Consistent Nonparametric Methods for Network Assisted Covariate Estimation

Xueyu Mao

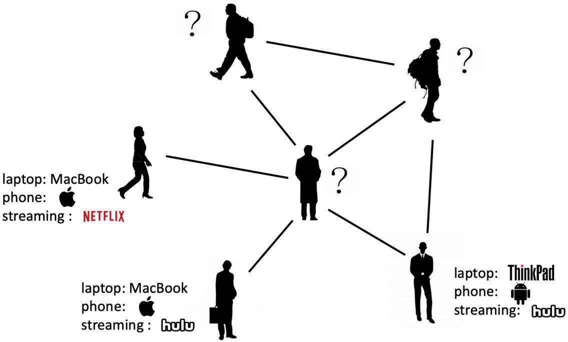
Department of Computer Science
The University of Texas at Austin

International Conference on Machine Learning, 2021

Joint work with **Deepayan Chakrabarti** and **Purnamrita Sarkar**

Node Covariate Estimation

- Example: in a social network, we have information on some people's interests
  - desired products
  - preferred news topics
  - sporting interests



- **Problem: Can we infer such information for the rest from a few people's known interests and the structure of the social network?**
  - Application: content and ad targeting, friend and group recommendations, etc.

Model

Latent Variable Models:

- Each node $i \in [n]$ has latent vector $\mathbf{z}_i$
- Link probabilities:

$$\mathbf{P}_{ij} = \rho_n f(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\Theta}) \quad \text{for all } i \neq j$$

- Network:

$$\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{P}_{ij}) \quad \text{for all } i \neq j$$

- Node Covariate:

$$\mathbf{X}_i = g(\mathbf{z}_i) + \epsilon_i$$

- **Problem: given node covariates $\{\mathbf{X}_i, i \in S\}$ for a subset of nodes S and the adjacency matrix $\mathbf{A}$, infer the node covariates of the remaining nodes $\{\mathbf{X}_i, i \in [n] \setminus S\}$.**

Main Contribution

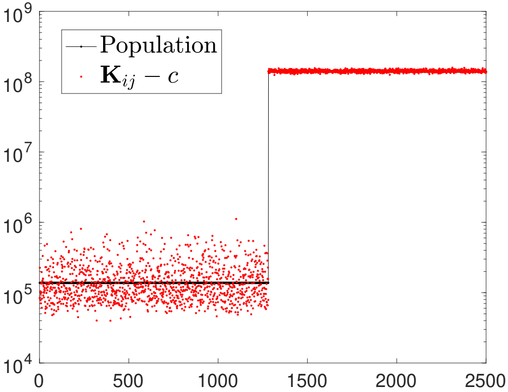
- With no information on link function f:
  - Propose a *model-agnostic* algorithm (CN-VEC) to estimate the node covairates by k nearest-neighbor regression, where nearest neighbors are chosen by a carefully designed similarity measure.
    - Provably consistent
    - Needs no fine-tuning
    - Much faster than embedding methods, with comparable or better performance
- When f makes $\mathbf{P}$ low rank:
  - Provide a nonparametric algorithm (SVD-RBF), which is provably consistent for sparser graphs.

Model-Agnostic Algorithm

- Construct a new similarity measure

$$\mathbf{K}_{ij} = \sum_{k \neq i;j} \left[(\mathbf{C}_{ik}^2 \quad 2)1(\mathbf{C}_{ik} \quad 2) + (\mathbf{C}_{jk}^2 \quad 2)1(\mathbf{C}_{jk} \quad 2) \quad 2\mathbf{C}_{ik}\mathbf{C}_{jk}\right].$$

  - $\mathbf{C}_{ij}$ is the number of common neighbors between i and j



Model-Agnostic Algorithm

- Algorithm summary:

CN-VEC

for $i \in [n] \setminus S$

- $dist(j) \leftarrow \mathbf{K}_{ij}$, for $j \in S$
- $top_k(i) \leftarrow k$ nodes with the smallest values of $dist(j)$
- $\hat{\mathbf{X}}_i \leftarrow \frac{1}{k} \sum_{j \in top_k(i)} \mathbf{X}_j$

- We prove weak consistency result on CN-VEC when average degree grows faster than $n^{1/3}$
  - $\mathbf{C}_{ij}$ only works when average degree grows faster than $\sqrt{n}$

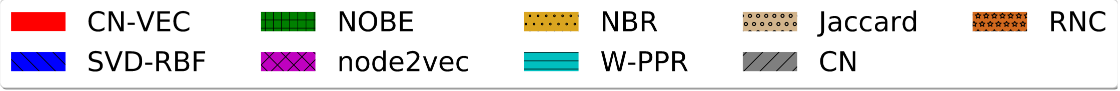
Algorithm for low-rank models

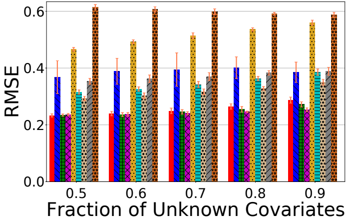
SVD-RBF

- $\hat{\mathbf{U}} \leftarrow$ top-d eigenvector matrix for $\mathbf{A}$
- $\hat{\phi}_i \leftarrow i^{th}$ row of $\hat{\mathbf{U}}|\hat{\mathbf{E}}|^{1/2}$
- for $i \in [n] \setminus S$
  - $dist(j) \leftarrow \|\hat{\phi}_i - \hat{\phi}_j\|$ for $j \in S$
  - $\hat{\mathbf{X}}_i \leftarrow \frac{\sum_{j \in S} K_\theta(\hat{\phi}_i, \hat{\phi}_j)\mathbf{X}_j}{\sum_{j \in S} K_\theta(\hat{\phi}_i, \hat{\phi}_j)}$, where $K_\theta(\mathbf{v}_1, \mathbf{v}_2) = \exp\left(-\frac{\|\mathbf{v}_1 - \mathbf{v}_2\|^2}{2\theta^2}\right)$

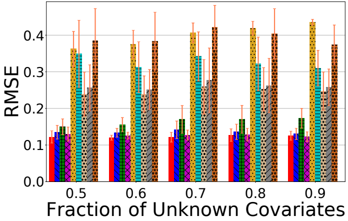
- We prove uniform consistency result on SVD-RBF when average degreee grows faster than $\mathcal{O}(\log n)$

Simulation Experiments

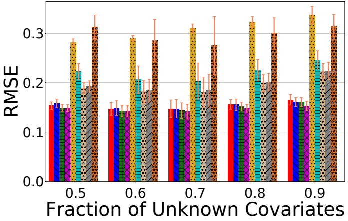




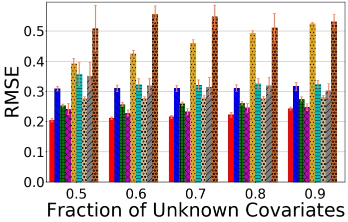
(a) LSM



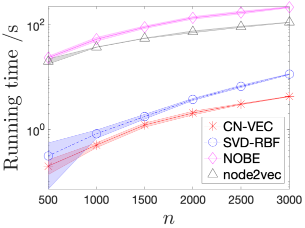
(b) SBM



(c) MMSB



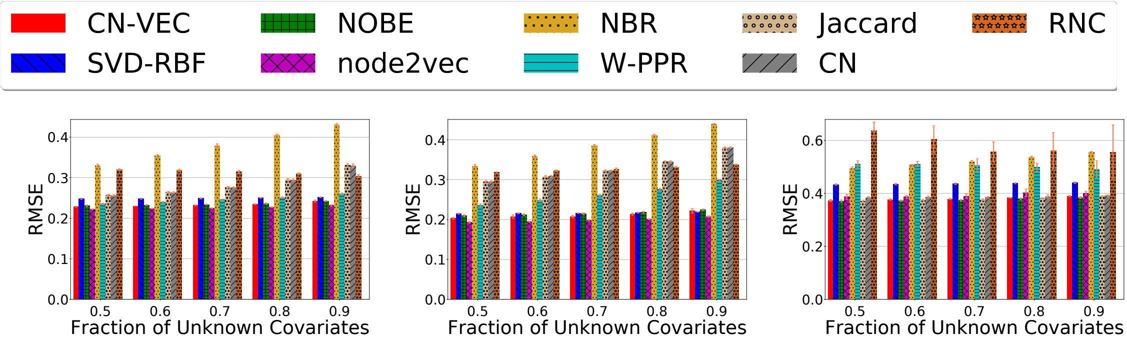
(d) RDPG



(e) Running time (log scale)

Real-world Network Results

- Citation networks (Cora and CiteSeer), and social network (Sinanet)
- Use topic distribution as node covariate



(a) Cora (b) CiteSeer (b) Sinanet

Thanks!