Learning Granger Graphical Models for Google Flu Trends Data

Pancheewa Arayacheeppreecha

Jitkomut Songsiri

Department of Electrical Engineering

Chulalongkorn University

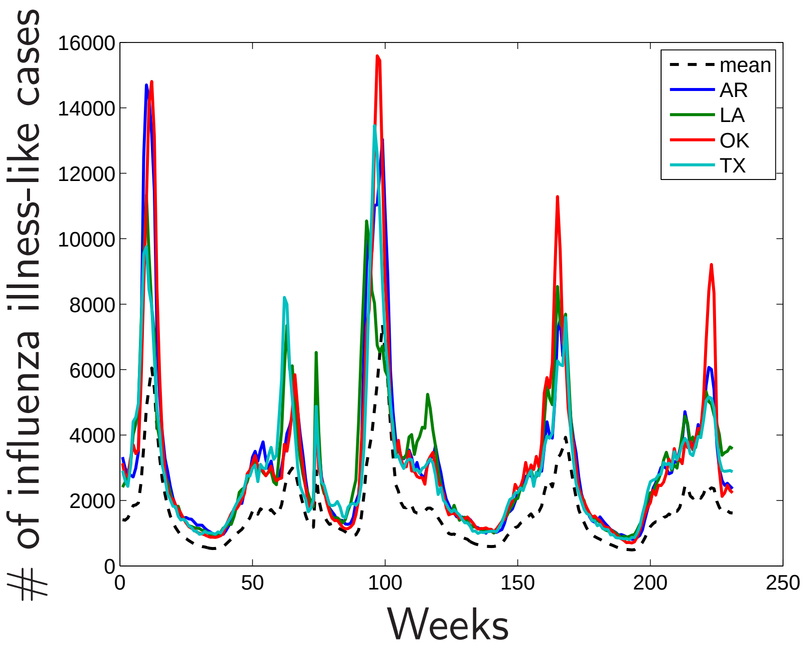
Joint Seminar on Control Systems

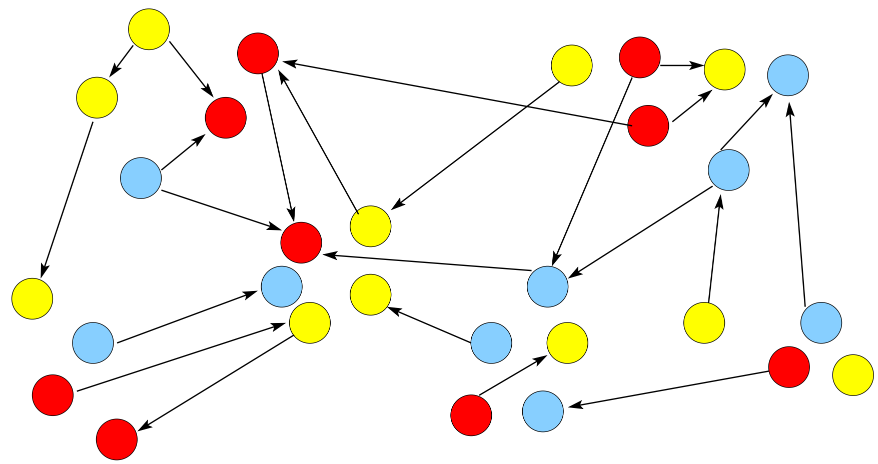
Fri Oct 26, 2012

Outline

- **Granger Graphical Models**
- Sparse multivariate autoregressive models
- Numerical examples

Graphical Models





a graphical model consists of

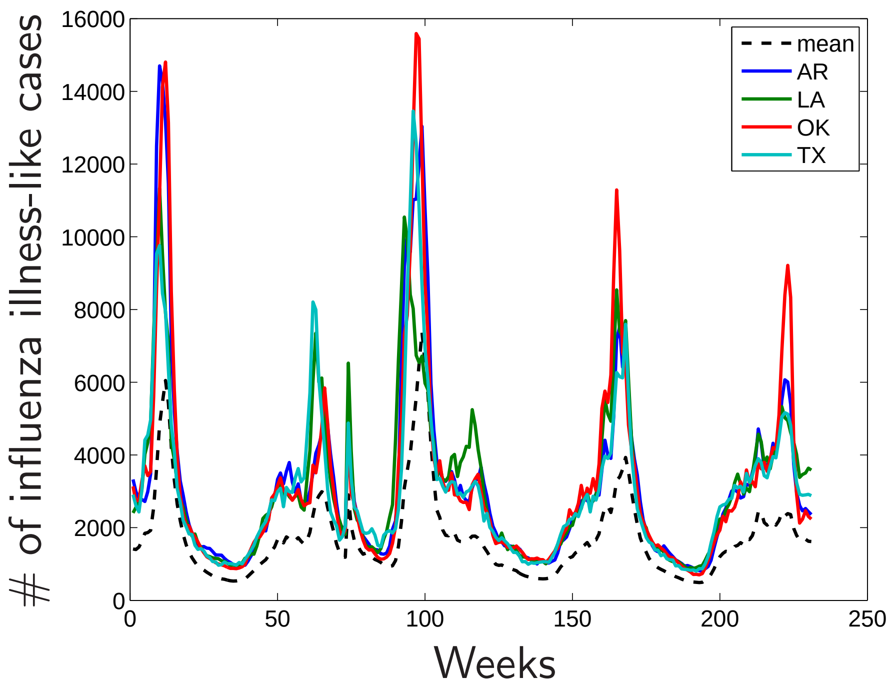
- **nodes:** represent variables of interest
  here the ith node is the number of cases in the ith state
- **edges:** explain relationships between variables

Autoregressive Models

explain a multivariate time series by a vector AR process of order p

$$y(t) = A_1 y(t-1) + A_2 y(t-2) + \cdots + A_p y(t-p) + \nu(t)$$

$y(t) \in \mathbf{R}^n$, $A_k \in \mathbf{R}^{n \times n}$, $k = 1, 2, \ldots, p$, $\nu(t)$ is noise



$n = 51$ (51 states in the U.S.)

y_1	the number of patients in AK
y_2	the number of patients in LA
$\vdots$	$\vdots$
y_{51}	the number of patients in WA

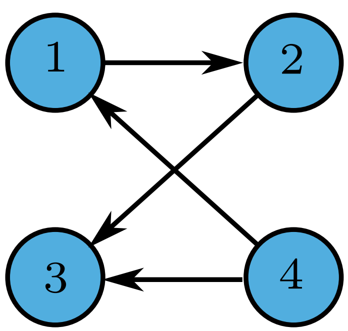
Granger Graphical Models (Granger 1969)

sparsity in coefficients A_k

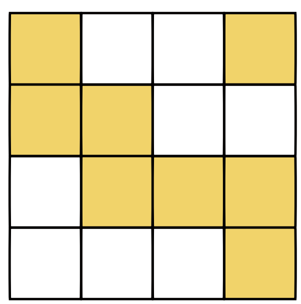
$$(A_k)_{ij} = 0, \quad \text{for } k = 1, 2, \ldots, p$$

is the characterization of **Granger causality** of AR models

- y_i is not *Granger-caused* by y_j
- knowing y_j does not help to improve the prediction of y_i



granger graphical model



zero patterns in A_k

for example, 4-dimensional AR

y_2 is Granger caused by y_1

y_4 is NOT Granger caused by y_2

Outline

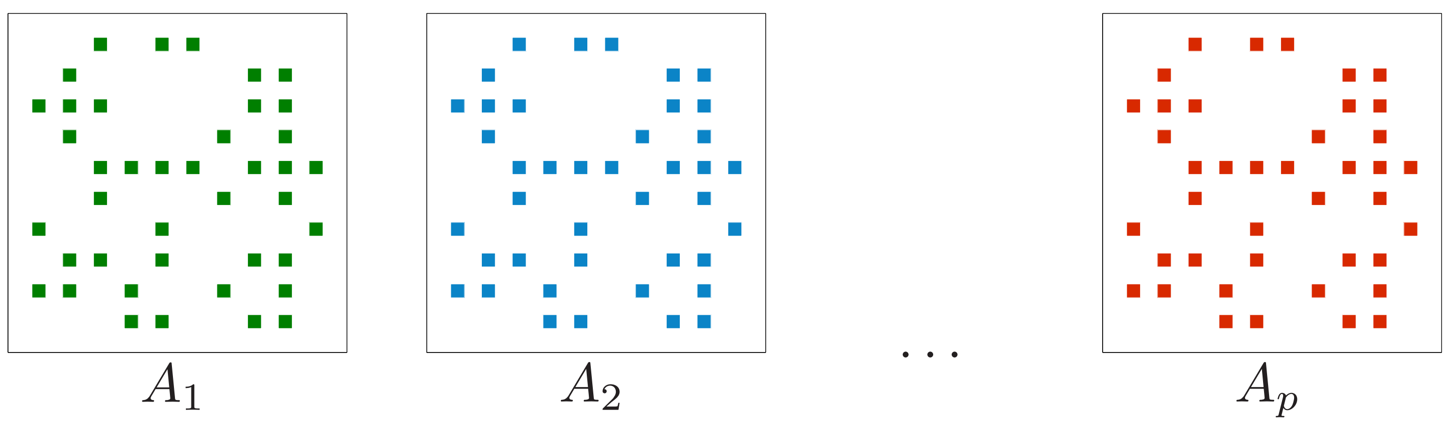
- Granger Graphical Models
- **Sparse multivariate autoregressive models**
- Numerical examples

Sparse Autoregressive (AR) Models

Problem: find A_k's that minimize the sum-square error

$$\sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t-k)\|_2^2$$

- A_k's contain many zeros (to infer Granger causality among variables)
- $A_1, A_2, \ldots, A_p$ have a common zero pattern



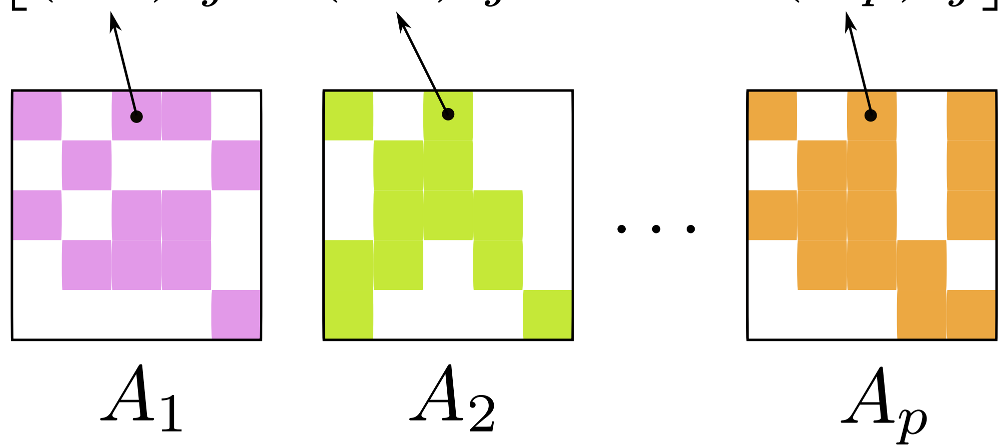
this formulation finds many applications in neuroscience and system biology

(Salvador et al. 2005, Valdes-Sosa et al. 2005, Fujita et al. 2007, ...)

Group sparsity

stack the (i,j) entries of all A_k's in vector $B_{ij} \in \mathbf{R}^p$

$$\begin{bmatrix} (A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix} \triangleq B_{ij}$$



$$\|B_{ij}\|_2 = 0 \quad \Longrightarrow \quad (A_1)_{ij} = (A_2)_{ij} = \cdots = (A_p)_{ij} = 0$$

obtain a group sparsity in A_k's if we can enforce

$$\|B_{ij}\|_2 = 0, \quad \text{or} \quad \| \begin{bmatrix} (A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix} \|_2 = 0$$

for some (i,j)

Estimation problem

given the measurements $y(1), y(2), \ldots, y(N)$

$$\text{minimize} \quad \sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t-k)\|^2 + \rho \sum_{i \neq j} \left\| \begin{bmatrix} (A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix} \right\|_2$$

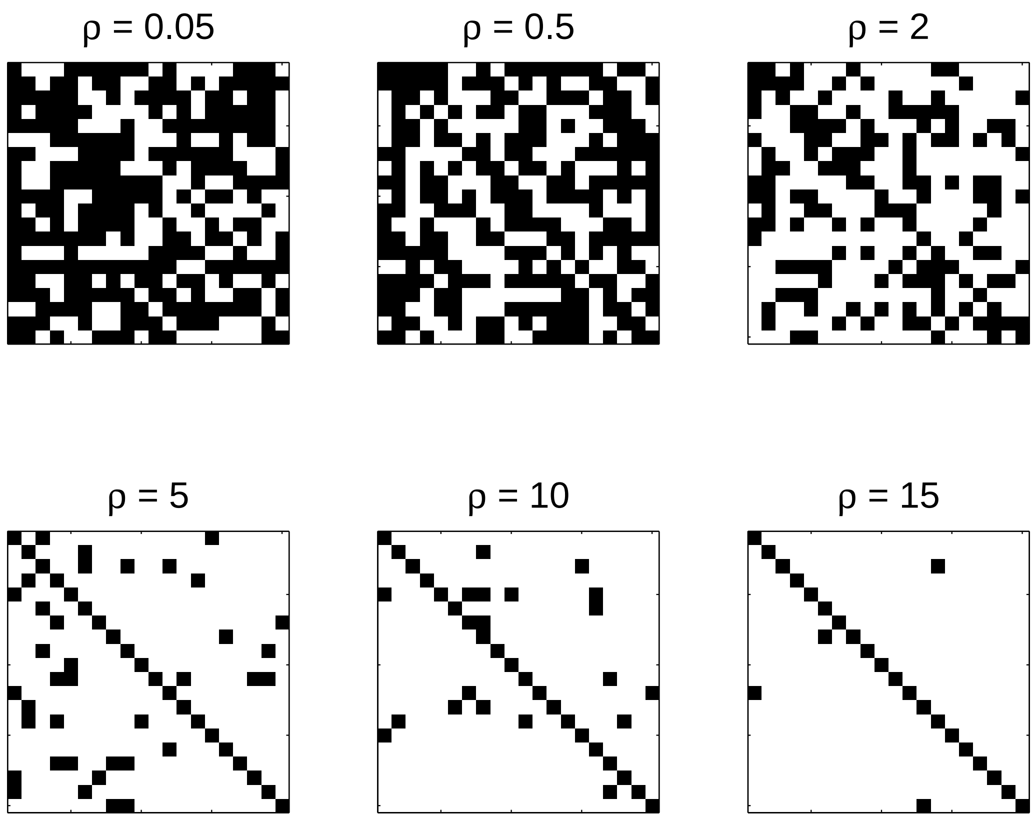
with variables $A_k \in \mathbf{R}^{n \times n}$ for $k = 1, 2, \ldots, p$

- regarded as an ℓ_1-regularized least-squares problem
- summation over (i, j) plays a role of ℓ_1-type norm
- using the ℓ_2 norm of p-tuple of $(A_k)_{ij}$ yields a *group sparsity*
- ρ is called a regularization parameter ($\rho > 0$)

a heuristic convex approach to obtain sparse AR coefficients

Example: 20-dimensional AR of order 3

a common zero pattern of a solution A_1, A_2 and A_3



as ρ increases, A_k's contain more zeros

Numerical method

the estimation problem can be expressed as

$$\text{minimize} \quad f(x) + \rho\|x\|_1$$

- uncontrained convex problem
- *nonsmooth* problem; make it challenging to solve in *large scale*
- suitable for ADMM algorithm; simple and fast in practice

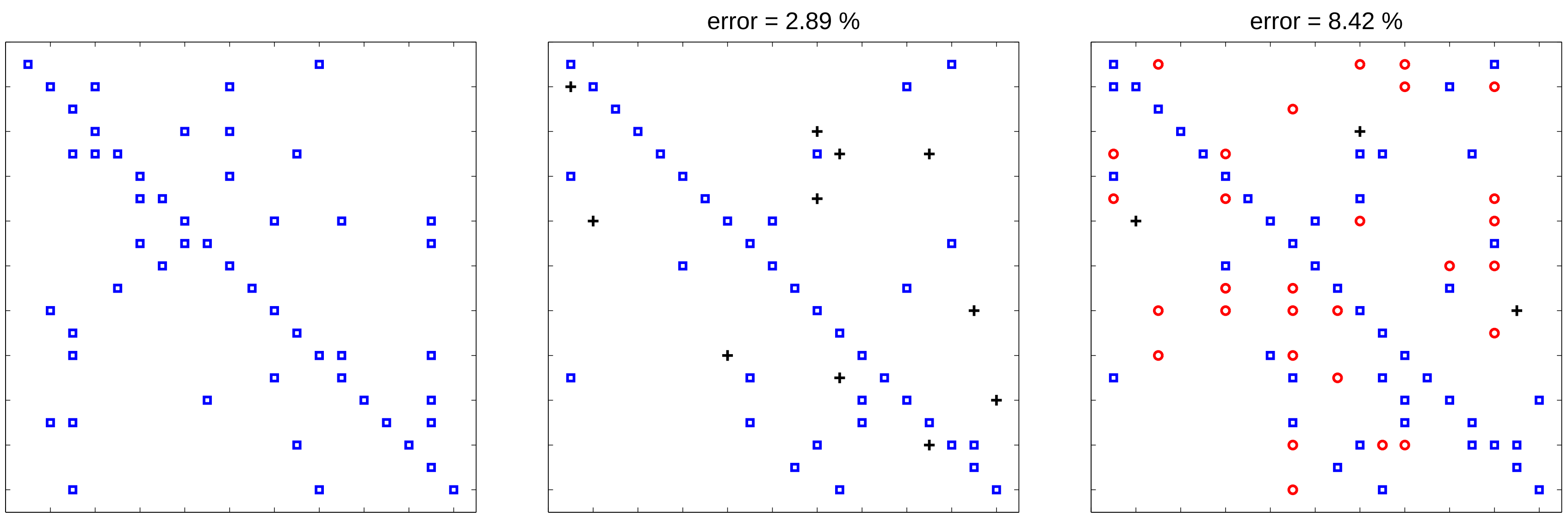
  (S. Boyd, et al. *Distributed optimization and statistical learning via the alternating direction method of multipliers*, 2010)
- many approaches on choosing ρ have been proposed; BIC, cross validation, etc.

Outline

- Granger Graphical Models
- Sparse multivariate autoregressive models
- **Numerical examples**

Sparse AR estimation

generate 1000 time points from a sparse AR process with $n = 20$ and $p = 3$

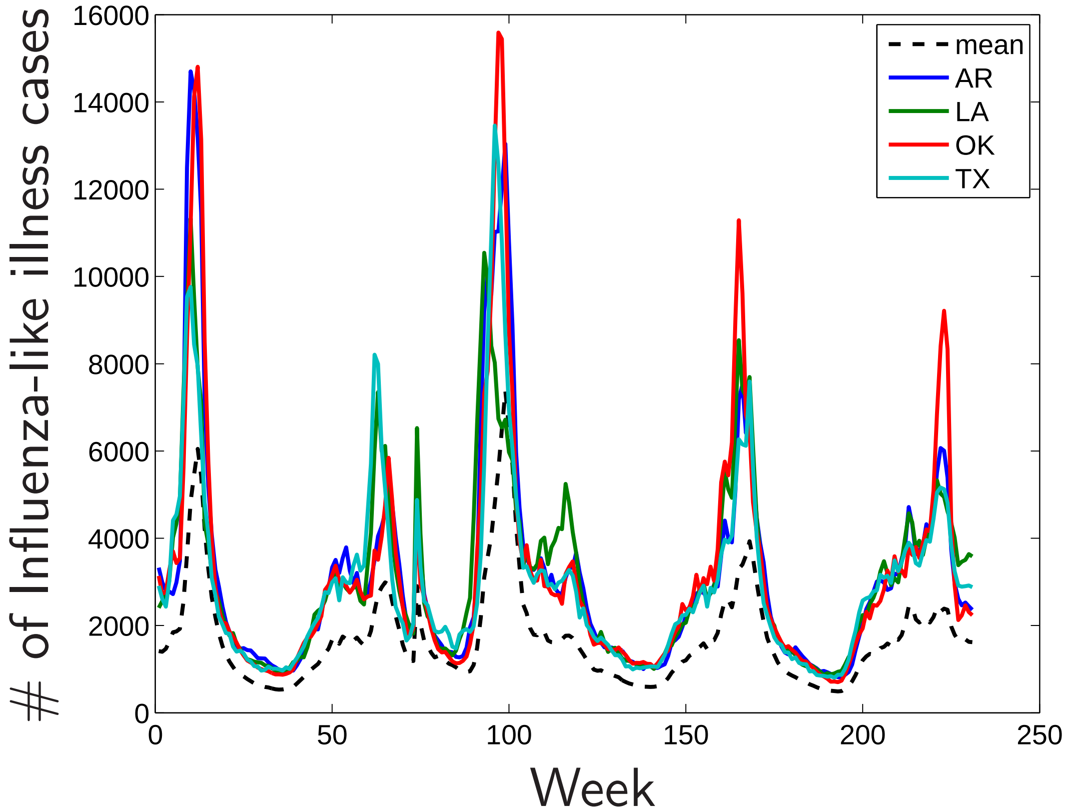


(a) True topology (b) ρ choosen by BIC (c) ρ choosen by Cross validation

- **blue** squares are the correctly identified nonzero entries
- **red** circles are misclassified entries as nonzeros
- **black** crosses are misclassified entries as zeros

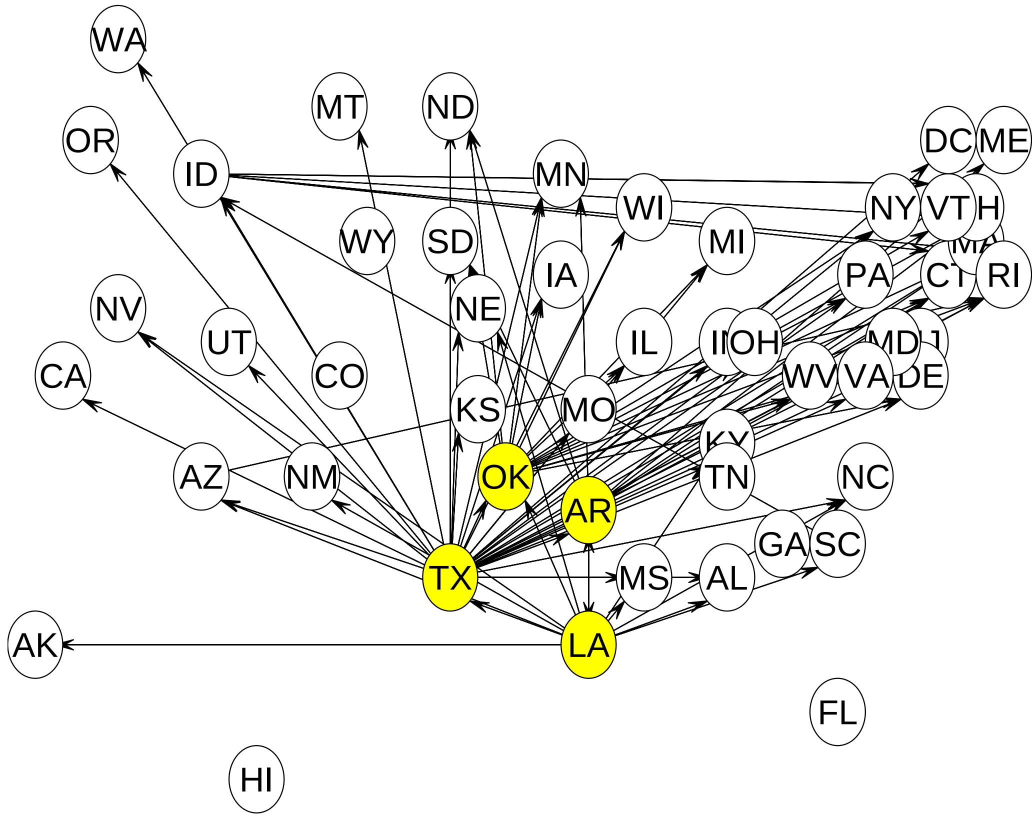
Google Flu Trends

http://www.google.org/flutrends/



- show the number of influenza-like illness (ILI) cases per 100,000 population (estimated by Google)
- Arkansas, Texas, Oklahoma and Louisiana are among the states that have *higher* numbers of ILI cases than the mean value

Graphical Models for Google Flu Trends Data



- TX, OK, LA, and AR have significant influences on many states
- factors such as climate, geography and public health policies can be taken into account to verify this result

Summary

- graphical models are useful for explaining relationships in complex systems

- a problem of learning graph topologies can be formulated as a sparse identification problem

- to obtain a sparse model, we add an ℓ_1-type regularization to the estimation problem

- the resulting problem is unconstrained convex but nondifferentiable

- solving the problem in large scale is done by ADMM algorithm (shown to be efficiently fast in many applications)