The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

Consistent Nonparametric Methods for Network Assisted Covariate Estimation

Xueyu Mao, Deepayan Chakrabarti, Purnamrita Sarkar

The University of Texas at Austin McCOMBS SCHOOL OF BUSINESS

The University of Texas at Austin
Department of Statistics and Data Sciences
College of Natural Sciences

Node Covariate Estimation

- Example: in a social network, each person has a vector of interests
  - desired products
  - preferred news topics
  - sporting interests
- Know a few people's interest vector from, say, their past tweets or comments. Unavailable or insufficient for others.
- **Problem: Can we infer their interests from a few people's known interests and the structure of the social network?**

Model

Latent Variable Models:

- Each node $i \in [n]$ has latent vector $\mathbf{z}_i \in \mathbb{R}^d$
- Network:

$$\mathbf{P}_{ij} := \mathrm{P}(\mathbf{A}_{ij} = 1 | \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n) = \rho_n f(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\Theta}) \quad \text{for all } i \neq j$$

  - $f(\cdot)$ is bounded in $[0,1]$ and has parameters $\boldsymbol{\Theta}$
  - $\rho_n = o(1)$ controls the sparsity of the graph
- Node Covariate:

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

  - $g : \mathbb{R}^d \to \mathbb{R}^p$ is bounded, and is Lipschitz
  - $\boldsymbol{\epsilon}_i$ are i.i.d. with uncorrelated elements, whose mean is 0 and variance is σ^2
- **Problem: given node covariates $\{\mathbf{X}_i \in \mathbb{R}^p; i \in S\}$ for a subset of nodes S, infer the node covariates of the remaining nodes $\{\mathbf{X}_i; i \in [n] \setminus S\}$.**

Related work

- Node similarity measures
  - Common neighbors and its weighted variants (Adamic/Adar), preferential attachment, resource allocation, Katz index, PageRank, SimRank, and graph neural networks, etc.
  - Only a few have consistency guarantees (**sarkar2011theoretical**; **sarkarchak2015**).
  - Our method constructs a similarity measure that provably works in sparser settings.
- Node classification
  - Methods based on random walks: label propagation, personalized PageRank, partially absorbing random walks, etc.
  - Node embeddings: represent nodes by vectors while retaining some network-based properties: DeepWalk, LINE, node2vec, spectral embedding etc.
    - Can train a classifier to predict the unseen labels with the embedding vectors.
    - Typically lack provable guarantees, but often work well in practice.
    - Spectral embedding is well studied and has provable guarantees, but limited to low-rank models.
    - Our method does not need low-rank assumption.

Model-Agnostic Algorithm

- Nonparametric estimator: $\hat{\mathbf{X}}_i = \dfrac{\sum_{j \in top_k(i)} \mathbf{W}_{ij} \cdot \mathbf{X}_j}{\sum_{j \in top_k(i)} \mathbf{W}_{ij}}$
  - $\mathbf{W}_{ij}$: measure of similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$
    - Adjacency matrix $\mathbf{A}$: average of neighbors, but cannot distinguish in-cluster and out-of-cluster edges for stochastic blockmodel (SBM)
    - Common neighbor matrix $\mathbf{C}$: works only when average degree larger than $\sqrt{n}$
    - Distances between rows of $\mathbf{C}$: goes beyond common neighbors and can work for sparser graph – our proposed work
  - $top_k(i)$: set of nodes $j \in S$ with the largest $\mathbf{W}_{ij}$ values.
- Construct a new similarity measure

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

  - We prove that when average degreee grows faster than $n^{1/3}$, it concentrates to $\left(\sum_{k \neq i,j} \left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2\right)$, and can be used to pick nearest neighbors.
  - Proof sketch:
    - when $n\rho^2 \to 0$, $\mathrm{E}[\mathbf{C}_{ik}] \approx 0$ and $\mathbf{C}_{ik}$ does not concentrate
    - $\mathbf{C}_{ik}$ is well-approximated by a Poisson random variable with rate $\lambda_{ik} = (\mathbf{P}^2)_{ik} = O(n\rho^2)$.
    - indicator $1(\mathbf{C}_{ik} \geq 2)$ is true when $\mathbf{C}_{ik} = 2$ with probability $\approx \lambda_{ik}^2/2$, and $\mathbf{C}_{ik} > 2$ can be ignored
    - $\mathbf{C}_{ik}\mathbf{C}_{jk} = 1$ with probability $\approx \lambda_{ik}\lambda_{jk}$, with higher values having probabilities of a lower order.
    - we expect $\mathbf{K}_{ij} \approx \sum_k (2 \cdot (\lambda_{ik}^2/2 + \lambda_{jk}^2/2) - 2\lambda_{ik}\lambda_{jk}) = \sum_k (\lambda_{ik} - \lambda_{jk})^2$, which gives the desired concentration result.
- 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 degreee grows faster than $n^{1/3}$

Algorithm for low-rank models

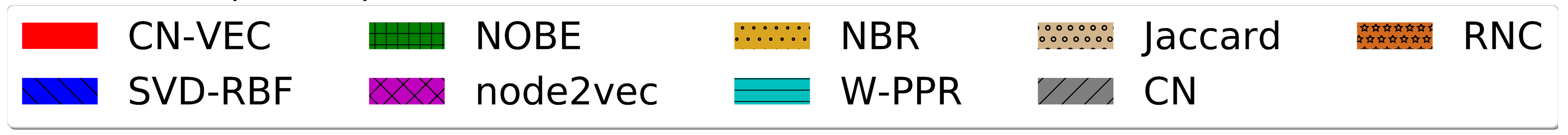
SVD-RBF

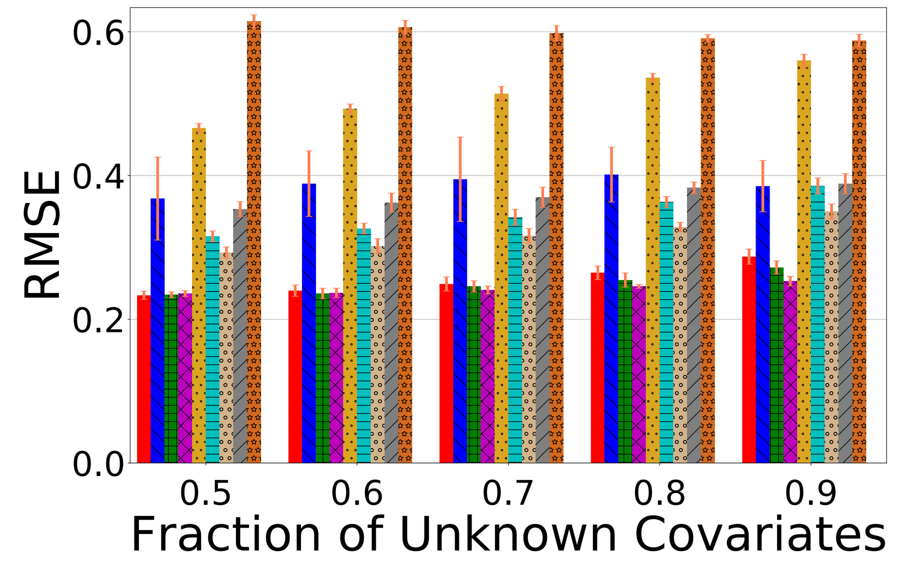
- $\hat{\mathbf{U}} \leftarrow$ top-d eigenvector matrix for $\mathbf{A}$
- $\hat{\mathbf{v}}_i \leftarrow i^{th}$ row of $\hat{\mathbf{U}}|\hat{\mathbf{E}}|^{1/2}$
- for $i \in [n] \setminus S$
  - $dist(j) \leftarrow \|\hat{\mathbf{v}}_i - \hat{\mathbf{v}}_j\|$ for $j \in S$
  - $\hat{\mathbf{X}}_i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

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

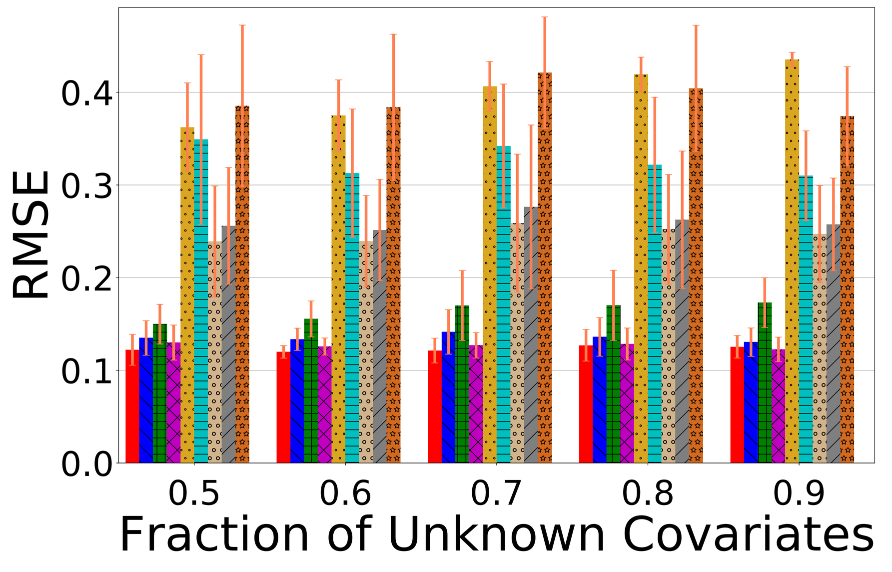
Simulation Experiments

- Comparing with embedding methods (NOBE, node2vec), neighborhood average (NBR), personalized pagerank (W-PPR), JACCARD, common neighbors (CN), regression with network cohesion (RNC)
- Simulate on latent space model (LSM), stochastic blockmodel (SBM), mixed-membership stochastic blockmodel (MMSB), rand random dot product model (RDPG)

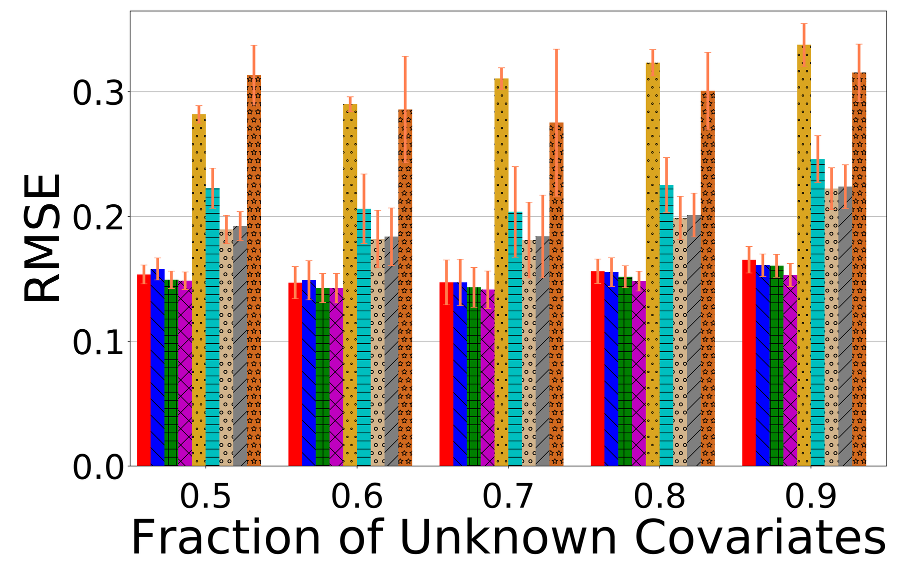




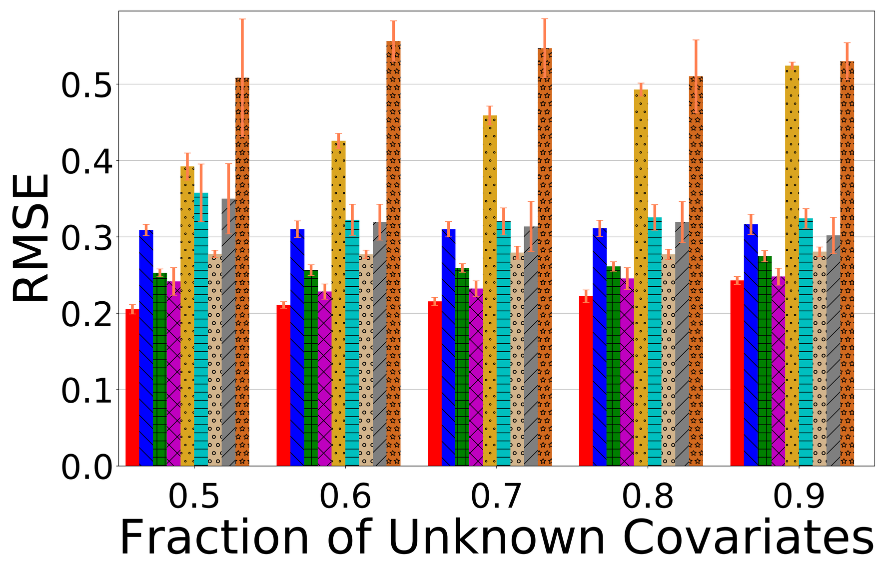
(a) LSM



(b) SBM



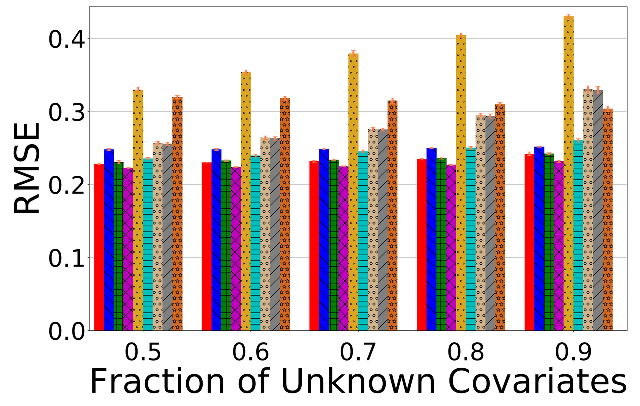
(c) MMSB

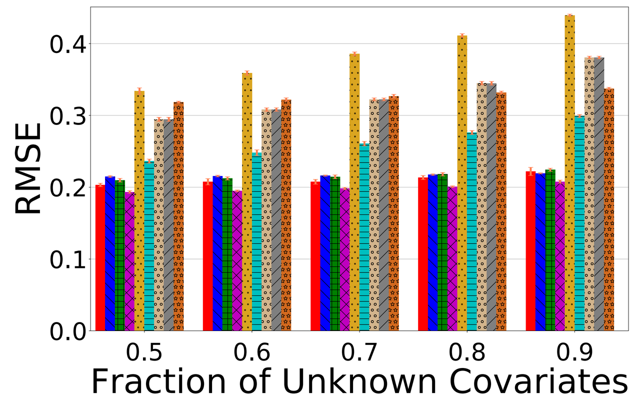


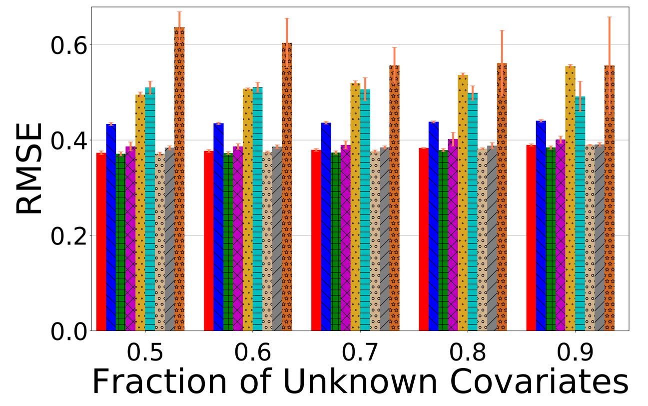
(d) RDPG

Real-world Network Results

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







{xmao@cs, deepay@, purna.sarkar@austin}.utexas.edu