock should have at least 950 reported closing price. We replace the missing return values by zeros.

3 Results

Here we describe the results of our analysis on the spectral structure, filtering and the risk measure.

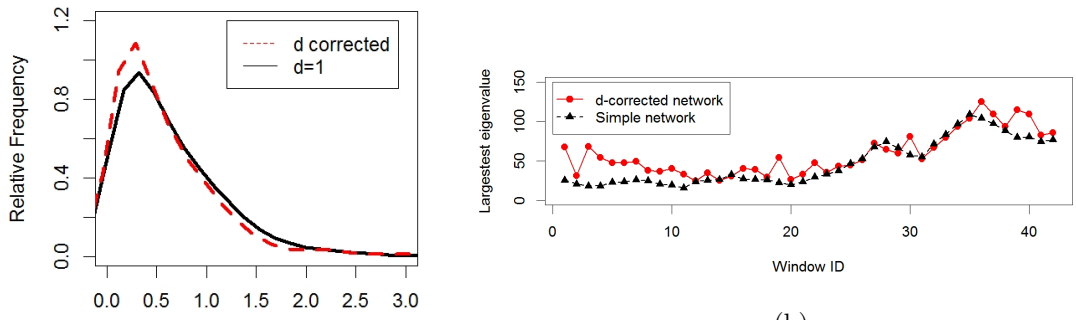
3.1 Spectral Structure

First we report the results from exercise on choice of optimal degree of differencing. The *difference parameter* d is estimated by GPH (Geweke and Porter-Hudak [12]). For 300 stocks in the time-window 2013-16 and 2015-18, the histograms of estimated difference parameter are shown in Fig. 1. We can see that there is considerable dispersion around 1. The question is whether taking unit difference for all the stocks to create the return would affect the analysis of data. To investigate it we first calculate the return series using those estimated parameters. In Fig. 2, we have plotted the d-corrected return series and the



(a) (b)

Figure 2: Visual comparison: d-corrected return series and simple return series of two stocks from 2014-17 snapshot. *Panel (a):* corresponds to a stock with $d > 1$ and *Panel (b):* refers to a stock with $d < 1$. Visually it is difficult to differentiate the top return series (d-corrected) from the bottom return series ($d = 1$). However, the correlation structure changes quite substantially as we will demonstrate in the following.



(a) (b)

Figure 3: (Color online) Illustration of the similarities and differences in spectral structure for simple and d-corrected return series. *Panel (a):* The bulk of the spectra for 2015-18 data, *Panel (b):* Evolution of the largest eigenvalues corresponding to the d-corrected and uncorrected correlation matrices over time (1972-2018). Each point represent the largest eigenvalues evaluated within a single window. While the left panel shows that the differences in the bulk of the eigenvalues is very minimal, the right panel shows that the dominant eigenvalue corresponding to the *market mode*, exhibits higher magnitudes in almost all windows for the d-corrected return data. However, that difference has gradually diminished over time.

return series obtained by unit differencing for two different stocks. These plots show that visual differences between the return series are not very prominent.

In Fig. 3a, we show the probability density function of the eigenvalues of the correlation matrices constructed from the d-corrected and $d = 1$ return series. The differences in the bulk of the empirical distribution are very minimal. In Fig. 3b, we plot the dominant eigenvalues of each of the overlapping 4 year-long windows across time from 1972 to 2018. We see that while there are substantial differences in magnitude in the first 12 windows, the difference becomes much less pronounced in the later periods. However, the dominant eigenvalue of the correlation matrix obtained from d-corrected network almost always dominates the dominant eigenvalue of the uncorrected correlation matrix.

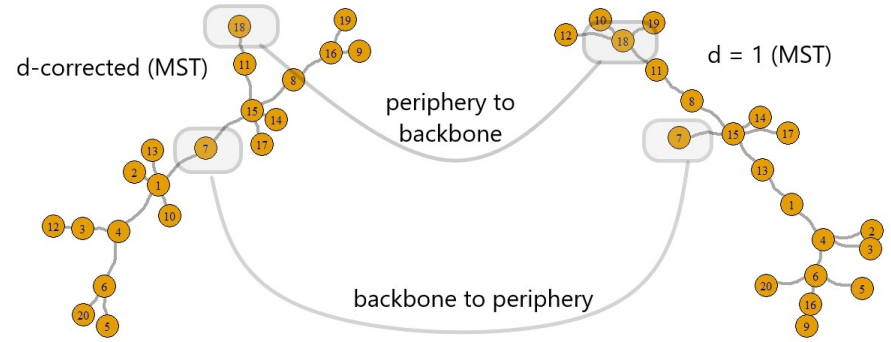
To summarize, the differences in the spectral structure seems to be existent, but minimal. Our later analysis on the other hand, shows that both filtering as well as risk measures calculations are substantially affected by the choice of optimal d.

3.2 Filtering: Dissimilarities across MSTs and TMFGs

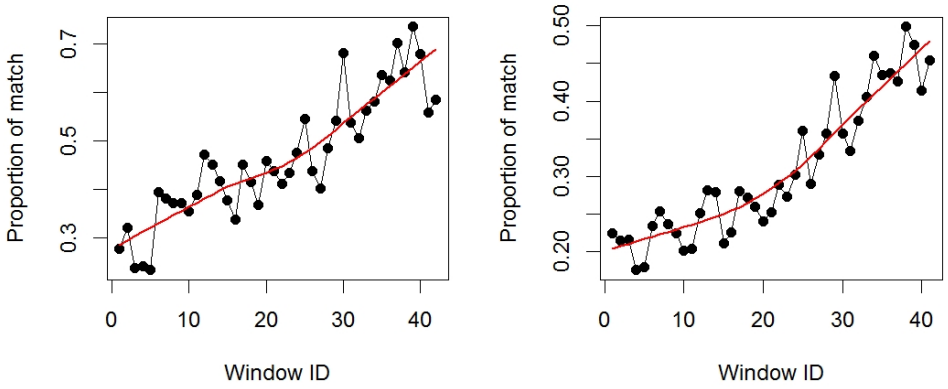
Next, we construct the correlation matrices from the return data and the construct the corresponding distance matrices (using the transformation given in Eqn. 5). Based on the distance matrices, we compute the MSTs. Our analysis shows that there are substantial differences in the MSTs.

First, we motivate our findings based on a simple, small-scale exposition with only 20 stocks. Fig. 4a shows a subgraph of a minimum spanning tree for $d = 1$ and d-corrected return series. We can see that node 18 was at the periphery of the graph before appropriate differencing. It becomes a more central node (right panel) after correcting for appropriate d. The opposite happens to node number 7. One can try to match the other nodes as well and confirm that their positions also often changes. This comparison illustrates that the MST structure can change quite substantially depending on the choice of difference parameter.

Next, we quantify this the degree of similarity across two MSTs and two TMFGs. Obviously when we plot the MSTs (or TMFGs) with 300 stocks, it is not visually possible to count all the changes. So we have created a measure of similarity by taking the ratio of the number of common edges to the total number of edges. It is useful to note that an MST with N number of nodes will have $N - 1$ edges. Therefore, in our case, the denominator would be 299 as the number of stocks is 300. For a TMFG with N number of nodes will have $3(N-2)$ edges and so the denominator will be 894. Fig. 4b and Fig. 4c show the proportion of identical edges in two MSTs and TMFGs respectively, across consecutive windows of four years' lengths. The smooth lines were obtained by a nonparametric locally weighted polynomial regression method, called LOWESS (locally weighted scatter-plot smoother), which combines several regression models into a meta-model. Each of those regression lines are estimated through weighted least square method on a subset of data, determined by nearest neighbors algorithm [8]. A smoothing parameter controls the flexibility of the model. A clear pattern is visible indicating an increasing trend in the degree of similarity over time.

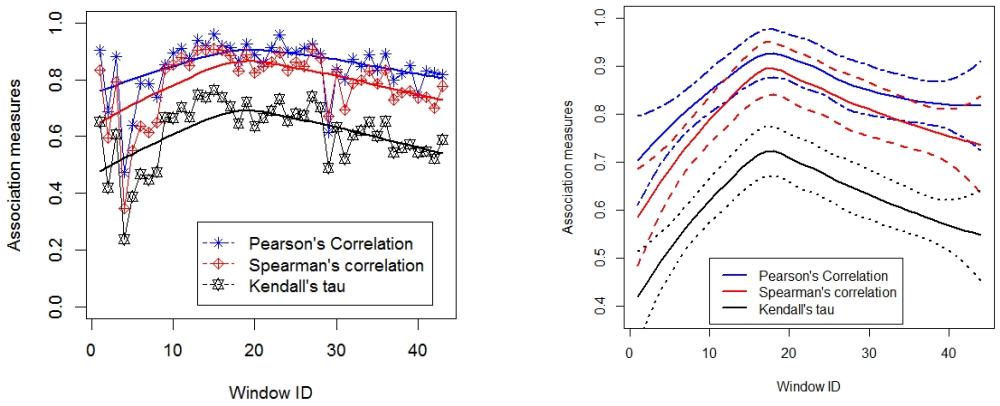


(a)



(b) (c)

Figure 4: (Color online) Visual exposition of instability of the MSTs. *Panel (a):* We have deliberately chosen a small sample size (only 20 stocks) to clearly identify the shifts in relative positions of the stocks on the MST. The MST in the left has been constructed from d-corrected return series and the MST in the right has been constructed from simple/uncorrected return series. As can be seen from the figures, node 18 appears in the backbone of the MST after correcting the return series whereas node number 7 goes to the periphery. *Panel (b):* Proportion of matching edges across MSTs constructed from d-corrected and d= 1 return series, across 47 windows over time (1972-2018). *Panel (c):* Proportion of matching edges across TMFGs constructed from d-corrected and d= 1 return series, across 47 windows over time (1972-2018).



(a) (b)

Figure 5: (Color online) Evolution of the association measures between risk measures calculated from d-corrected and $d = 1$ return series: *Panel (a):* Pearson's correlation, Spearman's correlation and Kendall's tau over three-year-window staring from 1972 till 2018. *Panel (b):* Trend in three measures of association and their 95% confidence intervals. We are observing an inverted U-shaped pattern across the full time period. The trends are plotted identically to those found in the Panel (a) for ease of comparison.

3.3 Risk Measure from Granger Causal Networks

Finally, we want to examine whether two sets of PageRanks (corresponding to two choices of differencing parameter) are significantly different from one other. A significant difference would suggest that the choice of differencing parameter is important for determining systemic risk. We first calculate the extent of linear association between these two series. Fig. 5a shows Pearson's correlation, Spearman's correlation and Kendall's tau for every four years window starting from 1973. We can see that at the initial period the correlation is a bit unstable. But with time it stabilizes. The smoothed lines in the plot obtained by LOWESS regression, suggest that the association increased approximately till 2000. During the first two decades of this century it decreases gradually over time.

3.4 Robustness: Subsample Analysis

To check for robustness of our results, we performed a subsample analysis consisting the top 100 stocks from our original data of 300 stocks in each window. The subsample analysis corroborates our earlier results. We have provided the results in the Appendix. In Fig. 6a and Fig. 6b, we observed that the same upward trend is present across time in the proportion of matched-edges for both MST and TMFG networks. As in the Fig. 3b of the time series of dominant eigenvalues obtained from the full-sample, Fig. 7a also represents the identical features for the subsample. Similarly, Fig. 7b confirms presence of an inverted U-shaped pattern in the similarity measures between two sets of PageRanks, obtained from

the subsample. In summary, these results validate our earlier results from the full-sample analysis and concurs on the effects of d-correction.

4 Summary

Analysis of spectral structure of large dimensional financial data is at the forefront of the literature on statistical finance, risk analysis and portfolio optimization (see for example [31], [5], [19]). Risk measures obtained from lagged comovement structure of financial data is a prominent area of research in the complementary domain of risk analytics and financial networks (see for example Ref. [4], [33]). One common feature across (almost) all of the relevant studies is the way return series is constructed from raw price data. It has been widely recognized in the literature that the raw price data is non-stationary in nature. Typically, the first stage of analysis is to construct log return, i.e. to take first difference of the log price series. The idea behind this operation is to get rid of the *unit root* in the raw price data by first differencing. However, it has also been noticed for quite some time that choosing the difference operator d to be exactly 1 may not be appropriate for all stocks [10]. Sometimes this leads to over-differencing, and sometimes it leads to under-differencing. Thus all of the following network studies that utilizes the return data as the primary building block, could potentially be susceptible to the order of differencing.

In this paper, we address precisely this question: Does the order of differencing matter for constructing return series in terms of the correlation-based studies analyzing collective dynamics, filtered networks and risk measures? We analyze this question on a large dimensional historical stock price data obtained from NASDAQ (1972-2018). We have shown that the impact of the choice of differencing parameter can be mild but significant in the spectral structure of the correlation matrices. After correcting for the difference parameter, the eigenspectra seem slightly heavier in the tail. Although the differences are very marginal, the dominant eigenvalue of the eigenspectra from d-corrected correlation matrix almost always seems to dominate the dominant eigenvalue obtained from the eigenspectra of $d = 1$ correlation matrix.

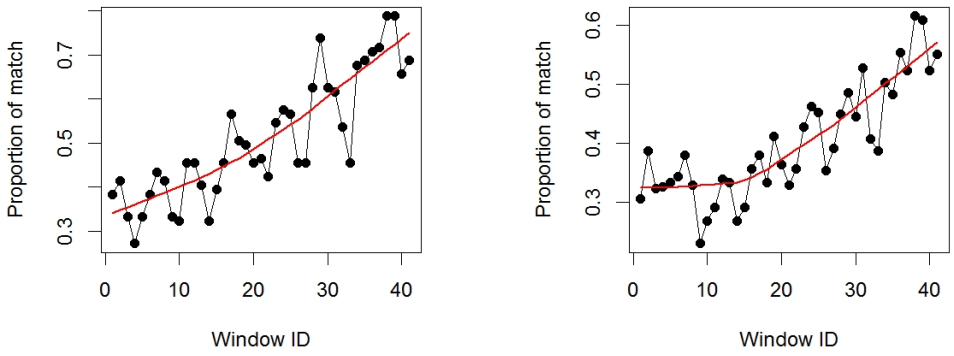
More prominent differences start showing up once we construct the minimum spanning trees (MST) from the correlation matrices, following Ref. [22]. We see that the degree of mismatch for the MSTs was quite high in the beginning of the period of analysis and the mismatch decreased almost monotonically over the years. However, even the latest value of the mismatch is in the order of around 30%. A similar feature is also observed in case of triangulated maximally filtered graphs (TMFG) which retain more information than MSTs. Empirical results show that the effects of d-corrections are substantial in the filtered asset graphs.

Finally, we analyze risk measures constructed from the Granger causal matrices, which accounts for interdependence of the stocks with a lag and therefore, provides a summary measure of propensity of shock spillovers. Following Ref. [35], we construct a PageRank-based measure of risk from the Granger causal matrices. Our analysis indicates that there is an inverted U-shaped relationship between the risk measures obtained from d-corrected and uncorrected returns. The degree of similarity increased for almost the first half of the period considered and then it started decreasing.

We can summarize the implications as follows. One, for individual time series the effects of d-correction are not substantial. However, the correlation matrices of multivariate return series exhibit subtle discrepancies manifested in minor changes in the eigenspectrum. Two, well known filtering algorithms like minimum spanning tree and triangulated maximally filtered graph show substantial change with respect to d-corrections. Since they are often inputs for further analysis (a prominent example being portfolio optimization, see e.g. [18], [29], [20]), caution need to exercised about identification of the true correlation structure and the resultant filtered graphs. Although, the analyzed data indicates that the d-corrected results for filtered networks are becoming more and more similar to the standard $d = 1$ result over time. Finally, for lagged comovements indicated by the Granger causal matrices, the fit between d-correction and $d = 1$ results is decreasing in the later half of the sample period (1972-2018). These last two features combined together indicate a chasm between the equal time and lagged comovement structures.

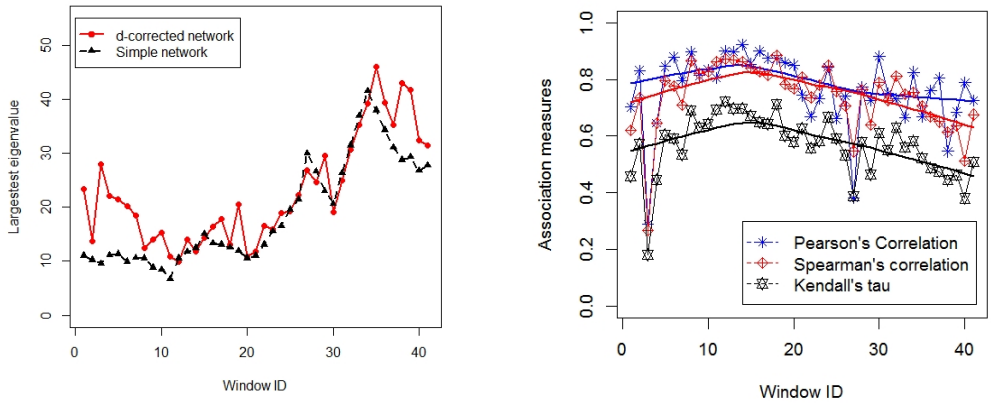
The present analysis is important from the point of view of management of risk in a complex, evolving and interdependent financial system. Further exploration is necessary to establish the extent to which the d-correction would impact filtering analysis and computation of risk on financial networks for stock price data from other markets (e.g. from developed countries, other asset classes and same asset class with high frequency data) and if the effects differ systematically across countries with different levels of financial maturity.

5 Appendix



(a) (b)

Figure 6: (Color online) Subsample analysis of MST and TMFG networks. *Panel (a):* Proportion of matching edges across MSTs constructed from d-corrected and d= 1 return series of 100 stocks, across 47 windows over time (1972-2018). *Panel (b):* Proportion of matching edges across TMFGs constructed from d-corrected and d= 1 return series of 100 stocks, across 47 windows over time (1972-2018). As in the full-sample analysis here also we see an upward trend in proportion of match across time.



(a) (b)

Figure 7: (Color online) Subsample analysis with top 100 stocks across 47 windows from 1972 till 2018. *Panel (a):* Evolution of the largest eigenvalues corresponding to the d-corrected and uncorrected correlation matrices. The plot shows that the subsample analysis concurs with the full-sample analysis. *Panel (b):* Association between risk measures calculated from d-corrected and $d = 1$ return series; Evolution of Pearson's correlation, Spearman's correlation and Kendall's tau.

References

[1] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608, 2015.

[2] T Adrian and MK Brunnermeier. Covar. new york. Technical report, Princeton University Working Paper, 2010.

[3] Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559, 2012.

[4] Giovanni Bonanno, Guido Caldarelli, Fabrizio Lillo, and Rosario N Mantegna. Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E*, 68(4):046130, 2003.

[5] Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge University Press, 2003.

[6] Markus Brunnermeier and Martin Oehmke. Complexity in financial markets. *Princeton University (working paper)*, 8, 2009.

[7] J Chan-Lau, Macro Espinosa, and Juan Solé. Co-risk measures to assess systemic financial linkages. *IMF working paper series*, 2009.

[8] William S Cleveland. Lowess: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, 35(1):54, 1981.

[9] Olivier De Bandt and Philipp Hartmann. Systemic risk: a survey. 2000.

[10] Marcos Lopez De Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.

[11] Kristin J Forbes and Roberto Rigobon. No contagion, only interdependence: measuring stock market comovements. *Journal of Finance*, 57(5):2223–2261, 2002.

[12] John Geweke and Susan Porter-Hudak. The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4(4):221–238, 1983.

[13] Tapio Heimo, Jari Saramäki, Jukka-Pekka Onnela, and Kimmo Kaski. Spectral and network methods in the analysis of correlation matrices of stock returns. *Physica A: Statistical Mechanics and its Applications*, 383(1):147–151, 2007.

[14] JRM Hosking. Fractional differencing. biometrika 68 165–176. *Mathematical Reviews (MathSciNet): MR614953 Zentralblatt MATH*, 464, 1981.

[15] Takashi Isogai. Building a dynamic correlation network for fat-tailed financial asset returns. *Applied Network Science*, 1(1):7, 2016.

[16] DeokJong Jeong and Sunyoung Park. The more connected, the better? impact of connectedness on volatility and price discovery in the korean financial sector. *Managerial Finance*, 2018.

[17] Chandrashekar Kuyyamudi, Anindya S Chakrabarti, and Sitabhra Sinha. Emergence of frustration signals systemic risk. *Physical Review E*, 99(5):052306, 2019.

[18] Yan Li, Xiong-Fei Jiang, Yue Tian, Sai-Ping Li, and Bo Zheng. Portfolio optimization based on network topology. *Physica A: Statistical Mechanics and its Applications*, 515:671–681, 2019.

[19] Giacomo Livan, Simone Alfarano, Mishael Milaković, and Enrico Scalas. A spectral perspective on excess volatility. *Applied Economics Letters*, 22(9):745–750, 2015.

[20] Giacomo Livan, Jun-ichi Inoue, and Enrico Scalas. On the non-stationarity of financial time series: impact on optimal portfolio selection. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(07):P07025, 2012.

[21] Benoit B Mandelbrot. Fractals: Form, chance and dimension. 1977.

[22] Rosario N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

[23] Guido Previde Massara, Tiziana Di Matteo, and Tomaso Aste. Network filtering for big data: Triangulated maximally filtered graph. *Journal of complex Networks*, 5(2):161–178, 2016.

[24] J-P Onnela, Anirban Chakraborti, Kimmo Kaski, Janos Kertesz, and Antti Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 68(5):056110, 2003.

[25] J-P Onnela, Kimmo Kaski, and Janos Kertész. Clustering and information in correlation based financial networks. *The European Physical Journal B*, 38(2):353–362, 2004.

[26] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[27] Epaminondas Panas. Estimating fractal dimension using stable distributions and exploring long memory through arfima models in athens stock exchange. *Applied Financial Economics*, 11(4):395–402, 2001.

[28] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A Nunes Amaral, and H Eugene Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters*, 83(7):1471, 1999.

[29] Francesco Pozzi, Tiziana Di Matteo, and Tomaso Aste. Spread of risk across financial markets: better to invest in the peripheries. *Scientific Reports*, 3:1665, 2013.

[30] Steven L Schwarcz. Systemic risk. *Geo. LJ*, 97:193, 2008.

[31] Sitabhra Sinha, Arnab Chatterjee, Anirban Chakraborti, and Bikas K Chakrabarti. *Econophysics: an introduction*. John Wiley & Sons, 2010.

[32] Ruey S Tsay. *Analysis of financial time series*, volume 543. John wiley & sons, 2005.

[33] Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization*, 75(1):40–58, 2010.

[34] Ganapathy Vidyamurthy. *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons, 2004.

[35] Tae-Sub Yun, Deokjong Jeong, and Sunyoung Park. "too central to fail" systemic risk measure using pagerank algorithm. *Journal of Economic Behavior & Organization*, 162:251–272, 2019.