Consistent Nonparametric Methods for Network Assisted Covariate Estimation

Xueyu Mao*, Deepayan Chakrabarti†, Purnamrita Sarkar‡
The University of Texas at Austin

Abstract

Networks with node covariates are commonplace: for example, people in a social network have interests, or product preferences, etc. If we know the covariates for some nodes, can we infer them for the remaining nodes? In this paper we propose a new similarity measure between two nodes based on the patterns of their 2-hop neighborhoods. We show that a simple algorithm (CN-VEC) like nearest neighbor regression with this metric is consistent for a wide range of models when the degree grows faster than $n^{1/3}$ up-to logarithmic factors, where n is the number of nodes. For "low-rank" latent variable models, the natural contender will be to estimate the latent variables using SVD and use them for non-parametric regression. While we show consistency of this method under less stringent sparsity conditions, our experimental results suggest that the simple local CN-VEC method either outperforms the global SVD-RBF method, or has comparable performance for low rank models. We also present simulated and real data experiments to show the effectiveness of our algorithms compared to the state of the art.

1 Introduction

Suppose we have a social network where each person has a vector of interests, such as desired products, or preferred news topics, or sporting interests. For some people, we know their interest vector from, say, their past tweets or comments. But for others, such data may be unavailable or insufficient. Can we infer their interests from a few people's known interests and the structure of the social network? This basic question is relevant for many applications, such as content and ad targeting, friend and group recommendations, and for investigating privacy in social networks, among others. Thus, a general solution to this problem would be useful in many contexts.

Formally, we consider a network where each node has a vector of node covariates. For some nodes, these covariates are known; we want to predict the covariates for all other nodes. Further, the predictions must have consistency guarantees. That is, the predicted covariates must converge to their actual values as the size of the network grows under some limiting process. In particular, we want consistency even for relatively "sparse" networks, often seen in real-world settings, where the average node degree grows slowly compared to the number of nodes.

To predict node covariates using the network structure, we use latent variable models. Here, the network and the node covariates are "generated" by latent variables associated with the nodes.

*Department of Computer Science. Email: xmao@cs.utexas.edu
†Department of Information, Risk, and Operations Management. Email: deepay@utexas.edu
‡Department of Statistics and Data Sciences. Email: purna.sarkar@austin.utexas.edu

Such models have been used before for community detection [77, 80, 71, 10, 82, 76]. But the node covariate prediction problem is less well-studied.

A seemingly simple solution is to take the average of the covariates of a node's neighbors in the network. But this is not effective in sparse networks. In sparse network models, with high probability, two nodes with identical latent values will not have an edge or even share a common neighbor. So, a node's neighbors may not be the nodes most similar to it. Random-walk heuristics go beyond a node's direct neighbors, but these lack provable guarantees except in special cases [43]. Thus, we need a more refined measure of similarity between nodes, which accurately reflects distances in latent space and can be estimated consistently from even sparse networks.

We propose a method (CN-VEC) to predict node covariates by a nearest-neighbor regression using the top-k nodes with the most similar two-hop neighborhoods. The similarity between nodes i and j depends on the number of common neighbors between the node pair (i, h), compared against (j, h), over all nodes h. This goes beyond a simple count of the common neighbors of i and j. Our carefully chosen similarity formula is provably consistent for a wide range of latent variable models and sparsities; to our knowledge, it is the first such algorithm. We do not need to know the function linking the probabilities to the latent variables. Also, the similarity measure has no parameters, so CN-VEC needs no fine-tuning.

If we have some prior knowledge of the underlying model, for example, if it is a "low-rank" model [68] like the Generalized Random Dot Product Graph (GRDPG) models [78, 59], which include many models like the Stochastic Blockmodel [27], the Mixed Membership Stochastic Blockmodel [4], it will be natural to first do singular value decomposition to estimate the latent variables, and then use those directly in non-parametric regression for estimating an unknown covariate. We denote this method by SVD-RBF.

For both CN-VEC and SVD-RBF, we provide consistency guarantees. For general models, CN-VEC is consistent when the average degree grows faster than $n^{1/3}$ up-to logarithmic factors, where n is the number of nodes. Note that CN-VEC depends on 2-hop connections, but the number of 2-hop paths between two nodes only concentrates when the average degree grows faster than $\sqrt{n}$ [60]. The better convergence guarantee of CN-VEC is due to its specially constructed similarity measure. This similarity measure concentrates even when 2-hop path counts do not concentrate. Thus, the analysis for CN-VEC is quite different to analysis of common neighbors [60, 61]. For low-rank models, we show that SVD-RBF is consistent when the average degree grows faster than polylog of n.

We compare CN-VEC with SVD-RBF and a variety of other methods, including random-walks, regression using Jaccard similarity, node2vec [24], a recent embedding-based algorithm called NOBE [31] and regression with network cohesion (RNC) [42, 39]. We run experiments using 4 simulated graph models and 3 real-world networks. Overall, SVD-RBF, CN-VEC, node2vec and NOBE outperform the rest. Among the four, CN-VEC is either the best or close to it. This is a surprising observation, since CN-VEC uses local statistics like 2-hop paths, whereas nearly all other methods use the whole network for inference. Also, CN-VEC is 10x-100x faster than node2vec and NOBE, depending on the sparsity of the network.

Our main contribution is the CN-VEC algorithm, which is based on a novel similarity measure. CN-VEC does not assume a low-rank model, needs no parameters for its similarity measure, and uses only local information, yet mostly outperforms the global SVD-RBF algorithm both in accuracy and time. We provide SVD-RBF mainly to show that CN-VEC does not lose much due to its weaker assumptions.

The paper is organized as follows. We review related work in Section 2. In Section 3, we present

our model and describe CN-VEC and SVD-RBF, and provide consistency guarantees. Section 4 shows the empirical results. We conclude in Section 5.

2 Related work

We will survey connections to node similarity measures, node classification, regression with network cohesion, and estimation in latent variable models.

Node similarity measures: There are many existing similarity measures, based on the number of common neighbors [61] and its weighted variants (Adamic/Adar) [2], preferential attachment [8], resource allocation [84], Katz index [36], PageRank [12], SimRank [28], and graph neural networks [79] (see [45] for a survey). While these often work well, only a few have consistency guarantees [61, 60]. Our CN-VEC method constructs a similarity measure that provably works in sparser settings.

Node classification: Here the goal is to predict labels of nodes based on the network structure. Many methods are based on **random walks**, such as label propagation [74], personalized PageRank [52, 38, 29], partially absorbing random walks [73], etc. A weighted version of personalized PageRank has provable guarantees under the degree-corrected Stochastic Blockmodel, but only when there are two communities of nodes [43]. Another direction is **node embeddings**, which aims to represent nodes by vectors while retaining some network-based properties [65, 9, 67, 24, 55, 31, 56]. With the embedding vectors, one can train a classifier to predict the unseen labels. These typically lack provable guarantees, but often work well in practice. As a special case of node embedding, SVD-RBF uses the eigenvectors and eigenvalues of the adjacency matrix to embed nodes, which is also known as the spectral embedding. It has been well studied in the statistics literature [66, 64, 59]. However theoretical consistency of spectral embedding is typically studied under low-rank GRDPG models, while CN-VEC does not need any such model restriction.

When only features are given, semi-supervised learning [85, 51, 19, 13] constructs the similarity matrix from the features. Note that features are analogous to the latent positions of nodes in our problem, and these are unknown for us.

Regression with network cohesion: In regression with network cohesion [42, 39, 35, 34], (x_i, y_i) pairs are observed for each node, and the network is used as a regularizer. In Network Lasso [26], the network structure is used to enforce smoothness much like [42]. Since node features are assumed to be observed in the latter, it cannot be applied directly to our setting. As Network Lasso requires edge weights, our similarity matrix can also be potentially used as the edge weight matrix.

Latent variable inference: Consistent latent inference algorithms have been developed for the Latent Space Model [46], Stochastic Blockmodel [58, 41, 1] and its degree-corrected version [83, 32, 21], Mixed Membership Stochastic Blockmodel [47, 53, 49] and its degree-corrected version [33, 48], Stochastic Blockmodel with Overlaps [37], Generalized Random Dot Product Graph models [64, 7, 59], and so on. However, one needs specialized algorithms for different models, and the true model may be unknown for real world networks. For low rank models, our SVD-RBF estimates the latent variables via a singular value decomposition. However, the low rank assumption is not required in our CN-VEC algorithm, which works for a broad range of latent variable models.

For latent variable models, there is also related work on estimating distances/dot-products in latent space. When the latent variables represent positions in a random geometric graph, spectral methods [5], shortest path lengths [6], and common neighbor counts [61] have been used. Recently

[54] recovers the shortest path metric from a noisy neighborhood graph. However, those methods are specially designed for different link functions, while CN-VEC does not require prior knowledge on the form of the link function. A more in depth discussion of related works on latent distance estimation can be found in [6].

Other related problems: In **matrix completion**, we try to fill in matrix entries given a partially observed noisy matrix. [63, 44] introduce a framework to estimate the missing values using nearest neighbors under a latent variable matrix generation model. In **graphon estimation**, we try to estimate underlying edge probabilities of a random graph from the observed adjacency matrix. Some recent work includes sorting-and-smoothing [14], Stochastic Blockmodel approximation [3], and neighborhood smoothing [81]; see also [20, 11, 75]. Our problem is different; we want to predict node covariates.

3 Proposed Work

We are given an undirected and unweighted network between n nodes, represented by a binary adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$. We are also given the node covariates $\{\mathbf{X}_i \in \mathbb{R}^p; i \in S\}$ for a subset of nodes S. Our goal is to infer the node covariates of the remaining nodes $\{\mathbf{X}_i; i \in [n] \setminus S\}$. We will present our notation and model, followed by our two algorithms for the model-agnostic and low-rank cases.

Model. We consider networks generated from general latent variable models. Each node $i \in [n]$ in the network has a latent vector $\mathbf{z}_i \in \mathbb{R}^d$, with $\|\mathbf{z}_i\|$ bounded by a constant C. The probability that there is an edge between node i and j depends solely on $\mathbf{z}_i$ and $\mathbf{z}_j$:

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

where $f(\cdot)$ is bounded in $[0,1]$ and has parameters $\mathbf{\Theta}$, and $\rho_n = o(1)$ controls the sparsity of the graph. For simplicity, we will drop the subscript on ρ_n for the rest of the paper. Thus, the matrix $\mathbf{P}$ denotes the conditional expectation of $\mathbf{A}$ given the latent variables; we set the diagonal of $\mathbf{A}$ to zero. The node covariates are also generated from the latent vectors:

$$\mathbf{X}_i = g(\mathbf{z}_i) + \boldsymbol{\epsilon}_i, \tag{2}$$

where $g : \mathbb{R}^d \to \mathbb{R}^p$ is bounded, and $\boldsymbol{\epsilon}_i$ are i.i.d. noise random vectors with uncorrelated elements, whose mean is 0 and variance is σ^2. We assume that $p = \Theta(1)$, $\sigma = \Theta(1)$, and $g(\cdot)$ is suitably smooth, which is a standard assumption in nonparametric regression [25, 72, 18]:

Assumption 3.1. $g(\cdot)$ is Lipschitz, that is, there is a constant $L_g > 0$ such that

$$\|g(\mathbf{v}_1) - g(\mathbf{v}_2)\| \leq L_g \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|.$$

We denote by the column vector $\mathbf{a}_i$ the i^{th} column of the adjacency matrix $\mathbf{A}$. We denote by $\mathbf{e}_i$ a vector such that $\mathbf{e}_i(j) = 1(i = j)$, with the vector size being evident from the context. We use the standard o, O and ω, Ω order notations, with $\tilde{O}$ hiding poly-logarithmic factors, and o_P and O_P probabilistic order notations [69].

3.1 Model-Agnostic Algorithm

The node covariate $\mathbf{X}_i$ of a node i depends on the latent $\mathbf{z}_i$ and the function $g(.)$, both of which are unknown. If we knew the latents but not $g(.)$, we could still estimate $\mathbf{X}_i$ using a non-parametric estimator:

$$\hat{\mathbf{X}}_i = \frac{\sum_{j \in top_k(i)} \mathbf{W}_{ij} \cdot \mathbf{X}_j}{\sum_{j \in top_k(i)} \mathbf{W}_{ij}}, \tag{3}$$

where $\mathbf{W}_{ij}$ is a measure of similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$, and $top_k(i)$ is a set of k nodes $j \in S$ with the largest $\mathbf{W}_{ij}$ values. Under a smooth $g(.)$ (Assumption 3.1) and mild conditions on $\mathbf{W}$, this is asymptotically consistent [25, 72]. But in our case, we do not know the latents.

Our approach is to use the network to find $top_k(i)$, and we will show how this is possible even without knowing the latents. The underlying idea is that the latents generate the $\mathbf{P}$ matrix, which in turn generates the adjacency matrix $\mathbf{A}$. So, similarities between latents should be reflected in the network structure. We will now present a series of methods of increasing complexity for finding $top_k(i)$, culminating in our proposed method.

Adjacency matrix: The simplest idea is to average the covariates of a node's neighbors in the network. For example, consider a Stochastic Blockmodel (SBM) [27]. Here $\mathbf{z}_i$ are latent memberships to r blocks and the network is generated such that the probability of connection of node i in block a and node j in block b is simply $\mathbf{B}_{ab}$, where $\mathbf{B}$ is a $r \times r$ matrix. For this simple example, assume that $\mathbf{A}$ is generated from an SBM and the node covariates are generated such that nodes in block i have i.i.d covariates from a distribution mean μ_i. The means of different blocks are different. Suppose $\mathbf{P}_{ij}$ is high if i and j belong to the same cluster ($\mathbf{z}_i = \mathbf{z}_j$), and low otherwise. Then, if we could set $\mathbf{W} = \mathbf{P}$, the nodes selected in $top_k(i)$ would be those in the same cluster as i. So they would have the same latent as i. Thus, averaging over the covariates of $top_k(i)$ would give a good prediction for $\mathbf{X}_i$. Now, we do not have $\mathbf{P}$, but the adjacency matrix $\mathbf{A}$, which is a stochastic version of $\mathbf{P}$. However, if we use $\mathbf{W} = \mathbf{A}$, there is no way to distinguish between in-cluster versus out-of-cluster neighbors of i. This leads to a biased prediction, so we cannot use the adjacency matrix as the $\mathbf{W}$ matrix.

Common neighbor matrix: The previous idea of using the adjacency matrix $\mathbf{A}$ failed because it did not accurately reflect the probability matrix $\mathbf{P}$. To remedy this, we can set $\mathbf{W} = \mathbf{C}$, where $\mathbf{C}_{ij} = \mathbf{a}_i^T \mathbf{a}_j$ is the number of common neighbors of nodes i and j (for $i \neq j$). The off-diagonal entries of $\mathbf{C}$ concentrate around those of $\mathbf{P}^2$ when the average degree of nodes grows faster than $\tilde{O}(\sqrt{n})$ [58, 60]. For the stochastic blockmodel under appropriate conditions, the nodes selected in $top_k(i)$ are again those in the same cluster as i. Thus, setting $\mathbf{W} = \mathbf{C}$ works in dense networks where nodes have high degree. However, this method will not work for sparse networks seen in real-world settings. For sparse matrices, one may need to use more complex similarity matrices like the personalized pagerank [29] matrix, which also uses information from long paths.

We will show experimentally that prediction accuracy matches the above discussion. Using the adjacency matrix directly is worse than the matrix of common neighbors, which in turn is worse than matrices based on personalized pagerank. However, we can do much better, and provably so, by extending the common-neighbors idea. We describe this next.

Distances between rows of C: Using $\mathbf{W} = \mathbf{C}$ allowed us to use the "rest of the network" in computing the similarity between i and j. However, only nodes that are common neighbors of both i and j contributed to this measure. Our key observation is that if i and j have similar latents, then we should also expect $\mathbf{P}_{i\ell} \approx \mathbf{P}_{j\ell}$ for any node $\ell \neq i, j$. If the same also holds for $\mathbf{P}^2$ (i.e., $(\mathbf{P}^2)_{i\ell} \approx (\mathbf{P}^2)_{j\ell}^2$), then $\mathbf{C}_{i\ell} \approx (\mathbf{P}^2)_{i\ell}^2 \approx (\mathbf{P}^2)_{j\ell}^2 \approx \mathbf{C}_{j\ell}$ by concentration. So, instead of just considering $\mathbf{C}_{ij}$ as the similarity between i and j, we should use a measure that compares $\mathbf{C}_{i\ell}$ to $\mathbf{C}_{j\ell}$ for all ℓ. In other

words, we set $\mathbf{W}_{ij}$ to be the similarity between rows i and j of the matrix $\mathbf{C}$. This goes beyond just the common neighbors of i and j, and hence can work even in sparse networks.

We need the following assumption:

Assumption 3.2. There exist positive constants ℓ and L, and $\Delta_n = o(1)$, such that

$$\ell\|\mathbf{z}_i - \mathbf{z}_j\| - \Delta_n \leq \frac{1}{n^{1.5}\rho^2}\|(\mathbf{e}_i - \mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I} - \mathbf{e}_i\mathbf{e}_i^T - \mathbf{e}_j\mathbf{e}_j^T\right)\| \leq L\|\mathbf{z}_i - \mathbf{z}_j\|.$$

The middle term equals the square root of $\sum_{k\neq i,j}\left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2$, normalized by its order. So, the assumption states that $\mathbf{z}_i$ is far from $\mathbf{z}_j$ *iff* $(\mathbf{P}^2)_{ik}$ is far enough from $(\mathbf{P}^2)_{jk}$ for one or more $k \in [n] \setminus \{i, j\}$. That is, for some nodes k in the 2-hop neighborhood of i or j, there should be significant differences.

Remark 3.1. The second inequality of Assumption 3.2 can be derived from the piece-wise Lipschitz condition that is commonly used in graphon estimation literature [3, 81, 14, 21, 75]. The LHS ensures that each node has enough nearest neighbors in latent space. While the condition looks technical, we show that it is satisfied for Generalized Random Dot Product Graph (GRDPG) models [78, 59] which include Stochastic Blockmodel and Mixed Membership Stochastic Blockmodel. The details are in the supplementary material.

By Assumption 3.2, the similarity between i and j can be inferred from $\sum_{k\neq i,j}\left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2$. But $\mathbf{P}$ is unknown. So, we need a statistic that converges to this quantity, up to a constant. Recall that $\mathbf{C}_{ij} = \mathbf{a}_i^T\mathbf{a}_j$ denotes the number of common neighbors between nodes i and j. Now, it is easily shown that $\mathbf{E}[\mathbf{C}_{ik}] = (\mathbf{P}^2)_{ik}$. So it may seem that $\sum_{k\neq i,j}\left(\mathbf{C}_{ik} - \mathbf{C}_{jk}\right)^2$ will work. But $\mathbf{C}_{ik}$ converges to $(\mathbf{P}^2)_{ik}$ only for "dense" networks, where the average degree grows faster than $\tilde{O}(\sqrt{n})$. In sparse networks, $\mathbf{C}_{ik} = 0$ for most (i, k) pairs. For a given k, it is very unlikely that both $\mathbf{C}_{ik} > 0$ and $\mathbf{C}_{jk} > 0$. So, paradoxically, $|\mathbf{C}_{ik} - \mathbf{C}_{jk}| = \mathbf{C}_{ik} + \mathbf{C}_{jk}$ for many k. This means that $\sum_{k\neq i,j}\left(\mathbf{C}_{ik} - \mathbf{C}_{jk}\right)^2$ may be large even if $\mathbf{z}_i = \mathbf{z}_j$ and $\sum_{k\neq i,j}\left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2 = 0$.

Instead, we propose the following statistic to measure the similarity of i and j:

$$\mathbf{K}_{ij} = \sum_{k\neq i,j}\big[(\mathbf{C}_{ik}^2 - 2)1(\mathbf{C}_{ik} \geq 2) + (\mathbf{C}_{jk}^2 - 2)1(\mathbf{C}_{jk} \geq 2) - 2\mathbf{C}_{ik}\mathbf{C}_{jk}\big]. \tag{4}$$

This statistic concentrates around the desired quantity (up to a constant) for both sparse and dense networks, as the next theorem shows.

Theorem 3.1. We have:

$$\mathbf{K}_{ij} = \left(\sum_{k\neq i,j}\left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2\right) + e + c,$$

where $e = O_P(n^{2.5}\rho^3\sqrt{\log n})$, $c = -4(n-2)$ if $n\rho^2 = \Omega(\log^\xi n)$, $\xi > 1$; and $e = O_P(n^4\rho^6 + n\rho\sqrt{\log^5 n})$, $c = 0$ if $n\rho^2 = o(1)$, $n^2\rho^3 = \Omega(\log^\xi n)$, $\xi > 2.5$.

Algorithm 1 CN-VEC: model-agnostic algorithm

Input: Adjacency matrix $\mathbf{A}$, Set S of nodes with known covariates, number of neighbors k
Output: Estimated node covariates $\hat{\mathbf{X}}_i$, $i \in [n] \setminus S$

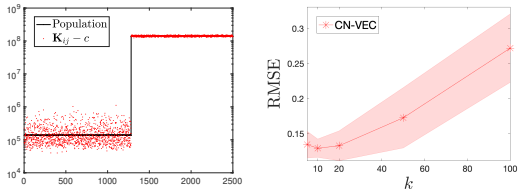
1: **for** $i \in [n] \setminus S$ **do**
2: $\quad dist(j) \leftarrow \mathbf{K}_{ij}$ (by Eq. (4)), for $j \in S$
3: $\quad top_k(i) \leftarrow k$ nodes with the smallest values of $dist(j)$
4: $\quad \hat{\mathbf{X}}_i \leftarrow \frac{1}{k}\sum_{j \in top_k(i)} \mathbf{X}_j$
5: **end for**

Proof Sketch. We may understand the intuition for $\mathbf{K}_{ij}$ by separately considering the cases of dense and sparse networks. In the case of a dense network ($n\rho^2 \to \infty$), we expect $\mathbf{C}_{ik}$ be large, so the indicators may be safely ignored. Thus, we expect $\mathbf{K}_{ij} \approx \sum_{k \neq i,j} (\mathbf{C}_{ik} - \mathbf{C}_{jk})^2 + c$. Now, $\mathbf{C}_{ik} = \sum_h \mathbf{A}_{ih}\mathbf{A}_{hk}$, so it is a sum of independent random variables. Hence, by Bernstein's inequality, $\mathbf{C}_{ik}$ concentrates around its expectation $(\mathbf{P}^2)_{ik}$. This leads to the desired concentration result in the dense case.

This reasoning does not hold for the sparse case ($n\rho^2 \to 0$) because $\mathrm{E}[\mathbf{C}_{ik}] \approx 0$ and $\mathbf{C}_{ik}$ does not concentrate. In this case, $\mathbf{C}_{ik}$ is well-approximated by a Poisson random variable with rate $\lambda_{ik} = (\mathbf{P}^2)_{ik} = O(n\rho^2)$. Thus, the 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 since its probability is of a lower order. Similarly, $\mathbf{C}_{ik}$ and $\mathbf{C}_{jk}$ can be treated as nearly independent since it is very unlikely that a node h is connected to i, j, and also k. So $\mathbf{C}_{ik}\mathbf{C}_{jk} = 1$ with probability $\approx \lambda_{ik}\lambda_{jk}$, with higher values having probabilities of a lower order. Thus, 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 again gives the desired concentration result. The detailed proof is more involved, and is presented in the supplementary material. □

Remark 3.2. When $\mathbf{z}_i = \mathbf{z}_j$, we have $\mathbf{e}_i^T\mathbf{P} = \mathbf{e}_j^T\mathbf{P}$, so $\mathbf{K}_{ij} - c = e$. But when $\|\mathbf{z}_i - \mathbf{z}_j\| \gg \Delta_n/\ell$, $\mathbf{K}_{ij} - c \approx \left(\sum_{k \neq i,j} \left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2\right) = \Omega(n^3\rho^4) \gg e$. So for both sparse and dense networks, the node pairs with small $\mathbf{K}_{ij}$ are also the node pairs with small $\|\mathbf{z}_i - \mathbf{z}_j\|$.

Remark 3.3. We also want to emphasize that the above theoretical result makes use of the fact that our common-neighbor based metric is looking at an *ensemble* of common neighbors, and hence it concentrates in a broader range of sparsity parameters compared to pairwise common neighbors [60]. Our analysis is also completely different from [60], and requires finer analysis.



(a) $\mathbf{K}_{ij}$ on SBM (log-scale). (b) Changing # of neighbors k.

Figure 1: Simulations.

Remark 3.4. Note that although the c term in Theorem 3.1 may be large, for any graph, it is a constant and does not affect the ordering of $\mathbf{K}_{ij}$. CN-VEC only needs this ordering to pick nearest neighbors for any node i. So, the goal of Theorem 3.1 is to show that ordering by $\mathbf{K}_{ij}$ matches the ordering by the population quantity $\sum_{k \neq i,j} \left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2$. The error term e is of a smaller order (Remark 3.2). Figure 1(a) shows this for a Stochastic Blockmodel (SBM) with two communities. The y-axis shows, in log-scale, the value of $(\mathbf{K}_{ij} - c)$ and its population counterpart for a random node i. The x-axis shows nodes grouped by their communities. The figure shows that $(\mathbf{K}_{ij} - c)$ concentrates well around the population. Varying the graph's sparsity yields qualitatively similar results.

Thus, given a node $i \in [n] \setminus S$, ordering the nodes $j \in S$ according to $\mathbf{K}_{ij}$ is equivalent to ordering them according to $\sum_{k \neq i,j} \left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2$. We find the top $\log(n)$ nodes among S with the smallest values of $\mathbf{K}_{ij}$ (call this set $top_k(i)$), and average their node covariates to estimate the covariates for node i. So the $\mathbf{W}$ matrix in Eq. (3) can be thought of as a binary matrix with $\mathbf{W}_{ij} = 1$, if $j \in top_k(i)$. We call this algorithm CN-VEC; Algorithm 1 shows the details. Theorem 3.1 coupled with the following theorem shows that the CN-VEC algorithm is consistent.

Theorem 3.2. Suppose in Eq. (2) each element of the random noise vector $\boldsymbol{\epsilon}_i$ has same variance σ^2, $|S| = \Theta(n)$, and Assumptions 3.1 and 3.2 hold, then for any sequence k_n such that $k_n \to \infty$, $k_n/n \to 0$, k_n-nearest-neighbors regression using $\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{P}^2 \left(\mathbf{I} - \mathbf{e}_i\mathbf{e}_i^T - \mathbf{e}_j\mathbf{e}_j^T\right)\|$ as the distance metric yields weakly consistent estimates for node covariates when ties occur with probability 0:

$$\mathrm{E}[\|\hat{\mathbf{X}}_i - g(\mathbf{z}_i)\|^2] = o(1) \quad \text{for } i \in [n] \setminus S.$$

Remark 3.5. It is possible to relax $|S|$ to be $o(n)$ as long as $|S| \to \infty$. Intuitively, to predict the covariates for node i, we need to only consider $\mathbf{K}_{ij}$ for $j \in S$. So we can apply Theorem 3.2 by replacing n by the effective size $|S|$.

Remark 3.6. Theorem 3.2 suggests that we can set the number of nearest neighbors in Algorithm 1 as $k = O(\log n)$, with the constant chosen by cross validation. Figure 1(b) shows the RMSE of simulations of Stochastic Blockmodel (SBM) when we change k. This shows a sweet spot for $k \in [10, 20]$.

3.2 Algorithm for Low-Rank Models

One popular class of network models assumes that the probability matrix $\mathbf{P}$ is low-rank. This results from a bilinear form for f, that is, $f(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T \boldsymbol{\Theta} \mathbf{z}_j$ for some model parameters $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$. For example, in the Stochastic Blockmodel (SBM) [27], the d-dimensional latent vector $\mathbf{z}_i$ for node i is of the form $\mathbf{z}_i = \mathbf{e}_a$ for some $a \in [d]$. Here, we say that node i belongs to "community" a. The probability of a link between nodes i and j is given by $\mathbf{P}_{ij} = \rho f(\mathbf{z}_i, \mathbf{z}_j) = \rho \mathbf{z}_i^T \boldsymbol{\Theta} \mathbf{z}_j = \rho \boldsymbol{\Theta}_{ij}$. That is, link probabilities are solely dependent on the community memberships of nodes, and $\boldsymbol{\Theta}$ represents the community interconnections. The Mixed Membership Stochastic Blockmodel (MMSB) [4] generalizes this to allow "soft" community memberships. Here, $\mathbf{z}_i$ is a probability vector, representing a distribution over communities for node i.

The Generalized Random Dot Product Graph (GRDPG) model [78, 59] allows for more general $\mathbf{z}_i$ and sets $\mathbf{P}_{ij} = \rho \mathbf{z}_i^T \mathbf{I}_{q,d-q} \mathbf{z}_j$, where $\mathbf{I}_{q,d-q}$ is a diagonal matrix with first q elements on the diagonal as 1 and the rest as -1. Let $\hat{\mathbf{U}}\hat{\mathbf{E}}\hat{\mathbf{U}}^T$ be the top-d eigen-decomposition of $\mathbf{A}$, where $\hat{\mathbf{E}}$ is a diagonal

Algorithm 2 SVD-RBF: nonparametric regression for low rank models with the RBF kernel $K_\theta(\mathbf{v}_1, \mathbf{v}_2)$

Input: Adjacency matrix $\mathbf{A}$, Set S of nodes with known covariates, bandwidth θ, rank of matrix d
Output: Estimated node covariates $\hat{\mathbf{X}}$

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

matrix, and both $\hat{\mathbf{U}}$ and $\hat{\mathbf{E}}$ have rank d (typically, $d \ll n$). Then for large enough n, the latent vectors $\mathbf{z}_i$ are arbitrarily close to a linear transformation of the rows of $\hat{\mathbf{U}}|\hat{\mathbf{E}}|^{1/2}$ (call them $\hat{\mathbf{v}}_i$) [59]. So if the Assumption 3.1 holds for $g(\mathbf{z}_i)$, then it also holds for $g(\hat{\mathbf{v}}_i)$. Hence, we can use $\hat{\mathbf{v}}_i$ as the latent positions in place of $\mathbf{z}_i$. In practice, the number of eigenvectors can be chosen via the USVT estimator [15]. For a node i with unknown covariates, we calculate its distances $\|\hat{\mathbf{v}}_i - \hat{\mathbf{v}}_j\|$ to other nodes $j \in S$ and put them to an RBF kernel to get the weights for nonparametric regression. The estimated covariates $\hat{\mathbf{X}}_i$ is then the weighted average of the covariates $\mathbf{X}_j$. We call this algorithm SVD-RBF; Algorithm 2 shows the details. We prove that Algorithm 2 gives consistent results.

Proposition 3.1. Consider a sequence of networks generated from a GRDPG model with bounded supported where the density exists and its infimum is positive. If Assumptions 3.1 holds, $|S| = \Theta(n)$, $\rho n = \omega(\log^{4\xi} n)$ for some constant $\xi > 0$, $d = \Theta(1)$, and the smallest singular value of $\mathbf{P}$ grows linearly with $n\rho$, then for bandwidth $\theta = \tilde{\Theta}(n^{-\frac{1}{2(d+1)}})$, and $\hat{\mathbf{X}}_i$ returned by Algorithm 2, we have, with probability tending to one,

$$\max_{i\in[n]\setminus S} \|\hat{\mathbf{X}}_i - g(\mathbf{z}_i)\| = o(1).$$

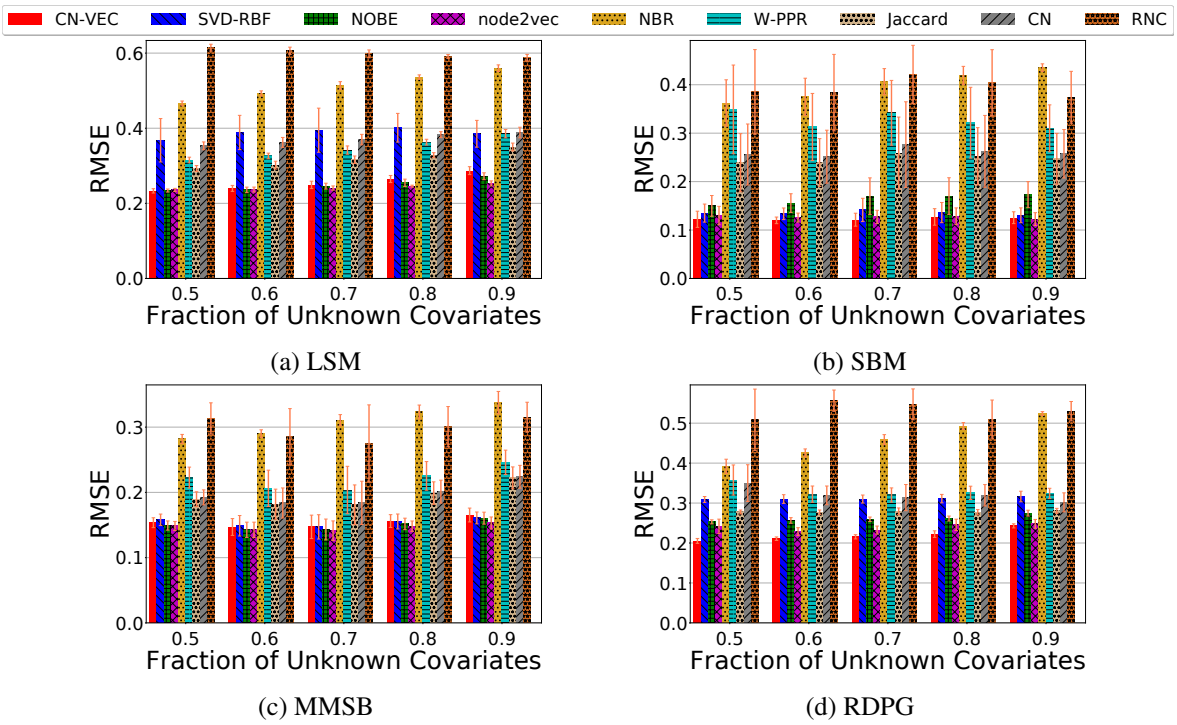
Proof Sketch. The proof follows from an analysis of the Nadaraya–Watson estimator using an RBF kernel. The bandwidth is chosen using the bound

$$\max_{i\in[n]} \|\mathbf{O}_n\hat{\mathbf{v}}_i - \sqrt{\rho}\mathbf{z}_i\| = O_P\left(\frac{(\log n)^\xi}{n^{1/2}}\right)$$

in [59], for some full rank matrix $\mathbf{O}_n \in \mathbb{R}^{d\times d}$ whose spectral norm is almost surely bounded. The details are presented in the supplementary material. □

Proposition 3.1 gives a guidance on choosing the bandwidth θ for Algorithm 2, e.g., setting $\theta = \Theta\left((\log n)^{\frac{3}{d}}/n^{\frac{1}{2(d+1)}}\right)$, while the constant can be fine-tuned by cross-validation.

Complexity: SVD-RBF needs $O((n^2+E)d)$ time to predict the covariates for all nodes in a network with n nodes, E edges, and rank d for $\mathbf{P}$. CN-VEC needs $O(nE)$ time to perform three matrix-matrix multiplications involving $\mathbf{A}$. Both have a space complexity of $O(n^2)$ to store the pairwise node similarities.



(a) LSM

(b) SBM

(c) MMSB

(d) RDPG

Figure 2: RMSE on recovering hidden node covariates.

4 Experiments

We evaluate the accuracy and speed of CN-VEC and SVD-RBF on several simulated and real-world networks. Since both CN-VEC and SVD-RBF are based on non-parametric regression using our proposed similarity measures, we mainly compare against other similarity measures. So, each method constructs a similarity $\mathbf{W}_{ij}$ between each pair of nodes i and j. Then, given a node i, it picks the top-10 most similar nodes according to the $\mathbf{W}$, and calculates the weighted average of their node covariates, with $\mathbf{W}$ as the weights. We consider the following similarity measures:

- NBR: This predicts the missing covariates for a node using the average of covariates of the neighbors of the node. This simply uses the adjacency matrix $\mathbf{A}$ as $\mathbf{W}$. We use all neighbors of a node instead of selecting top-10 neighbors.

- W-PPR: This is based on personalized pagerank, which can be interpreted as similarity based on random walks [29, 38, 43]. The similarity weights are given by $\mathbf{W} = (\mathbf{M} + \mathbf{M}')/2$, where $\mathbf{M} = (1-\gamma)(I - \gamma\mathbf{A}\mathbf{D}^{-1})^{-1}$, and $\mathbf{D}$ is the diagonal matrix of degrees. We set $\gamma = \exp(-.25)$ as recommended in [43].

- JACCARD: Here we use the Jaccard index as the similarity matrix $\mathbf{W}$. The Jaccard score between two nodes i and j is defined as $\mathbf{C}_{ij}/(d_i + d_j - \mathbf{C}_{ij})$, where d_i is the degree of node i.

- CN: Here we use the number of common neighbors $\mathbf{C}$ as the similarity matrix $\mathbf{W}$.

- NODE2VEC: This constructs a node embedding $\mathbf{u}_i$ for each node i in the graph [24]. We use

the default setting of the code[1] provided by the authors. The similarity between i and j is then constructed as for SVD-RBF. That is, we set $\mathbf{W}_{ij} = \exp(-\|\mathbf{u}_i - \mathbf{u}_j\|^2/(2\theta^2))$ for a bandwidth θ.

- NOBE: This is another recent node embedding algorithm [31]. We use the default setting of the code[2] provided by the authors and construct $\mathbf{W}$ similar to node2vec.

We also compare with a method that is not based on similarity measure: regression with network cohesion (RNC) [42, 39]. Their response variables are *linear* functions of *observed* independent variables $\mathbf{Z}$ and unobserved node-wise effects that are learned from a network-based regularizer. Unlike us, the network is *fixed and not random*. To apply it to our setting, we set $\mathbf{Z}$ to zero and predict the unobserved $\mathbf{X}$ values using their semi-supervised method.

All experiments are performed with Matlab R2018b on servers with 24-core Intel Xeon X5675 and 99GB RAM.

4.1 Simulations

We generate networks from Latent Space Model, Stochastic Blockmodel, Mixed-membership Stochastic Blockmodel, and Random Dot Product Graph model. Each network has $n = 2,500$ nodes and latent dimension $d = 5$ by default. The node covariates are generated by $X_i = \boldsymbol{\beta}^T\mathbf{z}_i + \mathcal{N}(0, .1)$, where $\boldsymbol{\beta}$ is sampled uniformly from the surface of a unit sphere and $\mathbf{z}_i \in \mathbb{R}^d$ is the latent vector of node i.

For each network, we vary the fraction of nodes with unknown covariates from 0.5 to 0.9. For each fraction, we randomly select the nodes with unknown covariates and predict their covariates using the various algorithms. We report the mean and variance of root mean square error (RMSE) of the predictions over 10 runs.

Latent Space Model (LSM): The latent vectors $\mathbf{z}_i$ are sampled independently and uniformly between 0 and 1, and $\mathbf{P}_{ij} = \rho \cdot (1 + \exp(2.5 \times (\|\mathbf{z}_i - \mathbf{z}_j\|)))^{-1}$ with $\rho = 1$. Figure 2(a) shows that CN-VEC, node2vec and NOBE outperform the other methods. Under LSM, the probability matrix $\mathbf{P}$ has full rank, so SVD-RBF is not suited for this model. Indeed, we find that SVD-RBF performs similarly to CN.

Stochastic Blockmodel (SBM): We split the set of n nodes into $d = 5$ equal-sized communities; $\mathbf{z}_i = \mathbf{e}_j$ for $j \in [5]$. The probability of forming a link between i and j is given by $\mathbf{P}_{ij} = \rho \cdot \mathbf{z}_i^T\boldsymbol{\Theta}\mathbf{z}_j$ with $\rho = 0.1$, where we sample each cell of $\boldsymbol{\Theta}$ uniformly from 0 to 1, and then symmetrize $\boldsymbol{\Theta}$ by $\boldsymbol{\Theta} = (\boldsymbol{\Theta} + \boldsymbol{\Theta}^T + 2 \cdot \mathbf{I})/4$. Since this is a low-rank model, we expect SVD-RBF to perform well. Indeed, Figure 2(b) shows that CN-VEC perform best, followed by SVD-RBF, node2vec and NOBE. The remaining methods are significantly worse.

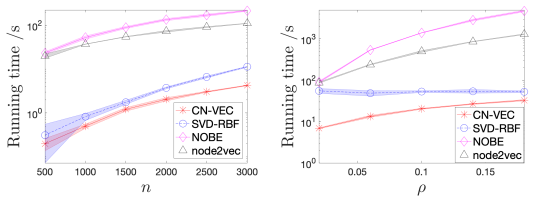
Mixed-membership Stochastic Blockmodel (MMSB): Here, each $\mathbf{z}_i$ is a 5-dimensional probability vector where the ℓ^{th} component is the probability that node i belongs to community ℓ. The latent variables $\mathbf{z}_i$ are sampled from a Dirichlet distribution that gives equal weight $1/5$ to each of the 5 communities. The link probabilities are given by $\mathbf{P}_{ij} = \rho \cdot \mathbf{z}_i^T\boldsymbol{\Theta}\mathbf{z}_j$ with $\rho = 0.1$, where $\boldsymbol{\Theta}$ has a unit diagonal and 0.1 on all off-diagonals. Thus, within-community links are preferred to across-community links. By construction, MMSB leads to a low-rank $\mathbf{P}$, so we expect SVD-RBF to do well. Figure 2(c) shows that CN-VEC, SVD-RBF, node2vec and NOBE are best.

Random Dot Product Graph model (RDPG): We sample the latent variables $\mathbf{z}_i$ from a mixture of d-dimensional Gaussians with means $\mathbf{e}_\ell$ ($\ell = 1, \ldots, 5$) and covariance $0.1 \cdot \mathbf{I}$. The link probabilities

[1] https://github.com/aditya-grover/node2vec

[2] https://github.com/Jafree/NOnBacktrackingEmbedding

are $\mathbf{P}_{ij} = \min(1, \max(0, \rho \cdot \mathbf{z}_i^T \mathbf{z}_j))$ with $\rho = 0.1$. Since $\mathbf{P}_{ij}$ is clipped to $[0, 1]$, $\mathbf{P}$ need not be low-rank. Figure 2(d) shows that CN-VEC significantly outperforms all other methods. Since $\mathbf{P}$ need not be low-rank, SVD-RBF is worse than CN-VEC, and is comparable to W-PPR and CN.



(a) Increasing n with $n\rho$ fixed.

(b) Increasing ρ with n fixed.

Figure 3: Running time (log scale).

To summarize, we find that CN-VEC, node2vec, NOBE and SVD-RBF perform better than the other methods. Among them, SVD-RBF works very well for low-rank models, as expected. The model-agnostic CN-VEC algorithm works well in most cases – it outperforms SVD-RBF by all but the MMSB model, NOBE on SBM and RDPG models, and performs comparably with node22ec on all models. However, node2vec and NOBE have no convergence guarantees and takes 10x longer time than CN-VEC, as can be seen from the wall-clock timing results in Figure 3. The timing results are for the SBM graph using Matlab implementations of all algorithms except node2vec, which is implemented in Python. For the first example, we increase n and set ρ such that $n\rho = 250$. In the second plot, we fix n and increase ρ. The same pattern is seen for other network models as well. Thus, for large networks, CN-VEC is more computationally feasible than node2vec and NOBE.

4.2 Real networks

We evaluated our method on two citation networks, namely Cora [50] and CiteSeer [22][3], and one social network, namely Sinanet [30][4]. The citation networks have roughly 3,000 nodes, with average degree 2-4. The nodes in citation networks represent publications and directed edges represents a who-cites-whom relationship. By training a topic model on the words associated with each publication, we obtain a topic distribution for each node, which are then used as node covariates. The number of topics range between 6-7. For this experiment, we remove the directionality of the edges to create an undirected network. Sinanet is a social network extracted from a microblog website[5] [30]. The nodes are users of the website, and the node covariates are the topic distributions published by Jia et al. [30]. It has roughly 3500 nodes, 10 topics and average degree 16. Table 1 shows the number of nodes, average degree and number of topics for all three datasets.

For all three datasets, the covariate for a node i is the topic distribution vector $\mathbf{X}_i$. So, our evaluation metric is the RMSE of our estimates $\hat{\mathbf{X}}_i$, measured as RMSE $= \sqrt{\frac{1}{|U|} \sum_{i \in U} \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|^2}$, where U is the set of unlabeled nodes.

In Figure 4, we see that CN-VEC, node2vec and NOBE are the best on all three datasets. SVD-RBF is comparable for Cora and CiteSeer, but much worse for Sinanet. Since real-world datasets

[3] https://linqs.soe.ucsc.edu/data

[4] https://github.com/smileyan448/Sinanet

[5] http://www.weibo.com

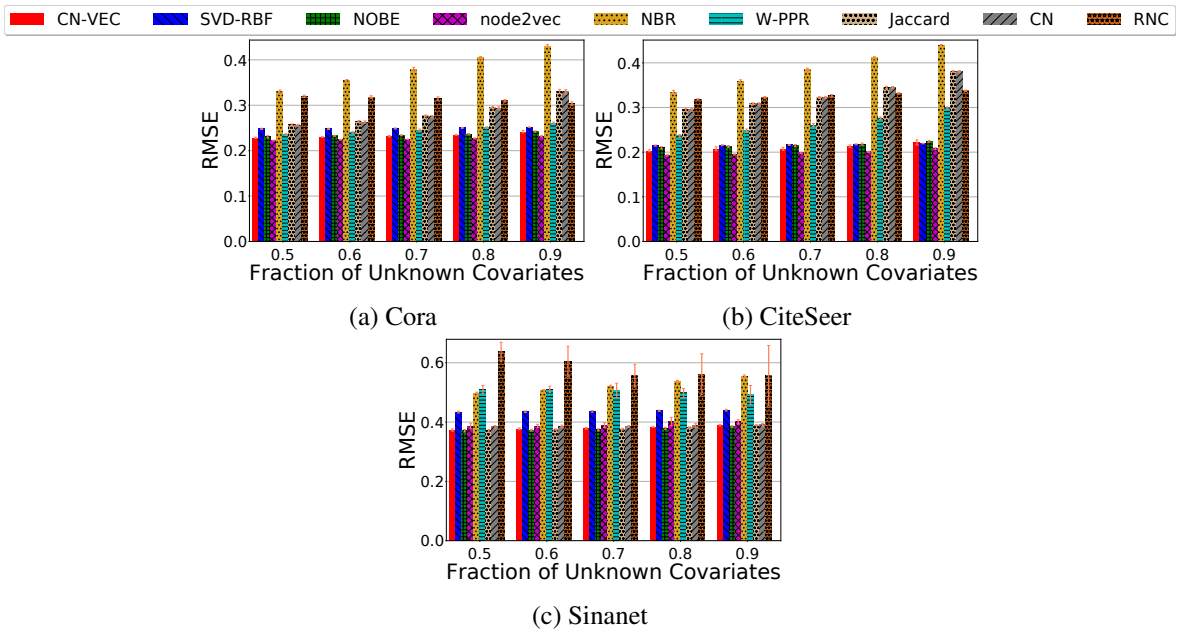


Figure 4: RMSE on recovering hidden topic distributions for each node.

Table 1: Network statistics.

DATASET	n	AVG. DEGREE	d
CORA	2,708	3.90	7
CITESEER	3,312	2.75	6
SINANET	3,490	16.4	10

may not follow low-rank models, it is not surprising that SVD-RBF fails in some cases. However, the model-agnostic CN-VEC works well everywhere.

Among the other methods, we find that CN and JACCARD have similar accuracies in all cases. For the citation networks, W-PPR is better than them. But for Sinanet, CN and JACCARD are better than W-PPR, and also SVD-RBF.

5 Conclusion

In this paper, we study the problem of estimating covariates for some nodes in a network, given the covariates for other nodes and the full network structure. This problem has applications in ad targeting and content recommendations, among others. We propose two provably consistent and computationally efficient algorithms. The first, called CN-VEC, applies without knowledge of the underlying model, which is the main contribution of our paper. The second, called SVD-RBF, is aimed at low-rank latent variable models, and works for a more flexible sparsity regime than CN-VEC. Both outperform several popular network statistics in simulated and real-world experiments, with CN-VEC being better overall. CN-VEC is also comparable or better than using a recent node-

embedding methods while being 10x-100x faster.

References

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[3] Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[4] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[5] Ernesto Araya Valdivia and De Castro Yohann. Latent distance estimation for random geometric graphs. In *Advances in Neural Information Processing Systems*, pages 8721–8731, 2019.

[6] Ery Arias-Castro, Antoine Channarond, Bruno Pelletier, and Nicolas Verzelen. On the estimation of latent distances using graph distances. *Electronic Journal of Statistics*, 15(1):722–747, 2021.

[7] Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, and Yichen Qin. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

[8] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999.

[9] Dimitris Berberidis and Georgios B Giannakis. Node embedding with adaptive similarities for scalable learning over graphs. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[10] Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

[11] Christian Borgs and Jennifer Chayes. Graphons: A nonparametric method to model, estimate, and design algorithms for massive networks. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 665–672, 2017.

[12] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. 1998.

[13] Jeff Calder and Dejan Slepčev. Properly-weighted graph laplacian for semi-supervised learning. *Applied Mathematics & Optimization*, pages 1–49, 2019.

[14] Stanley Chan and Edoardo Airoldi. A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216, 2014.

[15] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[16] Luc P Devroye. The uniform convergence of the nadaraya-watson regression function estimate. *Canadian Journal of Statistics*, 6(2):179–191, 1978.

[17] Partha Dey. Stein-chen method for poisson approximation, 2013.

[18] John Duchi. Nonparametric regression: minimax upper and lower bounds, 2019.

[19] Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin J Wainwright, and Michael I Jordan. Asymptotic behavior of\ell_p-based laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

[20] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43 (6):2624–2652, 2015.

[21] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018.

[22] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98, 1998.

[23] Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

[24] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[25] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[26] David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396, 2015.

[27] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, June 1983. ISSN 0378-8733.

[28] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, 2002.

[29] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279, 2003.

[30] Caiyan Jia, Yafang Li, Matthew B Carson, Xiaoyang Wang, and Jian Yu. Node attribute-enhanced community detection in complex networks. *Scientific reports*, 7(1):1–15, 2017.

[31] Fei Jiang, Lifang He, Yi Zheng, Enqiang Zhu, Jin Xu, and Philip S Yu. On spectral graph embedding: A non-backtracking perspective and graph approximation. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 324–332. SIAM, 2018.

[32] Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

[33] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*, 2017.

[34] Alexander Jung. Learning networked exponential families with network lasso. *arXiv preprint arXiv:1905.09056*, 2019.

[35] Alexander Jung and Nguyen Tran. Localized linear regression in networked data. *IEEE Signal Processing Letters*, 26(7):1090–1094, 2019.

[36] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[37] Emilie Kaufmann, Thomas Bonald, and Marc Lelarge. A spectral algorithm with additive clustering for the recovery of overlapping communities in networks. In *International Conference on Algorithmic Learning Theory*, pages 355–370. Springer, 2016.

[38] Isabel M Kloumann, Johan Ugander, and Jon Kleinberg. Block models and personalized pagerank. *Proceedings of the National Academy of Sciences*, 114(1):33–38, 2017.

[39] Can M Le and Tianxi Li. Linear regression and its inference on noisy network-linked data. *arXiv preprint arXiv:2007.00803*, 2020.

[40] Lucien Le Cam. An approximation theorem for the poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960.

[41] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[42] Tianxi Li, Elizaveta Levina, Ji Zhu, et al. Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132–164, 2019.

[43] Ting Li, Ningchen Ying, Xianshi Yu, and Bin-Yi Jing. Semi-supervised learning in unbalanced and heterogeneous networks. *arXiv preprint arXiv:1901.01696*, 2019.

[44] Yihua Li, Devavrat Shah, Dogyoon Song, and Christina Lee Yu. Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 2019.

[45] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[46] Zhuang Ma, Zongming Ma, and Hongsong Yuan. Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research*, 21(4):1–67, 2020.

[47] Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. On mixed memberships and symmetric nonnegative matrix factorizations. In *International Conference on Machine Learning*, pages 2324–2333. JMLR. org, 2017.

[48] Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Overlapping clustering models, and one (class) svm to bind them all. In *Advances in Neural Information Processing Systems*, pages 2126–2136, 2018.

[49] Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*, 0(0):1–13, 2020. doi: 10.1080/01621459.2020.1751645.

[50] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

[51] Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data. *Advances in neural information processing systems*, 22:1330–1338, 2009.

[52] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[53] Maxim Panov, Konstantin Slavnov, and Roman Ushakov. Consistent estimation of mixed memberships with successive projections. In *International Workshop on Complex Networks and their Applications*, pages 53–64. Springer, 2017.

[54] Srinivasan Parthasarathy, David Sivakoff, Minghao Tian, and Yusu Wang. A Quest to Unravel the Metric Structure Behind Perturbed Networks. In *33rd International Symposium on Computational Geometry (SoCG 2017)*, volume 77, pages 53:1–53:16, 2017. ISBN 978-3-95977-038-5.

[55] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

[56] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Chi Wang, Kuansan Wang, and Jie Tang. Netsmf: Large-scale network embedding as sparse matrix factorization. In *The World Wide Web Conference*, pages 1509–1520, 2019.

[57] Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *The Annals of Statistics*, 38 (5):2678–2722, 2010.

[58] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

[59] Patrick Rubin-Delanchy, Carey E Priebe, Minh Tang, and Joshua Cape. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*, 2020.

[60] Purnamrita Sarkar and Deepayan Chakrabarti. The consistency of common neighbors for link prediction in stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pages 3016–3024, 2015.

[61] Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W Moore. Theoretical justification of popular link prediction heuristics. In *Conference on Learning Theory*, 2010.

[62] Vinesh Solanki, Patrick Rubin-Delanchy, and Ian Gallagher. Persistent homology of graph embeddings. *arXiv preprint arXiv:1912.10238*, 2019.

[63] Dogyoon Song, Christina E Lee, Yihua Li, and Devavrat Shah. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2155–2163, 2016.

[64] Daniel L Sussman, Minh Tang, and Carey E Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57, 2014.

[65] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077, 2015.

[66] Minh Tang, Daniel L Sussman, and Carey E Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, 2013.

[67] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. Verse: Versatile graph embeddings from similarity measures. In *Proceedings of the 2018 World Wide Web Conference*, pages 539–548, 2018.

[68] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

[69] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

[70] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[71] Haolei Weng and Yang Feng. Community detection with nodal information. *arXiv preprint arXiv:1610.09735*, 2016.

[72] Dominik Wied and Rafael Weißbach. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53(1):1–21, 2012.

[73] Xiao-Ming Wu, Zhenguo Li, Anthony M So, John Wright, and Shih-Fu Chang. Learning with partially absorbing random walks. In *Advances in neural information processing systems*, pages 3077–3085, 2012.

[74] Zhu Xiaojin and Ghahramani Zoubin. Learning from labeled and unlabeled data with label propagation. *Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University*, 2002.

[75] Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, 2018.

[76] Bowei Yan and Purnamrita Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, 0(0):1–12, 2020. doi: 10.1080/01621459.2019.1706541.

[77] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *13th International Conference on Data Mining*, pages 1151–1156. IEEE, 2013.

[78] Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

[79] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, pages 5165–5175, 2018.

[80] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153–3178, 2016.

[81] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783, 2017.

[82] Yun Zhang, Kehui Chen, Allan Sampson, Kai Hwang, and Beatriz Luna. Node features adjusted stochastic block model. *Journal of Computational and Graphical Statistics*, 28(2):362–373, 2019.

[83] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

[84] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, 2009.

[85] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *International conference on Machine learning*, pages 912–919, 2003.

Supplementary Material

In this document, we present the technical proofs of the results in our main paper. To be specific, we first show the proof of Theorems 3.1 (concentration of $\mathbf{K}_{ij}$) and 3.2 (consistency of k-NN estimator) in Sections A.1 and A.2 respectively, which constitutes the whole proof for consistency of CN-VEC. Then we show that Assumption 3.2 holds for GRDPG/SBM/MMSB in Section B. Finally we show the proof of Proposition 3.1 (consistency of SVD-RBF) in Section C. We follow the numbering scheme of the main paper.

A Consistency of CN-VEC (Proof of Theorems 3.1 and 3.2)

We first show that $\mathbf{K}_{ij}$ concentrates around its expectation in Section A.1, and then show that using the expectation as distance metric on k-nearest-neighbors regression will give us consistent estimation in Section A.2.

A.1 Concentration of K_{ij} (Proof of Theorem 3.1)

Lemma A.1. Recall that $\mathbf{C}_{ij} = \sum_{k\neq i,j} \mathbf{A}_{ik}\mathbf{A}_{jk} = \sum_{k\neq i,j} \mathrm{Bernoulli}(\mathbf{P}_{ik}\mathbf{P}_{jk})$ is the number of common neighbors between node i and j. We have for $m > 0$,

$$\mathbf{P}(|\mathbf{C}_{ij} - \mathbf{E}[\mathbf{C}_{ij}]| \geq m) \leq 2\exp\left(-\frac{m^2/2}{n\rho^2 + m/3}\right).$$

Proof. Let $X_h = \mathbf{A}_{ih}\mathbf{A}_{jh} - \mathbf{P}_{ih}\mathbf{P}_{jh}$, then $|X_i| \leq 1$, and,

$$\begin{aligned}\mathrm{E}[X_h^2] &= \mathrm{E}[\mathbf{A}_{ih}\mathbf{A}_{jh} + \mathbf{P}_{ih}^2\mathbf{P}_{jh}^2 - 2\mathbf{A}_{ih}\mathbf{A}_{jh}\mathbf{P}_{ih}\mathbf{P}_{jh}] \\ &= \mathbf{P}_{ih}\mathbf{P}_{jh} - \mathbf{P}_{ih}^2\mathbf{P}_{jh}^2 \leq \rho^2.\end{aligned}$$

Applying Bernstein inequality, we have:

$$\mathbf{P}(|\mathbf{C}_{ij} - \mathbf{E}[\mathbf{C}_{ij}]| \geq m) \leq 2\exp\left(-\frac{m^2/2}{\sum_h \mathrm{E}[X_h^2] + m/3}\right) \leq 2\exp\left(-\frac{m^2/2}{n\rho^2 + m/3}\right).$$

□

Theorem A.2. When $n\rho^2 \geq c\log^\xi n$ for some constant c and $\xi > 1$,

$$\mathbf{K}_{ij} = \left\|(\mathbf{e}_i - \mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I} - \mathbf{e}_i\mathbf{e}_i^T - \mathbf{e}_j\mathbf{e}_j^T\right)\right\|^2 + O_P(n^{2.5}\rho^3\sqrt{\log n}) - 4(n-2).$$

Proof. By Lemma A.1 with $m = c_r\sqrt{\rho^2 n\log n}$ for some constant c_r, we have:

$$\mathbf{P}(|\mathbf{C}_{ij} - \mathbf{E}[\mathbf{C}_{ij}]| \geq m) \leq 2\exp\left(-\frac{m^2/2}{n\rho^2 + m/3}\right) \leq \frac{2}{n^r}.$$

Then we have,

$$\mathbf{C}_{ik} = \mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k + t_{ik},$$

where $t_{ik} = O_P(\sqrt{\rho^2 n\log n})$, with probability larger than $1 - 1/n^r$.

By definition of $\mathbf{K}_{ij}$, and noting that $\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k = O(n\rho^2)$, we have,

$$\begin{aligned}
\mathbf{K}_{ij} &= \sum_{k\neq i,j}\left[(\mathbf{C}_{ik}^2-2)1(\mathbf{C}_{ik}\geq 2)+(\mathbf{C}_{jk}^2-2)1(\mathbf{C}_{jk}\geq 2)-2\mathbf{C}_{ik}\mathbf{C}_{jk}\right]\\
&= \sum_{k\neq i,j}\left[(\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k+t_{ik})^2+(\mathbf{e}_j^T\mathbf{P}^2\mathbf{e}_k+t_{jk})^2-2(\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k+t_{ik})(\mathbf{e}_j^T\mathbf{P}^2\mathbf{e}_k+t_{jk})\right]-4(n-2)\\
&= \sum_{k\neq i,j}\left[(\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k)^2+(\mathbf{e}_j^T\mathbf{P}^2\mathbf{e}_k)^2-2(\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k)(\mathbf{e}_j^T\mathbf{P}^2\mathbf{e}_k)\right]+\sum_{k\neq i,j}\left[2(t_{ik}-t_{jk})(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\mathbf{e}_k+(t_{ik}-t_{jk})^2\right]-4(n-2)\\
&= \sum_{k\neq i,j}\left[(\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k)^2+(\mathbf{e}_j^T\mathbf{P}^2\mathbf{e}_k)^2-2(\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k)(\mathbf{e}_j^T\mathbf{P}^2\mathbf{e}_k)\right]+O_P(n^{2.5}\rho^3\sqrt{\log n})-4(n-2)\\
&= \left\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\right\|+O_P(n^{2.5}\rho^3\sqrt{\log n})-4(n-2).
\end{aligned}$$

□

Lemma A.3. When $n\rho^2 = o(1)$, and $n^2\rho^3 = \Omega(\log n)$, we have,

$$\mathrm{E}[\mathbf{K}_{ij}] = \left\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\right\|^2+O(n^4\rho^6).$$

Proof. Recall that $\mathbf{C}_{ij}=\sum_{k\neq i,j}\mathbf{A}_{ik}\mathbf{A}_{jk}=\sum_{k\neq i,j}\mathrm{Bernoulli}(\mathbf{P}_{ik}\mathbf{P}_{jk})$ is the number of common neighbors between node i and j. Then when $n\rho^2\to 0$, by [40] and simple algebraic manipulations using Theorem 4.1 of [17], $\mathbf{C}_{ij}$ can be approximated by a Poisson Distribution with parameter $\lambda=\bar{\mathbf{C}}_{ij}=\sum_{k\neq i,j}\mathbf{P}_{ik}\mathbf{P}_{jk}=\Theta(n\rho^2)$. Then, in the Poisson limit, we have

$$\begin{aligned}
\mathrm{E}[(\mathbf{C}_{ik}^2-2)1(\mathbf{C}_{ik}\geq 2)] =&\mathrm{E}[\mathbf{C}_{ik}^2]-\sum_{h<2}h^2\frac{\lambda^h e^{-\lambda}}{h!}-2(1-\sum_{h<2}\frac{\lambda^h e^{-\lambda}}{h!})\\
=&\mathrm{var}[\mathbf{C}_{ik}]+(\mathrm{E}[\mathbf{C}_{ik}])^2-\lambda e^{-\lambda}-2(1-e^{-\lambda}-\lambda e^{-\lambda})\\
=&\lambda+\lambda^2+\lambda e^{-\lambda}+2e^{-\lambda}-2\\
=&\lambda+\lambda^2+(\lambda+2)(1-\lambda+\frac{\lambda^2}{2}-\frac{\lambda^3}{6}+\cdots)-2\\
=&\lambda^2+O(\lambda^3).
\end{aligned}$$

Since $\mathbf{C}_{ik}=\sum_{u\neq i,k}\mathbf{A}_{ku}\mathbf{A}_{iu}$, we have, for $k\neq i,j$,

$$\begin{aligned}
\mathbf{C}_{ik}\mathbf{C}_{jk} &= \sum_{u\neq i,k}\mathbf{A}_{ku}\mathbf{A}_{iu}\sum_{v\neq j,k}\mathbf{A}_{kv}\mathbf{A}_{jv}\\
&= \sum_{u\neq v;u,v\neq i,j,k}\mathbf{A}_{ku}\mathbf{A}_{iu}\mathbf{A}_{kv}\mathbf{A}_{jv}+\mathbf{A}_{ij}\mathbf{A}_{ki}\sum_{u\neq i,j,k}\mathbf{A}_{ku}\mathbf{A}_{iu}+\mathbf{A}_{ij}\mathbf{A}_{kj}\sum_{u\neq i,j,k}\mathbf{A}_{ku}\mathbf{A}_{ju}\\
&\quad+\sum_{u\neq i,j,k}\mathbf{A}_{ku}^2\mathbf{A}_{iu}\mathbf{A}_{ju}+\mathbf{A}_{ij}^2\mathbf{A}_{ki}\mathbf{A}_{kj}\\
&= \sum_{u\neq v;u,v\neq i,j,k}\mathbf{A}_{ku}\mathbf{A}_{iu}\mathbf{A}_{kv}\mathbf{A}_{jv}+\mathbf{A}_{ij}\mathbf{A}_{ki}\sum_{u\neq i,j,k}\mathbf{A}_{ku}\mathbf{A}_{iu}+\mathbf{A}_{ij}\mathbf{A}_{kj}\sum_{u\neq i,j,k}\mathbf{A}_{ku}\mathbf{A}_{ju}\\
&\quad+\sum_{u\neq i,j,k}\mathbf{A}_{ku}\mathbf{A}_{iu}\mathbf{A}_{ju}+\mathbf{A}_{ij}\mathbf{A}_{ki}\mathbf{A}_{kj}
\end{aligned}$$

Then as different edges are independent, we have,

$$\begin{aligned}
\mathrm{E}[\mathbf{C}_{ik}\mathbf{C}_{jk}] &= \sum_{u\neq v;u,v\neq i,j,k} \mathbf{P}_{ku}\mathbf{P}_{iu}\mathbf{P}_{kv}\mathbf{P}_{jv} + \mathbf{P}_{ij}\mathbf{P}_{ki}\sum_{u\neq i,j,k}\mathbf{P}_{ku}\mathbf{P}_{iu} + \mathbf{P}_{ij}\mathbf{P}_{kj}\sum_{u\neq i,j,k}\mathbf{P}_{ku}\mathbf{P}_{ju} \\
&\quad + \sum_{u\neq i,j,k}\mathbf{P}_{ku}\mathbf{P}_{iu}\mathbf{P}_{ju} + \mathbf{P}_{ij}\mathbf{P}_{ki}\mathbf{P}_{kj} \\
&= \sum_{u\neq i,k}\mathbf{P}_{ku}\mathbf{P}_{iu}\sum_{v\neq j,k}\mathbf{P}_{kv}\mathbf{P}_{jv} - \sum_{u\neq i,j,k}\mathbf{P}_{ku}^2\mathbf{P}_{iu}\mathbf{P}_{ju} - \mathbf{P}_{ij}^2\mathbf{P}_{ki}\mathbf{P}_{kj} + \sum_{u\neq i,j,k}\mathbf{P}_{ku}\mathbf{P}_{iu}\mathbf{P}_{ju} + \mathbf{P}_{ij}\mathbf{P}_{ki}\mathbf{P}_{kj} \\
&= \sum_{u\neq i,k}\mathbf{P}_{ku}\mathbf{P}_{iu}\sum_{v\neq j,k}\mathbf{P}_{kv}\mathbf{P}_{jv} + O(n\rho^3).
\end{aligned}$$

Summing up, we have,

$$\begin{aligned}
\mathrm{E}[\mathbf{K}_{ij}] &= \mathrm{E}\left[\sum_{k\neq i,j}\left[(\mathbf{C}_{ik}^2-2)1(\mathbf{C}_{ik}\geq 2) + (\mathbf{C}_{jk}^2-2)1(\mathbf{C}_{jk}\geq 2) - 2\mathbf{C}_{ik}\mathbf{C}_{jk}\right]\right] \\
&= \sum_{k\neq i,j}\left[(\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k)^2 + (\mathbf{e}_j^T\mathbf{P}^2\mathbf{e}_k)^2 - 2\mathbf{e}_i^T\mathbf{P}^2\mathbf{e}_k\mathbf{e}_j^T\mathbf{P}^2\mathbf{e}_k + O(n^3\rho^6) + O(n\rho^3)\right] \\
&= \sum_{k\neq i,j}\left((\mathbf{P}^2)_{ik} - (\mathbf{P}^2)_{jk}\right)^2 + O(n^4\rho^6) \\
&= \left\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\right\|^2 + O(n^4\rho^6),
\end{aligned}$$

when $n^2\rho^3 = \Omega(\log n)$. □

Theorem A.4. When $n\rho^2 = o(1)$ and $n^2\rho^3 = \Omega(\log^\xi n)$, for some $\xi > 2.5$, we have,

$$\mathbf{K}_{ij} = \mathrm{E}[\mathbf{K}_{ij}] + O_P\left(n\rho\sqrt{\log^5 n}\right) = \left\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\right\|^2 + O_P\left(n^4\rho^6 + n\rho\sqrt{\log^5 n}\right).$$

Proof. Let $m = c\log n$ for some constant c to be decided, and $\mathbf{a}_i$ be the i^{th} column of $\mathbf{A}$. Define the following events:

$$\begin{aligned}
\mathcal{E}_0 &= \{\forall i\in[n] : \mathbf{a}_i^T\mathbf{1} = O(n\rho)\} \\
\mathcal{E}_1 &= \{\forall i,j\in[n] : C_{ij} = \mathbf{a}_i^T\mathbf{a}_j \leq m\}
\end{aligned}$$

By Chernoff bound when $n\rho = \Omega(\log n)$ we have $\mathrm{P}(\mathcal{E}_0)\geq 1-1/n^s$ for some constant s.

Give $\mathbf{a}_i$, let $X_h = (\mathbf{A}_{kh}-\mathbf{P}_{kh})\mathbf{A}_{ih}$, then we have $|X_h|\leq 1$ and $\sum_h \mathrm{var}(X_h|\mathbf{a}_i,\mathcal{E}_0) = \sum_h \mathbf{A}_{ih}^2\mathrm{var}(\mathbf{A}_{kh})|\mathcal{E}_0 = \sum_h \mathbf{A}_{ih}\mathbf{P}_{kh}(1-\mathbf{P}_{kh})|\mathcal{E}_0 \leq \sum_h \rho\mathbf{A}_{ih}|\mathcal{E}_0 = O(n\rho^2)$.

By Bernstein's inequality,

$$\mathrm{P}(|\mathbf{e}_k^T\mathbf{A}\mathbf{a}_i - \mathbf{e}_k^T\mathbf{P}\mathbf{a}_i| \geq t|\mathcal{E}_0) \leq 2\exp\left(-\frac{t^2/2}{2n\rho^2 + t/3}\right),$$

taking $t = c_r\log n$, we have with probability larger than $1-2/n^r$,

$$\mathbf{C}_{ik}|\mathbf{a}_i,\mathcal{E}_0 = \mathbf{e}_k^T\mathbf{A}\mathbf{a}_i|\mathcal{E}_0 \leq \mathbf{e}_k^T\mathbf{P}\mathbf{a}_i|\mathcal{E}_0 + t|\mathcal{E}_0 = O(n\rho^2) + c_r\log n = c\log n := m,$$

for some constant c. Then $\mathrm{P}(\mathcal{E}_1|\mathcal{E}_0)\geq 1-2/n^r$.

Recall

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

then $-8 < Y_k|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_0, \mathcal{E}_1 \leq 2m^2$. Under condition $\mathcal{E}_0$, there are at most $O(n\rho)$ neighbors of each node, so there are at most $n^2\rho^2$ 2-hop neighbors of each node. In that case, there are at most $O(n^2\rho^2)$ nonzero Y_k. Applying Hoeffding's inequality with $t = c'\rho n m^2\sqrt{\log n}$ for some constant c' to those nonzero Y_ks, we have

$$\mathrm{P}(|\mathbf{K}_{ij} - \mathrm{E}[\mathbf{K}_{ij}]| \geq t|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_0, \mathcal{E}_1) \leq 2\exp\left(-\frac{t^2}{\rho^2 n^2(2m^2+8)^2}\right) \leq \frac{1}{n^C},$$

for some constant C.

Now let use get the order of $\mathrm{E}[\mathbf{K}_{ij}|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_0, \mathcal{E}_1]$.

Consider sets:

$$\begin{aligned} I &:= \{u : \mathbf{A}_{iu} = 1, \mathbf{A}_{ju} = 0\} \\ J &:= \{u : \mathbf{A}_{iu} = 0, \mathbf{A}_{ju} = 1\} \\ B &:= \{u : \mathbf{A}_{iu} = 1, \mathbf{A}_{ju} = 1\}. \end{aligned}$$

Define $\tilde{C}_{kI} = \sum_{u\in I}\mathbf{P}_{ku}\mathbf{A}_{iu}$, similarly define $\tilde{C}_{kJ} = \sum_{u\in J}\mathbf{P}_{ku}\mathbf{A}_{ju}$, $\tilde{C}_{kB} = \sum_{u\in B}\mathbf{P}_{ku}\mathbf{A}_{iu}$. Then $\tilde{C}_{kI}|\mathcal{E}_0, \mathcal{E}_1 = O(n\rho^2)$, $\tilde{C}_{kJ}|\mathcal{E}_0, \mathcal{E}_1 = O(n\rho^2)$, $\tilde{C}_{kB}|\mathcal{E}_0, \mathcal{E}_1 = O(\rho\log n)$. As $\mathrm{E}[\mathbf{C}_{ki}|\mathbf{a}_i] = \sum_{u\neq i,k}\mathbf{P}_{ku}\mathbf{A}_{iu} = \tilde{C}_{kI} + \tilde{C}_{kB}$, $\mathrm{E}[\mathbf{C}_{kj}|\mathbf{a}_j] = \sum_{u\neq i,k}\mathbf{P}_{ku}\mathbf{A}_{ju} = \tilde{C}_{kJ} + \tilde{C}_{kB}$, we have,

$$\begin{aligned} &\mathrm{E}[\mathbf{C}_{ik}\mathbf{C}_{jk}|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_0, \mathcal{E}_1] \\ =&\mathrm{E}\left[(\sum_{u\in I}\mathbf{A}_{ku}\mathbf{A}_{iu} + \sum_{u\in B}\mathbf{A}_{ku}\mathbf{A}_{iu})(\sum_{u\in J}\mathbf{A}_{ku}\mathbf{A}_{ju} + \sum_{u\in B}\mathbf{A}_{ku}\mathbf{A}_{ju})|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_0, \mathcal{E}_1\right] \\ =&\left(\tilde{C}_{kI}\tilde{C}_{kJ} + \tilde{C}_{kI}\tilde{C}_{kB} + \tilde{C}_{kB}\tilde{C}_{kJ}\right)|\mathcal{E}_0, \mathcal{E}_1 + \mathrm{E}\left[\sum_{u\in B}\mathbf{A}_{ku}\mathbf{A}_{iu}\sum_{v\in B}\mathbf{A}_{kv}\mathbf{A}_{jv}|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_0, \mathcal{E}_1\right] \\ =&O(n^2\rho^4) + \mathrm{E}\left[\sum_{u\neq v; u,v\in B}\mathbf{A}_{ku}\mathbf{A}_{iu}\mathbf{A}_{kv}\mathbf{A}_{jv} + \sum_{u\in B}\mathbf{A}_{ku}\mathbf{A}_{iu}\mathbf{A}_{ku}\mathbf{A}_{ju}|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_0, \mathcal{E}_1\right] \\ =&O(n^2\rho^4) + O(\rho^2\log^2 n) + \sum_{u\in B}\mathbf{P}_{ku}\mathbf{A}_{iu}\mathbf{A}_{ju}|\mathcal{E}_0, \mathcal{E}_1 \\ =&O(n^2\rho^4) + O(\rho\log n). \end{aligned}$$

Now, conditioned on $\mathcal{E}_0$ and $\mathbf{a}_i$, we have $\lambda = \mathbf{e}_k^T\mathbf{P}\mathbf{a}_i = O(n\rho^2)$. With same arguments in Lemma A.3, we have

$$\mathrm{E}[(\mathbf{C}_{ik}^2 - 2)1(\mathbf{C}_{ik} \geq 2)] = \lambda^2 + o(\lambda^2) = O(\rho^4 n^2).$$

Similarly,

$$\mathrm{E}[(\mathbf{C}_{jk}^2 - 2)1(\mathbf{C}_{jk} \geq 2)] = O(n^2\rho^4).$$

Summing up, $\mathrm{E}[\mathbf{K}_{ij}|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_i, \mathcal{E}_{ij}] = O(n^3\rho^4) + O(\rho n\log n)$. So as long as $n^3\rho^4 \gg \rho n m^2\sqrt{\log n}$, we have $\mathbf{K}_{ij}|\mathbf{a}_i, \mathbf{a}_j, \mathcal{E}_0, \mathcal{E}_1$ concentrates around its expectation. Using a union bound on failure probabilities of $\mathcal{E}_0$, $\mathcal{E}_1$ and note that for any $\mathbf{a}_i$, $\mathbf{a}_j$ the equation holds, we can remove the above condition and have the same concentration. As

$$\begin{aligned} \mathrm{E}[\mathbf{K}_{ij}] &= \mathrm{P}(\mathcal{E}_0, \mathcal{E}_1)\cdot\mathrm{E}[\mathbf{K}_{ij}|\mathcal{E}_0, \mathcal{E}_1] + \mathrm{P}(\neg(\mathcal{E}_0, \mathcal{E}_1))\cdot\mathrm{E}[\mathbf{K}_{ij}|\neg(\mathcal{E}_0, \mathcal{E}_1)] \\ &\leq \mathrm{E}[\mathbf{K}_{ij}|\mathcal{E}_0, \mathcal{E}_1] + O\left(n^3\left(\frac{1}{n^s} + \frac{1}{n^r}\right)\right), \end{aligned}$$

and

$$\begin{aligned}\mathrm{E}[\mathbf{K}_{ij}] &= \mathrm{P}(\mathcal{E}_0,\mathcal{E}_1)\cdot\mathrm{E}[\mathbf{K}_{ij}|\mathcal{E}_0,\mathcal{E}_1] + \mathrm{P}(\neg(\mathcal{E}_0,\mathcal{E}_1))\cdot\mathrm{E}[\mathbf{K}_{ij}|\neg(\mathcal{E}_0,\mathcal{E}_1)]\\ &\geq \left(1-O\left(\frac{1}{n^s}+\frac{1}{n^r}\right)\right)\mathrm{E}[\mathbf{K}_{ij}|\mathcal{E}_0,\mathcal{E}_1],\end{aligned}$$

with r, s chosen sufficient large (>3), we have $\mathrm{E}[\mathbf{K}_{ij}] = \mathrm{E}[\mathbf{K}_{ij}|\mathcal{E}_0,\mathcal{E}_1](1+o(1))$, this concludes the proof combining with Lemma A.3.

□

Proof of Theorem 3.1. It is a clear combination of Theorems A.2 and A.4. □

A.2 Consistency of K-NN estimator (Proof of Theorem 3.2)

Let $\mathbf{s}_i = \frac{1}{n^{1.5}\rho^2}\mathbf{e}_i^T\mathbf{P}^2\left(I-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)$, $\forall i\in[n]$. For all $i,j\in[n]$, if $\|\mathbf{z}_i-\mathbf{z}_j\|\leq\epsilon$, by Assumption 3.2, we have $\|\mathbf{s}_i-\mathbf{s}_j\|\leq L\epsilon$.

Lemma A.5. For all $i,j\in[n]$, if $\|\mathbf{s}_i-\mathbf{s}_j\|\leq\delta$, then $\|\mathbf{z}_i-\mathbf{z}_j\|\leq\epsilon_\delta$ for $\epsilon_\delta=(\delta+\Delta_n)/\ell$.

Proof. By Assumption 3.2, we have,

$$\|\mathbf{z}_i-\mathbf{z}_j\|\leq\frac{1}{\ell}\left(\|\mathbf{s}_i-\mathbf{s}_j\|+\Delta_n\right)\leq\frac{1}{\ell}\left(\delta+\Delta_n\right)$$

□

Lemma A.6. Denote the probability measure of Z by μ. Let $B_{\mathbf{z},\epsilon}$ be the closed ball centered at $\mathbf{z}$ of radius $\epsilon>0$. The collection of all $\mathbf{z}$ with $\mu(B_{\mathbf{z},\epsilon})>0$ for all $\epsilon>0$ is called the support of Z or μ. Suppose $\forall i,j$ such that $\|\mathbf{z}_i-\mathbf{z}_j\|\leq\epsilon$, $\exists\delta_\epsilon$ such that $\|\mathbf{s}_i-\mathbf{s}_j\|\leq\delta_\epsilon$, then if $\mathbf{z}\in\text{support}(\mu)$ and $\lim_{n\to\infty}k_n/n=0$, we have $\|S_{(k_n,n)}(\mathbf{s})-\mathbf{s}\|\to0$, where $S_{(k_n,n)}(\mathbf{s})$ is the k_n-th nearest neighbor of $\mathbf{s}$.

Proof. We adapt a similar proof as that for Lemma 6.1 in [25]. Denote ν as the probability measure of S. $\forall\epsilon>0$, by definition, $\mathbf{z}\in\text{support}(\mu)$ implies $\mu(B_{\mathbf{z},\epsilon})>0$. Now as $\forall i,j$ such that $\|\mathbf{z}_i-\mathbf{z}_j\|\leq\epsilon$, $\exists\delta_\epsilon$ such that $\|\mathbf{s}_i-\mathbf{s}_j\|\leq\delta_\epsilon$, we have $\nu(B_{\mathbf{s},\delta_\epsilon})>0$, then following the proof in [25], we have $\|S_{(k_n,n)}(\mathbf{s})-\mathbf{s}\|\to0$. □

Proof of Theorem 3.2. First note that using $\mathbf{s}_i$ to calculate distances between node i and node j is same as using $\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\|$ as the distance metric.

Let $\eta(\mathbf{z})=g(\mathbf{z})+\boldsymbol{\epsilon}_\mathbf{z}\in\mathbb{R}^p$, $\eta_n(\mathbf{z},\mathbf{s})=\sum_{i=1}^n W_{ni}(\mathbf{s})\eta(\mathbf{z}_i)$, where $W_{ni}(\mathbf{s})$ is the weight for the i-th node, but this weighted is calculated with $\mathbf{s}$, for the k_n nearest neighbors, it is $1/k$, and 0 otherwise. Note we use n in the definition of $\eta_n(\mathbf{z},\mathbf{s})$ for simplicity of notation, it should actually be $|S|$, the number of nodes that have known covariates in the network. However, as we assume $|S|=\Theta(n)$, it will not affect the consistency result. Let $\bar{\boldsymbol{\epsilon}}=\sum_{i=1}^n W_{ni}(\mathbf{s})\boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ is a simpler notation of $\boldsymbol{\epsilon}_{\mathbf{z}_i}$. Then for each $\delta>0$,

$$\begin{aligned}&\mathrm{E}[\|\eta_n(\mathbf{z},\mathbf{s})-g(\mathbf{z})\|^2] = \mathrm{E}\left[\|\sum_{i=1}^n W_{ni}(\mathbf{s})(\eta(\mathbf{z}_i)-g(\mathbf{z}))\|^2\right] = \mathrm{E}\left[\|\sum_{i=1}^n W_{ni}(\mathbf{s})(g(\mathbf{z}_i)-g(\mathbf{z})+\boldsymbol{\epsilon}_i)\|^2\right]\\ &\leq \mathrm{E}\left[\|\sum_{i=1}^n W_{ni}(\mathbf{s})\boldsymbol{\epsilon}_i+\sum_{i=1}^n W_{ni}(\mathbf{s})(g(\mathbf{z}_i)-g(\mathbf{z}))\|^2\right]\\ &\leq \underbrace{\mathrm{E}[\|\bar{\boldsymbol{\epsilon}}\|^2]+\mathrm{E}\left[\bar{\boldsymbol{\epsilon}}^T\sum_{i=1}^n W_{ni}(\mathbf{s})(g(\mathbf{z}_i)-g(\mathbf{z}))\right]}_{R_1}+\underbrace{\mathrm{E}\left[\|\sum_{i=1}^n W_{ni}(\mathbf{s})(g(\mathbf{z}_i)-g(\mathbf{z}))\|^2\right]}_{R_2}.\end{aligned}$$

For R_1, recall that $\boldsymbol{\epsilon}_i$ are i.i.d. noise random vectors with uncorrelated elements, whose mean is 0 and variance is σ^2. By central limit theorem, $\forall i \in [p]$, $\mathbf{e}_i^T \bar{\boldsymbol{\epsilon}} \to \mathcal{N}(0, \frac{\sigma^2}{k})$ as $k_n \to \infty$, then

$$\mathrm{E}\left[\bar{\boldsymbol{\epsilon}}^T \sum_{i=1}^{n} W_{ni}(\mathbf{s})(g(\mathbf{z}_i) - g(\mathbf{z}))\right] = 0,$$

and,

$$\mathrm{E}[\|\bar{\boldsymbol{\epsilon}}\|^2] = \sum_i \mathrm{E}[(\mathbf{e}_i^T \bar{\boldsymbol{\epsilon}})^2] = \sum_i \mathrm{var}(\mathbf{e}_i^T \bar{\boldsymbol{\epsilon}}) = \frac{p\sigma^2}{k_n}.$$

For R_2, By Jensen's inequality,

$$\begin{aligned}
&\mathrm{E}\left[\|\sum_{i=1}^{n} W_{ni}(\mathbf{s})(g(\mathbf{z}_i) - g(\mathbf{z}))\|^2\right] \leq \mathrm{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{s})\|g(\mathbf{z}_i) - g(\mathbf{z})\|^2\right] \\
=&\mathrm{E}\left[\sum_{i=1}^{n} \frac{1}{k_n} 1(\|\mathbf{s}_i - \mathbf{s}_j\| > \delta)\|g(\mathbf{z}_i) - g(\mathbf{z})\|^2\right] + \mathrm{E}\left[\sum_{i=1}^{n} \frac{1}{k_n} 1(\|\mathbf{s}_i - \mathbf{s}_j\| \leq \delta)\|g(\mathbf{z}_i) - g(\mathbf{z})\|^2\right],
\end{aligned}$$

Now, using Lemma A.6, with similar argument in [25], we have as $n \to \infty$,

$$\mathrm{E}\left[\sum_{i=1}^{n} \frac{1}{k_n} 1(\|\mathbf{s}_i - \mathbf{s}_j\| > \delta)\|g(\mathbf{z}_i) - g(\mathbf{z})\|^2\right] \to 0$$

Note that from Lemmm A.5, $\forall i, j$, if $\|\mathbf{s}_i - \mathbf{s}_j\| \leq \delta$, we have $\|\mathbf{z}_i - \mathbf{z}_j\| \leq \epsilon_\delta = (\delta + \Delta_n)/\ell$, then

$$\mathrm{E}\left[\sum_{i=1}^{n} \frac{1}{k_n} 1(\|\mathbf{s}_i - \mathbf{s}_j\| \leq \delta)\|g(\mathbf{z}_i) - g(\mathbf{z})\|^2\right] \leq \mathrm{E}\left[\sum_{i=1}^{n} \frac{1}{k_n} 1(\|\mathbf{z}_i - \mathbf{z}_j\| \leq \epsilon_\delta)\|g(\mathbf{z}_i) - g(\mathbf{z})\|^2\right] \leq L_g^2 \epsilon_\delta^2 = \frac{L_g^2}{\ell^2}(\delta + \Delta_n)^2,$$

where the last inequality is due to Assumption 3.1.

Summing up, as $\delta \to 0$ and $n \to \infty$, we have $\Delta_n \to 0$, and by assumption $p = \Theta(1)$, $\sigma = \Theta(1)$, $k_n \to \infty$, so $p\sigma^2/k_n \to 0$, then

$$\mathrm{E}[\|\eta_n(\mathbf{z}, \mathbf{s}) - g(\mathbf{z})\|^2] \to 0.$$

□

B Assumption 3.2 holds for GRDPG/SBM/MMSB

For SBM/MMSB, we have $\mathbf{P}_{ij} = \rho \mathbf{Z}\boldsymbol{\Theta}\mathbf{Z}^T$. Note that if $\mathbf{z}_i = \mathbf{z}_j$, $\mathbf{e}_i\mathbf{P} = \mathbf{e}_j\mathbf{P}$, Assumption 3.2 automatically holds. Now consider the case where $\mathbf{z}_i \neq \mathbf{z}_j$. We have,

$$\begin{aligned}
&\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{P}^2 \left(\mathbf{I} - \mathbf{e}_i\mathbf{e}_i^T - \mathbf{e}_j\mathbf{e}_j^T\right)\| \leq 3\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{P}^2\| = 3\rho^2\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{Z}\boldsymbol{\Theta}\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Theta}\mathbf{Z}^T\| \\
\leq& 3\rho^2\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{Z}\|\|\boldsymbol{\Theta}\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Theta}\mathbf{Z}^T\| \leq 3\rho^2\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{Z}\|(\lambda_1(\boldsymbol{\Theta}))^2(\lambda_1(\mathbf{Z}^T\mathbf{Z}))^{1.5}.
\end{aligned}$$

Denote the i-th eigenvalue and i-th singular value of a matrix as $\lambda_i(\cdot)$ and $\sigma_i(\cdot)$ respectively. Let $\mathbf{P} = \mathbf{U}\mathbf{E}\mathbf{U}^T$ be the eigen-decomposition of $\mathbf{P}$, as shown in [49], we have $\mathbf{U} = \mathbf{Z}\mathbf{U}_P$ and $\lambda_d(\mathbf{U}_P\mathbf{U}_P^T) = 1/\lambda_1(\mathbf{Z}^T\mathbf{Z})$, where P is the set that for each $i \in P$, $\mathbf{z}_i$ only has 1 nonzero entry, and each $\mathbf{z}_i$ are different, $|P| = d$. Then,

$$\begin{aligned}
&\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{P}^2\| = \rho^2\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{Z}\mathbf{U}_P\mathbf{E}^2\mathbf{U}^T\| \geq \rho^2\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{Z}\|\sigma_d(\mathbf{U}_P\mathbf{E}^2\mathbf{U}^T) \\
\geq& \rho^2\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{Z}\|(\lambda_d(\mathbf{U}_P\mathbf{U}_P^T))^{0.5}\lambda_d(\mathbf{E}^2) \geq \rho^2\|(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{Z}\|(\kappa(\mathbf{Z}^T\mathbf{Z}))^{-0.5}(\lambda_d(\boldsymbol{\Theta}))^2(\lambda_d(\mathbf{Z}^T\mathbf{Z}))^{1.5},
\end{aligned}$$

where $\kappa(\cdot)$ is the condition number of a matrix.

For SBM with balanced community size, both $\lambda_1(\mathbf{Z}^T\mathbf{Z})$ and $\lambda_d(\mathbf{Z}^T\mathbf{Z})$ are of order n. Assume $\mathbf{\Theta}$ has constant eigenvalues, then $(\lambda_1(\mathbf{\Theta}))^2(\lambda_1(\mathbf{Z}^T\mathbf{Z}))^{1.5} = \Theta(n^{1.5})$ and $(\kappa(\mathbf{Z}^T\mathbf{Z}))^{-0.5}(\lambda_d(\mathbf{\Theta}))^2(\lambda_d(\mathbf{Z}^T\mathbf{Z}))^{1.5} = \Theta(n^{1.5})$. For MMSB with balanced Dirichlet parameter and assume $d = \Theta(1)$, the same result holds.

Also note that $\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\|^2 = \sum_{k\neq i,j}\left((\mathbf{P}^2)_{ik}-(\mathbf{P}^2)_{jk}\right)^2 = \sum_k\left((\mathbf{P}^2)_{ik}-(\mathbf{P}^2)_{jk}\right)^2 - \left((\mathbf{P}^2)_{ii}-(\mathbf{P}^2)_{ji}\right)^2 - \left((\mathbf{P}^2)_{ij}-(\mathbf{P}^2)_{jj}\right)^2 \geq \|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\|^2 - O(n^2\rho^4)$. Then we have,

$$\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\|^2 \geq \|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\|^2 - O(n^2\rho^4) \geq \Theta(n^3\rho^4)\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{Z}\|^2 - O(n^2\rho^4).$$

Summing up, we have:

$$\frac{1}{n^{1.5}\rho^2}\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(I-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\| \leq \Theta(1)\|\mathbf{z}_i-\mathbf{z}_j\|$$
$$\frac{1}{n^{1.5}\rho^2}\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(I-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\| \geq \sqrt{\Theta(1)\|\mathbf{z}_i-\mathbf{z}_j\|^2 - O(1/n)} = \Theta(1)\|\mathbf{z}_i-\mathbf{z}_j\| - O(1/n).$$

So there exists constants l and L and $\Delta_n = O(1/n)$ to make Assumption 3.2 hold for SBM/MMSB.

Now for GRDPG, we have $\mathbf{P}_{ij} = \rho\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T$. Similarly if $\mathbf{z}_i = \mathbf{z}_j$, $\mathbf{e}_i\mathbf{P} = \mathbf{e}_j\mathbf{P}$, Assumption 3.2 automatically holds. Now consider the case where $\mathbf{z}_i \neq \mathbf{z}_j$. We have,

$$\begin{aligned}&\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\left(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T-\mathbf{e}_j\mathbf{e}_j^T\right)\| \leq 3\rho^2\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\| = 3\rho^2\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T\| \\ \leq& 3\rho^2\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{Z}\|\|\mathbf{Z}^T\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T\| \leq 3\rho^2\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{Z}\|(\lambda_1(\mathbf{Z}^T\mathbf{Z}))^{1.5}.\end{aligned}$$

And,

$$\begin{aligned}&\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{P}^2\| = \rho^2\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T\| \geq \rho^2\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{Z}\|\sigma_d(\mathbf{Z}^T\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T) \\ \geq& \rho^2\|(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{Z}\|(\lambda_d(\mathbf{Z}^T\mathbf{Z}))^{1.5},\end{aligned}$$

where $\sigma_d(\mathbf{Z}^T\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T)$ is the smallest singular value of $\mathbf{Z}^T\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T$. Then if $\lambda_1(\mathbf{Z}^T\mathbf{Z}) = \Theta(n)$ and $\lambda_d(\mathbf{Z}^T\mathbf{Z}) = \Theta(n)$, there exists constants l and L and $\Delta_n = O(1/n)$ to make Assumption 3.2 hold for GRDPG, similar to the discussion for SBM/MMSB before.

C Consistency of SVD-RBF (Proof of Proposition 3.1)

For Generalized Random Dot Product Graph (GRDPG) model the network comes from a low rank probability matrix $\mathbf{P}$ with eigen-decomposition $\mathbf{P} = \rho\mathbf{Z}\mathbf{I}_{q,d-q}\mathbf{Z}^T = \mathbf{U}\mathbf{E}\mathbf{U}^T$, where rows of $\mathbf{Z} \in \mathbb{R}^{n\times d}$ are the latent positions of each node and $\mathbf{I}_{q,d-q}$ is a diagonal matrix with first q elements on the diagonal as 1 and the rest as -1. For ease of exposition in this proof we absorb $\sqrt{\rho}$ into $\mathbf{z}_i$, then for Assumption 3.1, L_g becomes $L_g\sqrt{\rho}$. Let $\hat{\mathbf{U}}\hat{\mathbf{E}}\hat{\mathbf{U}}^T$ be the top-d eigen-decomposition of $\mathbf{A}$, where $\hat{\mathbf{E}}$ is a diagonal matrix, and both $\hat{\mathbf{U}}$ and $\hat{\mathbf{E}}$ have rank d (typically, $d \ll n$). Denote the rows of $\hat{\mathbf{U}}|\hat{\mathbf{E}}|^{1/2}$ as $\hat{\mathbf{v}}_i$. By Theorem 5 of [59], if $\rho n = \omega(\log^{4\xi} n)$ for some constant $\xi > 0$, $d = \Theta(1)$,

$$\max_{i\in[n]}\|\mathbf{O}_n\hat{\mathbf{v}}_i - \mathbf{z}_i\| = O_P\left(\frac{(\log n)^\xi}{n^{1/2}}\right), \tag{5}$$

for some full rank matrix $\mathbf{O}_n \in \mathbb{R}^{d\times d}$. By Lemma 9 of [59] and Lemmas 14 and 16 of [62],both $\|\mathbf{O}_n\|$ and $\|\mathbf{O}_n^{-1}\|$ are bounded by some constants when the smallest singular value of $\mathbf{P}$ grows linearly with $n\rho$. This translates to that $\mathbf{z}_i = \mathbf{O}_n(\hat{\mathbf{v}}_i - \mathbf{r}_i)$, where $\max_{i\in[n]}\|\mathbf{r}_i\| = \tilde{O}_P(1/\sqrt{n})$.

Recall that node covariate is generated by $\mathbf{X}_i = g(\mathbf{z}_i) + \boldsymbol{\epsilon}_i$. Denote $\phi(\mathbf{x}) = g(\mathbf{O}_n\mathbf{x})$, then $\mathbf{X}_i = \phi(\mathbf{O}_n^{-1}\mathbf{z}_i) + \boldsymbol{\epsilon}_i$. We can see that if we set $\mathbf{v}_i = \mathbf{O}_n^{-1}\mathbf{z}_i$, we can use $\phi(\cdot)$ as the new function for generating node covariate and $\phi(\cdot)$ satisfies Assumption 3.1 with a different constant $L_\phi \leq \|\mathbf{O}_n^{-1}\|L_g$.

Now W.L.O.G. and for ease of exposition, in the following proof we only consider one dimension of $\mathbf{X}_i$, it can be easily extended to multi-dimension case by summing up the estimation errors across d dimensions.

In nonparametric regression literature [16, 70], it is well known that for Nadaraya-Watson Regression:

$$m(\mathbf{v}) := \frac{N(\mathbf{v})}{D(\mathbf{v})} := \frac{\sum_j K(\mathbf{v}, \mathbf{v}_j) X_i}{\sum_j K(\mathbf{v}, \mathbf{v}_j)} \longrightarrow \phi(\mathbf{v}).$$

We want to show that for an arbitrary point $\mathbf{v}$, if $\hat{\mathbf{v}}_i = \mathbf{v}_i + \mathbf{r}_i$, where $\max_i \|\mathbf{r}_i\| \leq \delta$ w.h.p.,

$$\hat{m}(\mathbf{v}) := \frac{\hat{N}(\mathbf{v})}{\hat{D}(\mathbf{v})} := \frac{\sum_j K(\mathbf{v}, \hat{\mathbf{v}}_j) X_i}{\sum_j K(\mathbf{v}, \hat{\mathbf{v}}_j)} \longrightarrow \phi(\mathbf{v}).$$

Consider $K(\mathbf{v}, \cdot)$ as an RBF kernel [70] that $K(\mathbf{v}, \mathbf{v}_j) = h_\theta(\|\mathbf{v} - \mathbf{v}_j\|)$, then

$$\begin{aligned} K(\mathbf{v}, \hat{\mathbf{v}}_j) - K(\mathbf{v}, \mathbf{v}_j) &= h_\theta(\|\mathbf{v} - \hat{\mathbf{v}}_j\|) - h_\theta(\|\mathbf{v} - \mathbf{v}_j\|) \\ &= (\|\mathbf{v} - \hat{\mathbf{v}}_j\| - \|\mathbf{v} - \mathbf{v}_j\|) h'(\xi_j) \end{aligned}$$

for $\xi_j \in [\min(\|\mathbf{v} - \hat{\mathbf{v}}_j\|, \|\mathbf{v} - \mathbf{v}_j\|), \max(\|\mathbf{v} - \hat{\mathbf{v}}_j\|, \|\mathbf{v} - \mathbf{v}_j\|)]$, following Mean Value Theorem.

Lemma C.1. Let $\delta = o(1)$ be the upper bound of $\max_i \|\mathbf{r}_i\|$, where $\hat{\mathbf{v}}_i = \mathbf{v}_i + \mathbf{r}_i$. Consider all points within $r = 2\delta^{\frac{1}{d+1}}$ distance to $\mathbf{v}$, denote this points set as $B_r(\mathbf{v})$, then for $j \notin B_r(\mathbf{v})$, $\xi_i \geq \|\mathbf{v} - \mathbf{v}_j\| - \delta \geq 2\delta^{1/(d+1)} - \delta > \delta^{1/(d+1)}$, and, by setting $\theta = \delta^{1/(d+1)}$, we have,

$$\sum_{j \notin B_r(\mathbf{v})} |K(\mathbf{v}, \hat{\mathbf{v}}_j) - K(\mathbf{v}, \mathbf{v}_j)| \leq 2\theta^d \sum_{j \notin B_r(\mathbf{v})} h_{2\sqrt{2}\theta} \left(\|\mathbf{v} - \mathbf{v}_j\|\right).$$

Proof. By triangle inequality, it is easy to get that if $j \notin B_r(\mathbf{v})$, $\xi_i \geq \|\mathbf{v} - \mathbf{v}_j\| - \delta \geq 2\delta^{1/(d+1)} - \delta > \delta^{1/(d+1)}$. For RBF kernel, $h(x) = \exp(-\frac{x^2}{2\theta^2})$, $x \geq 0$, so $h'(x) = -\frac{x}{\theta^2}\exp(-\frac{x^2}{2\theta^2})$. It is easy to show that $|h'(x)|$ increases monotonously in $[0, \theta]$ and decreases monotonously when $x \geq \theta$. As $\theta = \delta^{1/(d+1)}$, when $j \notin B_r(\mathbf{v})$, $|h'(\xi_j)| \leq |h'(\|\mathbf{v} - \mathbf{v}_j\| - \delta)|$,

$$\begin{aligned} |K(\mathbf{v}, \hat{\mathbf{v}}_j) - K(\mathbf{v}, \mathbf{v}_j)| &= (\|\mathbf{v} - \hat{\mathbf{v}}_j\| - \|\mathbf{v} - \mathbf{v}_j\|) h'(\xi_j) \\ &\leq \delta \frac{\|\mathbf{v} - \mathbf{v}_j\| - \delta}{\theta^2} \exp\left(-\frac{(\|\mathbf{v} - \mathbf{v}_j\| - \delta)^2}{2\theta^2}\right) \\ &\leq \theta^d \frac{\|\mathbf{v} - \mathbf{v}_j\|}{\theta} \exp\left(-\frac{\|\mathbf{v} - \mathbf{v}_j\|^2}{8\theta^2}\right), \end{aligned}$$

where the first inequality is due to triangle inequality and the second inequality holds by $\delta < \|\mathbf{v} - \mathbf{v}_j\|/2$. It is also easy to check that $x\exp(-\frac{x^2}{2}) < \exp(-\frac{x^2}{4})$ for $x \geq 0$, so

$$|K(\mathbf{v}, \hat{\mathbf{v}}_j) - K(\mathbf{v}, \mathbf{v}_j)| \leq 2\theta^d \exp\left(-\frac{\|\mathbf{v} - \mathbf{v}_j\|^2}{16\theta^2}\right) = 2\theta^d h_{2\sqrt{2}\theta}\left(\|\mathbf{v} - \mathbf{v}_j\|\right).$$

□

Proof of Proposition 3.1. When $j \in B_r(\mathbf{v})$, $|h'(\xi_j)| \leq 1/\theta$, so $|K(\mathbf{v}, \hat{\mathbf{v}}_j) - K(\mathbf{v}, \mathbf{v}_j)| \leq \delta/\theta$, then

$$|\Delta_D(\mathbf{v})| := |\hat{D}(\hat{\mathbf{z}}) - D(\hat{\mathbf{z}})| \leq |B_r(\mathbf{v})|\delta/\theta + 2\theta^d \sum_j h_{2\sqrt{2}\theta}\left(\|\mathbf{v} - \mathbf{v}_j\|\right).$$

Denote $B := \max_i \|X_i\|$, then

$$|\Delta_N(\mathbf{v})| := |\hat{N}(\hat{\mathbf{z}}) - N(\hat{\mathbf{z}})| \leq |B_r(\mathbf{v})|B\delta/\theta + 2\theta^d B \sum_j h_{2\sqrt{2}\theta}\left(\|\mathbf{v} - \mathbf{v}_j\|\right).$$

Now,

$$|\hat{m}(\mathbf{v}) - m(\mathbf{v})| = \left| \frac{N(\mathbf{v}) + \Delta_N(\mathbf{v})}{D(\mathbf{v}) + \Delta_D(\mathbf{v})} - \frac{N(\mathbf{v})}{D(\mathbf{v})} \right| = \left| \frac{D(\mathbf{v})\Delta_N(\mathbf{v}) - N(\mathbf{v})\Delta_D(\mathbf{v})}{D(\mathbf{v})(D(\mathbf{v}) + \Delta_D(\mathbf{v}))} \right|$$
$$= \left| \frac{\Delta_N(\mathbf{v}) - \Delta_D(\mathbf{v})\frac{N(\mathbf{v})}{D(\mathbf{v})}}{D(\mathbf{v})(1 + \frac{\Delta_D(\mathbf{v})}{D(\mathbf{v})})} \right| \tag{6}$$

Denote $b(\mathbf{v})$ as the density function of $\mathbf{v}$. By Proposition 9 of [57], which was introduced in [23], when $\theta \to 0$ as $n \to \infty$, and $\frac{n\theta^d}{|\log \theta^d|} \to \infty$, we have

$$\mathrm{P}\left(\sup_{\mathbf{v}} \left| \frac{1}{n\theta^d} D(\mathbf{v}) - b(\mathbf{v}) \right| > \epsilon_1 \right) = O(\theta^{d/2}),$$

for $\epsilon_1 = \Omega\left(\sqrt{\frac{\log \theta^{-d/2}}{n\theta^d}} \right) \to 0$. Thus $\frac{1}{n\theta^d} D(\mathbf{v}) \to b(\mathbf{v})$ uniformly for all $\mathbf{v}$. Now let,

$$\epsilon_2(\mathbf{v}) := \left| \frac{1}{n(2\sqrt{2}\theta)^d} \sum_j h_{2\sqrt{2}\theta}\left(\|\mathbf{v} - \mathbf{v}_j\| \right) - b(\mathbf{v}) \right|,$$

similarly $\epsilon_2(\mathbf{v}) \to 0$ uniformly for all $\mathbf{v}$.

Also when $r \to 0$, $\frac{|B_r(\mathbf{v})|}{n} \to 0$. So, for some positive universal constants C, c_1,

$$\frac{|\Delta_D(\mathbf{v})|}{n} \le \frac{\delta}{\theta} \frac{|B_r(\mathbf{v})|}{n} + c_1 C^d \theta^{2d} (b(\mathbf{v}) + \epsilon_2(\mathbf{v})) =: \epsilon_3(\mathbf{v})$$
$$\frac{|\Delta_N(\mathbf{v})|}{n} \le B\epsilon_3(\mathbf{v}).$$

We also have:

$$\left| \frac{\Delta_D(\mathbf{v})}{D(\mathbf{v})} \right| \le \frac{\epsilon_3(\mathbf{v})}{\theta^d (b(\mathbf{v}) - \epsilon_1)}.$$

From Equation (6), we have:

$$|\hat{m}(\mathbf{v}) - m(\mathbf{v})| \le \left| \frac{\Delta_N(\mathbf{v})/n}{(D(\mathbf{v})/n)(1 + \Delta_D(\mathbf{v})/D(\mathbf{v}))} \right| + \left| \frac{(\Delta_D(\mathbf{v})/n)N(\mathbf{v})/D(\mathbf{v})}{(D(\mathbf{v})/n)(1 + \Delta_D(\mathbf{v})/D(\mathbf{v}))} \right|$$
$$\le \frac{B\epsilon_3(\mathbf{v})}{\theta^d (d(\mathbf{v}) - \epsilon_1)\left(1 - \frac{\epsilon_3(\mathbf{v})/\theta^d}{d(\mathbf{v}) - \epsilon_1}\right)} + \frac{\epsilon_3(\mathbf{v})}{\theta^d} \frac{\phi(\mathbf{v}) + |m(\mathbf{v}) - \phi(\mathbf{v})|}{(d(\mathbf{v}) - \epsilon_1)\left(1 - \frac{\epsilon_3(\mathbf{v})/\theta^d}{d(\mathbf{v}) - \epsilon_1}\right)}$$
$$\le \epsilon_3(\mathbf{v}) \frac{\max(B, \phi(\mathbf{v}) + |m(\mathbf{v}) - \phi(\mathbf{v})|)}{(d(\mathbf{v}) - \epsilon_1)\theta^d - \epsilon_3(\mathbf{v})}$$
$$\sup_{\mathbf{v}} |\hat{m}(\mathbf{v}) - m(\mathbf{v})| \le \sup_{\mathbf{v}} \epsilon_3(\mathbf{v}) \frac{\max(B, \sup_{\mathbf{v}} \phi(\mathbf{v}) + \sup_{\mathbf{v}} |m(\mathbf{v}) - \phi(\mathbf{v})|)}{(\inf_{\mathbf{v}} b(\mathbf{v}) - \epsilon_1)\theta^d - \sup_{\mathbf{v}} \epsilon_3(\mathbf{v})}$$

Uniform consistency of $m(\mathbf{v}) - \phi(\mathbf{v})$ can be shown by checking Conditions 1 and 2 in [16]. Two key conditions to check are: 1) as scalar $u \to \infty$, $u^d \exp(-\frac{u^2}{2}) \to 0$, which is easy to check as $d = \Theta(1)$. 2) $\inf_{\|\mathbf{v}\| \le C} b(\mathbf{v}) > 0$, which holds by model assumption. Other parts of the conditions are satisfied by RBF kernel. We also have:

$$\sup_{\mathbf{v}} \frac{\epsilon_3(\mathbf{v})}{\theta^d} \le \frac{|B_r(\mathbf{v})|}{n} + c_1 C^d \theta^d \sup_{\mathbf{v}} (d(\mathbf{v}) + \epsilon_2(\mathbf{v})) = o_P(1).$$

Since $\phi(\mathbf{v})$ and B are bounded, and by using the above steps, we see that

$$\sup_{\mathbf{v}} |\hat{m}(\mathbf{v}) - m(\mathbf{v})| = o_P(1).$$

Now by Equation (5) we can set $\delta = \tilde{\Theta}(1/\sqrt{n})$. Let $|S| = \Theta(n)$, then $\theta = \delta^{1/(d+1)} \to 0$ as $n \to \infty$ and $\frac{|S|\theta^d}{|\log |S|\theta^d|} = \frac{|S|\delta}{\log ||S|\delta|} = \tilde{\Theta}(n^{\frac{1}{2}\frac{d+2}{d+1}}) \to \infty$, note that we use $|S|$ here because there are only $|S|$ data points available for nonparametric regression.

Then,

$$\begin{aligned}\sup_{\mathbf{v}} |\hat{m}(\mathbf{v}) - \phi(\mathbf{v})| &\leq \sup_{\mathbf{v}} |\hat{m}(\mathbf{v}) - m(\mathbf{v})| + \sup_{\mathbf{v}} |m(\mathbf{v}) - \phi(\mathbf{v})| \\ &\leq o_P(1) + o_P(1) = o_P(1),\end{aligned}$$

As $d = \Theta(1)$, summing up the estimation errors across d dimensions does not affect the result. This concludes the proof. □