ISP Research/EI project report

Makhneva Elizaveta, MSc DS 1 year

The main purpose of the project is to go through the paper Consistent Nonparametric Methods for Network Assisted Covariate Estimation and repeat the experiments authors held. Authors provided two algorithms which can be used to estimate unknown covariates of nodes in the network.

Idea of CN-VEC is to consider nodes with common neighbors as similar nodes and to use weighted average of their covariates to get the estimate but with more details, e.g. consider not exactly common neighbors in theory but to compare probabilities of having the edge with one of the nodes in the network for two other nodes.

Idea of SVD-RBF is to consider matrix of probabilities of existing edges between nodes as low-rank matrix and to use SVD for adjacency matrix to reconstruct initial probabilities and use similat idea as in CN-VEC.

Plan of the project:

1. find the oldest version of the paper, supplementary materials and code if possible
2. investigate authors' algorithms and implement them using Python
3. implement all other algorithms which authors compared their algorithms to
4. create graph generators using different approaches mentioned in the paper
5. conduct experiments and compare them to results from the paper

1 Results

Step	Results	Problems
1	Preprint of the paper in GitHub of one of the authors (Purnamrita Sarkar)	Preprint doesn't contain links to the supplementary materials and I couldn't find any additional files
	I wrote to researcher a letter about the code and supplementary materials	She didn't answer me
2	I read all the theory in the paper related to author's algorithms	Some steps of proofs are not very clear but I skipped them now since they are not important on the current step of research (e.g. inequality for Lipschitz function) – not a big problem
	I implemented both algorithms (CN-VEC and SVD-RBF), you can find it here	They are very slow now: very first algorithm needs about 20 hours for the network with 2500 nodes
3	I implemented NBR, W-PPR, Jaccard, CN, Node2Vec using description from the paper. For the last algorithm I also used code provided by authors of Node2Vec with additional functions needed for experiments. Almost all algorithms work quite fast for network with 2500 nodes: 337 ms (NBR), 1.94 s (W-PPR), 1min 3s (Jaccard), 59.6 s (CN)	Node2Vec is very slow. I didn't implement mentioned RNC algorithm

Step	Results	Problems
	There is also algorithm NOBE which is originally provided in MATLAB code and I rewrote some functions in Python for it	I couldn't completely implement NOBE since some of functions in MATLAB are not so clear
4	I implemented LSM, SBM, MMSB, RDPG. They work fast	I'm not really sure that RDPG is implemented correctly since in the paper latent vectors are generated from "mixture of d-dimensional Gaussians with means $e_l (l = 1, \ldots, 5)$ and covariance $0.1 \cdot I$. I understood it like that: we have 5 independent d-dimensional random vectors distributed with e_l and $0.1 \cdot I$ as parameters and we sum up them to get latent vectors. I'm not sure because maybe authors meant that these vectors should be dependent and because in my implementation adjacency matrix contains only ones (besides diagonal elements)
5	I tried to conduct experiments for algorithms which executed in reasonable time. My results have similar charts (see 2)	My results have different scale: RMSE $\sim 500 - 1000$ times higher :)

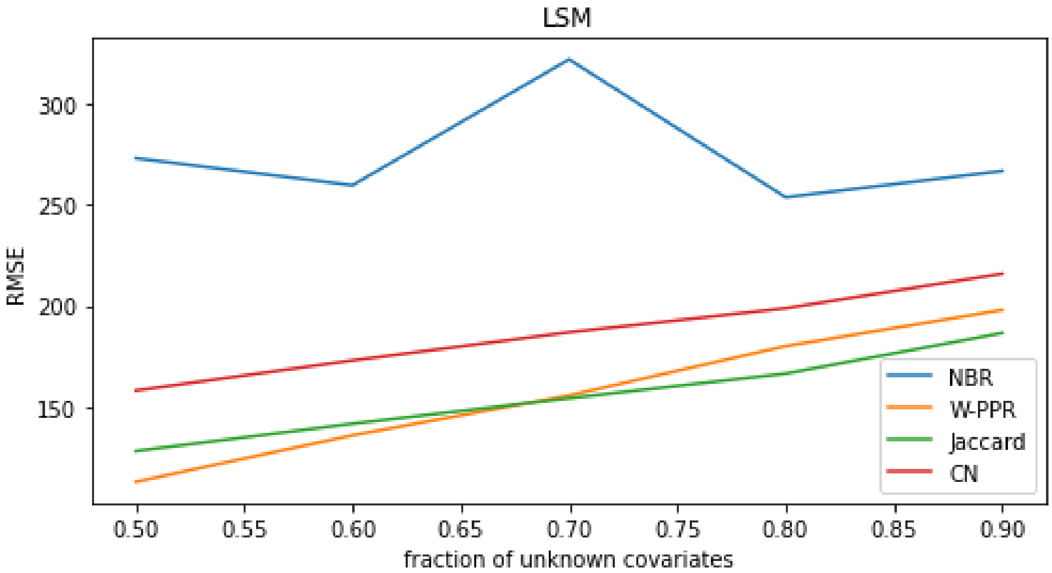
Further steps:

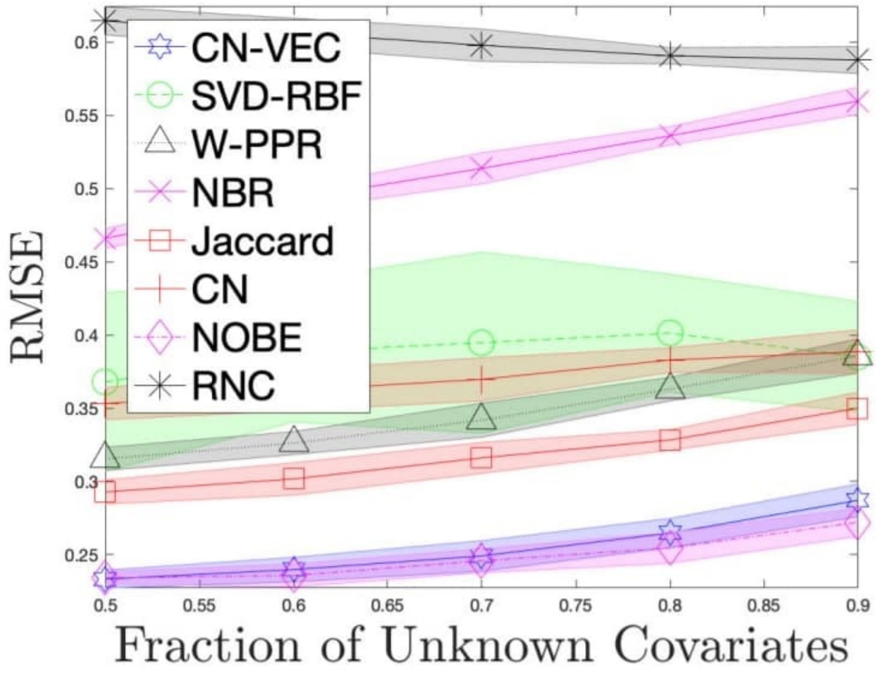
1. Find more optimal way to write algorithms from paper and implement it
2. implement RNC algorithm
3. implement NOBE algorithm
4. check whether RDPG is correct
5. check all the algorithms since the error is too high
6. conduct experiments for all algorithms and compare them to algorithms from paper

2 Appendix

Latent Space Model (LSM):

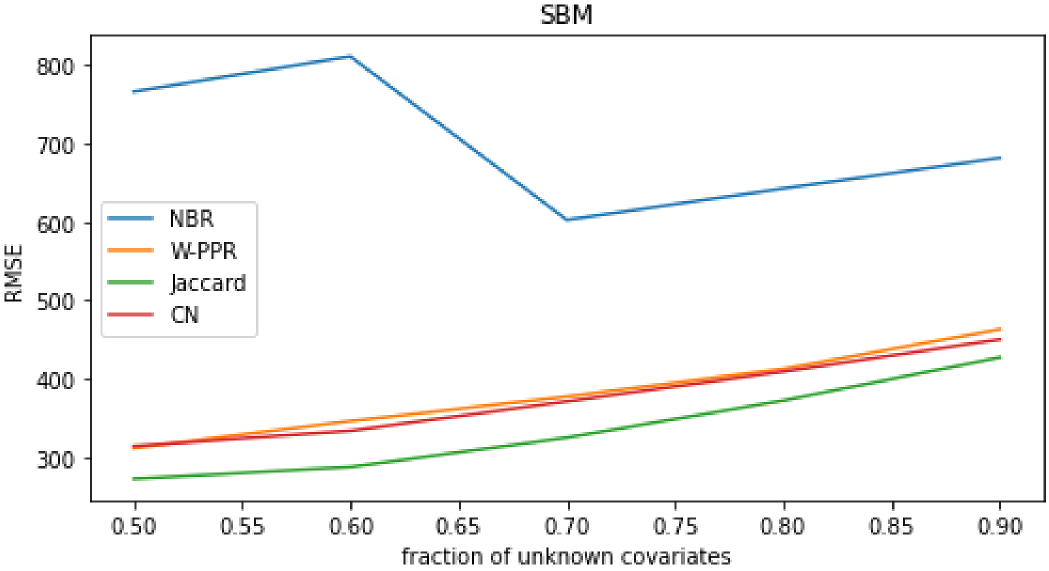
(1) – my experiments; (2) – authors' experiments

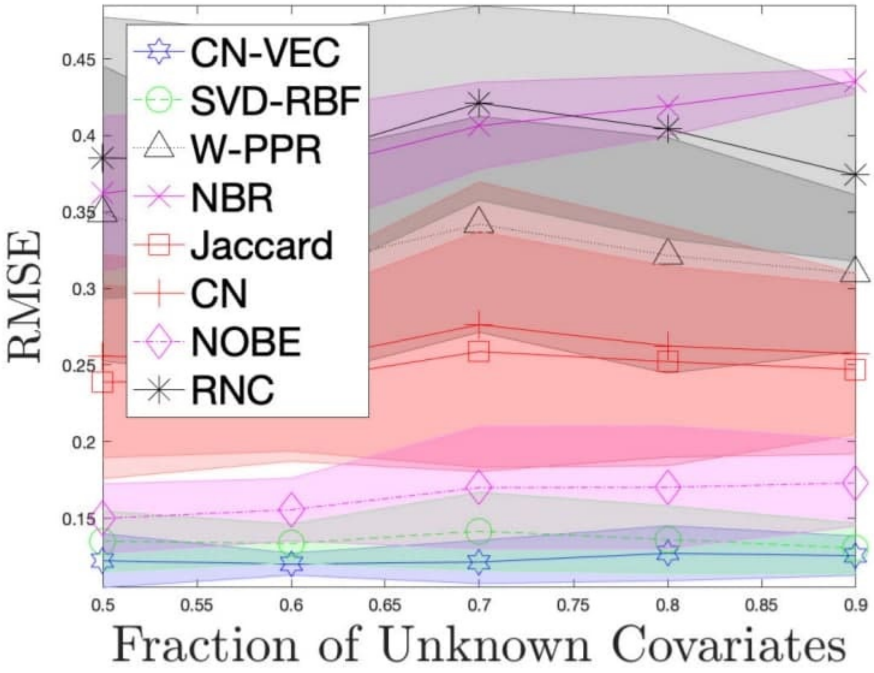




Stochastic Blockmodel (SBM):

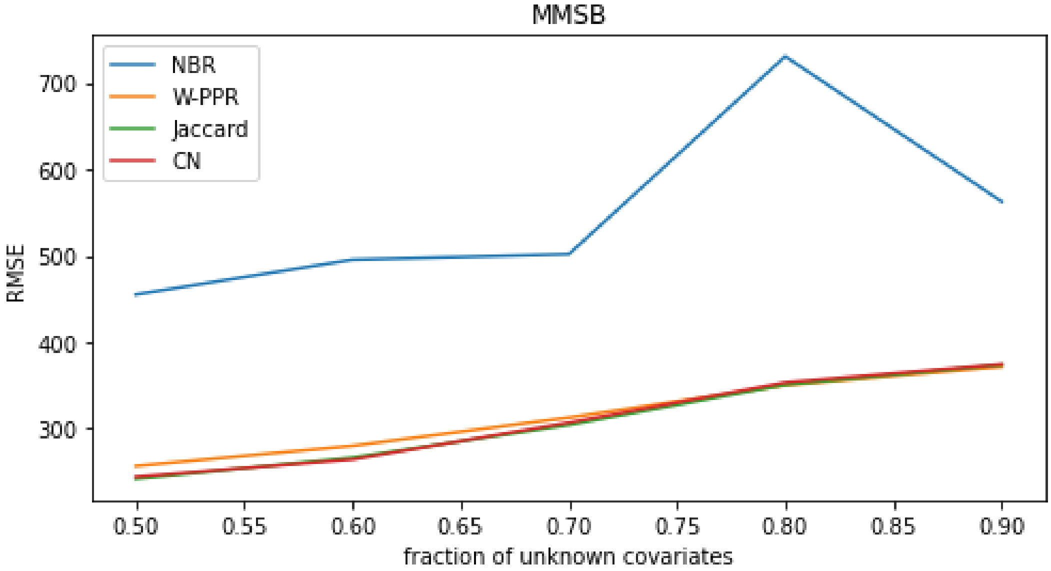
(1) – my experiments; (2) – authors' experiments

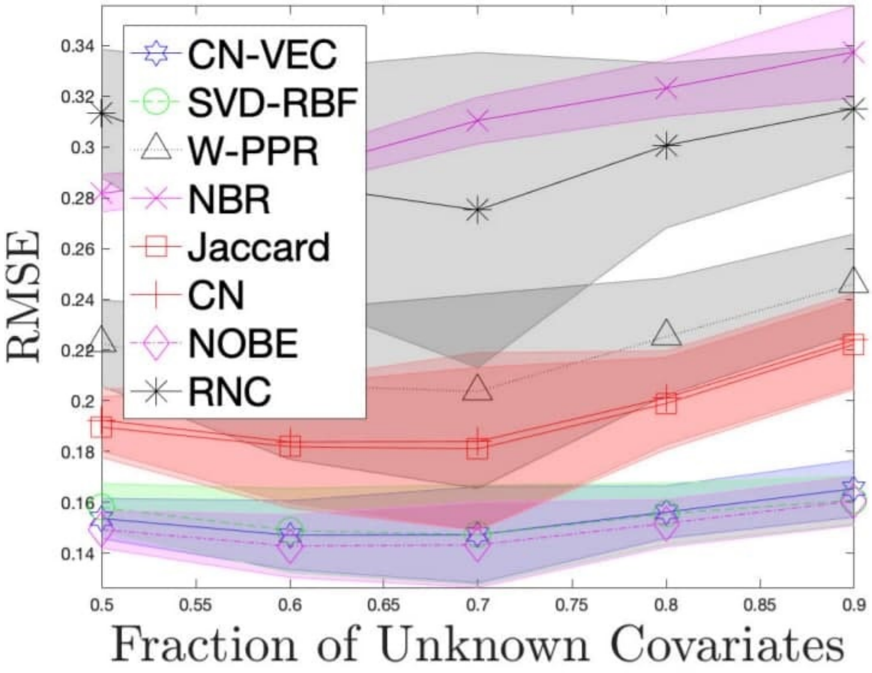




Mixed-membership Stochastic Blockmodel (MMSB):

(1) – my experiments; (2) – authors' experiments


MMSB
NBR
W-PPR
Jaccard
CN
RMSE
fraction of unknown covariates


CN-VEC
SVD-RBF
W-PPR
NBR
Jaccard
CN
NOBE
RNC
RMSE
Fraction of Unknown Covariates