Research Article
Special Issue on Econometrics and Business Analytics

Oleg Shirokikh, Grigory Pastukhov, Alexander Semenov, Sergiy Butenko, Alexander Veremyev, Eduardo L. Pasiliao, and Vladimir Boginski*

# Networks of causal relationships in the U.S. stock market

https://doi.org/10.1515/demo-2022-0110
received September 9, 2021; accepted April 2, 2022

**Abstract:** We consider a network-based framework for studying causal relationships in financial markets and demonstrate this approach by applying it to the entire U.S. stock market. Directed networks (referred to as "causal market graphs") are constructed based on publicly available stock prices time series data during 2001–2020, using Granger causality as a measure of pairwise causal relationships between all stocks. We consider the dynamics of structural properties of the constructed network snapshots, group stocks into network-based clusters, as well as identify the most "influential" market sectors via the PageRank algorithm. Interestingly, we observed drastic changes in the considered network characteristics in the years that corresponded to significant global-scale events, most notably, the financial crisis of 2008 and the COVID-19 pandemic of 2020.

**Keywords:** network analysis, graph theory, causal market graph, Granger causality, *k*-core, PageRank

**MSC 2020:** 91G45, 90B10

## 1 Introduction

Stock markets are complex interconnected systems, where various "local" factors can cause "global" changes in the behavior of the entire market. For instance, favorable or unfavorable economic conditions in certain market segments, or in certain countries, may affect other countries and industries and potentially cause positive or negative fluctuations that span the entire U.S. and international markets. The idea of describing causal relationships between different components of the market system has been addressed in several recent studies. For instance, the survey [24] discussed the concept of *contagion* in financial markets, which essentially implies the propagation of impact (such as risk) between different components of the market. A network-based model is a natural way to mathematically represent these "contagion" processes; however, the principles for constructing the networks that reflect certain types of processes may vary depending on the respective goals and assumptions of a study.

*** Corresponding author: Vladimir Boginski,** Department of Industrial Engineering & Management Systems, University of Central Florida, Orlando, FL, USA, e-mail: vladimir.boginski@ucf.edu
**Oleg Shirokikh:** Frontline Solver, Reno, NV, USA, e-mail: olegshirokikh@gmail.com
**Grigory Pastukhov:** CSX Transportation, Jacksonville, FL, USA, e-mail: grigoriypas@gmail.com
**Alexander Semenov:** Department of Industrial & Systems Engineering, University of Florida, Gainesville, FL, USA, e-mail: asemenov@ufl.edu
**Sergiy Butenko:** Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX, USA, e-mail: butenko@tamu.edu
**Alexander Veremyev:** Department of Industrial Engineering & Management Systems, University of Central Florida, Orlando, FL, USA, e-mail: alexander.veremyev@ucf.edu
**Eduardo L. Pasiliao:** Munitions Directorate, Air Force Research Laboratory, Eglin AFB, FL, USA, e-mail: eduardo.pasiliao@us.af.mil

A simple and intuitive technique for constructing a network-based (or, graphical) model of the market is to represent its elements (e.g., stocks) as nodes and connect the nodes by links (arcs) based on pairwise correlations between the corresponding entities (i.e., the correlations between stock price fluctuations over a certain period of time). Such an approach was studied in [5,6,23] in the context of identifying large correlated clusters and diversified portfolios in the U.S. stock market. Although pairwise correlation-based similarity measures have merit in certain applications, a substantial drawback of such measures is in the inability to produce *directed* links between entities, that is, to establish the direction of "contagion" (i.e., the propagation from node $i$ to node $j$ vs the propagation from node $j$ to node $i$).

In this work, we construct and analyze a directed network model, which describes causal relationships between all pairs of stocks in the U.S. stock market using the concept of Granger causality [15,16]. It should be noted that Granger causality (which will be formally defined later in the article) can be used to determine whether the time series describing stock $i$ contains useful information for predicting the behavior of the time series of stock $j$. This should not be confused with the statement "the increase/decrease in the price of stock $i$ causes the increase/decrease in the price of stock $j$," which is not necessarily true.

There are several previous studies constructing networks based on Granger causality, such as [12,26]; however, there has been no thorough analysis of the resulting networks. Also, the current literature contains little discussion about the influence of market sectors. Further, little attention has been paid to approaches widely used in network science, such as PageRank and $k$-core-based methods.

As it will be discussed in the next sections, networks constructed using Granger causality appear to capture certain structural properties of the stock market that reflect overall tendencies in its behavior. In particular, we investigate various aspects of connectivity patterns and the evolution of structural properties of the constructed network snapshots. In addition, the considered network representation is used to group stocks into network-based clusters and to identify the most "influential" market entities (sectors/ industries).

# 2 Basic concepts, data description, network construction

## 2.1 Relevant graph-theoretic concepts

Let $G = (N, A)$ be a simple *directed* graph with the set of nodes $N$ and the set of arcs $A = \{(i, j) : i, j \in N\}$, where the head $j$ and tail $i$ of each arc $(i, j)$ are specified. *Arc density* of $G$ is defined as the ratio of the number of arcs in the graph to the maximum possible number of arcs: $\rho(G) = \frac{|A|}{|N|(|N|-1)}$, where $|A|$ is the number of arcs and $|N|$ is the number of nodes in graph $G$. Obviously, $\rho(G) \in [0, 1]$.

Given a node $n \in N$, its *in-degree* $\mathbf{deg}_G^{\text{in}}(n)$ is the number of incoming arcs and its *out-degree* $\mathbf{deg}_G^{\text{out}}(n)$ is the number of outgoing arcs. Extensive empirical studies show that degree distributions of many real-life graphs representing diverse datasets follow the well-known *power-law* model [2–4, 8,19]. According to this model, the probability that a node has a degree $k$ (in- or out-degree for directed graphs) is $\mathbb{P}(k) \propto k^{-\gamma}$, or $\log\mathbb{P}(k) = -\gamma \log k + \text{const}$ in the log–log scale, which can be described by a straight line with the slope equal to the parameter $\gamma$ of the power-law degree distribution. One of the notable characteristics of such networks (known as the scale-free property) is that their power-law structure should not depend on the network's size.

A directed graph $G = (N, A)$ is called *strongly connected* if there is a directed path from each node to every other node in the set $N$. A disconnected graph can be decomposed into strongly connected subgraphs, which are referred to as *strongly connected components* of $G$. Distinct components can be interpreted as *clusters* in the corresponding dataset. Several algorithms exist for the efficient identification of strongly connected components in a directed graph. In our study, we use the popular Tarjan's algorithm based on the depth-first search technique [25].

In some situations, clusters based on strongly connected components can be extremely large and comparable with the size of the whole graph (which is in fact the case for the considered graphs, as it will be shown later). Therefore, the clustering approach based on connected components may not be necessarily appropriate for drawing meaningful conclusions regarding specific groups of nodes within a graph. There is a variety of definitions for "tighter" structures that may be interpreted as clusters that have specific cohesive properties of their connectivity patterns. In this study, we utilize the concepts of *$k$-degenerate graph* and *$k$-cores* for undirected graphs, introduced by [22], and modify them for the case of directed graphs. A simple undirected graph is called $k$-degenerate if every its subgraph has a vertex of degree at most $k$. The *degeneracy* of a simple undirected graph $G$, denoted by $\delta^*(G)$, is the smallest value of $k$ such that $G$ is $k$-degenerate. A $k$-core in a simple undirected graph is a subset of vertices that induces a subgraph with the minimum degree at least $k$. Alternatively, one can define $k$-cores as connected components that are left after all nodes of degree less than $k$ have been removed from the graph $G$; therefore, $\delta^*(G)$ is the maximum $k$ for which $G$ contains a nonempty $k$-core.

An extension of the notion of a $k$-core was proposed in [14] by introducing the concept of a $D$-core in a directed graph. The authors consider *min-in-degree* and *min-out-degree* of a graph $G$ defined as $\delta^{\text{in}}(G) = \min_{x\in N}\{\mathbf{deg}_G^{\text{in}}(x)\}$ and $\delta^{\text{out}}(G) = \min_{x\in N}\{\mathbf{deg}_G^{\text{out}}(x)\}$, respectively. Then, for two positive integers $k, l$, a $(k,l)$-$D$-core is a maximal size subgraph $G'$ of $G$, where $\delta^{\text{in}}(G') \geq k$ and $\delta^{\text{out}}(G') \geq l$. The intuition behind this notion is to find a subset of the graph, where all the nodes have sufficient out- and in-degrees in order to form a "tight" cluster. For the reasons that will become clear later in the article, we introduce a slightly different structure referred to as a *$k$-out-core*, where each node is only required to have an out-degree of at least $k$, i.e., the condition for the in-degree is relaxed. Therefore, for a positive integer $k$, a $k$-out-core of the graph $G$ is defined as a subgraph $G'$ of $G$, where $\delta^{\text{out}}(G') \geq k$. As in the case of undirected graphs, we can define $k$-out-cores as connected components that are left after all nodes of out-degree less than $k$ have been removed from the graph $G$; therefore, $\delta^*_{\text{out}}(G)$ is the maximum $k$ for which $G$ contains a nonempty $k$-out-core. Although the original definition of "degeneracy" differs from the definition of $\delta^*_{\text{out}}(G)$, for simplicity, we will use the same notation.

For a given $k$, $k$-out-cores can be easily found using a greedy algorithm, which recursively removes the nodes with out-degree less than $k$ one by one from the graph, until all the remaining nodes have sufficiently large out-degrees. Then one can decompose the resulting network into connected components, which are $k$-out-cores by definition. The degeneracy of $G$ can be found using a simple binary search technique.

## 2.2 Granger causality

In the aforementioned previous studies of the network-based model of the U.S. stock market [5,6], the *market graph* was constructed in such a way that a given pair of nodes is connected by an undirected edge if the corresponding stocks exhibit a similar behavior over a certain period of time. The similarity was measured by Pearson's correlation between the time series representing the returns of corresponding stocks. In this study, we propose a different technique for constructing the set of arcs: the similarity between stocks is measured by *Granger causality* [15,16], which is extensively used across many application areas because of its simplicity, robustness, and flexibility [9,13]. The details of the network construction will be presented in Section 2.3, whereas here we introduce the definition of causality and the procedure for conducting Granger causality test between two time series.

Consider two scalar-valued, stationary time series $\{x_t : t = 1, \dots, T\}$ and $\{y_t : t = 1, \dots, T\}$ corresponding to the returns $x$, $y$ of a pair of stocks. The basic idea behind the notion of causality is very general in its nature: one can state that $x$ causes $y$, denoted by $x \Rightarrow y$, if $x$ contains some unique information about $y$, so that $y$ can be better predicted using this information than in the absence of this information. In practice, Granger causality is often tested using the following linear autoregressive model:

$$y_t = \sum_{i=1}^{k} a_i y_{t-i} + \sum_{j=1}^{k} b_j x_{t-j} + \varepsilon_t, \tag{1}$$

where $k$ is the maximal time lag and $\varepsilon_t$ is a regression error. Then, $x$ does not cause $y$ if and only if

$$\mathbf{H}_0 : b_j = 0, \quad j = 1, \dots, k.$$

To test this hypothesis, one can apply the $F$-test, and rejecting $\mathbf{H}_0$ implies that $x$ "Granger causes" $y$. The procedure of testing for the presence of causality in the other direction ($y \Rightarrow x$) is similar.

It should be noted that the Granger causality test is valid only if the time series are covariance (or weak) stationary. In this article, we used the Augmented Dickey-Fuller test [20] to check the stationarity of time series. Further, we assume homoscedasticity, i.e., constant variance of $\varepsilon_t$.

## 2.3 Network construction

In the constructed directed unweighted network, the nodes are stocks represented as "ticker" symbols. We used all the stocks listed at NYSE, NASDAQ, and AMEX as of December 31, 2020: There were 7,240 stock symbols in total. The list of stock symbols was obtained from EODdata.[1] We obtained historical stock prices data from Yahoo Finance using *yfinance*[2] Python library.

The adjusted close prices data were transformed into the time series of daily returns, since returns possess scalability property (i.e., the values in time series representing each stock returns have the same order of magnitude) and thus are easily comparable. Furthermore, the logarithms of returns were calculated, due to the fact that log-returns have more attractive statistical properties [11], including weak stationarity, which was verified for all considered time series. If $P_i(t)$ and $P_i(t-1)$ are the adjusted close prices of stock $i$ on days $t$ and $t-1$, respectively, then the log-return time series for each stock $i$ are defined as follows:

$$r_i(t) = \ln \frac{P_i(t)}{P_i(t-1)}, \quad t = 2, \ldots, T,$$

where $T$ is the number of trading days in each of the considered calendar years (2001–2020).

A directed network (referred to as a *causal market graph*) was constructed for each time period (calendar year) to reflect the causal relationships between stocks. It should be noted that a network constructed for each time period contains only those stocks that were present in the market during that entire time period; therefore, the cardinality and composition of the sets of nodes change from period to period. Every stock is represented by a node, and the existence of an arc $(i, j)$ means that the time series of stock $i$ causes the time series of stock $j$ in the sense of Granger causality. Recall that Granger causality test checks the hypothesis that coefficients $b_j = 0, j = 1, \ldots k$. The null hypothesis (all $b_j$ are equal to zero) is rejected in favor of alternative if the $p$-value of $F$-test is less than a certain threshold. Hence, an arc between stocks $i$ and $j$ is constructed if the corresponding $p$-value is less than a chosen threshold. We picked this threshold to be 0.001, which means that Granger causality holds with 99.9% confidence. The motivation behind this threshold choice is to ensure that the constructed networks are *sparse* enough, so that it would be possible to observe significant changes in connectivity patterns over time (as opposed to the situation where each network contains close to the maximum possible number of arcs, which makes it difficult to detect any changes in connectivity patterns). Thus, only the most "meaningful" connections are reflected in the constructed networks. The summary of statistics of autoregression coefficients in Eq. (1) for edges kept in the networks is shown in Table 1. It is interesting that the values of $b$ are often close to zero, even if the null hypothesis is rejected.

The Granger causality test can be performed with different numbers of lags. In our preliminary computations, we found that in many cases the Bayesian information criterion (BIC) [21] produced the optimal quantity of one or two lags. Moreover, the corresponding $p$-values were very close for both cases. Since it is computationally expensive to check the BIC for every pair of time series, one lag was chosen for all the Granger causality tests, as it was optimal or near-optimal for most pairs of time series. This choice can also

[1] https://eoddata.com/.

[2] https://pypi.org/project/yfinance/.

**Table 1:** Summary statistics for autoregressive model from Eq. (1), all networks

| | Mean | Std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|
| $b$ | −0.000160 | 0.001923 | −0.040823 | −0.000692 | 0.000108 | 0.000621 | 0.046510 |

be justified by the widely used assumption in financial mathematics that stock returns possess the Markovian property [18].

# 3 Dynamics of structural properties of causal market graph

To reveal the long-term evolution of causal market graph characteristics over time, we consider 20 non-overlapping 1-year periods spanning the most recent two decades. We consider the dynamics of characteristics of the causal market graph, including the number of nodes, arc density, node degrees, connectivity, and degree distribution. In addition, we compute strongly connected components, $k$-out-cores and propose a structural decomposition of the causal market graph.

## 3.1 Basic characteristics

The set of stocks traded on NASDAQ, NYSE, and AMEX has undergone significant changes during 2001–2020. As it is shown in Table 2, the number of nodes (stocks) increased from 2087 in period 1 to 7240 in period 20. The number of publicly traded stocks increased by 246% despite the fact that many companies present in the market in earlier periods ceased to exist in later periods.

**Table 2:** Basic characteristics of networks corresponding to each time period

| Year | #Nodes | #Arcs | Max. o.d. | Max i.d. | Arc density (%) | GCC size (%) | In-in assort. | Out-out assort. |
|---|---|---|---|---|---|---|---|---|
| 2001 | 2,087 | 30,082 | 176 | 456 | 0.69 | 95.35 | −0.028 | 0.111 |
| 2002 | 2,253 | 33,625 | 292 | 641 | 0.66 | 94.85 | −0.048 | 0.115 |
| 2003 | 2,352 | 15,230 | 84 | 153 | 0.28 | 89.71 | −0.020 | 0.092 |
| 2004 | 2,482 | 27,947 | 154 | 242 | 0.45 | 93.03 | 0.160 | 0.244 |
| 2005 | 2,656 | 22,133 | 103 | 257 | 0.31 | 93.34 | −0.084 | 0.100 |
| 2006 | 2,841 | 33,701 | 128 | 878 | 0.42 | 94.65 | −0.067 | 0.177 |
| 2007 | 3,084 | 134,188 | 791 | 855 | 1.41 | 98.51 | −0.028 | 0.053 |
| 2008 | 3,418 | 651,729 | 1,649 | 2,591 | 5.58 | 99.44 | −0.214 | 0.060 |
| 2009 | 3,558 | 110,200 | 997 | 2,212 | 0.87 | 98.37 | −0.069 | −0.009 |
| 2010 | 3,700 | 95,185 | 1,432 | 1,820 | 0.70 | 95.62 | 0.009 | 0.023 |
| 2011 | 3,944 | 185,331 | 2,004 | 2,495 | 1.19 | 98.07 | −0.091 | −0.006 |
| 2012 | 4,130 | 82,420 | 450 | 1,598 | 0.48 | 97.34 | −0.011 | 0.168 |
| 2013 | 4,410 | 93,113 | 598 | 978 | 0.48 | 96.67 | −0.063 | 0.175 |
| 2014 | 4,697 | 121,270 | 557 | 1,630 | 0.55 | 98.59 | −0.025 | 0.163 |
| 2015 | 5,061 | 224,480 | 999 | 2,120 | 0.88 | 99.19 | −0.017 | 0.144 |
| 2016 | 5,403 | 205,611 | 558 | 2,106 | 0.70 | 99.44 | −0.034 | 0.173 |
| 2017 | 5,720 | 101,240 | 335 | 1,742 | 0.31 | 99.28 | −0.069 | 0.058 |
| 2018 | 6,147 | 407,644 | 1,198 | 3,081 | 1.08 | 99.74 | −0.017 | 0.172 |
| 2019 | 6,701 | 273,246 | 1,101 | 2,417 | 0.61 | 99.69 | −0.003 | 0.184 |
| 2020 | 7,240 | 3,416,051 | 4,720 | 5,685 | 6.52 | 99.90 | −0.118 | −0.091 |

(max. o.d. and max i.d. are maximum out-degree and maximum in-degree, respectively; GCC size is the size of the giant connected component as the percentage of the total number of nodes; the last two columns show the respective in- and out-degree assortativity).

The threshold value used to identify whether two nodes are connected controls the total number of arcs in the graph. Although the threshold specified in the previous section was chosen to be rather conservative, one can see that the number of arcs can still be large; however, it varies greatly: from 15,230 arcs in 2003 to over 3.4 million arcs in 2020. Due to the difference in the number of nodes in the networks corresponding to different time periods, it makes sense to calculate the *arc density* (i.e., the ratio of the number of arcs to the maximum possible number of arcs), which is a unit-less measure; thus, it can be used to compare graphs with different numbers of nodes. Table 2 summarizes basic characteristics of the networks corresponding to all considered time periods.

In the case of correlation-based (undirected) graph instances constructed over a shorter timeframe, the arc density steadily increased over time [5]. However, the causal market graph does not have this property: Table 2 presents the nonmonotonic dynamics of the number of arcs and the arc density, the latter being also visualized in Figure 1. One can interpret the arc density of the causal market graph as a proportion of ordered pairs of stocks, such that the data corresponding to returns of one stock can be potentially used in order to forecast the future return values of the other. Table 2 presents two other fields related to the network structure: maximum out- and in-degrees. Based on the model of causality, the stocks with high out-degrees are the most “informative” in the sense that their statistics could be used for investigating the behavior of a large number of adjacent stocks (successor nodes in the causal market graph). The in-degree of a node can be treated as the property reflecting the number of stocks containing unique information about this stock. Although this characteristic may be meaningful in certain contexts, in this part of the study, we concentrate mainly on *out-degrees* of nodes due to the aforementioned considerations.

The evolution of the density in the causal market graph is shown in Figure 1. One can observe that it has relatively small values during 2001–2006, but it starts to increase in 2007. Further, the arc density attains its highest values in 2008 and 2020. Many economists associate 2007 with the beginning of the worst financial downfall since the Great Depression (started with the U.S. subprime mortgage crisis). The most significant economic event of 2008 is the collapse of the stock market when Dow Jones and S&P500 endured their worst year since 1930. In 2009, although the US economy was still weak, the stock market started to slowly recover after hitting the bottom in March 2009. As one can see, the values of arc density fell drastically compared to 2008, and they stayed relatively stable until 2020, when COVID-19 pandemic started. It can also be observed that in-in assortativity has its lowest values in 2008 and 2020.

Although the analysis of these basic properties of the constructed networks may not by itself be sufficient to draw comprehensive conclusions, it can be seen that extreme values of arc density and maximum out-degree of the causal market graph correspond to extreme events in the stock market, and the trends can be noted for the transition periods as well. The different nature of events that impacted the market between 2008 and 2020 may explain the difference in the magnitude of these metrics. In particular, it can be observed that two drastic “spikes” of arc density of the causal market graph (in 2008 and 2020)

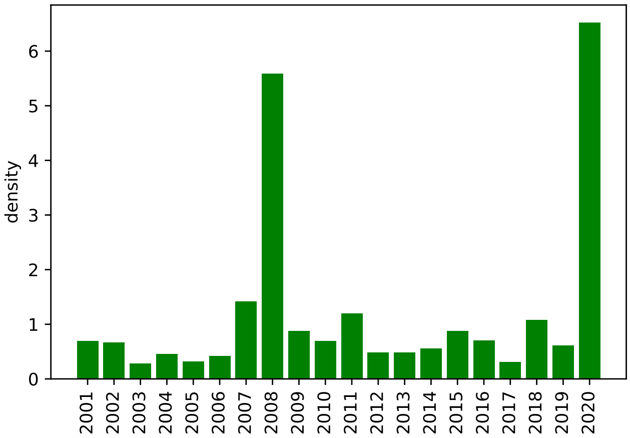

**Figure 1:** Evolution of arc density.

appear to be inherently different: the 2008 spike was preceded by a smaller yet still significant increase of arc density in 2007 (in fact, the arc density of the 2007 graph is the third largest among the considered time periods), whereas the 2020 spike was not preceded by such an increase. The difference between the respective underlying events that affected the market in 2008 and 2020 is that the 2008 crisis was anticipated by experts based on market trends that started during 2007, but the crisis associated with the 2020 COVID-19 pandemic was not anticipated during 2019.

In addition, we consider the specific nodes (stocks/companies) that are most "influential" in the sense that their time series data contain useful information about a large number of other stocks. Figure 2 presents the aggregate distribution of highest out-degree stocks by sector for all considered periods. As one may intuitively expect, the top sectors in this diagram are *Funds* (that corresponds to Funds, Trusts, and Tracking Stocks) and *Financial Services*, followed by several other important sectors of the market.

## 3.2 Degree distribution

As mentioned in Section 2.1, many previous studies have shown that the power-law distribution of out- and in-degrees appears to be a common property for many real-world networks. The degree distribution of most of the constructed causal market graphs also appears to follow a power law, although the quality of power-law fit varies between different network snapshots. Table 3 summarizes the evolution of the power-law parameter $\gamma$ and the respective $R^2$ value (which reflects the quality of a least-square fit of a straight line to the log–log data). One can observe from Table 3 that the $R^2$ is only about 63% for 2008 and about 71% for 2020 out-degree distribution fit, but it is significantly higher for other years. Thus, it appears that more substantial deviations from power-law degree distributions coincide with significant events affecting the market. For illustrative purposes, we present the out- and in-degree distributions for causal market graph instances (plotted in the log–log scale) for 2008 and 2019 (see Figures 3 and 4).

Although the value of the parameter $\gamma$ is rather stable for most of the considered time periods, there is a visible decrease for out- and in-degree distributions corresponding to 2008 and 2020, which is consistent with the aforementioned observations of other metrics, since a smaller value of $\gamma$ implies a heavier tail of the distribution (i.e., more nodes with high degrees).

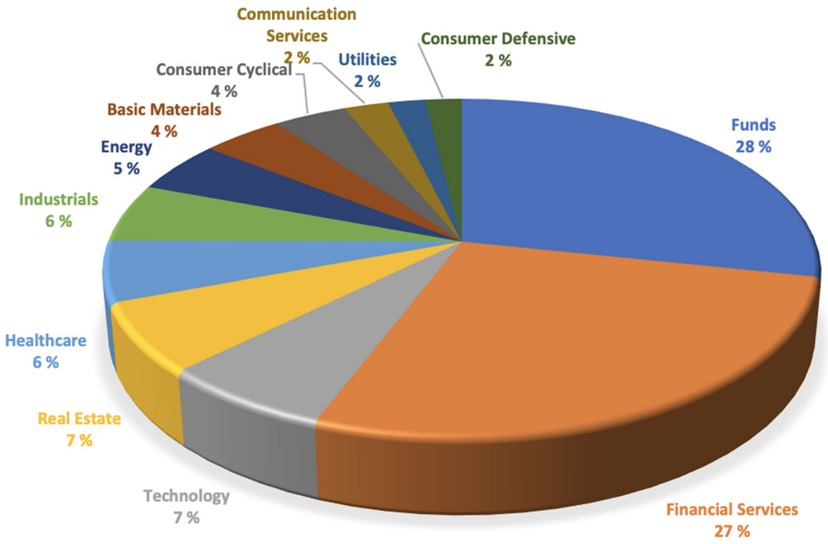

**Figure 2:** Distribution of highest out-degree stocks by sectors for all considered years.

**Table 3:** Power-law fit results for in-degree and out-degree distributions

| Year | $\gamma_{(\text{in})}$ | $R^2_{(\text{in})}$ | $\gamma_{(\text{out})}$ | $R^2_{(\text{out})}$ |
|---|---|---|---|---|
| 2001 | 1.3114 | 0.8517 | 1.4595 | 0.8856 |
| 2002 | 1.2130 | 0.8418 | 1.4457 | 0.8497 |
| 2003 | 1.5744 | 0.8433 | 1.9149 | 0.8822 |
| 2004 | 1.4431 | 0.8929 | 1.5107 | 0.8469 |
| 2005 | 1.4516 | 0.8614 | 1.7908 | 0.8578 |
| 2006 | 1.1817 | 0.7709 | 1.6346 | 0.8683 |
| 2007 | 1.0375 | 0.8090 | 1.1608 | 0.8565 |
| 2008 | 0.7791 | 0.7800 | 0.7257 | 0.6368 |
| 2009 | 1.0251 | 0.7584 | 1.2947 | 0.8085 |
| 2010 | 1.1015 | 0.8113 | 1.1790 | 0.7953 |
| 2011 | 1.0251 | 0.7867 | 1.1766 | 0.8145 |
| 2012 | 1.0573 | 0.7613 | 1.4694 | 0.8593 |
| 2013 | 1.0987 | 0.8123 | 1.5229 | 0.8362 |
| 2014 | 1.0875 | 0.7789 | 1.3902 | 0.8391 |
| 2015 | 1.0006 | 0.7807 | 1.2824 | 0.8642 |
| 2016 | 0.9873 | 0.7449 | 1.3971 | 0.7651 |
| 2017 | 1.3792 | 0.8172 | 1.6743 | 0.7321 |
| 2018 | 0.8402 | 0.6907 | 1.1444 | 0.7974 |
| 2019 | 1.1093 | 0.7716 | 1.3782 | 0.8196 |
| 2020 | 0.7239 | 0.7070 | 0.7314 | 0.7109 |

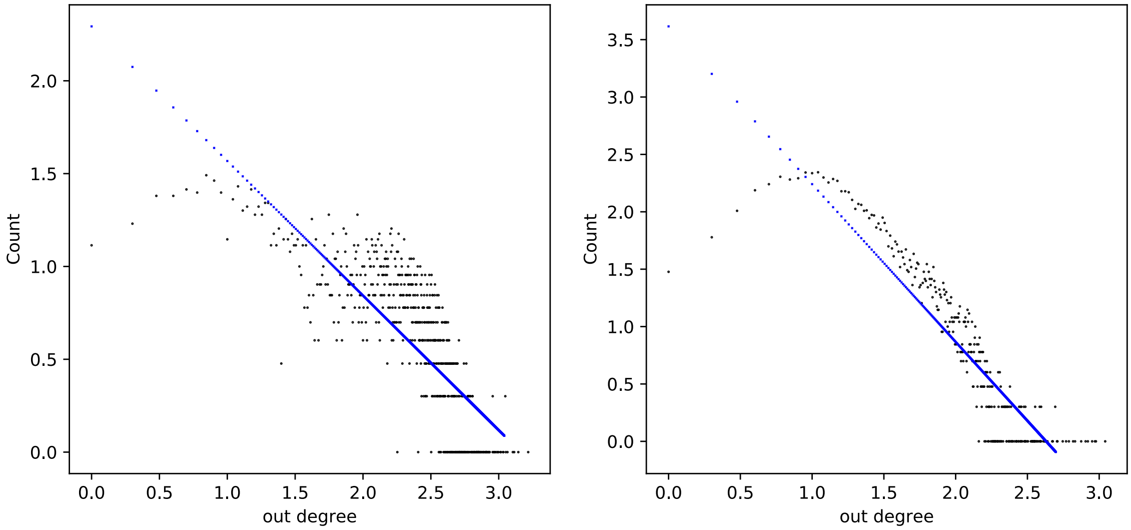

**Figure 3:** Out-degree distributions for 2008 (left) and 2019 (right).

## 3.3 Strongly connected components

Another interesting question concerning the causal market graph is whether it is strongly connected. If the answer is "yes," then it would mean that each stock $i$ has some relationship with every other stock $j$ via a directed path of causal relationships connecting nodes $i$ and $j$. To address this question, we have identified the largest strongly connected component in each considered network snapshot. We observed that every considered causal market graph had a "giant" component containing almost all of the nodes. In particular, the smallest size of a giant strongly connected component among the considered networks, which was observed for the 2003 network snapshot, contained almost 90% of the total number of nodes, whereas in many other instances, the relative size of the giant connected component was close to 100%.

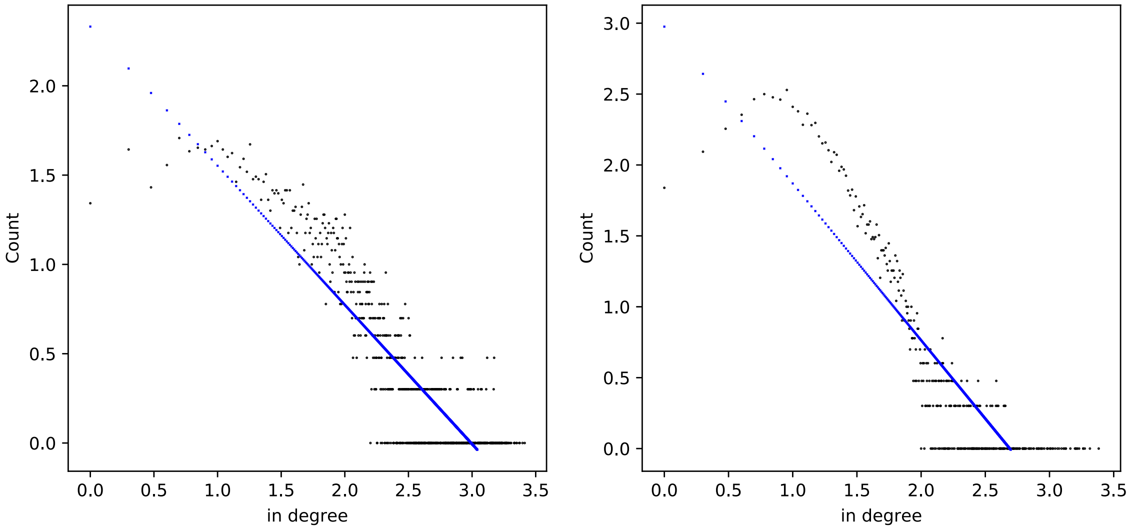

**Figure 4:** In-degree distributions for 2008 (left) and 2019 (right).

Returning to Table 3, it can be seen that the parameter $\gamma$ of the power-law distribution fluctuates between 0.7 and 1.9 for both out- and in-degrees. Most of these values of $\gamma$ are consistent with the range corresponding to the existence (with high probability) of a giant connected component in a power-law random graph, which has been theoretically proven to be $(1, 3.4785)$ for the undirected version of the power-law model in [1].

## 3.4 Identifying cohesive clusters based on $k$-out-cores

Due to the presence of a giant strongly connected component discussed in the previous subsection, strongly connected components cannot be used for clustering (i.e., partitioning a graph into subgraphs according to some similarity criterion), since one cluster would contain virtually all nodes in the graph. Therefore, in this section, we focus our attention on $k$-out-cores, which are more “cohesive” network clusters compared to connected components.

Recall from Section 2.1 that a $k$-out-core is a highly interconnected set of nodes with out-degrees of at least $k$ within this set. Therefore, in the context of the causal market graph, this structure represents a group of stocks, where each stock has causal relationships with at least $k$ other stocks within the group. To find out how large the number $k$ can be, we compute the graph degeneracy for each time period, as described in Section 2.1. Table 4 presents the degeneracy ($\delta^*_{\text{out}}(G)$), and $k$-out-core size ($|C|$) for $k = \delta^*_{\text{out}}(G)$, and the proportion of the $k$-out-core size to the number of nodes ($|C|/|N|$) in the causal market graph for all considered periods.

Taking a closer look at the $k$-out-core found in the 2008 network snapshot, one can see that 923 stocks form a connected cohesive structure, in which every stock has an out-degree of at least 92. This is a rather interesting observation, taking into account that Granger causality links were constructed using a very conservative threshold value. An intuitive explanation of this fact is that the crisis of 2007–2008 impacted a large portion of the market, which in turn substantially increased the number of statistically significant causal relationships between stocks. An even “denser” $k$-out-core (with out-degree of each node at least 256!) was found in the 2020 network, which was affected by COVID-19 pandemic. Overall, the $k$-out-core decomposition approach confirms the observations reported earlier; moreover, it allows one to observe “amplified” trends corresponding to significant events affecting the market.

**Table 4:** *k*-out-cores in causal market graphs for 2001–2020

| Year | Degeneracy | *k*-out-core size | Proportion (%) |
|---|---|---|---|
| 2001 | 6 | 519 | 24.87 |
| 2002 | 7 | 656 | 29.12 |
| 2003 | 2 | 1,754 | 74.57 |
| 2004 | 5 | 172 | 6.93 |
| 2005 | 3 | 1,735 | 65.32 |
| 2006 | 5 | 48 | 1.69 |
| 2007 | 17 | 952 | 30.87 |
| 2008 | 92 | 923 | 27.00 |
| 2009 | 9 | 493 | 13.86 |
| 2010 | 9 | 402 | 10.86 |
| 2011 | 17 | 365 | 9.25 |
| 2012 | 7 | 164 | 3.97 |
| 2013 | 9 | 394 | 8.93 |
| 2014 | 9 | 1,415 | 30.13 |
| 2015 | 29 | 173 | 3.42 |
| 2016 | 11 | 2,713 | 50.21 |
| 2017 | 7 | 3,774 | 65.98 |
| 2018 | 17 | 2,475 | 40.26 |
| 2019 | 19 | 524 | 7.82 |
| 2020 | 256 | 1,797 | 24.82 |

## 4 Identifying influential market sectors using pagerank

While a stock's out-degree appears to be a reasonable quantitative measure of the stock's importance, it treats all links as equal and does not take into account the difference in importance of out-neighbors. The PageRank method, which was proposed in [7] for ranking webpages in Google's search engine, is a simple yet very effective technique that overcomes this drawback. It can be applied to rank nodes in a directed network according to their importance or "centrality" expressed by a certain score. [10] describes the PageRank method as a "democracy," with links interpreted as votes in favor of the webpages they are directed to. Each webpage can vote for other webpages, and its score is divided evenly over the set of webpages it is voting for. In the realm of a causal market graph, webpages are replaced with stocks and hyperlinks – with causality relations. In addition, we reverse the directions of arcs in the causal market graph to reflect the idea that stock $i$ causing stock $j$ corresponds to stock $j$ "voting" for stock $i$. We call the resulting network a *reverse causal graph* and denote it by $G_r = (N, A_r)$. Then, a stock's score can be viewed as a weight $w_i$ assigned to the stock $i$, which is uniformly distributed among its out-neighbors in the reverse causal graph, and is computed as the sum of the corresponding proportional weights of in-neighbors, i.e.,

$$w_i = \sum_{j:(j,i)\in A_r} \frac{w_j}{\mathbf{deg}^{\text{out}}_{G_r}(j)}, \quad i = 1,\ldots,|N|, \tag{2}$$

or, in the matrix form, $w = Bw$, where $B = [b_{ij}]_{i,j=1}^{|N|}$ is given by

$$b_{ij} = \begin{cases} \dfrac{1}{\mathbf{deg}^{\text{out}}_{G_r}(j)}, & \text{if } (j,i) \in A_r; \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

Hence, the problem of finding the scores reduces to computing the eigenvector of the column-stochastic matrix $B$ that corresponds to the eigenvalue equal to 1. As soon as the scores are computed, we can rank the stocks by ordering the scores from highest to lowest. To overcome technical shortcomings arising when the network has nodes of out-degree 0 or is not connected, the original PageRank method is based on solving the system $w = (\mathrm{d}B + (1 - d)S)w$ instead of $w = Bw$, where $d = 0.85$ and $S$ is an $|N| \times |N|$ matrix with all entries equal to $1/|N|$. More detail on PageRank method and related literature are provided in [10].

In our experiments, we use PageRank to identify market sectors and industries within a given sector that are most important with respect to aggregated causal relationships. To rank the market sectors over a certain time period, we apply PageRank to the newly introduced *causal market sector graph* $G_s = (N_s, A_s)$ that is obtained from a causal graph $G_r = (N, A_r)$ by merging a subset of nodes $I_r$ representing stocks from the same market sector into a single node $i_s$ (referred to as *sector node*). In addition, for any two-sector nodes $i_s$ and $j_s$ in $N_s$, we assign a weight $l(i_s, j_s)$ to the arc between them as follows:

$$l(i_s, j_s) = \sum_{i \in I_r, j \in J_r} 1_{A_r}((i, j)),$$

where $I_r$ and $J_r$ are subsets of all nodes in $N_r$ that were used to define $i_r$ and $j_r$, respectively; and $1_{A_r}((i, j))$ is the indicator function for $A_r$, which yields 1 if $(i, j) \in A_r$ and 0 otherwise. To apply the PageRank method to the edge-weighted graph $G_s$, we need to solve the system

$$w^s = (\mathrm{d}B^s + (1 - d)S^s)w^s, \tag{4}$$

where $B^s = [b^s_{pq}]^{|N_s|}_{p,q=1}$ is given by

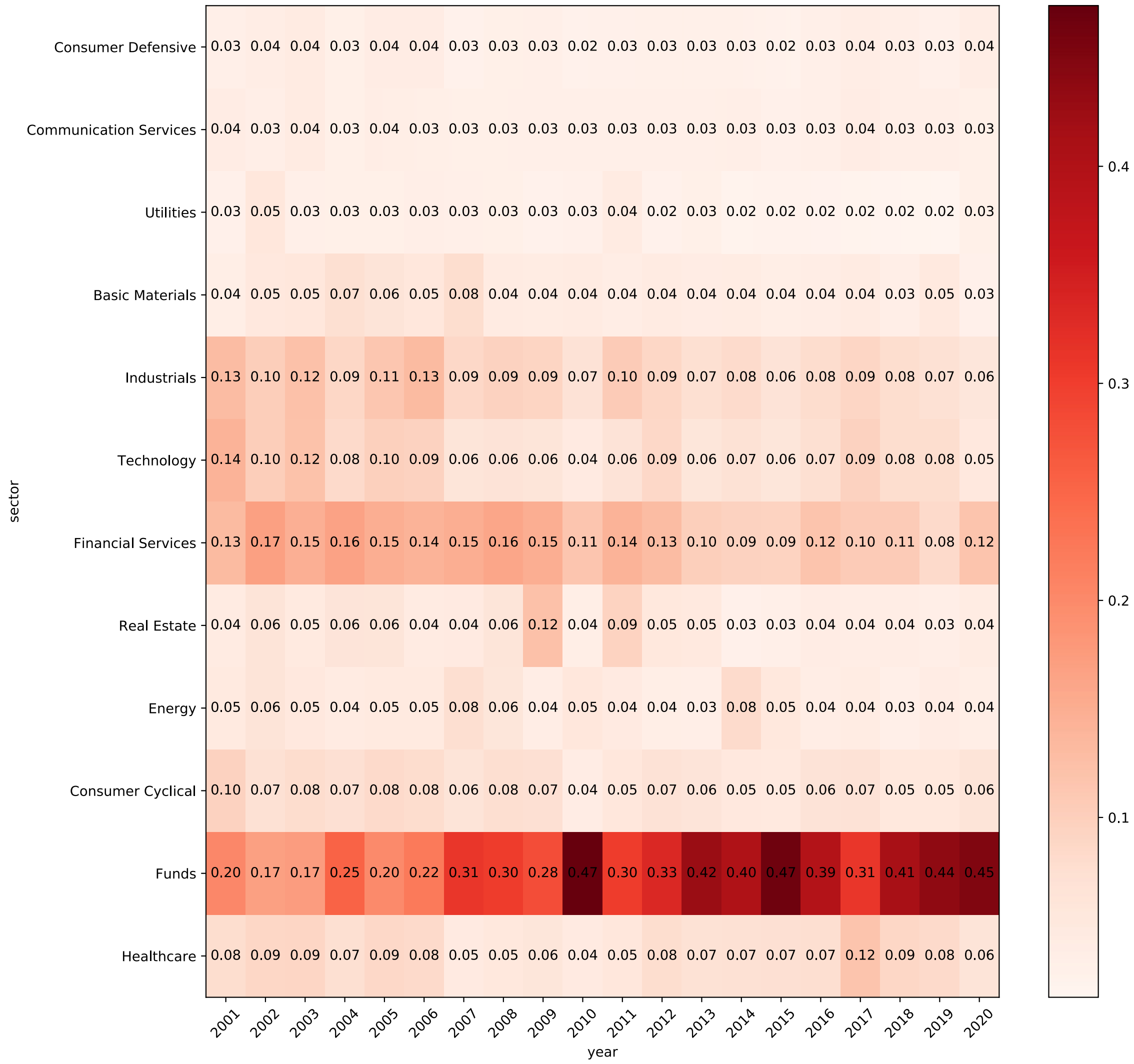

**Figure 5:** Breakdown of the most influential market sectors for each time period based on the PageRank method.

$$b^s_{pq} = \begin{cases} \dfrac{l(q,p)}{\sum_{p':(q,p')\in A_s} l(q,p')}, & \text{if } (q,p) \in A_s; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

$d = 0.85$, and $S^s$ is an $|N_s| \times |N_s|$ matrix with all entries equal to $1/|N_s|$.

Figure 5 shows the breakdown of most influential market sectors for each time period according to their PageRank scores. One can observe that Funds, Trusts, and Tracking Stocks is the top-ranked sector in all time periods except 2002, when Financial Services sector had the same PageRank score. The fact that Funds, Trusts, and Tracking Stocks is the most influential market sector is not surprising, since many stocks in this sector are by definition reflective of the behavior of the entire market. The fact that Financial Services is the second-most influential sector in most of the considered time periods is also somewhat expected; however, it is interesting to observe that the PageRank scores of Financial Services, Industrial, and Technology sectors have decreased in the most recent years. Although the PageRank-based approach has limitations since it takes into account only the respective network topology, these observations may be worth investigating further from more traditional economics, and finance-based perspectives.

## 5 Conclusion

In this article, we constructed a network-based map of *causal relationships* in the entire U.S. stock market. The considered network-based model of the stock market is based on publicly available stock prices data and a quantitative causality measure, which makes the model easily interpretable and reproducible. The proposed approach enables one to apply the rich arsenal of network analysis tools toward revealing market trends and investigating the properties of individual nodes and market clusters that may not be apparent otherwise. We focused on studying the basic structural properties of the causal market graph and detecting its most influential entities. The considered network-based metrics are nonmonotonic, with an interesting observation that significant changes over time appear to coincide with global-scale events, such as COVID-19 pandemic and the 2008 financial crisis. In addition, the proposed PageRank-based technique for identifying "influential" market sectors revealed interesting observations that may be worth investigating further.

In terms of other possible methods for constructing the respective networks, another potential direction of further research would be to analyze networks constructed using other connectedness computation methods such as [17,27]. It would also be of interest to consider heteroscedasticity in Granger causality and see its effect on the resulting networks. Future research may also include the investigation of the possibility of constructing a market index solely based on Granger causality metrics. The implication of the presence of power-law degree distribution in many of the networks is that a relatively small number of stocks have a large number of strong causal links to a large remaining portion of the market. Further, this observation suggests that the set of stocks comprising the $k$-out-cores can be potentially used to create a conceptually new network-based market index.

A limitation of this study, which may be addressed in future research, is the problem of multiple comparisons. In order to construct the edges, we do pair-wise Granger causality tests between each pair of nodes. For each pair-wise comparison, the employed statistical tests may result in incorrect rejection of the null hypothesis and adding a wrong edge with 0.1% chance. Even though the probability of adding a "wrong" edge is low, the networks analyzed in this article contain thousands of nodes, and considering independent tests, these networks may contain a few "wrong" edges. Despite the fact that these potential effects cannot be completely ruled out, the results presented in the article networks still contain interesting properties, such as the presence of power-law degree distributions, patterns of arc density changes corresponding to financial crises, and other observations, which are unlikely to appear solely due to statistical anomalies.

The considered approaches can potentially be applied in a wider variety of settings. One interesting future research direction would be to consider networks of causal relationships that span stock markets of multiple countries. Another potential area of interest would be applying these techniques to shorter time periods, possibly with smaller time increments between data points (e.g., one could consider hourly, or minute-by-minute stock prices data over a time period of several days or weeks). In particular, although this article focused mainly on a descriptive rather than predictive/prescriptive analysis of stock market data, it would be interesting to see if the considered network-based approaches (perhaps with some modifications) could be used in the context of predictive models of market trends.

**Acknowledgments:** Preliminary results related to this research were presented at the 7th International Conference on Computational Data and Social Networks (CSoNet 2018, Shanghai, China, December 2018) and at the 1st International Conference on Econometrics and Business Analytics (iCEBA 2021, Saint Petersburg, Russia, July 2021).

**Conflict of interest**: The authors declare no conflict of interest.

## References

[1] Aiello, W., Chung, F., & Lu, L. (2000). A random graph model for power law graphs. *Experimental Mathematics*, *10*, 53–66.
[2] Aiello, W., Chung, F., & Lu, L. (2001). Random evolution in massive graphs. *Annual Symposium on Foundations of Computer Science*, 42, 510–521.
[3] Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.
[4] Barabasi, A., Albert, R., & Jeong, H. (1999). Scale-free characteristics of random networks: The topology of the world wide web. *Physica*, *A 272*, 173–187.
[5] Boginski, V., Butenko, S., & Pardalos, P. (2006). Mining market data: A network approach. *Computers and Operations Research*, *33*, 3171–3184.
[6] Boginski, V., Butenko, S., & Pardalos, P. M. (2005). Statistical analysis of financial networks. *Computational Statistics and Data Analysis*, *48*, 431–443.
[7] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, *30*, 107–117.
[8] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., … Wiener, J. (2000). Graph structure in the web. *Computer Networks*, *33*, 309–321.
[9] Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., & Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 9849–9854.
[10] Bryan, K., & Leise, T. (2006). The $25,000,000,000 eigenvector: The linear algebra behind Google. *SIAM Review*, *48*, 569–581.
[11] Campbell, J., Lo, A., & MacKinlay, C. (1997). *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.
[12] Corsi, F., Lillo, F., Pirino, D., & Trapin, L. (2018). Measuring the propagation of financial distress with granger-causality tail risk networks. *Journal of Financial Stability*, *38*, 18–36.
[13] Diks, C., & Panchenko, V. (2004, August). Modified Hiemstra-Jones test for Granger non-causality. *Computing in Economics and Finance*, *192*. Society for Computational Economics.
[14] Giatsidis, C., Thilikos, D., & Vazirgiannis, M. (2011). D-cores: Measuring collaboration of directed graphs based on degeneracy. In *IEEE 11th International Conference on Data Mining* (*ICDM*) (pp. 201–210).
[15] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*, 424–438.
[16] Granger, C. W. J. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamic and Control*, *2*, 329–352.
[17] Hecq, A., Margaritella, L., & Smeekes, S. (2021, Nov). Granger causality testing in high-dimensional VARs: A post-double-selection procedure. *Journal of Financial Econometrics*, *11*, nbab023.
[18] Hull, J. (2008). *Options, futures, and other derivatives* (7th ed.). Lebanon, Indiana, USA: Prentice Hall.
[19] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.
[20] Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*, 599–607.
[21] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
[22] Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, *5*, 269–287.

[23] Shirokikh, O., Pastukhov, G., Boginski, V., & Butenko, S. (2013). Computational study of the us stock market evolution: A rank correlation-based network model. *Computational Management Science*, *10*(2–3), 81–103.

[24] Sowers, R., & Giesecke, K. (2011, October 18). Contagion. *SIAM News*.

[25] Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, *2*, 146–160.

[26] Výrost, T., Lyócsa, Š., & Baumöhl, E. (2015). Granger causality stock market networks: Temporal proximity and preferential attachment. *Physica A: Statistical Mechanics and its Applications*, *427*, 262–276.

[27] Wu, F., Zhang, D., & Zhang, Z. (2019). Connectedness and risk spillovers in china's stock market: A sectoral analysis. *Economic Systems*, *43*(3), 100718.