Transfer Learning for Latent Variable Network Models

Akhil Jalan

Department of Computer Science, UT Austin

November 11, 2024

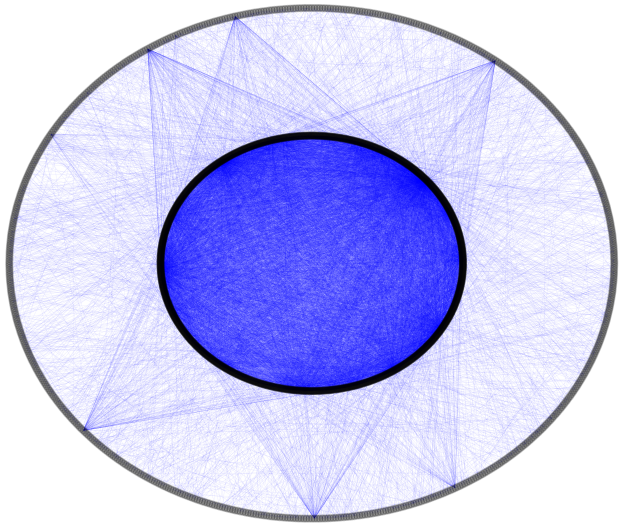


Joint with A. Mazumdar, S. Mukherjee & P. Sarkar.

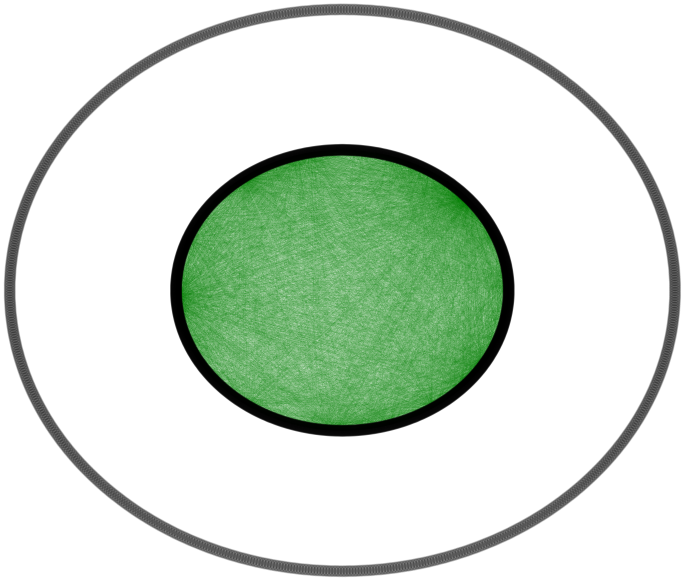
Motivation

Motivation: Biological Networks

Escherichia coli W, $n = 1973$



Pseudomonas putida, $n = 1973$



In *metabolic networks*, can only test edges between prepared set of metabolites. [1]

[1] Christensen, Bjarke, and Jens Nielsen. "Metabolic network analysis: a powerful tool in metabolic engineering." Bioanalysis and Biosensors for Bioprocess Monitoring (2000): 209-231.

Introduction

Latent Variable Model

Goal. Estimate a *target function* $f_Q : \mathcal{X} \times \mathcal{X} \to [0,1]$ for $\mathcal{X} \subset \mathbb{R}^d$ compact. (Call f_Q a **latent variable network model**.)
These generalize:

- Stochastic Block Models (SBMs)
- Mixed-Membership Stochastic Block Models
- Generalized Random Dot Product Graphs
- Graphons

Ordinary Network Estimation

For latents $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} \mathcal{X}$, observe:

$$\forall i, j : f_Q(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

And output $\hat{Q} \in [0,1]^{n \times n}$.

Limited Target Data

Our Setting. For $S \subset \{1, 2, \ldots, n\}$ with $|S| := n_Q << n$, observe:

$$\forall i, j \in S : f_Q(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Notice: We cannot do better than $\Omega(1)$ error without additional information.

Transfer Setting

Our Setting (Formal)

For source f_P, target f_Q, and $S \subset \{1, 2, \ldots, n\}$ chosen uniformly at random with $|S| := n_Q << n$, observe:

$$\forall i, j \in S : \text{Bernoulli}(f_Q(x_i, x_j))$$
$$\forall i, j \in [n] : \text{Bernoulli}(f_P(x_i, x_j))$$

Let $Q \in \mathbb{R}^{n \times n}$, $Q_{ij} = f_Q(x_i, x_j)$. Output $\hat{Q}$ to minimize:

$$\text{Mean Squared Error}(Q, \hat{Q}) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (Q_{ij} - \hat{Q}_{ij})^2$$

(All graphs are undirected so $f(x, y) = f(y, x)$ always.)

Results

Relation between source and target

Rankings Assumption at quantile h_n

(P, Q) satisfy the rankings assumption at quantile $h_n = o(1)$ if $\forall i, \forall j \neq i$

$$j \in \{ \text{ bottom } h_n\text{-quantile of } P\text{'s graph distance}(i, \cdot)\}$$

$\Rightarrow$

$$j \in \{ \text{ bottom } O(h_n)\text{-quantile of } Q\text{'s graph distance}(i, \cdot)\}$$

Topologically, we require 2-hop neighborhoods to be similar.

Algorithm

Data for our algorithm.

- Source $A_P \in \{0,1\}^{n \times n}$, with $A_{P;ij} \sim \text{Bernoulli}(P_{ij})$
- Target $A_P \in \{0,1\}^{n_Q \times n_Q}$, with $A_{Q;ij} \sim \text{Bernoulli}(Q_{ij})$ for $i, j \in S$.

Idea. Use the rankings relationship to compute neighborhoods in P, then do regression over Q.

Theorem

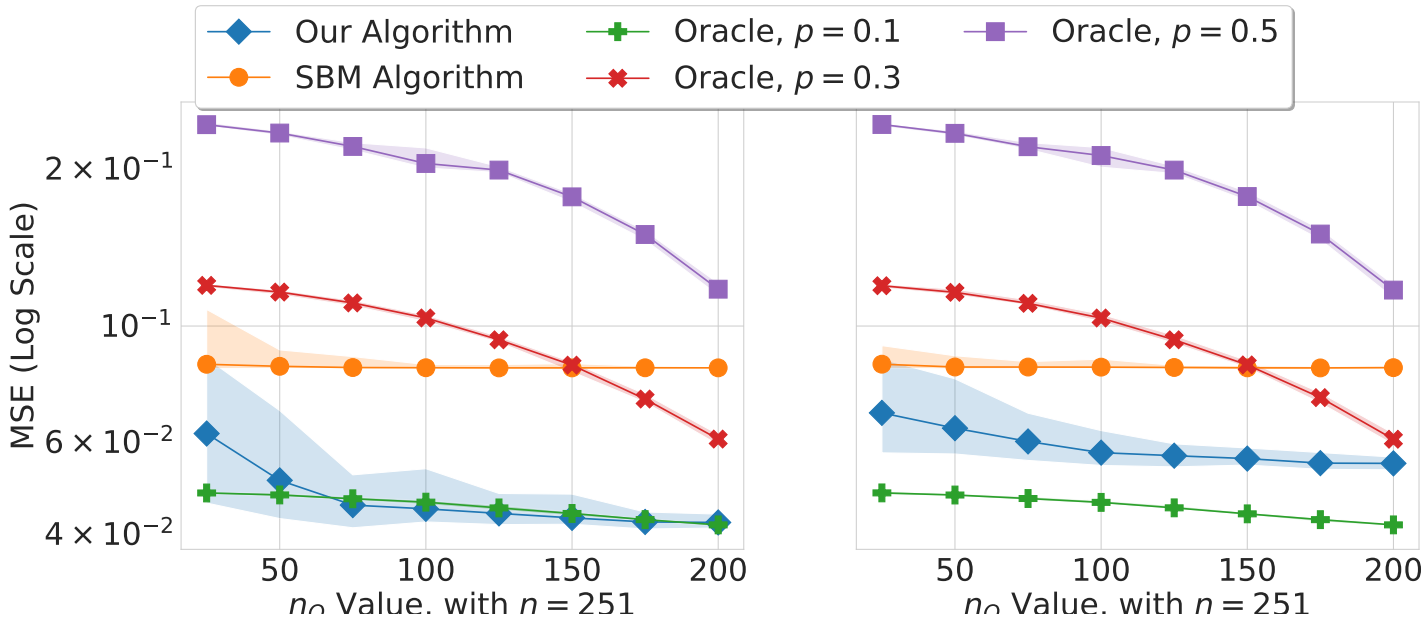
Theorem (Informal)

There exists an efficient algorithm such that, if P, Q satisfy rankings assumption and f_P, f_Q are Hölder smooth, outputs $\hat{Q} \in \mathbb{R}^{n \times n}$ such that:

$$\mathbb{P}\left[\text{Mean Squared Error}(Q, \hat{Q}) \leq \frac{1}{n_Q^{\Omega(1)}}\right] \geq 1 - \frac{1}{n_Q^{\Omega(1)}}$$

Error rates depend on: Hölder smoothness of P, Q, dimension d, and $\log n$.

Metabolic Network Estimation



Estimating metabolic network of iJN1463 (Pseudomonas putida).

- Left: Source iWFL1372 (Escherichia coli W)
- Right: Source iPC815 (Yersinia pestis).

Conclusion

Future Work

Directions for future work:

- Very sparse input graphs (need a different graph distance at edge density $n^{-1/2}$)
- Multiple sources with different guarantees
- Minimax lower bounds for latent variable models
- Incorporating side information in specific applications (e.g. bioinformatics)

Thank you!

akhiljalan@utexas.edu