Topology ID and Learning over Graphs: Accounting for Nonlinearities and Dynamics

Acknowledgments: Drs. B. Baingana, J.-A. Bazerque, P. Forero, G. Mateos, K. Rajawat, D. Romero; and V. Ioannidis, G.-V. Karanikolas, M. Ma, Y. Shen, and P. Traganitis

NSF 1514056, 1500713, 1711471
Networks as graphs ... everywhere

Social networks

Internet

Energy grids

Financial markets

Brain networks

Gene/protein-regulatory nets

Network Science


Challenges and opportunities

- Network-level challenges
  - Massive scale

- Desiderata
  - Parsimonious models of network structure
  - Efficient inference algorithms over networks

- Framework: Cross-roads of machine learning, statistical SP, optimization, networking
Linear structural equation models

- Setting: \( N \) nodes over which “observable” processes propagate [Goldberger’72,Kaplan’09]

\[
y_{it} = \sum_{j \neq i} a_{ij} y_{jt} + b_{ii} x_{it} + e_{it}
\]

\( i = 1, \ldots, N \)
\( t = 1, \ldots, T \)

**Goal:** Given \( \{x_{it}, y_{it}\} \), estimate \( \{a_{ij}, b_{ii}\} \) to capture directed dependencies \( a_{ij} \neq a_{ji} \)

**Theorem 1.** If \( a_{ii} = 0, b_{ii} \neq 0, \forall i, b_{ij} = 0, \forall i \neq j \), then \( A \) and \( B \) identifiable if \( kr_X^T > 2 \max_{\text{deg}}(G) \), \( X := [x_1 \ldots x_T] \)

OK if \( T > 2N \)

From single- to multi-layer SEMs

- Generalizes linear SEM to multi-layer settings
- Layers can model exogenous variables and time snapshots or lags

Multilayer linear SEMs

- Per node

\[ y_{it}^{(\ell)} = \sum_{j \neq i} a_{ij}^{(\ell)} y_{jt}^{(\ell)} + \sum_{\ell' \neq \ell} \sum_{j} a_{ij}^{(\ell',\ell)} y_{kt}^{(\ell')} + e_{it}^{(\ell)} \]

**Intra-layer term** \[ y_{it}^{(\ell)} = \sum_{j \neq i} a_{ij}^{(\ell)} y_{jt}^{(\ell)} \]

**Inter-layer term** \[ \sum_{\ell' \neq \ell} \sum_{j} a_{ij}^{(\ell',\ell)} y_{kt}^{(\ell')} + e_{it}^{(\ell)} \]

- Matrix form

\[ Y^{(\ell)} = Y^{(\ell)} A^{(\ell)} + \sum_{\ell' \neq \ell} Y^{(\ell')} A^{(\ell',\ell)} + E^{(\ell)}, \quad \ell = 1, \ldots, L \]

**Goal:** Given \( \{Y^{(\ell)}\}_{\ell=1}^{L} \) find \( \{A^{(\ell)}\}_{\ell=1}^{L}, \{A^{(\ell',\ell)}\}_{\ell \neq \ell'} \)

- Identifiability

**Theorem 2.** If \( Y = [Y^{(1)} \ldots Y^{(L)}] \), and \( \text{kr}_Y > 2 \max_{\text{deg}(G)} \), then \( \{A^{(\ell',\ell)}\}_{\ell \neq \ell'}, \{A^{(\ell)}\}_{\ell=1}^{L} \) identifiable

- Estimation via e.g., ordinary or regularized least-squares (LS)

Simulated and real data tests

- **Synthetic network**, $N=40$, $L = 4$ (each layer corresponds to a block diagonal)

- **US economic sectors**, $N=40$ industries, $L = 7$ sectors (textiles, automotive …)
Topology tracking from network cascades

Contagions

Popular news stories

Infectious diseases

Buying patterns

Network topologies:
Unobservable, dynamic, sparse

Topology inference vital:
Viral advertising, healthcare policy

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

Linear dynamic SEMs

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

$$y_{ic}^t = \sum_{j \neq i} a_{ij}^t y_{jc}^t + b_{ii}^t x_{ic}^t + e_{ic}^t$$

$$Y_t = A_t Y_t + B_t X_t + E_t, \quad t = 1, \ldots, T$$

**Goal:** Given data $\{Y_t, X_t\}$, track topology $\{A_t\}$ and external influences $\{B_t\}$

- **Q:** How do network topologies evolve?
  - **Switching** among discrete states $\sigma(t) \in \{1, \ldots, S\}$
  - **Slow-varying** network topologies $A_t$ changes slowly; e.g., Facebook

Tracking slowly-varying topologies

- Structural spatio-temporal properties
  - Slowly time-varying topology
  - Sparse edge connectivity, \#edges = \mathcal{O}(\#\text{nodes})

- Sparsity-promoting exponentially-weighted LS estimator (EWLSE)

\[
\{\hat{A}_t, \hat{B}_t\} = \arg\min_{A,B} \left(1/2\right) \sum_{\tau=1}^{t} \beta^{t-\tau} \|Y_\tau - AY_\tau - BX_\tau\|_F^2 + \lambda_t \|A\|_1 \quad \text{s.t. } a_{ii} = 0, \quad b_{ij} = 0, \quad \forall \ i \neq j
\]

- **Edge sparsity** encouraged by $\ell_1$-norm regularization with $\lambda_t > 0$
- **Tracking** dynamic topologies possible if $\beta < 1$ ($\beta \in (0, 1]$)

- **Solver**: proximal-splitting optimization methods [Daubechies et al’04]
The rise of Kim Jong-un

- Web mentions of “Kim Jong-un” tracked from Mar.’11 to Feb.’12

- $N = 360$ websites, $C = 466$ cascades, $T = 45$ weeks

**Increased media frenzy following Kim Jong-un’s ascent to power in 2011**

*Data: SNAP’s “Web and blog datasets”* http://snap.stanford.edu/infopath/data.html
Identifiability of SEMs with input statistics?

- Limited access to input $x$
  - Privacy concerns
  - Not explicitly available

Goal: Given statistics of $\{y_t, x_t\}$ identify and track hidden directed network topology

$$y_t = Ay_t + Bx_t \quad \Rightarrow \quad y_t = (I - A)^{-1}Bx_t = Ax_t$$

- Covar. over segment $m$ $R_y^m := \mathbb{E}\{y_t y_t^\top\}$, $t \in [\tau_m, \tau_{m+1} - 1]$, $m = 1 \ldots M$

How much did you invest in Apple stock yesterday? Uh ... not gonna say, man .... It's private!
Network snapshots as tensor slabs

Tensor slab

\[ R^y_m = A \text{Diag}(\rho^x_m) A^\top \]

Partially symmetric PARAFAC

\[ R^y = \sum_{i=1}^{N} \alpha_i \circ \alpha_i \circ r^x_i \]

**Proposition 1:** If \( a_{ii} = 0, b_{ii} \neq 0, \forall i, b_{ij} = 0, \forall i \neq j \), then \( A \) and \( B \) are uniquely expressible in terms of \( A \) as \( B = (\text{Diag}[A^{-1}])^{-1} \) and \( A = I - (\text{Diag}(A^{-1}))^{-1} A^{-1} \)

**Theorem 2a.** If \( kr^x_R > 1 \), and \( R^x \) available, then \( A \) is identifiable.

**Theorem 2b.** If \( kr^x_R > 1 \), but \( R^x \) unknown, \( A \) identifiable within permutations (finite!)

Real stock networks

- Dec. 23, 2011 to Sep. 30, 2016 (1,200 days), $M = 12$ time segments
- 100 runs each with random initialization

- Strong connectivity among major technology companies
- Stronger connectivity between Macy’s and Nordstrom
Linear structural vector autoregressive models

- Edge weights capture directed causal dependencies.
- Edge sparsity: only a few $a_{i,j}^\ell$ are nonzero.

Endogenous variables here played by lagged exogenous.

$$y_{jt} = \sum_{i \neq j} a_{i,j}^0 y_{it} + \sum_{i=1}^N \sum_{\ell=1}^L a_{i,j}^\ell y_{j(t-\ell)} + e_{jt}$$

From linear to nonlinear SVARMs

\[ y_{jt} = \tilde{f}_j(y_{-jt}, \{y_{t-\ell}\}_{\ell=1}^L) + e_{jt}, \quad j = 1, \ldots, N \]

**Idea:** Reduce complexity using a generalized additive model

\[ \tilde{f}_j(y_{-jt}, \{y_{t-\ell}\}_{\ell=1}^L) = \sum_{i \neq j} \tilde{f}_{ij}^0(y_{it}) + \sum_{i=1}^N \sum_{\ell=1}^L \tilde{f}_{ij}^\ell(y_{i(t-\ell)}) \]

\[ \tilde{f}_{ij}^\ell(y) := a_{ij}^\ell f_{ij}^\ell(y) \quad a_{ij}^\ell \in \{0, 1\} \]

- Linear SVARM is special case

- Draw each univariate function from a reproducing kernel Hilbert space (RKHS)

\[ \mathcal{H}_i^\ell := \{ f_{ij}^{\ell} \mid f_{ij}^{\ell}(y) = \sum_{t=1}^\infty \beta_{ij}^{\ell} \kappa_i^{\ell}(y, y_{i(t-\ell)}) \} \]

\[ \{ \hat{f}_{ij}^{\ell} \} = \arg \min_{\{f_{ij}^{\ell} \in \mathcal{H}_i^\ell\}} \frac{1}{2} \sum_{t=1}^T \left[ y_{jt} - \sum_{i \neq j} \alpha_{ij}^0 f_{ij}^0(y_{it}) - \sum_{i=1}^N \sum_{\ell=1}^L \alpha_{ij}^\ell f_{ij}^\ell(y_{it}) \right]^2 + \lambda \sum_{i=1}^N \sum_{\ell=0}^L \Omega(\|a_{ij}^\ell f_{ij}^\ell\|_{\mathcal{H}^\ell}) \]

Edge sparsity leads to group-sparsity

- Representer theorem [Wahba et al’90]
  \[ \hat{f}_{ij}(y) = \sum_{t=1}^{T} \beta_{ijt}^{\ell} \kappa_i^{\ell}(y, y_{i(t-\ell)}) \]

- \( \alpha_{ij}^{\ell} := a_{ij} \beta_{ij}^{\ell} \), \( \beta_{ij}^{\ell} := [\beta_{ij1}^{\ell}, \ldots, \beta_{ijT}^{\ell}]^\top \), \( [K_i^\ell]_{t,\tau} = \kappa_i^{\ell}(y_{it}, y_{i(\tau-\ell)}) \)

\[
\{\hat{\alpha}_{ij}^{\ell}\} = \arg\min_{\hat{\alpha}_{ij}^{\ell} = 0, \{\alpha_{ij}^{\ell}\}} \frac{1}{2} \left\| Y - \sum_{l=1}^{L} K^l W_{\alpha}^l \right\|_F^2 + \lambda \sum_{\ell=0}^{L} \sum_{j=1}^{N} \sum_{i=1}^{N} \sqrt{(\alpha_{ij}^{\ell})^\top K_i^\ell \alpha_{ij}^{\ell}}
\]

- \( Y := [y_1, \ldots, y_N] \in \mathbb{R}^{T \times N} \), \( \bar{K}^\ell := [K_1^\ell \ldots K_N^\ell] \)

- Edge sparsity \( \Rightarrow \) group sparsity of \( W_{\alpha}^\ell \)

\[ W_{\alpha}^\ell := \begin{bmatrix} \alpha_{11}^\ell & \cdots & \alpha_{1N}^\ell \\ \vdots & \ddots & \vdots \\ \alpha_{N1}^\ell & \cdots & \alpha_{NN}^\ell \end{bmatrix} \]

**Bottomline:** Nonzero \( \{\alpha_{ij}^{\ell}\} \) reveal edges; ADMM solver

- Multi-kernels can choose optimal kernel combination from a prescribed dictionary of kernels
Simulated test

- Synthetic graph via Erdős–Rényi model, $N=20$, $T=40$
  - $p = 0.3$  \[ \alpha_{ij}^\ell \sim \mathcal{N}(0, I) \]
  - $\sigma_e = 0.1$

Graphs:
- Polynomial SVARM
- Linear SVARM

![Graphs showing performance metrics for Polynomial and Linear SVARM models.]
Brain is densely networked

- **Data**: electrocorticography (ECoG) data for epilepsy [Kramer et al’ 08]
  - $\mathbf{Y}$: ECoG data samples
  - $N = 76$ electrodes, $T = 200$ samples, $L=1$

- Diffusion of information is inhibited after the onset of an epileptic seizure

Identifying connectivity of meshed power grids

- Grid of $N=14$ buses; nodal vectors are voltage angle time courses
  - IEEE-14 bus benchmark; voltage angles obtained using MATPOWER

- Real load data from 2012 Global Energy Forecasting Competition

- Nonlinear predictors improve ID of mesh grid connectivity

Interpolating and extrapolating over networks

- Undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = N$

- Process $y_{nt}$ per node $n$, timeslot $t$
  - Only measure $M < N$ nodes (e.g., link counts, delays)

  $$z_t = M_t y_t + \epsilon_t$$

  $$z_t := [z_{1t}, \ldots, z_{Nt}]^\top, \quad y_t := [y_{1t}, \ldots, y_{Mt}]^\top$$

  - Rows of $M_t$ selected from $I_{N \times N}$

**Goal:** Impute misses and predict $y_t$ from selected node observations $z_t$

**Motivating application:** Estimate network delays

---

Graph-regularized dictionary learning approach

- Generalizes low-rank based matrix completion that cannot predict!

- Adopt and learn basis and expansion coefficients \((y_t = Bs_t)\)

\[
\text{arg min}_{S, B: \{\|b_q\|_2 \leq 1\}^Q_{q=1}} \sum_{t=1}^{T} \left[ \|z_t - M_t Bs_t\|_2^2 + \lambda_s \|s_t\|_1 + \lambda_w s_t^\top B^\top L Bs_t \right]
\]

Dictionary: \(N \times Q\)  
Sparse coefficients  
Graph Laplacian  
Smoothness over graph

- With adjacency matrix \(A\), graph Laplacian \(L := \text{Diag}(A1_N) - A\)

\[
s_t^\top B^\top L Bs_t = \left(1/2\right) \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} (y_{it} - y_{jt})^2
\]

promotes smoothness
Test case: Internet2

- Link count measurements: $L=54$, $T=2,000$
  (other features possible, e.g., delays)

**Training phase – 30 links measured**

**Operational phase – 30 links predicted**

- Prediction improves as link load increases
Performance comparisons

- Normalized prediction error: \( \text{NPE} := \frac{1}{L t_0} \sum_{\tau=1}^{t_0} \| y_\tau - \hat{y}_\tau \|_2^2 \)

  - \( Q = \) number of columns of \( B; \ t_0 = 2,000 \)

- Gravity-based [Zhang et al’05]; Diffusion wavelets [Coifman-Maggioni’07]

- Graph-regularized DL with semi-supervised predictor outperforms alternatives
Graph-adaptive kernel-based interpolation

\[ z_t = M_t y_t + \epsilon_t \quad \text{with} \quad M_t \in \{0, 1\}^{M \times N}, \quad M < N \]

**Goal:** Given \( z_t, M_t, \) and \( G_t \), estimate \( y_t \)

- **RKHS model:** \( y_t \in \mathcal{H}_{K_t} \iff y_t = K_t \alpha_t, \quad \alpha_t \in \mathbb{R}^N \)
  - Graph-dependent symmetric \( K_t \geq 0 \)
  - Ex. Laplacian \( (L_t) \) family \( K_t := r^{-1}(L_t) \)

- **Kernel ridge regression (KRR)** [Smola-Kondor '03]

\[
\hat{y}_t = \arg \min_{y} \frac{1}{M} \|z_t - M_t y\|_2^2 + \mu \|y\|_{K_t}^2 \\
= K_t M_t^T (M_t K_t M_t^T + \mu M I_M)^{-1} z_t
\]

Spatio-temporal processes on graphs

\[ \mathbf{z}_t = \mathbf{M}_t \mathbf{y}_t + \mathbf{\epsilon}_t \]

- **Superimposed state model** for e.g., packet delays, stock values, temperature, …

\[ \mathbf{y}_t = \mathbf{y}_t^{(\nu)} + \mathbf{y}_t^{(\chi)}, \quad \mathbf{y}_t^{(\chi)} = \mathbf{A}_{t,t-1} \mathbf{y}_{t-1}^{(\chi)} + \mathbf{\eta}_t, \quad t = 1, 2, \ldots \]

- Spatial \( \mathbf{y}_t^{(\nu)} \) temporally uncorrelated ('fast' dynamics across slots)

- Spatio-temporal \( \mathbf{y}_t^{(\chi)} \) VARM ('slow' dynamics; trend)

- **Space-time kriging ridge regression (KRR)**

\[
\arg \min_{\{\mathbf{y}_t^{(\chi)}, \mathbf{y}_t^{(\nu)}\}_{t'}} \sum_{t'=1}^{t} \frac{1}{M_{t'}} \| \mathbf{z}_{t'} - \mathbf{M}_{t'} \mathbf{y}_{t'}^{(\chi)} - \mathbf{M}_{t'} \mathbf{y}_{t'}^{(\nu)} \|^2_2 + \mu_1 \sum_{t'=1}^{t} \| \mathbf{y}_{t'}^{(\chi)} - \mathbf{A}_{t',t'-1} \mathbf{y}_{t'-1}^{(\chi)} \|^2_{K_{t'}^{(\eta)}} \\
+ \mu_2 \sum_{t'=1}^{t} \| \mathbf{y}_{t'}^{(\nu)} \|^2_{K_{t'}^{(\nu)}}
\]

**Result.** KeKriKF produces the sequence of filtered \( \{\hat{\mathbf{y}}_{t'}^{(\chi)}, \hat{\mathbf{y}}_{t'}^{(\nu)}\}_{t'=1}^{t} \)

---

Temperature reconstruction

- Temperature sensor network $N=109$
  - Compare reconstruction NMSE per day

![Graph showing NMSE over time for different methods](image)

GDP prediction

- Financial network between $N=127$ countries
  - Track gross domestic product (GDP) at an unobserved country

Joint ID of topologies and signals on graphs

Q: What if topology unknown and just a subset of data available due to privacy/large-scale?

- Linear SEM: \( \mathbf{y}_l = \mathbf{A}\mathbf{y}_l + \mathbf{\eta}_l \)
- Measurements: \( \mathbf{z}_l = \mathbf{M}_l\mathbf{y}_l + \mathbf{\epsilon}_l \)
- \( \mathbf{M}_l \times N \quad l = 1, \ldots, L \)

Goal: Given \( \{\mathbf{z}_l, \mathbf{M}_l\}_{l=1}^L \), identify \( \mathbf{A} \) and \( \{\mathbf{y}_l\}_{l=1}^L \)

- Joint inference of signals and (directed) graphs (JISG)

\[
\min_{\mathbf{A} \in \mathcal{A}, \{\mathbf{y}_l\}_{l=1}^L} \sum_{l=1}^L \|\mathbf{y}_l - \mathbf{A}\mathbf{y}_l\|_2^2 + \sum_{l=1}^L \|\mathbf{z}_l - \mathbf{M}_l\mathbf{y}_l\|_2^2 + \lambda_1\|\mathbf{A}\|_1 + \lambda_2\|\mathbf{A}\|_F^2
\]

- BCD/ADMM solver: Guaranteed convergence at reduced complexity (separable per \( l \))

- Generalizable to nonlinear SEM; multi-layer; and dynamic signals and graphs

Testing JISG on gene-regulatory networks

- $N=39$ immune-related genes; $L=69$ unrelated individuals; $y$: gene expression level
- SEM oracle observes all genes $M=39$ (left); **JISG** with $M=31$ (right)

- $NMSE$ for $\{\hat{y}_l\}_{l=1}^L = 0.017$
- JISG-based recovery similar to that of the oracle

**Goal**: Given subset of user-item ratings, rank ‘N-best’ candidates of unavailable ratings

- Sparse linear model (SLIM) of ratings: SEM followed by interpolation and ranking

\[
\min \left\{ a_{ii'} \right\} \left\| r_i - \sum_{i'} a_{ii'} r_{i'} \right\|^2_2 + \lambda \sum_{i'} |a_{ii'}| \\
\text{s.t. } a_{ii} = 0, \quad a_{ii'} \geq 0 \quad \forall i'.
\]

**Our idea**: Instead of SLIM, employ sparse nonlinear SVARM with \( L=0 \) (SNIM)

---

Movielens dataset

- 3,706 users; 6,040 movies; and 1M ratings

  - **Training set**: 97% randomly sampled ratings; **Probe set**: All 5 star ratings in testing set
  - # hits: number of ratings in probe set that also appear in the recommendation list

  ![Graphs showing recall and precision](image)

  Recall \((N_r) = \text{#hits} / \text{#probe}\)

  Precision \((N_r) = \text{recall}(N_r) / N_r\)

  - Here SNIM outperforms SLIM by a slim margin

Current research and outlook

- **Topology identification** – a “deterministic” RKHS-based approach
  - Directed and linear multi-layer graphs are allowed with identifiability guarantees
  - Nonlinear dependencies accommodated through multi-kernel regression
  - Slow-varying and switching dynamics can be afforded

- **Learning of processes on graphs**
  - Interpolation and extrapolation from partially-observed processes on graphs
  - Topology can be known or unknown
  - Kernel Kriged Kalman Filtering for inference of dynamic processes

- **Ongoing research and future directions**
  - Graph-adaptive dimensionality reduction/manifold learning
  - Tracking and identifiability of nonlinear and dynamic topologies
  - RKHS-principled multi-kernel learning vis-à-vis DNNs

Thank you!