Dynamic Structural Equation Models for Tracking Cascades Over Social Networks

Brian Baingana *Gonzalo Mateos* *Georgios B. Giannakis*

Dept. of ECE and Digital Technology Center, Univ. of Minnesota, USA
{baing011, mate0058, georgios}@umn.edu

NIPS 2013 workshop
Frontiers of Network Analysis: Methods, Models, and Applications

Motivation

- Cascading processes such as web events, infectious diseases, product adoption propagate over implicit networks [Easley10].

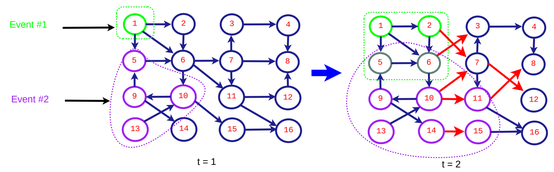

Figure 1: Cascades over a social network

- Although node "infection times" by cascades are observable, network topologies are **unknown**, **time-varying**, and exhibit **edge sparsity**.
- **Goal:** Track the time-varying network topology using node infection times.
- **Benefits:** Network topology is vital for meaningful web advertising, healthcare policy formulation, product promotions etc.

Contributions and Related Work

- Node infection times depend on:
  1. Causal interactions among nodes (topological influences)
  2. Susceptibility to cascades (external influences)
- Structural equation models (SEM) provide a general statistical framework for capturing causal interactions in psychometrics, social sciences, and gene regulation [Goldberger72][Cai13]

Contributions:

1. Dynamic SEM for tracking time-varying networks
2. Accounting for external (non-topological) influences in cascades

Related work:

1. Maximum likelihood estimation (MLE) for static network inference [Rodriguez11]
2. MLE-based stochastic gradient descent for dynamic network inference [Rodriguez13]
3. Time-invariant SEM for gene network inference [Cai13]

Model and Problem Statement

Consider a dynamic network of N nodes, over which C cascades propagate during T time intervals. The postulated dynamic SEM for infection time of node i by cascade c during time interval t is

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

Let $\mathbf{Y}^t := [y_{ic}^t]$, $\mathbf{X} := [x_{ic}]$, $\mathbf{E}^t := [e_{ic}^t]$, and $\mathbf{B}^t := \text{diag}(b_{11}, \ldots, b_{NN})$, collecting observations for N nodes and C contagions yields the dynamic matrix SEM

$$\mathbf{Y}^t = \mathbf{A}^t \mathbf{Y}^t + \mathbf{B}^t \mathbf{X} + \mathbf{E}^t \qquad t = 1, \ldots, T. \tag{1}$$

The model captures both **topological** ($\mathbf{A}^t$) and **external influences** ($\mathbf{X}$).

Problem Statement:
Given $\{\mathbf{Y}^t\}_{t=1}^T$ and $\mathbf{X}$ adhering to (1), track the underlying network topology $\{\mathbf{A}^t\}_{t=1}^T$ and the effect of external influences $\{\mathbf{B}^t\}_{t=1}^T$.

Sparse exponentially-weighted Least squares estimator

Assuming the network topology changes **slowly** and has **sparse** edge connectivity, the estimator

$$\{\hat{\mathbf{A}}^t, \hat{\mathbf{B}}^t\} = \arg\min_{\mathbf{A},\mathbf{B}} \ \frac{1}{2}\sum_{\tau=1}^{t} \beta^{t-\tau} \|\mathbf{Y}^\tau - \mathbf{A}\mathbf{Y}^\tau - \mathbf{B}\mathbf{X}\|_F^2 + \lambda_t \|\mathbf{A}\|_1$$
$$\text{s. to} \quad a_{ii} = 0, \ b_{ij} = 0, \ \forall i \neq j \tag{2}$$

tracks $\mathbf{A}^t$ and $\mathbf{B}^t$ where $\beta \in (0, 1]$, $\lambda_t \geq 0$, and $\|\mathbf{A}\|_1 := \sum_{i,j} |a_{ij}|$.

Merits of the estimator:

1. Edge sparsity is encouraged via the penalty term $\|\mathbf{A}\|_1$
2. Tracking time-varying topologies is possible if $\beta < 1$

Leveraging **proximal gradient (PG)** iterations [Parikh13] and ignoring equality constraints, solve

$$\mathbf{V}[k] := \arg\min_{\mathbf{V}} \left\{ \frac{L_f}{2} \|\mathbf{V} - (\mathbf{V}[k-1] - (1/L_f)\nabla f(\mathbf{V}[k-1]))\|_F^2 + \lambda_t \|\mathbf{A}\|_1 \right\} \tag{3}$$

per iteration k, where $\mathbf{V} := [\mathbf{A}\,\mathbf{B}]$, $f(\mathbf{V}) := \frac{1}{2}\sum_{\tau=1}^{t} \beta^{t-\tau} \|\mathbf{Y}^\tau - \mathbf{A}\mathbf{Y}^\tau - \mathbf{B}\mathbf{X}\|_F^2$, and L_f is a *Lipshitz* constant.

Proximal Gradient Algorithm

PG iterations with equality constraints yield the (pseudo) real-time tracking algorithm:

Require: $\{\mathbf{Y}^t\}_{t=1}^T$, $\mathbf{X}$, β.

Initialize $\hat{\mathbf{A}}^0 = \mathbf{0}_{N\times N}$, $\hat{\mathbf{B}}^0 = \mathbf{\Sigma}^0 = \mathbf{I}_N$, $\bar{\mathbf{Y}}^0 = \mathbf{0}_{N\times C}$, λ_0.
for $t = 1, \ldots, T$ **do**
  $\mathbf{\Sigma}^t = \beta\mathbf{\Sigma}^{t-1} + \mathbf{Y}^t(\mathbf{Y}^t)^\top$
  $\bar{\mathbf{Y}}^t = \beta\bar{\mathbf{Y}}^{t-1} + \mathbf{Y}^t$
  Initialize $\mathbf{A}[0] = \hat{\mathbf{A}}^{t-1}$, $\mathbf{B}[0] = \hat{\mathbf{B}}^{t-1}$, and set $k = 0$.
  while not converged **do**
    for $i = 1 \ldots N$ (in parallel) **do**
      $\mathbf{a}_{-i}[k+1] = \mathcal{S}_{\lambda_t/L_f}\left(\mathbf{a}_{-i}[k] - (1/L_f)\nabla_{\mathbf{a}_{-i}} f[k]\right)$
      $b_{ii}[k+1] = b_{ii}[k] - (1/L_f)\nabla_{b_{ii}} f[k]$
      $\mathbf{a}_i^\top[k+1] = [a_{-i,1}[k+1] \ldots a_{-i,i-1}[k+1]\ 0\ a_{-i,i}[k+1] \ldots a_{-i,N}[k+1]]$
    end for
    $k = k + 1$.
  end while
  return $\hat{\mathbf{A}}^t = \mathbf{A}[k]$, $\hat{\mathbf{B}}^t = \mathbf{B}[k]$.
end for

Attractive features of the algorithm:

1. Provably guaranteed convergence
2. Parallelizable iterations
3. Recursive updates ensure minimal past data storage

Numerical Results

Synthetic dataset:
Cascade data generated from $\mathbf{Y}^t = (\mathbf{I}_N - \mathbf{A}^t)^{-1}(\mathbf{B}^t\mathbf{X} + \mathbf{E}^t)$ where $x_{ij} \sim \text{unif}(0, 3)$, $\{e_{ij}^t, b_{ii}^t\} \sim \mathcal{N}(0, 1)$, $N = 100$, $C = 150$, $t = 1, \ldots, 1000$. Edge weights were varied as i) $a_{ij}^t \sim \text{Bernoulli}(0.5)$ ii) $a_{ij}^t \sim \text{unif}\{0.5 + 0.5\sin(0.1t),\ 0.5 + 0.5\cos(0.1t),\ e^{-0.01t}\}$, and iii) Non-smooth variations (Fig. 2).

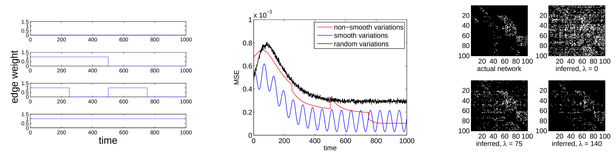

Figure 2: (Left) Non-smooth edge variations. (Center) MSE performance. (Right) Varying λ.

Real datasets:
Popular "memes" on the web were tracked between March 2011 and February 2012 [Rodriguez13]. Two datasets pertaining to the following phrases were used:

1. "Kim Jong-un" ($N = 360$ websites, $C = 466$ cascades, $T = 45$ weeks)
2. "Reid Hoffman" ($N = 125$ websites, $C = 85$ cascades, $T = 41$ weeks)

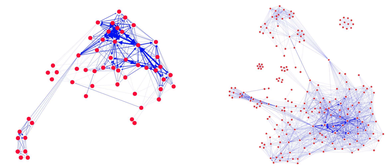

Figure 3: Inferred network for "Kim Jong-un" at (left) $t = 10$, and (right) $t = 40$ weeks.

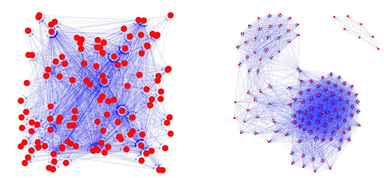

Figure 4: Inferred network for "Reid Hoffman" at (left) $t = 10$, and (right) $t = 40$ weeks.

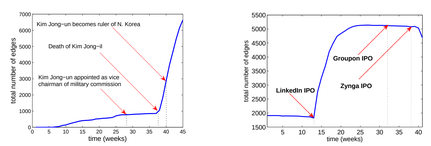

Figure 5: Evolution of total number of inferred edges among websites propagating cascades pertaining to i) (left) "Kim Jong-un", and ii) (right) "Reid Hoffman".

References

[Baingana13] B. Baingana, G. Mateos and G. B. Giannakis, "Dynamic structural equation models for tracking cascades over social networks," *Proc. of NIPS Workshop on Frontiers of Network Analysis*, Lake Tahoe, NV Dec 9, 2013.

[Cai13] X. Cai, J. A. Bazerque, and G. B. Giannakis, "Gene network inference via sparse structural equation modeling with genetic perturbations," *PLoS Comp. Biology*, vol. 9, May 2013.

[Easley10] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. New York, NY: Cambridge University Press, 2010.

[Goldberger72] A. S. Goldberger, "Structural equation methods in the social sciences," *Econometrica*, vol. 40, pp. 979–1001, Nov. 1972.

[Parikh13] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optimization*, vol. 1, pp. 123–231, 2013.

[Rodriguez11] M. G. Rodriguez, D. Balduzzi, and B. Scholkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proc. of 28th Intern. Conf. on Machine Learning*, Bellevue, WA, Jul. 2011.

[Rodriguez13] M. G. Rodriguez, J. Leskovec, and B. Scholkopf, "Structure and dynamics of information pathways in online media," in *Proc. of 6th ACM Intern. Conf. on Web Search and Data Mining*, Rome, Italy, Feb. 2013.

Work supported by the NSF ECCS Grants No. 1202135 and No. 1343248, and the NSF AST Grant No. 1247885.