IEEE GlobalSIP

Switched Dynamic Structural Equation Models for Tracking Social Network Topologies

Brian Baingana and Georgios B. Giannakis

Acknowledgments:

NSF-EARS Grant No. ECCS 1343248, MURI Grant No. AFOSR FA9550-10-1-0567

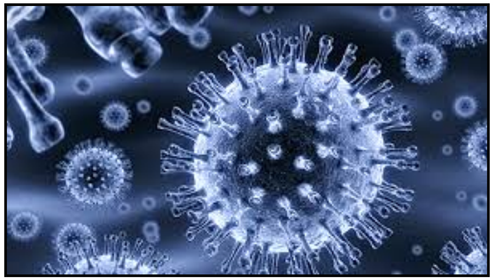
University of Minnesota
Driven to Discover℠

December 15, 2015

Context and motivation

Contagions

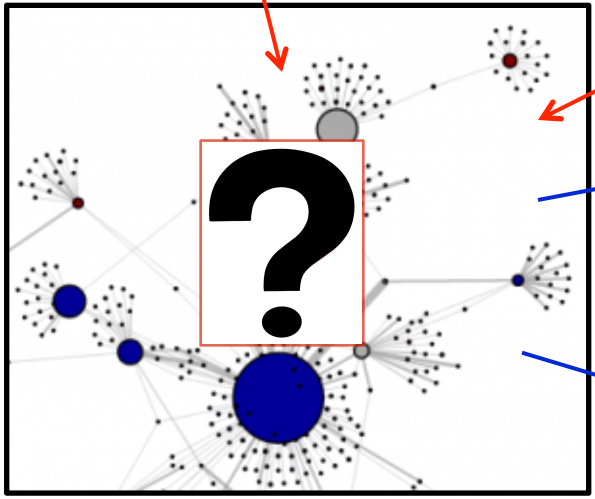
Popular news stories



Infectious diseases

Buying patterns

Propagate in cascades over social networks



Network topologies:

Unobservable, dynamic, sparse

Topology inference vital:

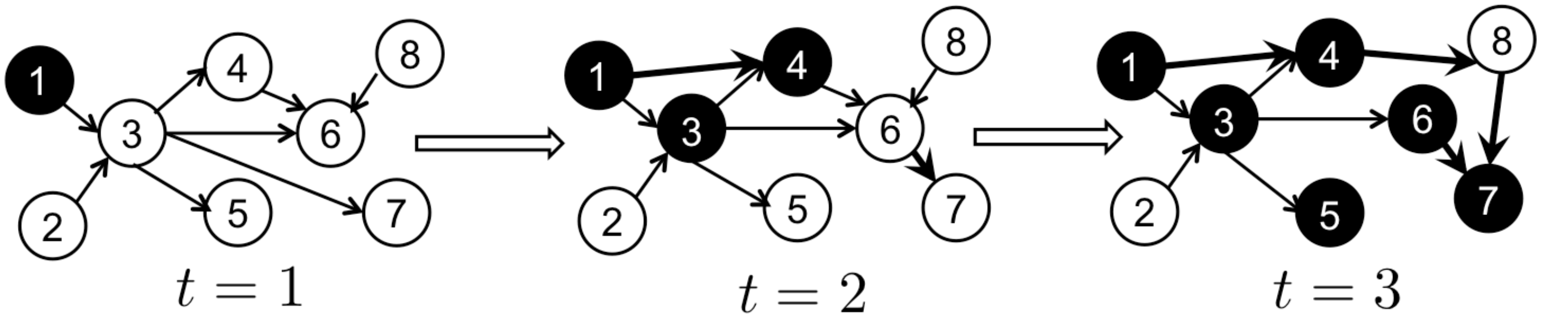
Viral advertising, healthcare policy

Goal: track unobservable time-varying network topology from cascade traces

B. Baingana, G. Mateos, and G. B. Giannakis, ``Dynamic structural equation models for social network topology inference,'' *IEEE J. of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 563-575, Aug. 2014.

Information cascades over dynamic networks

- **Example:** spread of 1 cascade over 3 time intervals



- Measurable/observable quantities:
  - **Infection time** of node by cascade (e.g., first appearance of news item on blog)
  - **Node susceptibility** to infection (e.g., politicians blog politics)

- Cascade infection times depend on:
  - Causal interactions among nodes (topological/endogenous influences)
  - Susceptibility to infection (non-topological/exogenous influences)

Contextual framework

- **Static structural equation models (SEM) for network inference**
  - Undirected topology inference [Gardner-Faith'05][Friedman et al'07]
  - Sparse SEMs for directed genetic networks [Cai-Bazerque-GG'13]

- **Causal inference from time-varying processes**
  - Graphical Granger causality and VAR models [Shojaie-Michailidis'10]
  - MLE-based dynamic network inference [Rodriguez-Leskovec'13]

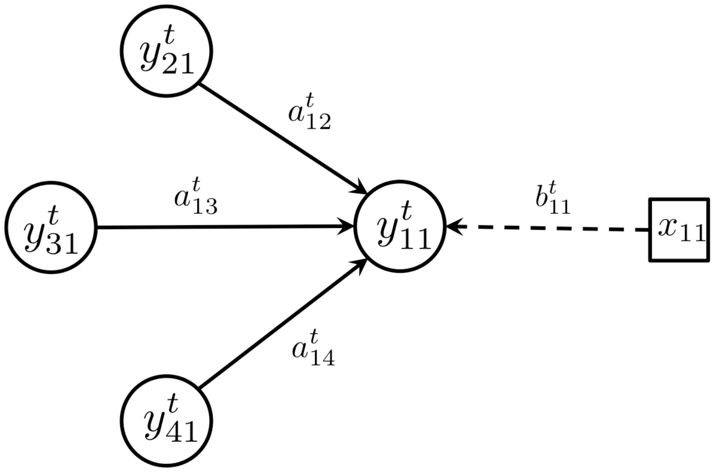
- **Contributions**
  - **Dynamic SEMs** for tracking dynamic and sparse networks
  - Accounting for external influences – **identifiability** [Bazerque-Baingana-GG'13]
  - **First-order** topology inference algorithms

J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd Ed., Cambridge Univ. Press, 2009

Model and problem statement

- **Data:** Infection time of node *i* by contagion *c* during interval *t*

$$y_{ic}^t = \underbrace{\sum_{j \neq i} a_{ij}^t y_{jc}^t}_{\textit{topological influence}} + \underbrace{b_{ii}^t x_{ic}}_{\textit{external influence}} + e_{ic}^t$$



  - $i = 1, \ldots, N, \; c = 1, \ldots, C, \; t = 1, \ldots, T$

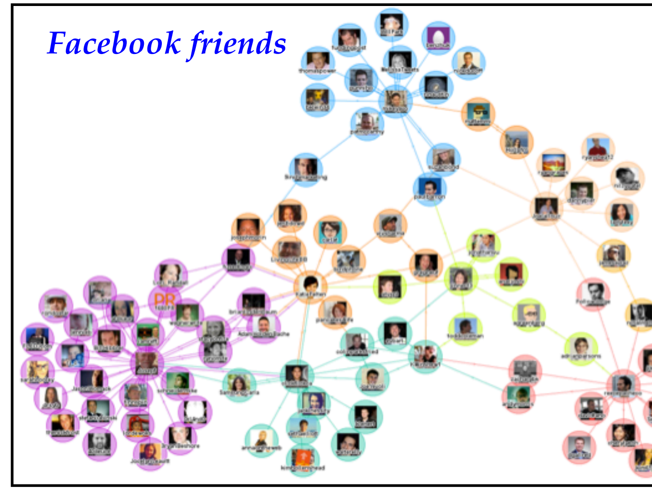
- Dynamic matrix SEM with $[\mathbf{A}^t]_{ij} = a_{ij}^t$ and $\mathbf{B}^t := \mathrm{Diag}\left(b_{11}^t, \ldots, b_{NN}^t\right)$

$$\boxed{\mathbf{Y}_t = \mathbf{A}^t \mathbf{Y}_t + \mathbf{B}^t \mathbf{X} + \mathbf{E}_t, \quad t = 1, \ldots, T}$$

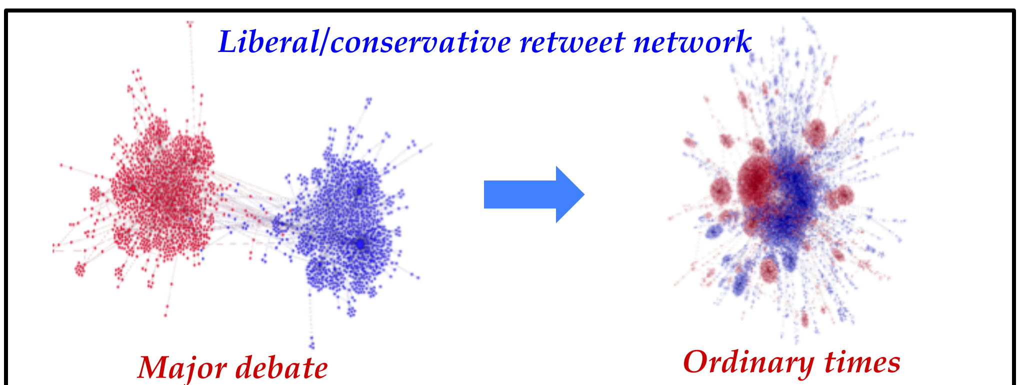
- Problem statement:
  - **Given:** Cascade data $\{\mathbf{Y}_t\}$ and $\mathbf{X}$
  - **Goal:** Track network topologies $\{\mathbf{A}^t\}$ and external influences $\{\mathbf{B}^t\}$

How do network topologies evolve?

- Slowly-varying network topologies
  - Entries of $\mathbf{A}^t$ do not suddenly change
  - Examples: ***Facebook*** friendships, web page links



- Switch between discrete network states [**This talk**]
  - $\mathbf{A}^t = \mathbf{A}^{\sigma(t)}$ where $\sigma(t) \in \{1, \ldots, S\}$
  - Example: ***Twitter*** influence network during major political/sports events
  - **Task:** identify states $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$ and switching sequence $\{\sigma(t)\}_{t=1}^T$



Tracking switched network topologies

- $\mathbf{A}^t$ and $\mathbf{B}^t$ switch between S states $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$ with **dynamic SEM**

$$\boxed{\mathbf{Y}_t = \mathbf{A}^{\sigma(t)}\mathbf{Y}_t + \mathbf{B}^{\sigma(t)}\mathbf{X} + \mathbf{E}_t} \quad \sigma(t) \in \{1, \ldots, S\}, \quad t = 1, \ldots, T$$

- **Model assumptions:**
  - **(as1)** All cascades are generated by some pair $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$ (S known)
  - **(as2)** $\{\mathbf{A}^s\}_{s=1}^S$ are sparse and $\{\mathbf{B}^s\}_{s=1}^S$ are diagonal
  - **(as3)** No two states can be jointly active during a given interval

$$\|\mathbf{Y}_t - \mathbf{A}^s\mathbf{Y}_t - \mathbf{B}^s\mathbf{X}\|_F = \|\mathbf{Y}_t - \mathbf{A}^{s'}\mathbf{Y}_t - \mathbf{B}^{s'}\mathbf{X}\|_F \implies s = s'$$

B. Baingana and G. B. Giannakis, ``Switched dynamic structural equation models for network topology tracking,'' *Proc. of GLOBALSIP*, Orlando, FL, Dec. 2015.

Sparsity-promoting estimator

- Constrained sparsity-promoting least-squares (LS) estimator

$$\arg\min_{\substack{\{\mathbf{A}^s,\mathbf{B}^s\}_{s=1}^S \\ \{\{\chi_{ts}\}_{s=1}^S\}_{t=1}^T}} \quad (1/2)\sum_{t=1}^T\sum_{s=1}^S \chi_{ts}\|\mathbf{Y}_t - \mathbf{A}^s\mathbf{Y}_t - \mathbf{B}^s\mathbf{X}\|_F^2 + \underbrace{\sum_{s=1}^S \lambda_s\|\mathbf{A}^s\|_1}_{\textit{promotes edge sparsity}}$$

$$\text{s.t.} \quad \sum_{s=1}^S \chi_{ts} = 1 \;\; \forall t, \chi_{ts} \in \{0,1\} \;\; \forall s,t$$

$$a_{ii}^s = 0, b_{ij}^s = 0, \;\; \forall s, i \neq j$$

  - $\chi_{ts} = 1$ if $\sigma(t) = s$ otherwise $\chi_{ts} = 0$, and $\|\mathbf{A}^s\|_1 := \sum_{ij} |a_{ij}^s|$

- **Caveats**
  - **NP-hard** mixed integer program
  - **Batch estimator** unsuitable for streaming cascade data

Sequential state estimation

- Setting: $\{\mathbf{Y}_t\}$ acquired sequentially
  - **Idea:** Adopt two-step sequential estimation strategy

- **S1.** Estimate active state $\hat{\sigma}(t)$ using most recent $\{\hat{\mathbf{A}}^s, \hat{\mathbf{B}}^s\}_{s=1}^S$

$$\hat{\sigma}(t) = \arg\min_{s \in \{1,\ldots,S\}} \|\mathbf{Y}_t - \hat{\mathbf{A}}^s \mathbf{Y}_t - \hat{\mathbf{B}}^s \mathbf{X}\|_F$$

  - Set $\hat{\chi}_{ts} = 1$ if $\hat{\sigma}(t) = s$ else $\hat{\chi}_{ts} = 0$

- **S2.** With $\{\{\hat{\chi}_{\tau s}\}_{s=1}^S\}_{\tau=1}^t$ known, solve decoupled problem per *t* and *s*

$$\begin{aligned} &\arg\min_{\mathbf{A}^s, \mathbf{B}^s} \quad (1/2) \sum_{\tau=1}^{t} \hat{\chi}_{\tau s} \|\mathbf{Y}_\tau - \mathbf{A}^s \mathbf{Y}_\tau - \mathbf{B}^s \mathbf{X}\|_F^2 + \lambda_s \|\mathbf{A}^s\|_1 \\ &\text{s.t.} \quad a_{ii}^s = 0, \;\; b_{ij}^s = 0, \;\; \forall i \neq j \end{aligned}$$

convex

Solving S2: First order algorithm

- **Iterative shrinkage-thresholding algorithm (ISTA)** [Parikh-Boyd'13]
  - Ideal for **convex + non-smooth** cost
- Let $\mathbf{V}^s := [\mathbf{A}^s \;\; \mathbf{B}^s]\,; \quad f(\mathbf{V}^s) := (1/2)\sum_{\tau=1}^{t} \hat{\chi}_{\tau s}\|\mathbf{Y}_\tau - \mathbf{A}^s\mathbf{Y}_\tau - \mathbf{B}^s\mathbf{X}\|_F^2$

$$\underbrace{\mathbf{V}^s[k] = \arg\min_{\mathbf{V}} \; (L_f/2)\|\mathbf{V} - \overbrace{(\mathbf{V}^s[k-1] - (1/L_f)\nabla f(\mathbf{V}^s[k-1]))}^{\textit{gradient descent}}\|_F^2 + \lambda_s\|\mathbf{A}\|_1}_{\textit{solvable by soft-thresholding operator [cf. Lasso]}}$$

- **Attractive features**
  - Provably convergent, closed-form updates
  - Recursive hence fixed computational and memory cost per *t*
  - Scales to large datasets (no matrix inversions)

B. Baingana and G. B. Giannakis, ``Switched dynamic structural equation models for network topology tracking,'' *Proc. of GLOBALSIP*, Orlando, FL, Dec. 2015.

Simulation setup

- $S = 4$ Kronecker graphs with adj. matrices $\{\mathbf{A}^s \in \mathbb{R}^{64\times 64}\}_{s=1}^4$ [Leskovec et al'10]

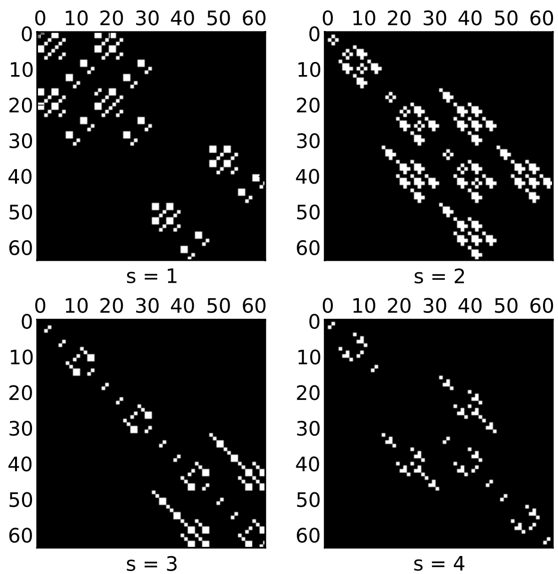
- $[\mathbf{X}]_{ij} \sim \mathcal{U}[0,3]$ and $\{\mathbf{B}^s = \mathbf{B}\}_{s=1}^4$
  - $\mathbf{B} = \mathrm{Diag}(b_{11}, \ldots, b_{NN}), \;\; b_{ii} \sim \mathcal{U}[0,1]$

- Synthetic cascade generation
  - $N = 64$ nodes, $C = 80$ cascades, and $T = 1{,}000$ intervals
  - $\sigma(t)$ sampled uniformly from $\mathcal{S} = \{1,2,3,4\}$ and $[\mathbf{E}_t]_{ij} \sim \mathcal{N}(0, 0.01)$
  - $\mathbf{Y}_t = (\mathbf{I}_N - \mathbf{A}^{\sigma(t)})^{-1}(\mathbf{B}^{\sigma(t)}\mathbf{X} + \mathbf{E}_t)$

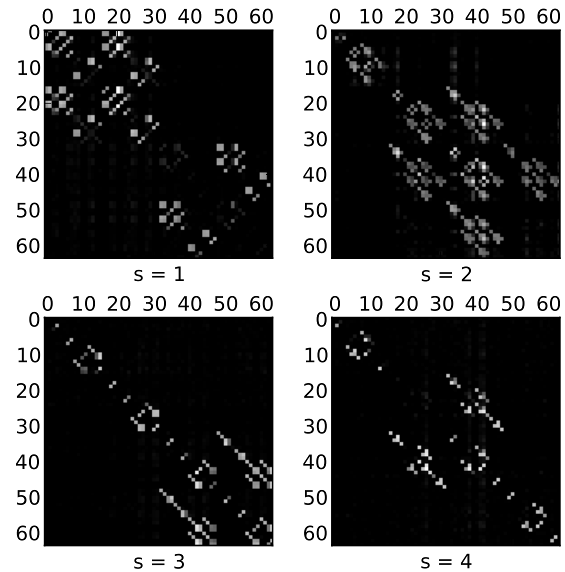
- Initialization by batch estimator

B. Baingana and G. B. Giannakis, ``Switched dynamic structural equation models for network topology tracking,'' *Proc. of GLOBALSIP*, Orlando, FL, Dec. 2015.

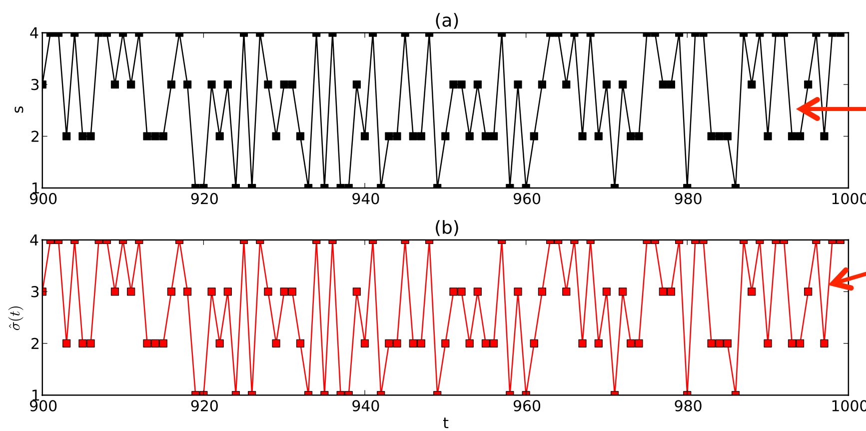
Simulation results



actual topologies



estimated topologies



actual state sequence

estimated state sequence

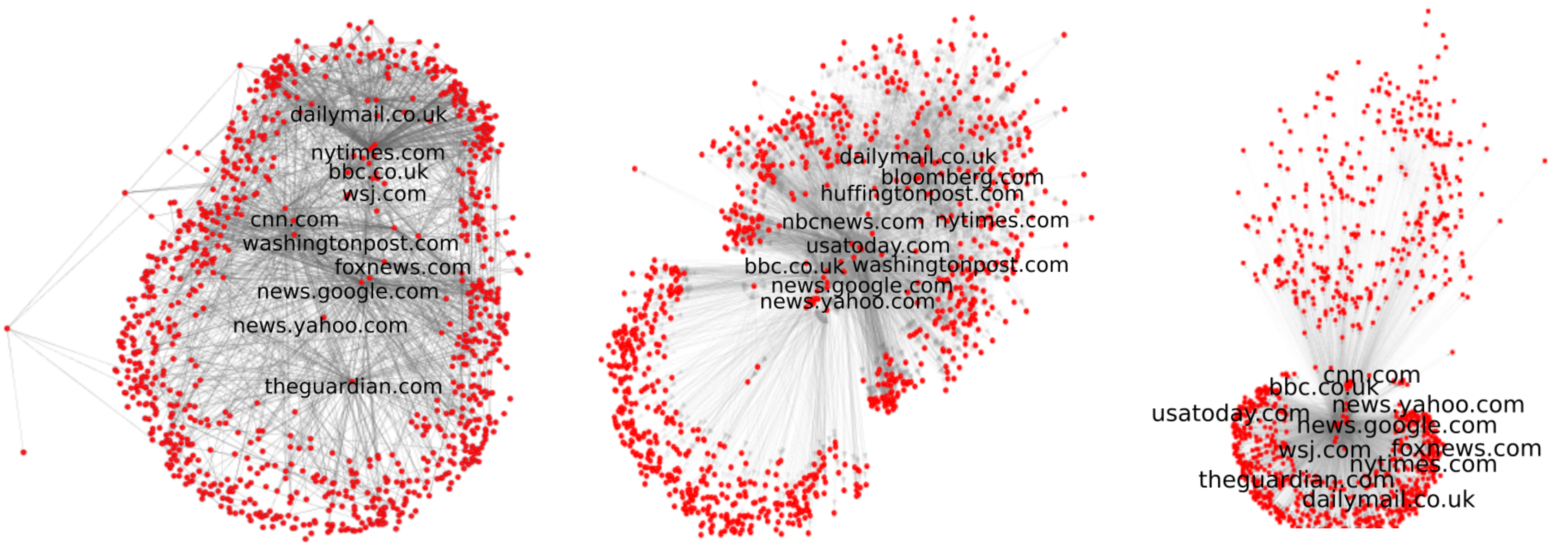
$$\lambda_s = 0.95 \quad \text{for all} \quad s$$

B. Baingana and G. B. Giannakis, ``Switched dynamic structural equation models for network topology tracking,'' *Proc. of GLOBALSIP*, Orlando, FL, Dec. 2015.

Tests on real information cascades

- Web mentions of popular ***memes*** tracked from Mar. '11 to Feb. '12
  - **Examples:** *Fukushima, Kim Jong-un, Osama, Steve Jobs, Arab spring*
  - N = **1,131** websites, C = **625** cascades, T = **180** intervals (approx. 2 days per t)

- Resulting network states with 10 most "central" websites labeled



s = 1 s = 2 s = 3

Data: SNAP's "Web and blog datasets" http://snap.stanford.edu/infopath/data.html

Conclusions

- **Switched dynamic SEM** for modeling node infection times due to cascades
  - Topological influences and external sources of information diffusion
  - Accounts for **edge sparsity** typical of social networks

- Proximal gradient algorithm for tracking switching sequence
  - Corroborating tests with simulated data
  - Real cascades of online social media revealed interesting patterns

- Ongoing and future research
  - Identifiability results for switched dynamic SEMs
  - Large-scale implementations using MapReduce/GraphLab platforms
  - Modeling nonlinearities via kernel methods

Thank You!