Econometric Modeling of Systemic Risk: A Time Series Approach

Jalal Etesami
Bosch Center for Artificial Intelligence
jalal.etesami@de.bosch.com

Ali Habibnia
Department of Economics, Virginia Polytechnic Institute and State University
a.habibnia@lse.ac.uk

Negar Kiyavash
Dep. of Industrial & Enterprise Systems Eng. Dep. of Electrical & Computer Eng, University of Illinois Urbana-Champaign
kiyavash@illinois.edu

ABSTRACT

Financial instability can lead to financial crises due to its contagion or spillover effects to other parts of the economy. Having an accurate measure of systemic risk gives central banks and policy makers the ability to take proper actions in order to stabilize financial markets. Much work is currently being undertaken on the feasibility of identifying and measuring systemic risk. In principle, there are two main approaches for learning interlinkages between financial institutions. First approach constructs a mathematical model of financial market participant relations as a network/graph by using a combination of information extracted from financial statements such as the market value of liabilities of counterparties. Second approach learns the network via an econometric model that estimates those relations from financial data. In this paper, we develop a data-driven econometric framework that promotes an understanding of the relationship between financial institutions using a nonlinearly modified Granger-causality network. Unlike existing literature, it is not limited to a linear pairwise estimations. The method allows for nonlinearity and has predictive power over future economic activity through a time-varying network of relationships. Moreover, it can quantify the interlinkages between financial institutions. We also show how the model improves the measurement of systemic risk and explain its link to generalized variance decompositions networks. We apply the method to the daily returns of U.S. financial Institutions including banks, broker and insurance companies to identify the level of systemic risk in the financial sector and the contribution of each financial institution.

KEYWORDS

Systemic risk; Risk Measurement; Financial Linkages and Contagion; Nonlinear Granger Causality; Directed Information Graphs

ACM Reference Format:
Jalal Etesami, Ali Habibnia, and Negar Kiyavash. 2018. Econometric Modeling of Systemic Risk: A Time Series Approach. In *MiLeTS '18, August 2018, London, United Kingdom.* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Understanding the interconnection between the financial institutions especially in the context of systemic risk is of great importance. In principle, there are two main approaches to measure such interconnections between the institutions in the literature which mainly represented as a directed graph. One is based on a mathematical model of financial market participant relations derived from a combination of information extracted from financial statements like the market value of liabilities of counterparties. The other is the approach adopted herein as well which is based on statistical analysis of financial series related to the institutions of interest. Regardless, most of the existing approaches in the literature are built on pairwise comparisons or assumption of linearity. For instance, [7] propose several measures of systemic risk to capture the connections between the monthly returns of different financial institutions (hedge funds, banks, brokers, and insurance companies) based on Granger-causality tests. [7] use principle component analysis to estimate the number and importance of common factors driving the returns of financial institutions, and they use *pairwise* Granger-causality tests to identify the network of Granger-causal relations among those institutions.

Another related work is [10]. In this work, the authors propose a connectedness measure based on generalized variance decomposition (GVD) and consequently, define a weighted, directed network. The measure introduced in this work is limited to linear dynamical systems, specifically, those formulated by data-generating processes (DGPs). Beside the restrictive linearity assumption, as we will discuss later in Section 3.2, this measure also suffers from disregarding the entire network akin to pairwise analysis commonly used in the literature.

[4] focus on one particular network structure: the long-run variance decomposition network (LVDN). Similar to [10], the LVDN defines a weighted and directed graph where the weight that is associated with edge (i, j) represents the proportion of h-step-ahead forecast error variance of variable i which is accounted for by the innovations in variable j. LVDNs are characterized by the infinite vector moving average (VMA) and as such are limited to linear systems.

Connectedness measures based on correlation remain widespread. They also measure only pairwise association and are mainly studied for linear Gaussian dynamics. This makes them of limited value in financial-market contexts. Different approaches have been developed to relax these conditions. For example, equi-correlation approach in [11] uses average correlations across all pairs. The CoVaR approach of [2] measures the value-at-risk (VaR) of financial institutions conditional on other institutions experiencing financial

MiLeTS '18, August 2018, London, United Kingdom

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

distressand. The marginal expected shortfall (MES) approach of [1] measures the expected loss to each financial institution conditional on the entire set of institutions' poor performance. Although these measures rely less on linear Gaussian methods and are certainly of interest, they measure different things, and a general framework that can be used to capture the connectedness in different networks remains elusive. Introducing such measure is the main purpose of this work.

Herein, we develop a framework for learning the interaction network that allows for incorporating nonlinearities in the data and go beyond pairwise relationships among time series. We also show how the model improves the measurement of systemic risk and explain how it relates to concept of Granger-causality and variance decompositions method.

2 CAUSAL NETWORK

In order to investigating the dynamic of systemic risk, it is important to measure the causal relationship between financial institutions. In this section, we propose a statistical approach to learn such causal interconnections using Granger causality [14].

2.1 Graphical Models and Granger Causality

Researchers from different fields have developed various graphical models suitable for their application of interest to encode interconnections among variables or processes. [18] define Markov Networks, Bayesian networks (BNs), and [25] introduces Dynamic Bayesian networks (DBNs). These are three examples of such graphical models that have been used extensively in the literature. In these particular graphical models, nodes represent random variables.

Markov networks are undirected graphs that represent the conditional independence between the variables. On the other hand BNs and DBNs are directed acyclic graphs (DAGs) that encode conditional dependencies in a reduced factorization of the joint distribution.

Since the size of such graphical models depends on the time-homogeneity and the Markov order of the random processes. Therefore, in general, the graphs can grow with time. As an example, the DBN graph of a vector autoregressive (VAR) introduced by [9] with m processes each of order L requires mL nodes. As such they are not suitable for succinct visualization of relationships between the time series such as systemic risks.

In this work, similar to [28] and [23], we use directed information graphs (DIGs) to represent interconnections among the financial institutions in which each node represents a time series . Below, we formally introduce this type of graphical models. We use an information-theoretical generalization of the notion of Granger causality to determine the interconnection between time series. The basic idea in this framework was originally introduced by Wiener [33], and later formalized by Granger [14]. The idea reads as follows: "we say that X is causing Y if we are better able to predict the future of Y using all available information than if the information apart from the past of X had been used."

Despite broad philosophical viewpoint of [15], his formulation for practical implementation was done using multivariate autoregressive (MVAR) models and linear regression. This version has been widely adopted in econometrics and other disciplines. More precisely, in order to identify the influence of X_t on Y_t in a MVAR comprises of three time series $\{X, Y, Z\}$, Granger's idea is to compare the performance of two linear regressions: the first predictor is non-nested that is it predicts Y_t given $\{X^{t-1}, Y^{t-1}, Z^{t-1}\}$, where X^{t-1} denotes the time series X up to time $t-1$ and the second predictor is nested that is it predicts Y_t given $\{Y^{t-1}, Z^{t-1}\}$. Clearly, the performance of the second predictor is bounded by the first predictor. If they have the same performance, then we say X does not Granger cause Y.

We will introduce directed information (DI), an information-theoretical measure that generalized Granger causality beyond linear models. [29] used this measure to infer causality in dynamical systems. DI has been used in many applications to infer causal relationships. For example, [30] and [17] used it for analyzing neuroscience data and [12] applied for market data .

2.2 Directed Information Graphs (DIGs)

In the rest of this section, we describe how the DI can capture the interconnections in causal[1] dynamical systems (linear or non-linear) and formally define DIGs.

Consider a dynamical system comprised of three time series $\{X, Y, Z\}$. To answer whether X has influence on Y or not over time horizon $[1, T]$, we compare the average performance of two particular predictors with predictions p and q over this time horizon. The first predictor uses the history of all three time series while the second one uses the history of all processes excluding process X. On average, the performance of the predictor with less information (the second one) is upper bounded by the performance of the predictor with more information (the first one). However, when the prediction of both predictors, i.e., p and q are close over time horizon $[1, T]$, then we declare that X does not cause Y in this time horizon; otherwise, X causes Y.

In order to measure the performance of a predictor, we consider a nonnegative loss function, $\ell(p, y)$, which defines the quality of the prediction. This loss function increases as the prediction p deviates more from the true outcome y. Although there are many candidate loss functions, e.g. the squared error loss, absolute loss, etc, for the purpose of this work we consider the logarithmic loss.

Moreover, in our setting, the prediction p lies in the space of probability measures over y. More precisely, we denote the past of all processes up to time $t-1$ by $\mathcal{F}^{t-1}$ that is the σ-algebra generated by $\{X^{t-1}, Y^{t-1}, Z^{t-1}\}$, where X^{t-1} represents the time series X up to time $t-1$, and denote the past of all processes excluding process X, up to time $t-1$ by $\mathcal{F}_{-X}^{t-1}$.

The prediction of the first predictor that is non-nested at time t is given by $p_t := P(Y(t)|\mathcal{F}^{t-1})$ that is the conditional distribution of $Y(t)$ given the past of all processes and the second predictor which is nested is given by $q_t := P(Y_t|\mathcal{F}_{-X}^{t-1})$.

Given a prediction p for an outcome $y \in \mathcal{Y}$, the log loss is defined as $\ell(p, y) := -\log p(y)$. This loss function has meaningful information-theoretical interpretations. The log loss is the Shannon code length, i.e., the number of bits required to efficiently represent

[1] In causal systems, given the full past of the system, the present of the processes become independent. In other words, there are no simulations relationships between the time series.

a symbol y drawn from distribution p. Thus, it may be thought of the description length of y.

When the outcome y_t is revealed for Y_t, the two predictors incur losses $\ell(p_t, y_t)$ and $\ell(q_t, y_t)$, respectively. The reduction in the loss (description length of y_t), known as regret is defined as

$$r_t := \ell(q_t, y_t) - \ell(p_t, y_t) = \log \frac{p_t}{q_t} = \log \frac{P(Y_t = y_t|\mathcal{F}^{t-1})}{P(Y_t = y_t|\mathcal{F}^{t-1}_{-X})} \geq 0.$$

Note that the regrets are non-negative. The average regret over the time horizon $[1, T]$ given by $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[r_t]$, where the expectation is taken over the joint distribution of X, Y, and Z is called *directed information* (DI). This will be our measure of causation and its value determines the strength of influence. If this quantity is close to zero, it indicates that the past values of time series X contain no significant information that would help in predicting the future of time series Y given the history of Y and Z. This definition may be generalized to more than 3 processes as follows,

Definition 1. *Consider a network of m time series $\underline{R} := \{R_1, ..., R_m\}$. We declare R_i influences R_j over time horizon $[1, T]$, if and only if*

$$I(R_i \rightarrow R_j||\underline{R}_{-\{i,j\}}) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\log \frac{P(R_{j,t}|\mathcal{F}^{t-1})}{P(R_{j,t}|\mathcal{F}^{t-1}_{-\{i\}})}\right] > 0, \quad (1)$$

where $\underline{R}_{-\{i,j\}} := \underline{R} \setminus \{R_i, R_j\}$. $\mathcal{F}^{t-1}$ denotes the sigma algebra generated by $\underline{R}^{t-1} := \{R_1^{t-1}, ..., R_m^{t-1}\}$, and $\mathcal{F}^{t-1}_{-\{i\}}$ denotes the sigma algebra generated by $\{R_1^{t-1}, ..., R_m^{t-1}\} \setminus \{R_i^{t-1}\}$.

Definition 2. *Directed information graph (DIG) of a set of m processes $\underline{R} = \{R_1, ..., R_m\}$ is a weighted directed graph $G = (V, E, W)$, where nodes represent processes ($V = \underline{R}$) and arrow $(R_i, R_j) \in E$ denotes that R_i influences R_j with weight $I(R_i \longrightarrow R_j||\underline{R}_{-\{i,j\}})$. Consequently, $(R_i, R_j) \notin E$ if and only if its corresponding weight is zero.*

Remark 1. *Pairwise comparison has been applied in the literature to identify the causal structure of time series. The works by [7], [6], and [3] are three such examples. Pairwise comparison is not correct in general and fails to capture the true underlying network. For more details see the work by [28].*

A causal model allows a factorization of the joint distribution in some specific ways. It was shown in [28] that under a mild assumption, the joint distribution of a causal discrete-time dynamical system with m time series can be factorized as follows,

$$P_{\underline{R}} = \prod_{i=1}^{m} P_{R_i||\underline{R}_{B_i}}, \quad (2)$$

where $B_i \subseteq -\{i\} := \{1, ..., m\} \setminus \{i\}$ is the minimal[2] set of processes that causes process R_i, i.e., parent set of node i in the corresponding DIG. Such factorization of the joint distribution is called minimal generative model. In Equation (2), $P(\cdot||\cdot)$ is called causal conditioning and defined as follows $P_{R_i||\underline{R}_{B_i}} := \prod_{t=1}^{T} P_{R_{i,t}|\mathcal{F}^{t-1}_{B_i\cup\{i\}}}$, and $\mathcal{F}^{t-1}_{B_i\cup\{i\}} = \sigma\{\underline{R}^{t-1}_{B_i\cup\{i\}}\}$.

It is important to emphasize that learning the causal network using DI does not require any specific model for the system. There are different methods that can estimate (1) given i.i.d. samples of the time series such as plug-in empirical estimator, k-nearest neighbor estimator, etc. For more details related to theses estimators see the works by [16], [13], and [19].

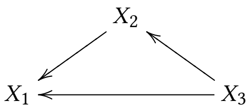


Figure 1: Corresponding DIG of the system in (3).

In general, estimating DI in (1) is a complicated task and has high sample complexity. However, knowing some side information about the system can simplify the learning task. In the following section, we describe learning the causal network of linear systems. Later in Section 4, we discuss generalization to non-linear models.

2.3 Strength of Causal Relationships

In this section, we show that the DI introduced in (1) can capture the strength of causal relationships in a network. We do so using a simple linear example and then generalize it to more general systems.

Consider a network of three time series $\vec{X}_t = (X_{1,t}, X_{2,t}, X_{3,t})^T$ with the following joint dynamics

$$\vec{X}_t = \begin{pmatrix} 0 & 0.1 & 0.3 \\ 0 & 0 & -0.2 \\ 0 & 0 & 0 \end{pmatrix} \vec{X}_{t-1} + \vec{\epsilon}_t, \quad (3)$$

where $\vec{\epsilon}_t$ denotes a vector of exogenous noises that has normal distribution with mean zero and covariance matrix $\mathbf{I}$. Figure 1 shows the corresponding DIG of this network. Note that in this particular example where the relationships are linear, the support of the coefficient matrix encodes the corresponding DIG of the network.

In order to compare the strength of causal relationships $X_2 \longrightarrow X_1$ and $X_3 \longrightarrow X_1$ over a time horizon $[1, T]$, we compare the performance of two linear predictors of $X_{1,t}$ over that time horizon. The first predictor (L_1) predicts $X_{1,t}$ using $\{X_1^{t-1}, X_3^{t-1}\}$ and the other predictor (L_2) uses $\{X_1^{t-1}, X_2^{t-1}\}$. If L_1 shows better performance compared to L_2, it implies that X_3 contains more relevant information about X_1 compared to X_2. In other words, X_3 has stronger influence on X_1 compared to X_2. To measure the performance of L_1 and L_2, we consider the mean squared errors of the prediction over the time horizon $[1, T]$.

$$L_1 : \ e_1 := \frac{1}{T}\sum_{t=1}^{T} \min_{y_t \in \mathcal{A}_t} \mathbb{E}||X_{1,t} - y_t||^2, \ \ \mathcal{A}_t := \text{span}\{X_1^{t-1}, X_3^{t-1}\},$$

$$L_2 : \ e_2 := \frac{1}{T}\sum_{t=1}^{T} \min_{z_t \in \mathcal{B}_t} \mathbb{E}||X_{1,t} - z_t||^2, \ \ \mathcal{B}_t := \text{span}\{X_1^{t-1}, X_2^{t-1}\}.$$

It is easy to show that $e_1 = 1 + 0.1^2$ and $e_2 = 1 + 0.3^2$. Since $e_1 < e_2$, we infer that X_3 has stronger influence on X_1 compared to X_2.

Analogous to the directed information graphs, we can generalize the above framework to non-linear systems. Consider a network of m time series $\underline{R} = \{R_1, ..., R_m\}$ with corresponding DIG $G = (V, E, W)$. Suppose (R_i, R_j) and (R_k, R_j) belong to E, i.e., R_i and R_k both are parents of R_j. We say R_i has stronger influence on R_j compared to R_k over a time horizon $[1, T]$ if $P(R_{j,t}|\mathcal{F}^{t-1}_{-\{k\}})$ is a better predictor for $R_{j,t}$ compared to $P(R_{j,t}|\mathcal{F}^{t-1}_{-\{i\}})$ over that time horizon. In other words, R_i has stronger influence on R_j compared

[2] Minimal in terms of its cardinality.

to R_k, if

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\log\frac{P(R_{j,t}|\mathcal{F}^{t-1}_{-\{k\}})}{P(R_{j,t}|\mathcal{F}^{t-1}_{-\{i\}})}\right] > 0.$$

The above inequality holds if and only if $I(R_i \rightarrow R_j||\underline{R}_{-\{i,j\}}) > I(R_k \rightarrow R_j||\underline{R}_{-\{k,j\}})$. Thus, the DI in (1) can quantify the causal relationships in a network. For instance, looking again at the system in (3), we obtain $I(X_2 \rightarrow X_1||X_3) = \frac{1}{2}\log(1.0104) < \frac{1}{2}\log(1.09) = I(X_3 \rightarrow X_1||X_2)$.

3 DIG OF LINEAR MODELS

Herein, we study the causal network of linear systems. Consider a set of m stationary time series, and for simplicity assume they have zero mean, such that their relationships are captured by the following model:

$$\vec{R}_t = \sum_{k=1}^{p}\mathbf{A}_k\vec{R}_{t-k} + \vec{\epsilon}_t, \tag{4}$$

where $\vec{R}_t = (R_{1,t}, ..., R_{m,t})^T$, and $\mathbf{A}_k$s are $m{\times}m$ matrices. Moreover, we assume that the exogenous noises, i.e., $\epsilon_{i,t}$s are independent and also independent from $\{R_{j,t}\}$. For simplicity, we assume that the $\{\epsilon_{i,t}\}$ have mean zero. For the model in (4), it was shown in [12] that $I(R_i \rightarrow R_j||\underline{R}_{-\{i,j\}}) > 0$, if and only if $\sum_{k=1}^{p}|(\mathbf{A}_k)_{j,i}| > 0$, where $(\mathbf{A}_k)_{j,i}$ is the (j,i)th entry of matrix $\mathbf{A}_k$. Thus, to learn the corresponding causal network (DIG) of this model, instead of estimating the DIs in (1), we can check whether the corresponding coefficients are zero or not. To do so, we use the Bayesian information criterion (BIC) as the model-selection criterion to learn the parameter p as descibed by [31], and use F-tests to check the null hypotheses that the coefficients are zero as descibed by [22].

[24] use Wiener filtering as another alternative approach to estimate the coefficients and consequently learn the DIG. The idea of this approach is to find the coefficients by solving the following optimization problem,

$$\{\hat{\mathbf{A}}_1, ..., \hat{\mathbf{A}}_p\} = \arg\min_{\mathbf{B}_1,...,\mathbf{B}_p}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}||\vec{R}_t - \sum_{k=1}^{p}\mathbf{B}_k\vec{R}_{t-k}||^2\right].$$

This leads to a set of Yule-Walker equations that can be solved efficiently by Levinson-Durbin algorithm introduced by [26].

3.1 DIG of GARCH models

The relationship between the coefficients of the linear model and the corresponding DIG can easily be extended to the financial data in which the variance of $\{\epsilon_{i,t}\}_{t=1}^{T}$ are no longer independent of $\{R_{i,t}\}$ but due to the heteroskedasticity, they are $\mathcal{F}_i^{t-1}$-measurable. More precisely, in financial data, the returns are modeled by GARCH that is given by

$$\begin{aligned} R_{i,t}|\mathcal{F}^{t-1} &\sim \mathcal{N}(\mu_{i,t}, \sigma^2_{i,t}),\\ \sigma^2_{i,t} &= \alpha_0 + \sum_{k=1}^{q}\alpha_k(R_{i,t-k} - \mu_{i,t})^2 + \sum_{l=1}^{s}\beta_l\sigma^2_{i,t-l}, \end{aligned} \tag{5}$$

where α_ks and β_ls are nonnegative constants.

PROPOSITION 1. *Consider a network of time series whose dynamic is given by (5). In this case, there is no arrow from R_j to R_i in its corresponding DIG, i.e., R_j does not cause R_i if and only if*

$$\mathbb{E}[R_{i,t}|\mathcal{F}^{t-1}] = \mathbb{E}[R_{i,t}|\mathcal{F}^{t-1}_{-\{j\}}],\ \forall t. \tag{6}$$

Multivariate GARCH models are a a generalization of (5) in which the variance of $e_{i,t}$ is $\mathcal{F}^{t-1}$-measurable. In this case, not only $\mu_{i,t}$ but also $\sigma^2_{i,t}$ capture the interactions between the returns. More precisely, in multivariate GARCH, we have

$$\begin{aligned} &\vec{R}_t|\mathcal{F}^{t-1} \sim \mathcal{N}(\vec{\mu}_t, \mathbf{H}_t),\\ &vech[\mathbf{H}_t] = \Omega_0 + \sum_{k=1}^{q}\Omega_k vech[\vec{\epsilon}_{t-k}\vec{\epsilon}^T_{t-k}] + \sum_{l=1}^{p}\Gamma_l vech[\mathbf{H}_{t-l}], \end{aligned}$$

where $\vec{\mu}_t$ is an $m{\times}1$ array, $\mathbf{H}_t$ is an $m{\times}m$ symmetric positive definite and $\mathcal{F}^{t-1}$-measurable matrix, and $\vec{\epsilon}_t = \vec{R}_t - \vec{\mu}_t$. Note that $vech$ denotes the vector-half operator, which stacks the lower triangular elements of an $m \times m$ matrix as an $(m(m+1)/2) \times 1$ array.

PROPOSITION 2. *Consider a network of time series whose dynamic is captured by a multivariate GARCH model. In this case, there is no arrow from R_j to R_i in its corresponding DIG, i.e., R_j does not influence R_i if and only if both the condition in Proposition 1 and the following condition hold*

$$\mathbb{E}[(R_{i,t} - \mu_{i,t})^2|\mathcal{F}^{t-1}] = \mathbb{E}[(R_{i,t} - \mu_{i,t})^2|\mathcal{F}^{t-1}_{-\{j\}}],\ \forall t. \tag{7}$$

REMARK 2. *Recall that as we mentioned in Remark 1, the pairwise Granger-causality calculation, in general, fails to identify the true causal network. It was proposed by [7] that the returns of the ith institution linearly depend on the past returns of the jth institution, when $\mathbb{E}[R_{i,t}|\mathcal{F}^{t-1}] = \mathbb{E}[R_{i,t}|R_{j,t-1}, R_{i,t-1}, \{R_{j,\tau} - \mu_{j,\tau}\}_{\tau=-\infty}^{t-2}, \{R_{i,\tau} - \mu_{i,\tau}\}_{\tau=-\infty}^{t-2}]$. This test is obtained based on pairwise Granger-causality calculation and does not consider non-linear causation through the variance of $\{\epsilon_i\}$. For instance, if the returns of two institutions R_j and R_k cause the returns of the ith institution, then the above equality does not hold, because R_k cannot be removed from the conditioning.*

3.2 DIG of Moving-Average (MA) Models

[27] show that the model in (4) can be represented as an infinite moving average (MA) or data-generating process (GDP), as long as $\vec{R}(t)$ is covariance-stationary, i.e., all the roots of $|\mathbf{I} - \sum_{k=1}^{p}\mathbf{A}_k z^k|$ fall outside the unit circle : $\vec{R}_t = \sum_{k=0}^{\infty}\mathbf{W}_k\vec{\epsilon}_{t-k}$, where $\mathbf{W}_k = 0$ for $k < 0$, $\mathbf{W}_0 = \mathbf{I}$, and $\mathbf{W}_k = \sum_{l=1}^{p}\mathbf{W}_{k-l}\mathbf{A}_l$. In this representation, $\{\epsilon_i\}$s are called shocks and if they are independent, they are also called orthogonal.

In this section, we study the causal structure of a MA model of finite order p. Consider a moving average model with orthogonal shocks given by

$$\vec{R}_t = \sum_{k=0}^{p}\mathbf{W}_k\vec{\epsilon}_{t-k}, \tag{8}$$

where $\mathbf{W}_i$s are $m \times m$ matrices such that $\mathbf{W}_0$ is non-singular with nonzero diagonals and without loss of generality, we can assume that $diag(\mathbf{W}_0)$ is the identity matrix. Equation (8) can be written as $\vec{R}_t = \mathbf{W}_0\vec{\epsilon}_t + \mathcal{P}(L)\vec{\epsilon}_{t-1}$, where $\mathcal{P}(L) := \sum_{k=1}^{p} \mathbf{W}_k L^{k-1}$. Subsequently, we have

$$\mathbf{W}_0^{-1}\vec{R}_t = \vec{\epsilon}_t + \sum_{k=1}^{\infty}(-1)^{k-1}\left(\mathbf{W}_0^{-1}\mathcal{P}(L)\right)^k \mathbf{W}_0^{-1}\vec{R}_{t-k}. \tag{9}$$

This representation is equivalent to an infinite AR model. Hence, using the result by [12], we can conclude the following corollary.

COROLLARY 1. *Consider a MA model described by (8) with orthogonal shocks such that $\mathbf{W}_0$ is non-singular and diagonal. In this case, R_j does not influence R_i if and only if the corresponding coefficients of $\{R_{j,t-k}\}_{k>0}$ in R_i's equation are zero.*

We studied the DIG of a MA model with orthogonal shocks. However, the shocks are rarely orthogonal in practice. To identify the causal structure of such systems, we can apply the whitening transformation to transform the shocks into a set of uncorrelated variables. More precisely, suppose $\mathbb{E}[\vec{\epsilon}_t\vec{\epsilon}_t^T] = \Sigma$, where the Cholesky decomposition of Σ is $\mathbf{V}\mathbf{V}^T$. Hence, $\mathbf{V}^{-1}\vec{\epsilon}_t$ is a vector of uncorrelated shocks. Using this fact, we can transform (8) with correlated shocks into

$$\vec{R}_t = \sum_{k=0}^{p} \tilde{\mathbf{W}}_k \vec{\tilde{\epsilon}}_{t-k}, \tag{10}$$

with uncorrelated shocks, where $\vec{\tilde{\epsilon}}_t := \mathbf{V}^{-1}\vec{\epsilon}_t$ and $\tilde{\mathbf{W}}_k := \mathbf{W}_k\mathbf{V}$.

REMARK 3. *[10] applied the generalized variance decomposition (GVD) method to identify the population connectedness or in another word the causal structure of a MA model with correlated shocks. Using this method, they monitor and characterize the network of major U.S. financial institutions during 2007-2008 financial crisis. In this method, the weight of R_j's influence on R_i in (8) was defined to be proportional to $d_{i,j} = \sum_{k=0}^{p}\left((\boldsymbol{W}_k\Sigma)_{i,j}\right)^2$, where $(\boldsymbol{A})_{i,j}$ denotes the (i,j)-th entry of matrix $\boldsymbol{A}$. Recall that $\mathbb{E}[\vec{\epsilon}_t\vec{\epsilon}_t^T] = \Sigma$. This method also seems to suffer from disregarding the entire network akin to pairwise analysis commonly used in traditional application of the Granger-causality.*

4 DIG OF NON-LINEAR MODELS

DIG as defined in Definition 2 does not require any assumptions on the underlying model. But, similar to the analysis by [6], side information about the model class can simplify computation of (1). For instance, let us assume that $\underline{R}$ is a first-order Markov chain with transition probabilities: $P(\underline{Y}_t|\underline{R}^{t-1}) = P(\underline{R}_t|\underline{R}_{t-1})$. In this setup, $I(R_i \to R_j||\underline{R}_{-\{i,j\}}) = 0$ if and only if $P(R_{j,t}|\underline{R}_{t-1}) = P(R_{j,t}|\underline{R}_{-\{i\},t-1}), \forall t$. Recall that $\underline{R}_{-\{i\},t-1}$ denotes $\{R_{1,t-1},...,R_{m,t-1}\} \setminus \{R_{i,t-1}\}$. Furthermore, suppose that the transition probabilities are represented through a logistic function similar to the work by [6]. More specifically, for any subset of processes $\mathcal{S} := \{R_{i_1},...,R_{i_s}\} \subseteq \underline{R}$, we have

$$P(R_{j,t}|R_{i_1,t-1},...,R_{i_s,t-1}) := \frac{\exp(\vec{\alpha}_{\mathcal{S}}^T\vec{U}_{\mathcal{S}})}{1+\exp(\vec{\alpha}_{\mathcal{S}}^T\vec{U}_{\mathcal{S}})},$$

where $\vec{U}_{\mathcal{S}}^T := \bigotimes_{i\in\mathcal{S}}(1,R_{i,t-1}) = (1,R_{i_1,t-1}) \otimes (1,R_{i_2,t-1}) \otimes \cdots \otimes (1,R_{i_s,t-1})$, $\otimes$ denotes the Kronecker product, and $\vec{\alpha}_{\mathcal{S}}$ is a vector of dimension $2^s \times 1$. Under these assumptions, the causal discovery in the network reduces to the following statement: R_i does not influence R_j if and only if all the terms of $\vec{U}_{\underline{R}}$ depending on R_i are equal to zero. More precisely:

$$\vec{U}_{\underline{R}} = \vec{U}_{\underline{R}_{-\{i\}}} \otimes (1,R_{i,t-1}) = (\vec{U}_{\underline{R}_{-\{i\}}}, \vec{U}_{\underline{R}_{-\{i\}}}R_{i,t-1}).$$

Let $\vec{\alpha}_{\underline{R}}^T = (\vec{\alpha}_1^T, \vec{\alpha}_2^T)$, where $\vec{\alpha}_1$ and $\vec{\alpha}_2$ are the vectors of coefficients corresponding to $\vec{U}_{\underline{R}_{-\{i\}}}$ and $\vec{U}_{\underline{R}_{-\{i\}}}R_{i,t-1}$, respectively. Then $R_i \nrightarrow R_j$ if and only if $\vec{\alpha}_2 = 0$.

Multiple chain Markov switching model (MCMS)-VAR of [5] are a family of non-linear models, in which the relationship between time series $\underline{Y}_t$ is given by

$$Y_{i,t} = \mu_i(S_{i,t}) + \sum_{k=1}^{p}\sum_{j=1}^{m}(\mathbf{B}_k(S_{i,t}))_{i,j}Y_{j,t-k} + \epsilon_{i,t}, \tag{11}$$

and $\vec{\epsilon}_t := (\epsilon_{1,t},...,\epsilon_{m,t}) \sim \mathcal{N}(0,\Sigma(\vec{S}_t))$, where the mean, the lag matrices, and the covariance matrix of the error terms all depend on a latent random vector $\vec{S}_t$ known as the state of the system. $S_{i,t}$ represents the state variable associated with $Y_{i,t}$ that can take values from a finite set $\mathcal{S}$. The random sequence $\{\vec{S}_t\}$ is assumed to be a time-homogenous first-order Markov process with one-step ahead transition probability $P(\vec{S}_t|\underline{S}^{t-1},\underline{Y}^{t-1}) = P(\vec{S}_t|\underline{S}_{t-1})$. Furthermore, given the past of the states, the presents are independent, i.e., $P(\vec{S}_t|\underline{S}_{t-1}) = \prod_j P(S_{j,t}|\underline{S}_{t-1})$. Next result stresses a set of conditions under which by observing only the time series $\underline{Y}_t$, we are able to identify the causal relationships between the processes.

PROPOSITION 3. *Consider a MCMS-VAR in which $\Sigma(\vec{S}_t)$ is diagonal for all $\vec{S}_t$. In this case, $I(Y_j \to Y_i||\underline{Y}_{-\{i,j\}}) = 0$ if*

- $(\boldsymbol{B}_k(s_{i,t}))_{i,j} = 0$ *for all realizations* $s_{i,t}$,
- $(\Sigma(\vec{S}_t))_{i,i} = (\Sigma(S_{i,t}))_{i,i}$,
- $P(S_{k,t}|\underline{S}^{t-1},\underline{S}_{-\{k\},t}) = P(S_{k,t}|S_{k,t-1})$ *for every* k.

Note that the conditions introduced in this proposition are slightly different from the ones of [5]. But notice that [5] study the causal relationships between the time series given the state variables. Assuming the state variables are given is not realistic as they are hidden.

5 EXPERIMENTAL RESULT

Herein, we apply the proposed methods to identify and monitor the evolution of connectedness among major financial institutions during 2006-2016. We obtained the data for individual banks, broker/dealers, and insurers, from which we selected the daily returns of all companies listed in Table 1.

5.1 Non-linearity Test

We applied a non-linearity test on the data to determine whether the underlying structure within the recorded data is linear or nonlinear. The non-linearity test applied is based on non-linear principle component analysis (PCA) of [20]. This test works as follows, the range of recorded data is divided into smaller disjunct regions; and accuracy bounds are determined for the sum of the discarded eigenvalues of each region. If this sum is within the accuracy bounds for each region, the process is assumed to be linear. Conversely, if at least one of these sums is outside, the process is assumed to be nonlinear.

Banks			
1	FNMA US	16	BNS US
2	AXP US	17	STI US
3	FMCC US	18	C US
4	BAC US	19	MS US
5	WFC UN	20	SLM US
6	JPM US	21	BBT US
7	DB US	22	USB US
8	NTRS US	23	TD US
9	RY US	24	HSBC US
10	PNC US	25	BCS US
11	STT US	26	GS US
12	COF US	27	MS US
13	BMO US	28	CS US
14	CM US		
15	RF UN		

Insurances			
1	MET US	16	PFG US
2	ANTM US	17	LNC US
3	AET US	18	AON US
4	CNA US	19	HUM US
5	XL US	20	MMC US
6	SLF US	21	HIG US
7	MFC US	22	CI US
8	GNW US	23	ALL US
9	PRU US	24	BRK/B US
10	AIG US	25	CPYYY US
11	PGR US	26	AHL US
12	CB US		
13	BRK/A US		
14	UNH US		
15	AFL US		

Brokers			
1	MS US	16	WDR US
2	GS US	17	EV US
3	BEN US	18	ITG UN
4	MORN US	19	JNS US
5	LAZ US	20	SCHW US
6	ICE US	21	ETFC US
7	AINV US	22	AMTD US
8	SEIC US		
9	FII US		
10	RDN US		
11	TROW US		
12	AMP US		
13	GHL US		
14	AMG US		
15	RJF US		

Table 1: . List of companies in our experiment.

More precisely, the second step in this test requires computation of the correlation matrix for each of the disjunct regions. Since the elements of this matrix are obtained using a finite dataset, applying t-distribution and χ^2-distribution establish confidence bounds for both estimated mean and variance, respectively. Subsequently, these confidence bounds can be utilized to determine thresholds for each element in the correlation matrix. Using these thresholds, the test calculates maximum and minimum eigenvalues relating to the discarded score variables, which in turn allows the determination of both a minimum and a maximum accuracy bound for the variance of the prediction error of the PCA model. This is because the variance of the prediction error is equal to the sum of the discarded eigenvalues. If this sum lies inside the accuracy bounds for each disjunct region, a linear PCA model is then appropriate over the entire region. Alternatively, if at least one of these sums is outside the accuracy bounds, the error variance of the PCA model residuals then differs significantly for this disjunct region and hence, a nonlinear model is required.

We divided the operating region into 3 disjunct regions. The accuracy bounds for each disjuct region and also sum of the discarded eigenvalues were computed. These bounds were based on thresholds for each element of the correlation matrix corresponding to confidence level of 95%. Note that the processes were normalized with respect to the mean and variance of the regions for which the accuracy bounds were computed. Figure 2 shows the accuracy bounds and the sum of the discarded eigenvalues. As figures 2-(a) and 2-(b) illustrate, the recorded financial data is nonlinear.

5.2 Estimating the DIs

As we mentioned earlier, there are different methods that can be used to estimate (1) given i.i.d. samples of the time series such as plug-in empirical estimator and k-nearest neighbor estimator. For our experimental results, we used k-nearest method since it shows relatively better performance compared to the other non-parametric estimators. To do so, we used the fact that

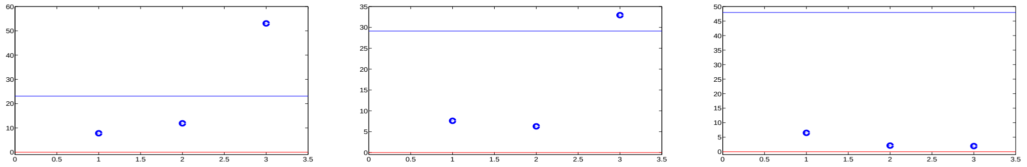
$$I(R_i \rightarrow R_j||\underline{R}_{-\{i,j\}}) = \frac{1}{T}\sum_{t=1}^{T} I(R_{j,t}; R_i^{t-1}|\underline{R}_{-\{i,j\}}^{t-1}, R_j^{t-1}),$$

where $I(X;Y|Z)$ denotes conditional mutual information between X and Y given Z. For more details see the book by [8]. Then, we estimated each of the above conditional mutual information using k-nearest method of [32]. Below, we describe the steps of k-nearest method to estimate $I(X;Y|Z)$.

Suppose that $N+M$ i.i.d. realizations $\{\mathbf{X}_1, ..., \mathbf{X}_{N+M}\}$ are available from $P_{X,Y,Z}$, where $\mathbf{X}_i$ denotes the ith realization of (X,Y,Z). The data sample is randomly divided into two subsets S_1 and S_2 of N and M points, respectively. In the first stage, an k-nearest density estimator $\widehat{P}_{X,Y,Z}$ at the N points of S_1 is estimated using the M realizations of S_2 as follows: Let $d(\mathbf{x},\mathbf{y})$ denote the Euclidean distance between points $\mathbf{x}$ and $\mathbf{y}$ and $d_k(\mathbf{x})$ denotes the Euclidean distance between a point $\mathbf{x}$ and its k-th nearest neighbor among S_2. The k-nearest region is $S_k(\mathbf{x}) := \{\mathbf{y} : d(\mathbf{x},\mathbf{y}) \leq d_k(\mathbf{x})\}$ and the volume of this region is $V_k(\mathbf{x}) := \int_{S_k(\mathbf{x})} dn$. The standard k-nearest density estimator of [32] is defined as $\widehat{P}_{X,Y,Z}(\mathbf{x}) := (k-1)/MV_k(\mathbf{x})$. Similarly, we obtain k-nearest density estimators $\widehat{P}_{X,Z}, \widehat{P}_{Y,Z}$, and $\widehat{P}_Z$. Subsequently, the N samples of S_1 is used to approximate the conditional mutual information: $\widehat{I}(X;Y|Z) := \frac{1}{N}\sum_{i\in S_1} \log \widehat{P}_{X,Y,Z}(\mathbf{X}_i) + \log \widehat{P}_Z(\mathbf{X}_i) - \log \widehat{P}_{X,Z}(\mathbf{X}_i) - \log \widehat{P}_{Y,Z}(\mathbf{X}_i)$. For more details corresponding this estimator including its bias, variance, and confidence, please see the works by[32] and [21].

5.3 DIG of the Financial Market

We learned the DIG of the financial institutions by estimating the directed information quantities in (1). To do so, we divided the data into four sectors each of length almost 36 months, 2006-2008, 2009-2011, 2011-2013, and 2013-2016. We assumed that the DIG of the network did not change over each of these time periods. Furthermore, the data collected per working day are assumed to be i.i.d.. Hence, in this experiment the length of each time series was



(a) Region 1.

(b) Region 2.

(c) Region 3.

Figure 2: Benchmarking of the residual variances against accuracy bounds of each disjunct region.

almost 36 and for each time instance we had nearly 19 independent realizations.

As we discussed in Section 2.2, in order to identify the influence from node i on node j, we need to estimate $I(R_i \to R_j || \underline{R}_{-\{i,j\}})$, which in this experiment, required estimating a joint distribution of dimension 76. In general, without any knowledge about the underlying distribution, estimating such object requires a large amount of independent samples. Unfortunately, in this experiment, we had limited number of independent samples. Thus, we reduced the dimension by instead of conditioning on $\underline{R}_{-\{i,j\}}$ that is a set of size 74, we conditioned on a smaller subset $\underline{K}_{i,j}$ of $\underline{R}_{-\{i,j\}}$ with size 7. This set contained only those institutions with highest correlation with R_j. In another words, we ordered the institutions in $\underline{R}_{-\{i,j\}}$ based on their correlation value with R_j, and picked the first 7 of them. Afterward, we estimated $I(R_i \to R_j || \underline{K}_{i,j})$ to identify the connection between R_i and R_j.

Figures 3 and 4 show the resulting graphs. Note that the type of institution causing the relationship is indicated by color: green for brokers, red for insurers, and blue for banks.

In order to compare our results with other methods in the literature, we also learned the causal network of these financial institutions by assuming linear relationships between the institutions and applying linear regression. Similarly, we reduced the dimension of the regressions by bounding the number of incoming arrows of each node to be a subset of size 18. More precisely, we picked 18 most correlated institutions with node i, let say $\{R_{j_1}, ..., R_{j_{18}}\}$ and obtained the parents of i by solving $\min_{a_j} \sum_t |R_{i,t} - \sum_{k=1}^{18} a_k R_{j_k,t-1}|^2$. The resulting graphs are depicted in Figures 5 and 6.

From these networks, we constructed the following network-based measures of systemic risk.

[7] introduced the degree of Granger causality (DGC) as a measure of the risk of a system event. DGC is defined as the fraction of statistically significant Granger causality relationships among all pairs of financial institutions Table 2 presents the DGC values and total number of connections of the DIGs and the networks obtain by linear regression.

DIGs		
2006-2008	0.1225	698
2009-2011	0.1114	635
2011-2013	0.1065	607
2013-2016	0.0930	530

Linear Models		
2006-2008	0.1453	828
2009-2011	0.1288	734
2011-2013	0.1174	669
2013-2016	0.1216	693

Table 2: . DGC values and total number of connections.

Tables 3 and 4 represent the average number of connections between the sectors e.g., 0.1719 fraction of connections are from Banks to Insurances during 2006-2008 in the DIG.

6 CONCLUSION

In this work, we developed a data-driven econometric framework to understand the relationship between financial institutions using a non-linearly modified Granger-causality. Unlike existing literature, it is not focused on a linear pairwise estimation. The proposed method allows for nonlinearity and it does not suffer from pairwise comparison to identify the causal relationships between financial institutions. We also show how the model improve the measurement of systemic risk and explain the link between Granger-causality and variance decomposition. We apply the model to the monthly returns of U.S. financial Institutions including banks, broker, and insurance companies to identify the level of systemic risk in the financial sector and the contribution of each financial institution.

REFERENCES

[1] Viral V Acharya, Lasse Heje Pedersen, Thomas Philippon, and Matthew P Richardson. 2010. Measuring systemic risk. (2010).

[2] Tobias Adrian and M CoVaR Brunnermeier. 2008. Staff Report No348. *Federal Reserve Bank of New York* (2008).

[3] Franklin Allen, Ana Babus, and Elena Carletti. 2010. *Financial connections and systemic risk*. Technical Report. National Bureau of Economic Research.

[4] Matteo Barigozzi and Marc Hallin. 2016. A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (2016).

[5] Monica Billio and Silvio Di Sanzo. 2015. Granger-causality in Markov switching models. *Journal of Applied Statistics* 42, 5 (2015), 956–966.

[6] Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. 2010. Measuring systemic risk in the finance and insurance sectors. (2010).

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

[8] Thomas M Cover and Joy A Thomas. 2012. *Elements of information theory*. John Wiley & Sons.

[9] Rainer Dahlhaus and Michael Eichler. 2003. Causality and graphical models in time series analysis. *Oxford Statistical Science Series* (2003), 115–137.

[10] Francis X Diebold and Kamil Yılmaz. 2014. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics* 182, 1 (2014), 119–134.

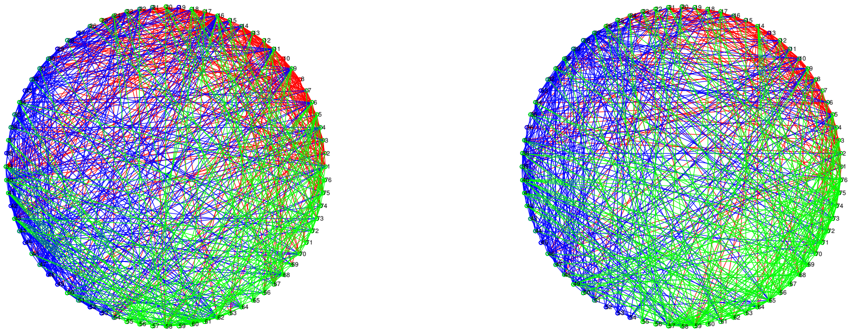
[11] Robert Engle and Bryan Kelly. 2012. Dynamic equicorrelation. *Journal of Business & Economic Statistics* 30, 2 (2012), 212–228.

[12] Jalal Etesami and Negar Kiyavash. 2014. Directed Information Graphs: A generalization of Linear Dynamical Graphs. In *American Control Conference (ACC), 2014*. IEEE, 2563–2568.

[13] Stefan Frenzel and Bernd Pompe. 2007. Partial mutual information for coupling analysis of multivariate time series. *Physical review letters* 99, 20 (2007), 204101.

[14] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.

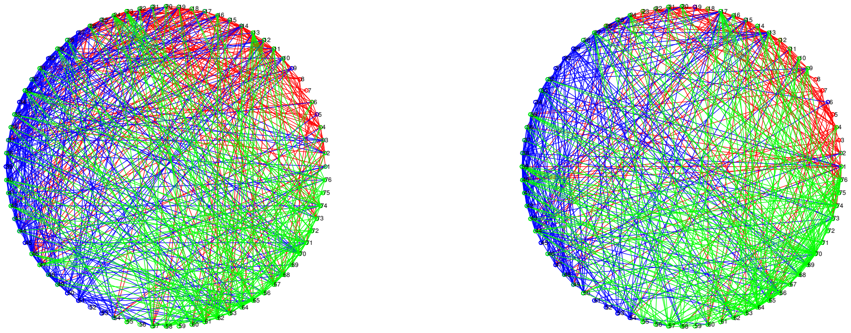
[15] Clive William John Granger. 1963. Economic processes involving feedback. *Information and control* 6, 1 (1963), 28–48.



(a) January 2006 to December 2008

(b) January 2009 to December 2011

Figure 3: Recovered DIG of the daily returns of the financial companies in Table 1. The type of institution causing the relationship is indicated by color: green for brokers, red for insurers, and blue for banks.



(a) January 2011 to December 2013

(b) January 2013 to June 2016

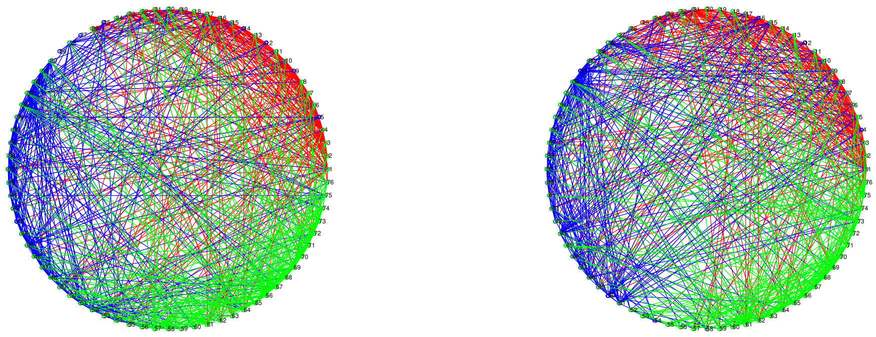
Figure 4: Recovered DIG of the daily returns of the financial companies in Table 1. The type of institution causing the relationship is indicated by color: green for brokers, red for insurers, and blue for banks.

2006-2008				**2009-2011**			**2011-2013**			**2013-2016**		
	Ins.	Ba.	Br.	Ins.	Ba.	Br.	Ins.	Ba.	Br.	Ins.	Ba.	Br.
Insurance	.1390	.1719	.1074	.1291	.1575	.1213	.1054	.1301	.1104	.1075	.1151	.1340
Bank	.1361	.1332	.0702	.0866	.1402	.1039	.1417	.1631	.1021	.0774	.1830	.1302
Broker	.0774	.1017	.0630	.0740	.929	.0945	.0906	.0873	.0692	.0774	.0774	.0981

Table 3: . Average number of connections between different sectors in the DIGs.

2006-2008				**2009-2011**			**2011-2013**			**2013-2016**		
	Ins.	Ba.	Br.	Ins.	Ba.	Br.	Ins.	Ba.	Br.	Ins.	Ba.	Br.
Insurance	.1896	.0688	.0737	.1785	.1076	.0640	.2033	.0792	.1016	.2107	.0851	.0678
Bank	.0906	.1872	.0809	.1322	.1431	.0899	.1136	.1226	.1001	.1010	.1515	.1053
Broker	.0857	.1063	.1171	.0790	.0708	.1349	.1226	.0673	.0897	.1082	.0895	.0808

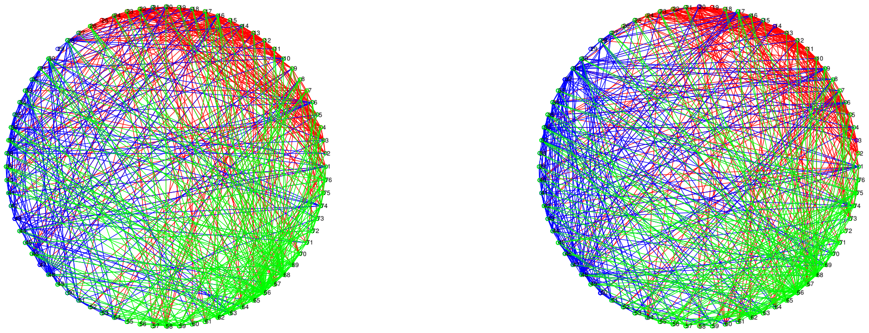
Table 4: . Average number of connections between different sectors in the networks obtained using regression.



(a) January 2006 to December 2008

(b) January 2009 to December 2011

Figure 5: Recovered network of the daily returns of the financial companies in Table 1 using linear regression. The type of institution causing the relationship is indicated by color: green for brokers, red for insurers, and blue for banks.



(a) January 2011 to December 2013

(b) January 2013 to June 2016

Figure 6: Recovered network of the daily returns of the financial companies in Table 1 using linear regression. The type of institution causing the relationship is indicated by color: green for brokers, red for insurers, and blue for banks.

[16] Jiantao Jiao, Haim H Permuter, Lei Zhao, Young-Han Kim, and Tsachy Weissman. 2013. Universal estimation of directed information. *Information Theory, IEEE Transactions on* 59, 10 (2013), 6220–6242.

[17] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown. 2011. A Granger causality measure for point process models of ensemble neural spiking activity. *PLoS computational biology* 7, 3 (2011), e1001110.

[18] Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques.* MIT press.

[19] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69, 6 (2004), 066138.

[20] Uwe Kruger, Junping Zhang, and Lei Xie. 2008. Developments and applications of nonlinear principal component analysis-a review. In *Principal manifolds for data visualization and dimension reduction.* Springer, 1–43.

[21] Don O Loftsgaarden, Charles P Quesenberry, et al. 1965. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics* 36, 3 (1965), 1049–1051.

[22] Richard G Lomax and Debbie L Hahs-Vaughn. 2013. *Statistical concepts: A second course.* Routledge.

[23] J Massey. 1990. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90).* Citeseer, 303–305.

[24] Donatello Materassi and Murti V Salapaka. 2012. On the problem of reconstructing an unknown topology via locality properties of the Wiener filter. *Automatic Control, IEEE Transactions on* 57, 7 (2012), 1765–1777.

[25] Kevin Patrick Murphy. 2002. *Dynamic bayesian networks: representation, inference and learning.* Ph.D. Dissertation. University of California, Berkeley.

[26] Bruce Ronald Musicus. 1988. *Levinson and fast Choleski algorithms for Toeplitz and almost Toeplitz matrices.* Citeseer.

[27] H Hashem Pesaran and Yongcheol Shin. 1998. Generalized impulse response analysis in linear multivariate models. *Economics letters* 58, 1 (1998), 17–29.

[28] Christopher Quinn, Negar Kiyavash, and Todd P Coleman. 2015. Directed information graphs. *Transactions on Information Theory* 61, 12 (2015), 6887–6909.

[29] Christopher J Quinn, Todd P Cole, and Negar Kiyavash. 2011. A generalized prediction framework for Granger causality. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on.* IEEE, 906–911.

[30] Christopher J Quinn, Todd P Coleman, Negar Kiyavash, and Nicholas G Hatsopoulos. 2011. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of computational neuroscience* 30, 1 (2011), 17–44.

[31] Gideon Schwarz et al. 1978. Estimating the dimension of a model. *The annals of statistics* 6, 2 (1978), 461–464.

[32] Kumar Sricharan, Raviv Raich, and Alfred O Hero. 2011. k-nearest neighbor estimation of entropies with confidence. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on.* IEEE, 1205–1209.

[33] Norbert Wiener. 1956. The theory of prediction. *Modern mathematics for engineers* 1 (1956), 125–139.