

Learning Sparse Markov Network Structure via Ensemble-of-Trees Models

Yuanqing Lin, Shenghuo Zhu

NEC Laboratories America



Empowered by Innovation

Daniel D. Lee^a, Ben Taskar^b

^a Department of Electrical and Systems Engineering

^b Department of Computer and Information Science

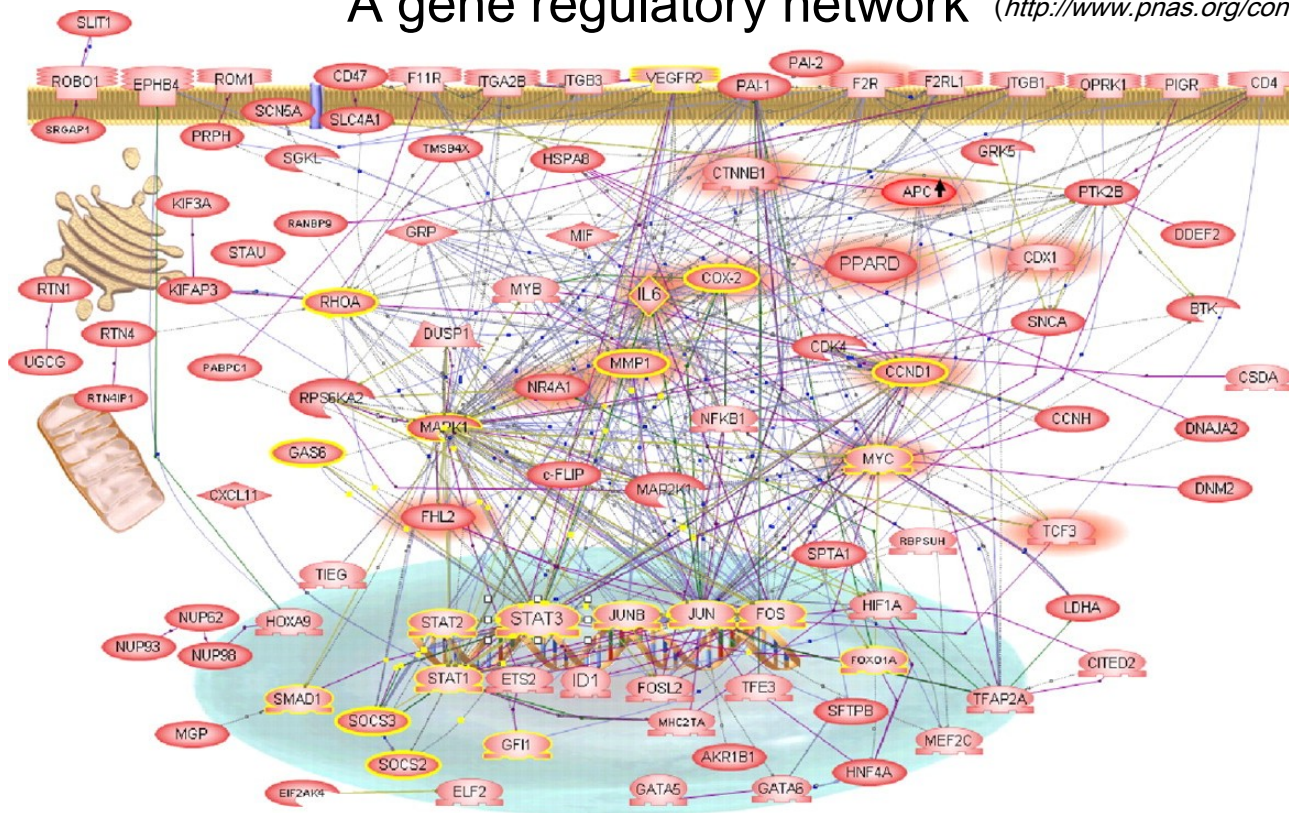
University of Pennsylvania



PENN

Introduction (1) -- why sparse Markov Network

A gene regulatory network (<http://www.pnas.org/content/104/31/12890/F2.large.jpg>)



◆ Sparse structure:

- 1) real world are sparsely connected (e.g. genes, people, words)
- 2) sparseness in Markov network encodes conditional independence
- 3) discovering sparse patterns can be important for knowledge discovery

Introduction (2) -- existing work

❖ low tree-width Markov network

- Tree Markov network (Chow & Liu, 1968)
- Thin tree width (Bach & Jordan, 2001; Srebro, 2001; Chechetka & Guestrin, 2007)

😊 Exact inference (but learning is approximate and very expensive)

❖ Gaussian Markov networks

$$\mathbf{X}^* = \arg \min_{\mathbf{X} \succ 0} -\log \det(\mathbf{X}) + \text{Tr}(\mathbf{S}^T \mathbf{X}) + \lambda \sum_{u \neq v} |X_{uv}|$$

(Banerjee et al., 2006; Friedman et al., 2007)

😊 convex, easy to optimize (1000x1000 problem solved in minutes)

☹ limited in modeling practical data

❖ General Markov networks

see next two slides ...

Introduction (3) -- existing work

MAP estimation with Laplacian sparsity prior:

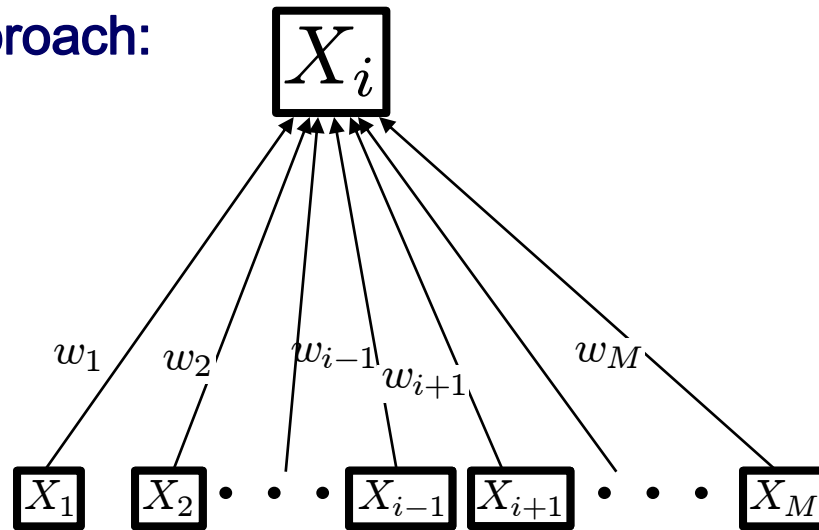
$$\alpha^* = \underbrace{- \sum_{i=1}^N \sum_{u,v} \alpha_{u,v} f(x_u^{(i)}, x_v^{(i)}) + N \log[Z(\alpha)]}_{\text{negative log-likelihood}} + \underbrace{\lambda \sum_{u,v} |\alpha_{u,v}|}_{l_1\text{-norm sparsity regularization}}$$

Lee, S.-I., Ganapathi, V., & Koller, D. (2007)

- ☺ Optimization is convex - general extension of Gaussian MRF
- ☹ Partition function (and inference) is intractable in general

Introduction (4) -- existing work

Pseudo-likelihood approach:



$$\mathbf{w}_{/i}^* = \arg \min_{\mathbf{w}_{/i}} \underbrace{-\log P(x_i | \mathbf{x}_{/i}, \mathbf{w}_{/i})}_{\text{pseudo negative log-likelihood}} + \underbrace{\lambda |\mathbf{w}_{/i}|}_{l_1\text{-norm sparsity regularization}}$$

pseudo negative log-likelihood

l_1 -norm sparsity regularization

Wainwright, M. J., Ravikumar, P., & Lafferty, J. (2006).

- ☺ Each individual optimization is convex (and simple)
- ☹ Pseudo-likelihood: requires a large amount of data for achieving good estimates

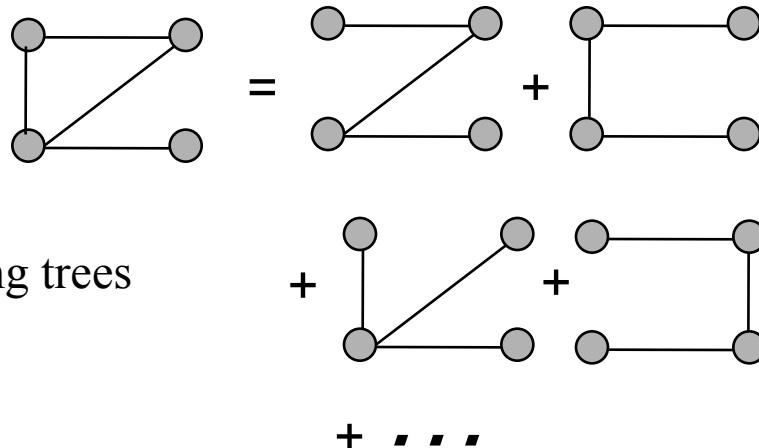
Ensemble-of-trees (ET) model (1)

-- probabilistic model

Data likelihood:

$$P(\mathbf{x}) = \sum_T P(\mathbf{x}|T)P(T)$$

sum over *all possible* spanning trees



Tree probability:

$$P(T) = \frac{1}{Z} \prod_{\{u,v\} \in T} \beta_{u,v}$$

$$\text{where } Z = \sum_T \prod_{\{u,v\} \in T} \beta_{u,v}$$

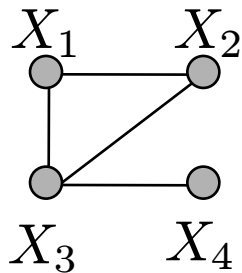
◆ ET model:

- ▶ mixture of (all possible spanning) trees model
- ▶ partition function and data likelihood in closed form

ET model (2) -- partition function

Partition function: $Z = \sum_T \prod_{\{u,v\} \in T} \beta_{u,v}$

Matrix tree theorem:



$$B(\beta) = \begin{bmatrix} \boxed{Q(\beta)} & & & & \\ \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 3 \end{bmatrix} & \begin{matrix} 0 \\ 0 \\ 1 \end{matrix} & & & \\ \hline 0 & 0 & -1 & 1 & \end{bmatrix}$$

$$\beta_{u,v} = \begin{cases} 1 & \text{if } \{u,v\} \text{ is an edge} \\ 0 & \text{otherwise} \end{cases}$$

$$B_{u,v} = \begin{cases} -\beta_{u,v} & \text{if } u \neq v \\ \sum_{k \neq u} \beta_{u,k} & \text{if } u = v \end{cases}$$

$$Z = \det[Q(\beta)]$$

◆ Generalized matrix tree theorem:

$\beta_{u,v}$ can be any nonnegative number

☺ Partition function: has a closed form

◆ β : encodes the structure of a graph

ET model (3) -- data likelihood

Data likelihood:

$$\begin{aligned} P(\mathbf{x}) &= \sum_T P(\mathbf{x}|T) P(T) \\ &= \sum_T \prod_{\{u,v\} \in T} w_{u,v} \prod_u w_{0u} \frac{1}{Z} \prod_{\{u,v\} \in T} \beta_{u,v} \\ &= \prod_u w_{0u} \frac{1}{Z} \sum_T \prod_{\{u,v\} \in T} w_{u,v} \beta_{u,v} \\ &= \frac{\det[\mathbf{Q}(\boldsymbol{\beta} \otimes \mathbf{w})]}{\det[\mathbf{Q}(\boldsymbol{\beta})]} \prod_u w_{0u} \end{aligned}$$

Data likelihood on a tree:

$$\begin{aligned} P(\mathbf{x}|T) &= \prod_{\{u,v\} \in T} \frac{P(X_u = x_u, X_v = x_v)}{P(X_u = x_u)P(X_v = x_v)} \prod_u P(X_u = x_u) \\ &= \prod_{\{u,v\} \in T} w_{u,v} \prod_u w_{0u} \end{aligned}$$

☺ Data likelihood: has a closed form

ET model (4) -- ML estimation

Given data $\{\mathbf{x}^i\}_{i=1}^N$

$$\beta^* = \arg \min_{\beta} - \sum_{i=1}^N \log \det[\mathbf{Q}(\beta \otimes \mathbf{w}_i)] + N \log \det[\mathbf{Q}(\beta)]$$

Subject to: $\beta_{u,v} \geq 0$

$$\sum_{u,v:u \neq v} \beta_{u,v} = 1$$

- ◆ **Marginals (and \mathbf{w}_i):** estimated directly from data
- ◆ **Sparse solutions:** tree number regularization

ET model (5) -- nonconvex optimization

Initialization (by solving a convex upper bound):

$$\beta^* = \arg \min_{\beta} - \sum_{i=1}^N \log \det[\mathbf{Q}(\beta \otimes \mathbf{w}_i)] + c$$

$$\text{Subject to: } \beta_{u,v} \geq 0$$

$$\sum_{u,v:u \neq v} \beta_{u,v} = 1$$

◆ **Optimizations:** solved by projected gradient descent (PGD)

ET model (6) -- nonconvex optimization (cont'd)

Projected gradient descent :

$$\begin{aligned}\tilde{\beta} &\longleftarrow \beta^k - \lambda \nabla \beta \\ \beta^{k+1} &\longleftarrow \mathcal{P}(\tilde{\beta})\end{aligned}$$

Step size λ : determined by Armijo's rule

Projection operation $\mathcal{P}(\tilde{\beta})$: need to solve a quadratic programming

$$\beta^{k+1} = \arg \min_{\beta} \|\beta - \tilde{\beta}\|_2^2$$

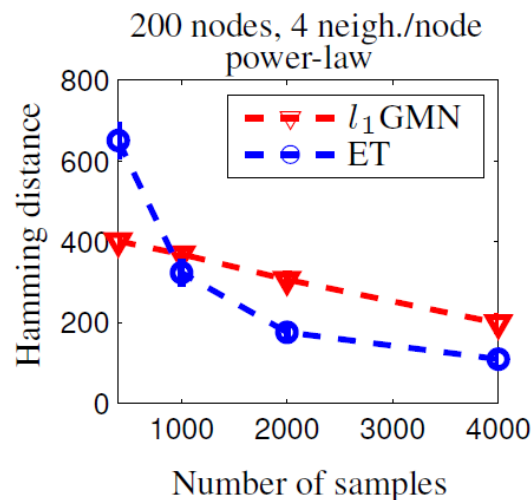
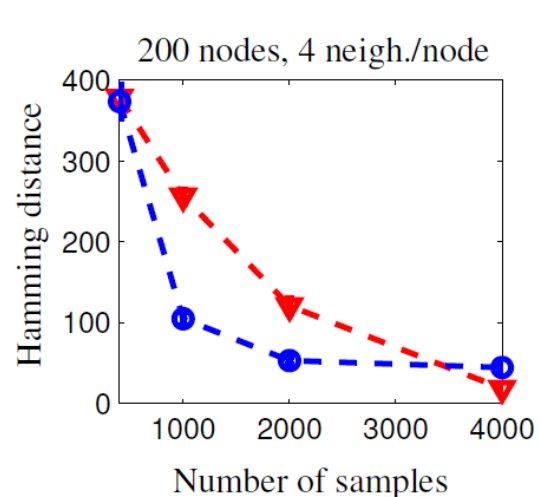
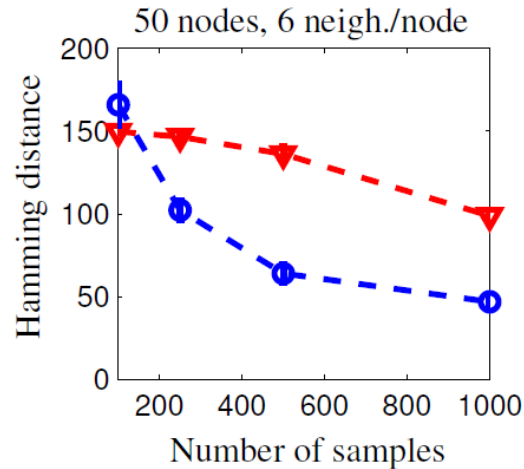
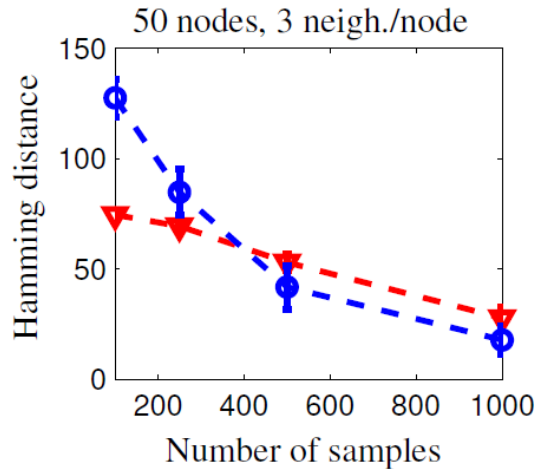
$$\text{Subject to: } \beta_{u,v} \geq 0$$

$$\sum_{u,v:u \neq v} \beta_{u,v} = 1$$

◆ **PGD (with Armijo's rule):**

-- guaranteed to converge to a local optimizer

Simulations (1) -- Gaussian Markov network



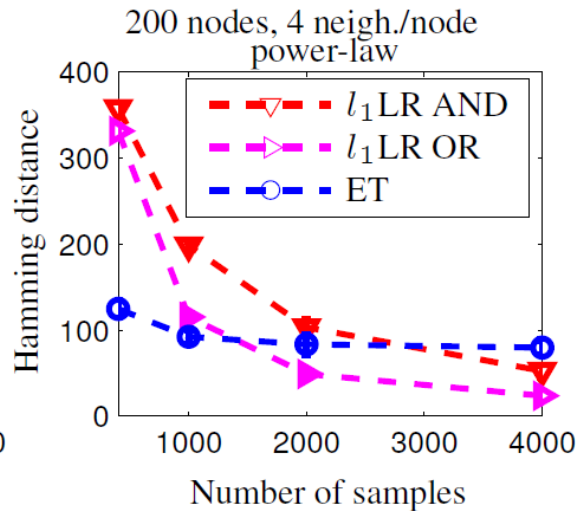
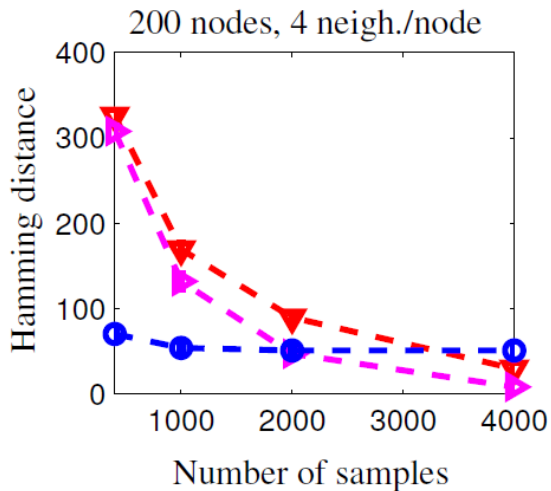
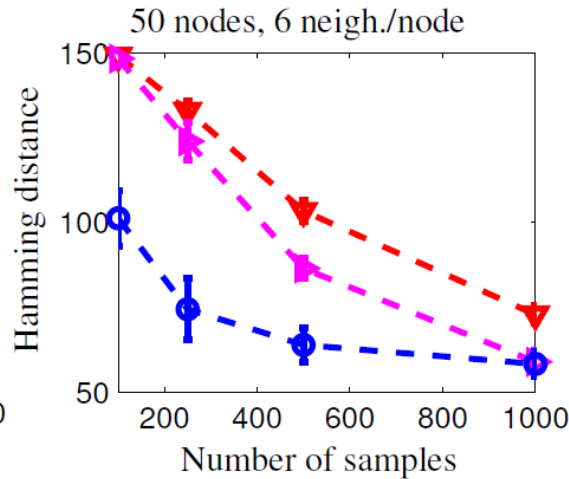
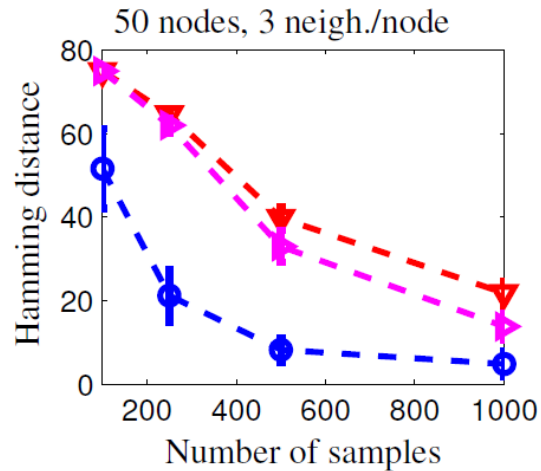
◆ l_1 -norm regularized GMN vs. ET model:

1) ET provides better performance when only a small number of data samples is available

2) ET does NOT need to assume Gaussian model

3) ET is free of l_1 -norm regularization

Simulations (2) -- binary Markov network



◆ ET model vs. pseudo-likelihood approach:

Again, ET provides better performance when only a small number of data samples is available

Results of learning a word network

apple	iphone	mac	ipod	itunes	store	nasdaq
iphone	apple	apple	iphone	apple	shop	newsletter
mac	ipod	os	apple	download	retail	e-mail
microsoft	mobile	pc	touch	music	apple	analyst
ipod	device	windows	itunes	ipod	purchase	blackberry
itunes	phone	macbook	music	amazon	music	expectation
store	rumor	user	device	podcast	price	trend
pc	blackberry	hardware		song	iphone	download
user	smartphone	computer		dvd	library	letter
device	pc	leopard		stream	item	print
hardware	gps	desktop			selling	network
computer	user	safari			amazon	directory
macbook	nokia	pro			storage	responsibility
price	store	graphics			backup	hp
upgrade	wireless	vista			package	portfolio
mobile	gadget					demonstration
os	app					apple
leopard	australia					microsoft
nasdaq	carrier					exchange
safari	samsung					
maker	touch					
apps	launch					
smartphone						
competitor						

◆ **ET model:** is able to discover meaningful word network from documents

Summary

- ✚ The ET model provides a novel paradigm for MRF structure learning
 - ❖ both partition function and data likelihood are tractable
 - ❖ ET model is versatile for different distributions: continuous or discrete, Gaussian or non-Gaussian, different tree-width ...
 - ❖ Empirical results show that ET model provides competitive performance compared to the state-of-art methods
- ✚ We developed a projected gradient algorithm for efficiently solving the optimization arising in the ET model
- ✚ Future work: theoretical analysis, more efficient algorithm (especially stochastic algorithm) for the ET model