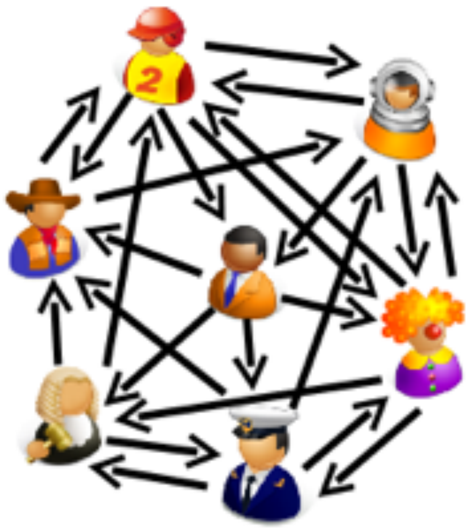
Econometric Modeling of Systemic Risk: Going Beyond Pairwise Comparison and Allowing for Nonlinearity

Negar Kiyavash
University of Illinois at Urbana-Champaign
joint work with
Jalal Etesami (UIUC), Ali Habibnia (London School of Economics)

2016

Goal



- **Goal:** to find casual influence structure
  - Graph:
    - nodes: random processes
    - directed edges: direction of causal influence
- **Observations:** all (or a subset) of time series of the realizations of processes

- Causal not necessary a case-effect relationship, rather a directional influence

Network inference via timing

- Passive learning: Observation only
- Intervention is not possible
- Requires a natural notion of time axis
- Advantages:
    - Going beyond correlation-only relations: causal inference
    - Existence of neighborhood and local operation concept: savings in complexity

Passive learning: Observation only

- Intimately related to time series analysis
- Econometrics and computational finance
- Financial networks:
  - risk of banking and insurance systems to determine of sovereign risk
  - influence structure of stock markets: inter or intra market

- Granger causality (by Nobel Laureate)
  - Limited in scope:
    - pairwise analysis
    - linear models

Passive learning: Observation only

- Interlinkages between financial institutions:

  1. Construct a **mathematical model** using a combination of information extracted from financial statements like the market value of liabilities of counterparties.

  2. Statistical analysis of financial series like **Granger-causality** network.

- Most of the existing approaches:
  - Pairwise comparison.
  - Assuming linear relationship between the time series.

- In this work, we develop a method that allows for **nonlinearity** of the data and does **not depend on pairwise** relationships among time series.

Granger causality

Clive Granger (1969):
"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."

Granger's Formulation: AR Model

$$Y_t = \sum_{\tau>0} a_\tau Y_{t-\tau} + b_\tau X_{t-\tau} + c_\tau Z_{t-\tau} + E_t$$

$$Y_t = \sum_{\tau>0} \tilde{a}_\tau Y_{t-\tau} + \tilde{c}_\tau Z_{t-\tau} + \tilde{E}_t$$

Past of **X** does not help prediction, if:

$$var(E_t) = var(\tilde{E}_t), \ \mathbf{X} \not\rightarrow \mathbf{Y}$$

$$G_{X\rightarrow Y} = \log \frac{var(\tilde{E})}{var(E)}$$

Revisiting Granger's viewpoint:

X causes **Y** if the future of **Y** given the past of both **Y** and **X** better predicts the future of **Y** given its past alone.

- Consider sequential predictors: $b_i = g_i(y^{i-1}, x^i), \qquad b_i = \tilde{g}_i(y^{i-1})$
- Outcome y is revealed, loss incurred: $l(y, b)$
- Reduction in loss (regret): $\frac{1}{n}\sum_{i=1}^{n} l(y_i, \tilde{b}_i) - l(y_i, b_i)$
- Non-negative; zero iff future of **Y** is independent of past and present **X** given past of **Y**
- Applicable to any modality, e.g. point process

Case:

Logarithmic loss: $l(y, b) = -\log b(y)$

Predictor: beliefs, then optimal predictors: conditional densities

Expected regret is directed information: $\frac{1}{n} I(X^n \to Y^n)$

Appears in information theory & control theory in context of communication with feedback

Revisiting Granger's viewpoint:

X causes **Y** if the future of **Y** given the past of both **Y** and **X** better predicts the future of **Y** given its past alone.

Directed Information:

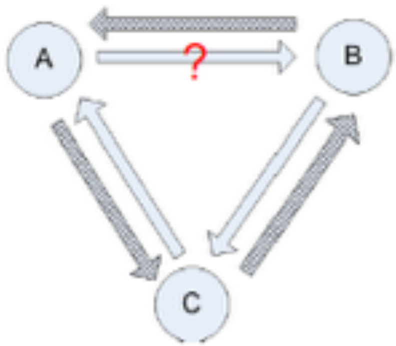
Logarithmic loss and optimal belief predictors: conditional densities: $l(y,b) = -\log b(y)$

Expected regret is directed information: $\frac{1}{n} I(X^n \to Y^n)$

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} l(Y_i,\tilde{B}_i) - l(Y_i,B_i)\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} \log \frac{P(Y_i|X^i,Y^{i-1})}{P(Y_i|Y^{i-1})}\right]$$

Directed information graph

- Pairwise measure can not distinguish cascading/proxy effects

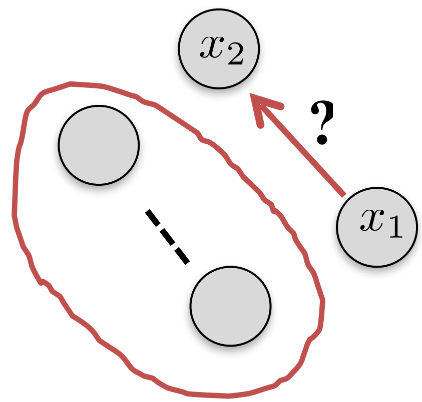


$$I(A \to B) > 0$$

- Causally condition on all other signals: $I(A \to B || C) > 0$

- There is an arrow from x_1 to x_2 if

$$I(X_1 \to X_2 || X_{-\{1,2\}}) > 0,$$



Resulting graph is a directed information graph

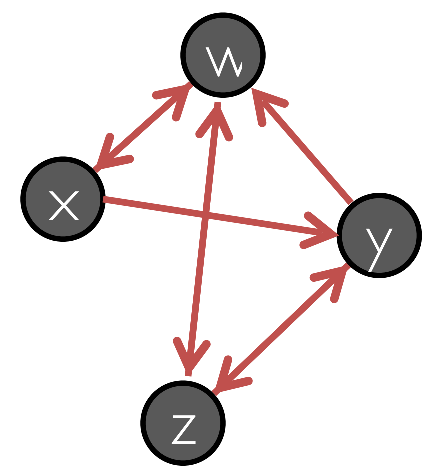
Generative Models and Causality

- Dynamical system (Deterministic):

$$\dot{w} = f_1(w, x, y, z) \qquad \dot{y} = f_3(x, y, z)$$
$$\dot{x} = f_2(w, x) \qquad \dot{z} = f_4(w, y, z)$$

- Dynamical system (Stochastic):

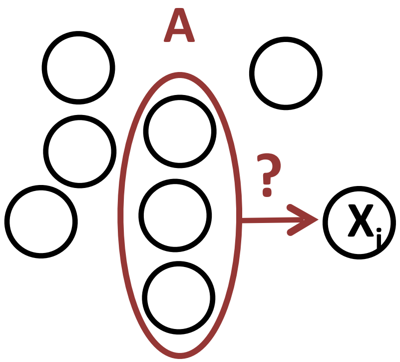
$$P_{\mathbf{w},\mathbf{x},\mathbf{y},\mathbf{z}} = \prod_{t=1}^{n} P_{w(t),x(t),y(t),z(t)|w^{t-1},x^{t-1},y^{t-1},z^{t-1}}$$
$$= P_{\mathbf{w}||\mathbf{x},\mathbf{y},\mathbf{z}} P_{\mathbf{x}||\mathbf{w}} P_{\mathbf{y}||\mathbf{x},\mathbf{z}} P_{\mathbf{z}||\mathbf{w},\mathbf{y}}$$



$$P_{\mathbf{x}||\mathbf{w}} := \prod_{t=1}^{n} P_{x(t)|x^{t-1},w^{t-1}}$$

Generative Model

- **Causal Markov blanket:**
  - For each process X_i, find parent set A:



$$P_{\mathbf{A}}(\underline{\mathbf{x}}) = \prod_{i=1}^{m} P_{\mathbf{x}_i || \underline{\mathbf{x}}_{\mathbf{A}(i)}}(\mathbf{x}_i || \underline{\mathbf{x}}_{\mathbf{A}(i)}),$$
$$D(P_{\underline{\mathbf{x}}} || P_{\mathbf{A}}) = 0.$$

- Resulting graph is a generative model graphs (factorization of joint).
- In a minimal generative for each process i , A(i) is of minimal cardinality.

- Unlike Bayesian networks, generative mode graph is unique as long as:

Assumption 1: The joint distributions satisfies spatial conditional independence:

$$P_{\mathbf{X}}(\underline{\mathbf{X}}) = \prod_{j=1}^{n} \prod_{i=1}^{m} P_{X_{i,j} | \underline{\mathbf{X}}_{(1:j-1)}}(X_{i,j} | \underline{\mathbf{X}}_{(1:j-1)})$$

and nontrivial conditional distributions are non-degenerate.

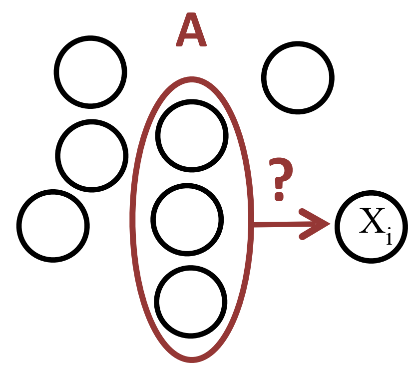
Quinn, Kiyavash, and Coleman "Directed information graphs," IEEE Trans. on Information Theory, 2015.

Generative Model vs Directed Information Graphs

- **GMGs: Identify causal Markov blanket:**
  - For each process X_i, find parent set A:

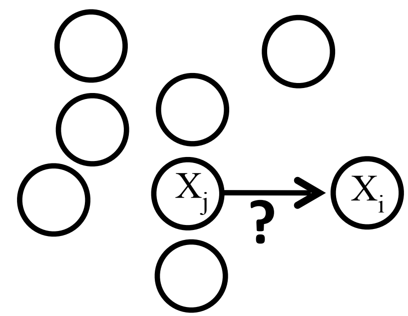
$$P_{\mathbf{A}}(\underline{\mathbf{x}}) = \prod_{i=1}^{m} P_{\mathbf{x}_i||\underline{\mathbf{x}}_{\mathbf{A}(i)}}(\mathbf{x}_i||\underline{\mathbf{x}}_{\mathbf{A}(i)}),$$

$$D(P_{\underline{\mathbf{x}}}||P_{\mathbf{A}}) = 0.$$



- **Directed Information (DI) graph definition**
  - Draw an edge from X to Y if

$$I(X_j \to X_1||\underline{\mathbf{X}}_{-\{X_i,X_j\}}) > 0,$$



Theorem: Two above graphs are equivalent.

Quinn, Kiyavash, and Coleman "Directed information graphs," IEEE Trans. on Information Theory, 2015.

Linear Models vs DIGs

DIGs:

- Based on statistical dependencies (as opposed to functional dependencies)
- Works for General models (say not confined to linear models)
- Learning algorithm for general graphs (no assumptions on the topology)
    - High complexity
    - Side information about the model class can help reduce the complexity.

Etesami, Kiyavash, "Directed Information Graphs: a Generalization of Linear Dynamical Graphs," ACC 2014.

DIG of Linear Models

- Consider the following Autoregressive (AR) model

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

where $\vec{R}_t = (R_{1,t}, ..., R_{m,t})^T$, the exogenous noises are independent, then,

$$I(R_i \to R_j || \underline{R}_{-\{i,j\}}) > 0, \quad \Longleftrightarrow \quad \sum_{k=1}^{p} |(\mathbf{A}_k)_{j,i}| > 0$$

- To learn the DIG, instead of estimating the DIs, can check whether the **corresponding coefficients** are zero or not.

- Wiener filtering is one approach to estimate the coefficients

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

Etesami, Kiyavash, "Directed Information Graphs: a Generalization of Linear Dynamical Graphs," ACC 2014.

DIG of Moving Average (MA) Models

- Consider a MA model:

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

- This can be written as an AR model

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

where $\mathcal{P}(L) := \sum_{k=1}^{p} \mathbf{W}_k L^{k-1}$.

- Corresponding DIG can be inferred the same way as the AR model.

DIG of Moving Average (MA) Models

- Example: MA(1)

$$\boldsymbol{W}_0 = \boldsymbol{I}, \quad \boldsymbol{W}_1 = \begin{pmatrix} 0.3 & 0 & 0.5 \\ 0.1 & 0.2 & 0.5 \\ 0 & 0.4 & 0.1 \end{pmatrix}.$$

- The corresponding AR model of this example is

$$\vec{R}_t = \vec{\epsilon}_t + \sum_{k=1}^{\infty} (-1)^{k-1} \boldsymbol{W}_1^k \vec{R}_{t-k}.$$

Because $\boldsymbol{W}_1^2$ has no non-zero entries, the DIG of this system is a complete graph.

$$\mathbf{W}_1^2 = \begin{pmatrix} 0.09 & 0.2 & 0.2 \\ 0.05 & 0.24 & 0.2 \\ 0.04 & 0.12 & 0.21 \end{pmatrix}.$$

Generalized Variance Decomposition for MA Models

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

- In GVD the weight of R_j's influence on R_i is proportional to:

$$d_{i,j} = \sum_{k=0}^{p} (\boldsymbol{W}_k \Sigma)_{i,j}^2 \,,$$

  where $\mathbb{E}[\vec{\epsilon}_t \vec{\epsilon}_t^T] = \Sigma.$

- Computing the GVD method to the Example, we obtain
  - R2 does not influence R1 and R1 does not influence R3.
  - This result is not consistent with the causal network (DIG) of this example, which is a complete graph, i.e., every node has influence on any other node.
  - Thus, GVD analysis seems to suffer from the pairwise analysis deficit commonly used in traditional application of the Granger-causality.

DIG of GARCH Models

- Consider the following model

$$R_{i,t}|\mathcal{F}^{t-1} \sim \mathcal{N}(\mu_{i,t}, \sigma^2_{i,t}),$$

$$\sigma^2_{i,t} = \alpha_0 + \sum_{k=1}^{q} \alpha_k (R_{i,t-k} - \mu_{i,t})^2 + \sum_{l=1}^{s} \beta_l \sigma^2_{i,t-l},$$

where $\mathcal{F}^{t-1}$ denotes the sigma algebra generated by $\underline{R}^{t-1} := \{R_1^{t-1}, ..., R_m^{t-1}\}$

- Only term that contains the effect of the other returns on the i-th return: $\mu_{i,t}$.
- Hence, R_j does not influence R_i iff

$$I(R_j \to R_i || \underline{R}_{-\{i,j\}}) = 0 \iff \mathbb{E}[R_{i,t}|\mathcal{F}^{t-1}] = \mathbb{E}[R_{i,t}|\mathcal{F}^{t-1}_{-\{j\}}].$$

where $\mathcal{F}^{t-1}_{-\{j\}}$ is sigma algebra generated by all processes except j-th return.

Multivariate DIG of GARCH Models

- Consider the following model

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

where $\vec{\epsilon}_t = \vec{R}_t - \vec{\mu}_t$.

- In this case: R_j does not influence R_i iff

$$I(R_j \to R_i || \underline{R}_{-\{i,j\}}) = 0 \iff \begin{array}{l} \mathbb{E}[(R_{i,t}-\mu_{i,t})^2|\mathcal{F}^{t-1}] = \mathbb{E}[(R_{i,t}-\mu_{i,t})^2|\mathcal{F}^{t-1}_{-\{j\}}]. \\ \mathbb{E}[R_{i,t}|\mathcal{F}^{t-1}] = \mathbb{E}[R_{i,t}|\mathcal{F}^{t-1}_{-\{j\}}]. \end{array}$$

- Pairwise Granger-causality calculation will fail to capture the true causal structure in this case.

Multivariate GARCH models vs DIGs

- Example:

$$\begin{pmatrix} R_{1,t} \\ R_{2,t} \end{pmatrix} = \begin{pmatrix} 0.2 & 0.3 \\ 0 & 0.2 \end{pmatrix} \begin{pmatrix} R_{1,t-1} \\ R_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix},$$

$$\begin{pmatrix} \sigma^2_{1,t} \\ \rho_t \\ \sigma^2_{2,t} \end{pmatrix} = \begin{pmatrix} 0 \\ 0.3 \\ 0.1 \end{pmatrix} + \begin{pmatrix} 0.2 & 0 & 0.3 \\ 0 & 0.2 & 0.7 \\ 0.1 & 0.4 & 0 \end{pmatrix} \begin{pmatrix} \epsilon^2_{1,t-1} \\ \epsilon_{1,t-1}\epsilon_{2,t-1} \\ \epsilon^2_{2,t-1} \end{pmatrix} + \begin{pmatrix} 0.3 & 0.5 & 0 \\ 0.1 & 0.2 & 0 \\ 0 & 0 & 0.4 \end{pmatrix} \begin{pmatrix} \sigma^2_{1,t-1} \\ \rho_{t-1} \\ \sigma^2_{2,t-1} \end{pmatrix},$$

where $\rho_t = \mathbb{E}[\epsilon_{1,t}\epsilon_{2,t}]$.

- The corresponding DIG of this model is $R_1 \leftrightarrow R_2$.
- This is because R2 influences R1 through the mean and variance and R1 influences R2 only through the variance.

Non-linear models vs DIGs

- DIG does not require any linearity assumptions on the model.

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

- Side information about the model class can simplify the calculation.
  - Markov switching model:

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

where $U_{\mathcal{S}}^T := (1, R_{i_1,t-1}) \otimes (1, R_{i_2,t-1}) \otimes \cdots \otimes (1, R_{i_s,t-1})$, $\otimes$ denotes the Kronecker product.

- R_i does not influence R_j if and only if all the terms of U_R depending on R_i are equal to zero.

Experimental Result

- We identified and monitored the evolution of connectedness among major financial institutions during 2006-2016.

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		

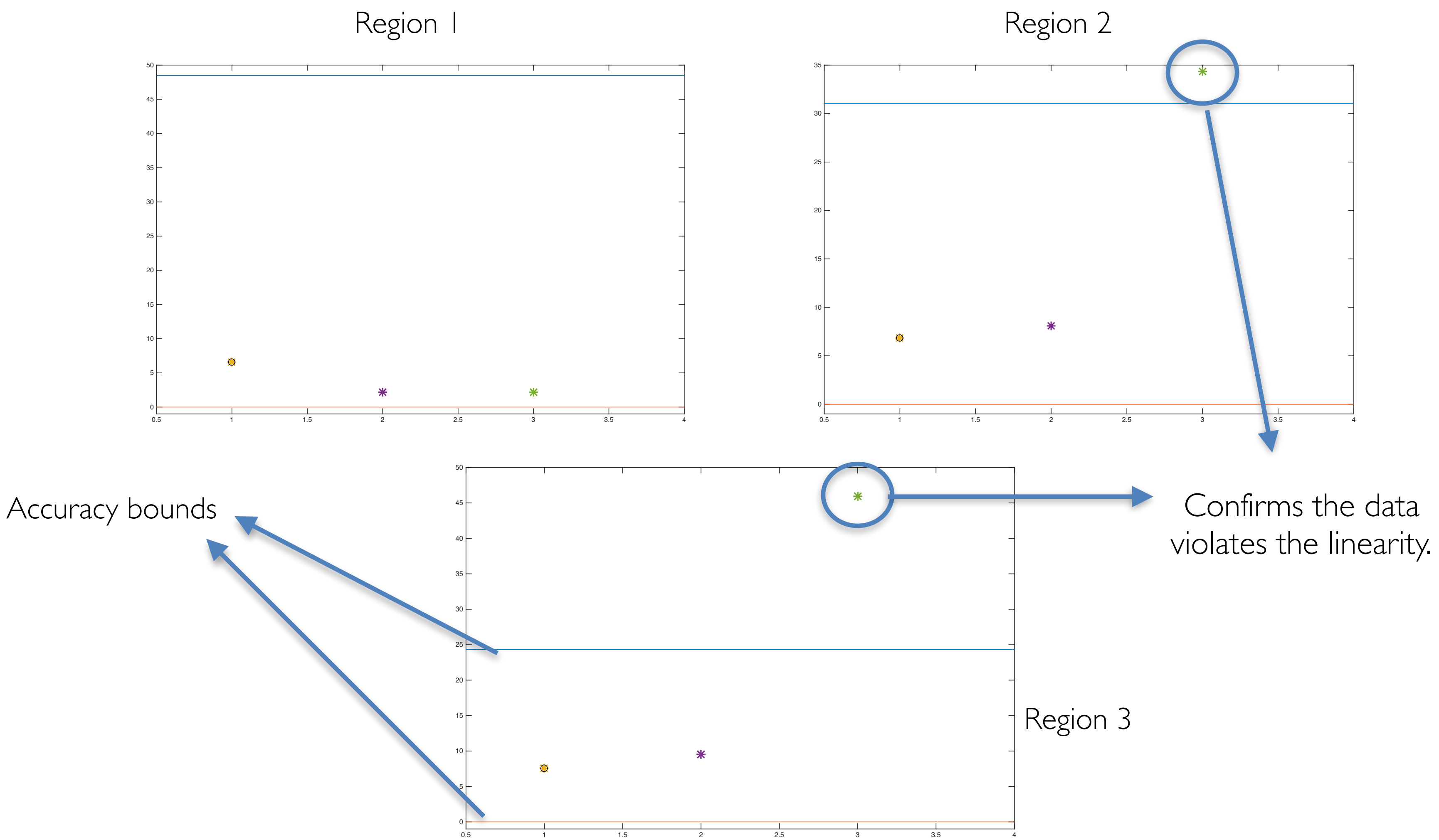
Test of Non-linearity

- Applied a non-linearity test based on nonlinear component analysis
- The test divides the operating region into several disjunct regions
  - Computes the accuracy bounds of the principle components in each region.
  - Calculates the residual variance of the remaining regions.
  - The data is said to be linear if the residual variances are within the accuracy bounds for all regions.

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

Test of Non-linearity

- We divided into 3 regions.



Resulting DIGs

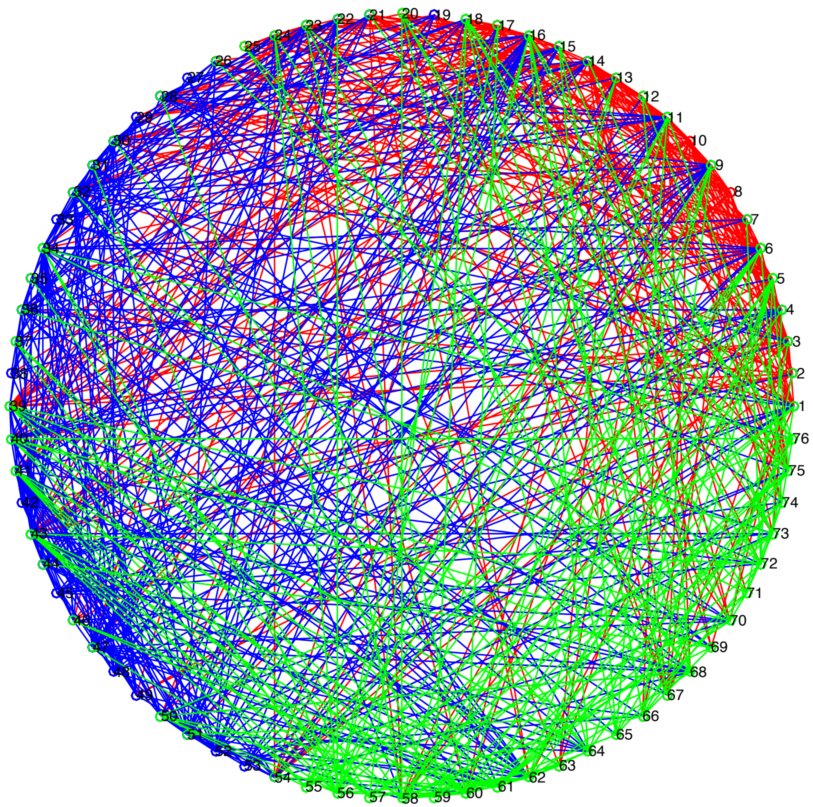
- We obtained the daily returns for individual banks, broker/dealers, and insurers.
- Estimated the DIs for each pair of companies using

$$I(X_j \to X_i || \underline{X}_{\mathcal{K}}) = \sum_{t \geq 1} H(X_i(t)|X_i^{t-1}, \underline{X}_{\mathcal{K}}^t) - H(X_i(t)|X_i^{t-1}, X_j^t, \underline{X}_{\mathcal{K}}^t).$$

where *H*( ) denotes the entropy.

- To reduce the complexity, instead of conditioning on all, we chose the ten most correlated ones.
- To decide whether the estimated DI was zero or positive, we set the threshold to be 1.16

Resulting DIGs



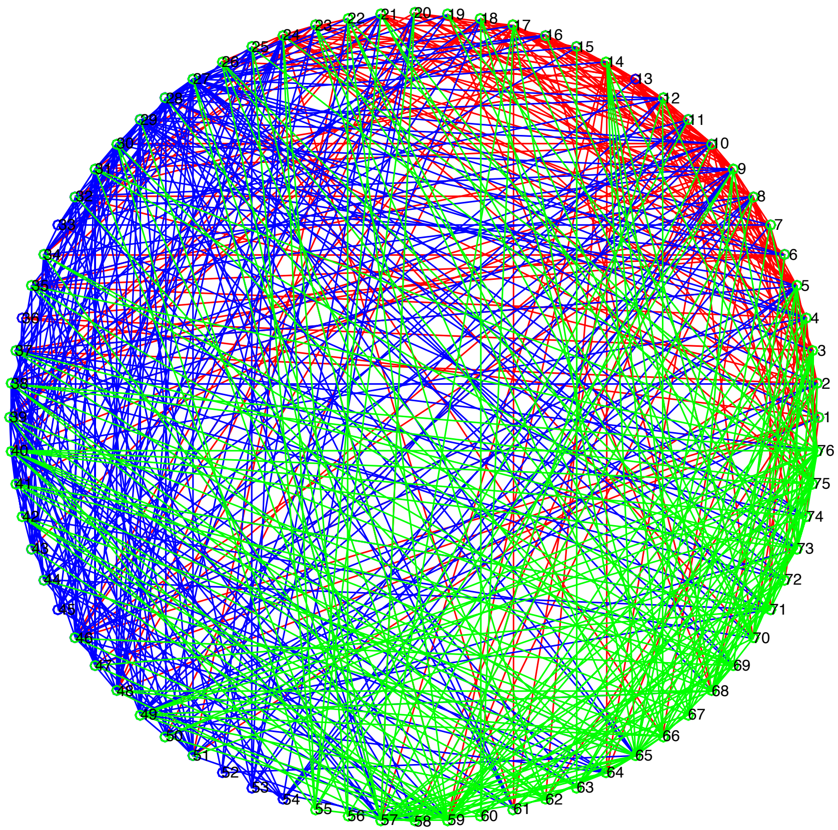
2006-2008

	Insurance	Bank	Broker
Insurance	97	120	75
Bank	95	93	49
Broker	54	71	44

\# total connections: 698

Green for brokers, red for insurers, and blue for banks

Resulting DIGs



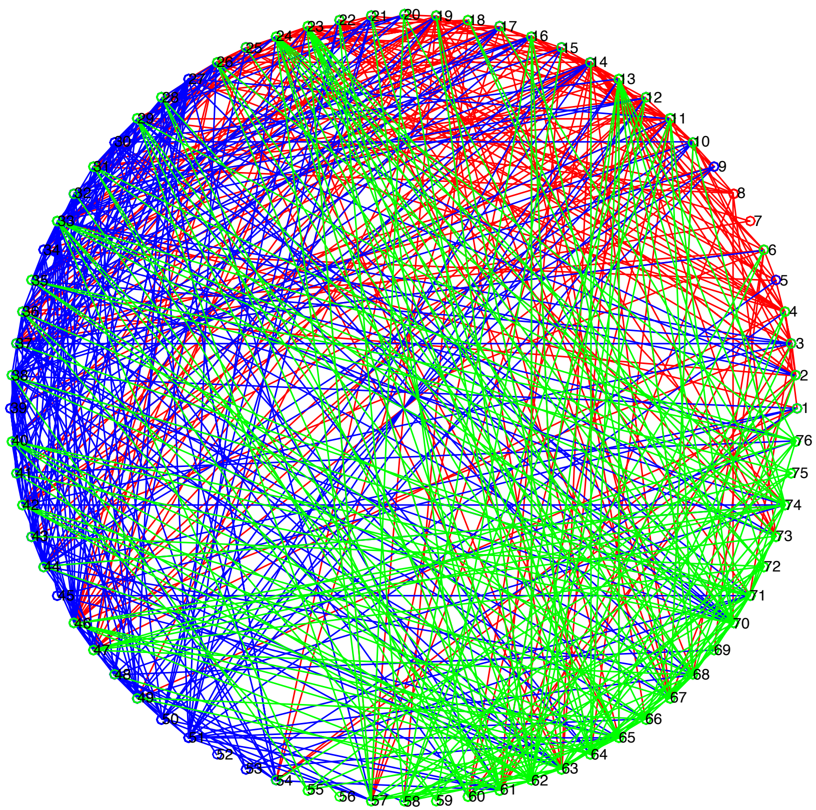
2009-2011

	Insurance	Bank	Broker
Insurance	82	100	77
Bank	55	89	66
Broker	47	59	60

total connections: 635

Green for brokers, red for insurers, and blue for banks

Resulting DIGs



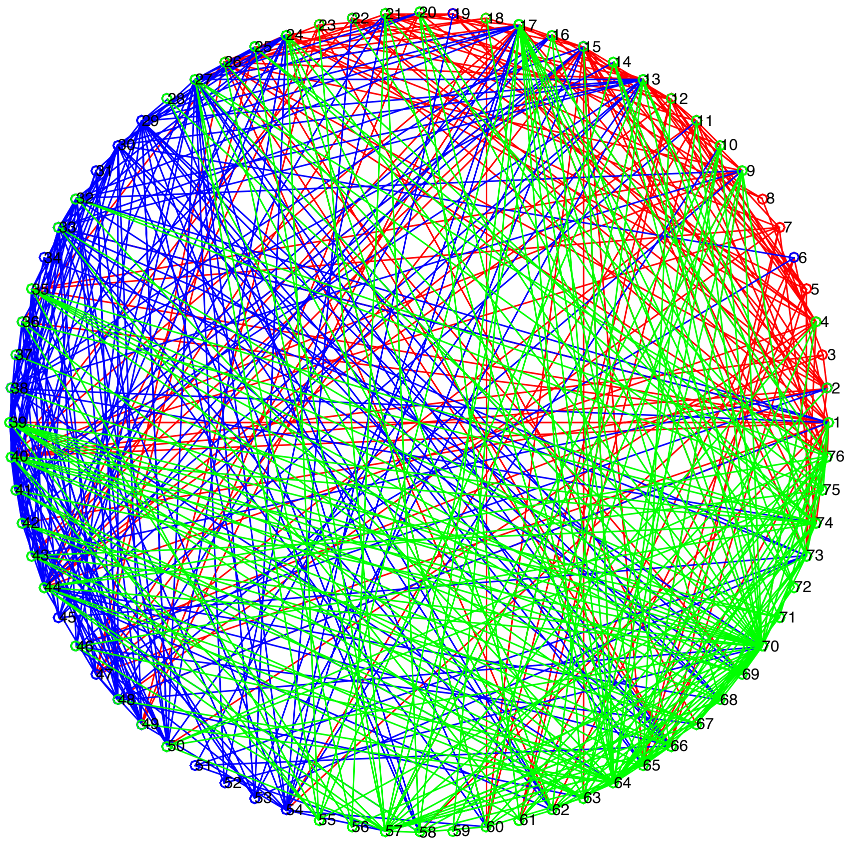
2011-2013

	Insurance	Bank	Broker
Insurance	64	79	67
Bank	86	99	62
Broker	55	53	42

total connections: 607

Green for brokers, red for insurers, and blue for banks

Resulting DIGs



2013-2016

	Insurance	Bank	Broker
Insurance	57	61	71
Bank	41	97	69
Broker	41	41	52

total connections: 530

Green for brokers, red for insurers, and blue for banks

Conclusion

- Went beyond linear pairwise estimation

- Developed a data-driven econometric framework to understand the relationship between financial institutions using non-linearly modified Granger-causality test

- applied the model to the monthly returns of U.S. financial Institutions including banks, broker, and insurance companies to see if crisis is detectable from network topology

Thank You...