Consistent Nonparametric Methods for Network Assisted Covariate Estimation

Anonymous Authors[1]

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. Such models have been used before for community detection (Yang et al., 2013; Zhang et al., 2016; Weng and Feng, 2016; Binkiewicz et al., 2017; Zhang et al., 2019; Yan and Sarkar, 2020). 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 (Li et al., 2019b). 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

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

needs no fine-tuning.

If we have some prior knowledge of the underlying model, for example, if it is a "low-rank" models like the Generalized Random Dot Product (GRDPG) models (Young and Scheinerman, 2007; Rubin-Delanchy et al., 2017; Udell and Townsend, 2019), which include many models like the Stochastic Blockmodel (Holland et al., 1983), the Mixed Membership Blockmodel (Airoldi et al., 2008), it will be natural to first do singular value decomposition to estimate the latent variables, and then use those directly in nonparametric 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 $\sqrt{n}$ (Sarkar and Chakrabarti, 2015). 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 (Sarkar and Chakrabarti, 2015; Sarkar et al., 2010). 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, and a recent embedding-based algorithm called NOBE (Jiang et al., 2018). We run experiments using 4 simulated graph models and 3 real-world networks. Overall, SVD-RBF, CN-VEC, and NOBE outperform the rest. Among the three, 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 both NOBE and SVD-RBF use the whole network to first estimate the latent positions. Also, CN-VEC is 10x-100x faster than 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 matrix, 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 (Sarkar et al., 2010) and its weighted variants (Adamic/Adar) (Adamic and Adar, 2003), preferential attachment (Barabási and Albert, 1999), resource allocation (Zhou et al., 2009), Katz index (Katz, 1953), PageRank (Brin and Page, 1998), SimRank (Jeh and Widom, 2002), and graph neural networks (Zhang and Chen, 2018) (see (Lü and Zhou, 2011) for a survey). While these often work well, only a few have consistency guarantees (Sarkar et al., 2010; Sarkar and Chakrabarti, 2015). 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 (Xiaojin and Zoubin, 2002), personalized PageRank (Page et al., 1999; Kloumann et al., 2017; Jeh and Widom, 2003), partially absorbing random walks (Wu et al., 2012), 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 (Li et al., 2019b). Another direction is **node embeddings**, which aims to represent nodes by vectors while retaining some network-based properties (Tang et al., 2015; Berberidis and Giannakis, 2019; Tsitsulin et al., 2018; Grover and Leskovec, 2016; Perozzi et al., 2014; Jiang et al., 2018; Qiu et al., 2019). 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 (Tang et al., 2013; Sussman et al., 2014; Rubin-Delanchy et al., 2017). 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 (Zhu et al., 2003; Nadler et al., 2009; El Alaoui et al., 2016; Calder and Slepčev, 2019) 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 (Li et al., 2019a; Le and Li, 2020; Jung and Tran, 2019; Jung, 2019), (x_i, y_i) pairs are observed for each node, and the network is used as a regularizer. In Network Lasso (Hallac et al., 2015), the network structure is used to enforce smoothness much like (Li et al., 2019a). 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 (Ma et al., 2020), Stochastic Blockmodel (Rohe et al., 2011; Lei and Rinaldo, 2015; Abbe, 2017) and its degree-corrected version (Zhao et al., 2012; Jin, 2015; Gao et al., 2018), Mixed-membership Stochastic Blockmodel (Mao et al., 2017; Panov et al., 2017; Mao et al., 2020) and its degree-corrected version (Jin et al., 2017; Mao et al., 2018), Stochastic Blockmodel with Overlaps (Kaufmann et al., 2016), Random Dot Product Graph model (Sussman et al., 2014; Athreya et al., 2017), 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 (Araya Valdivia and Yohann, 2019), shortest path lengths (Arias-Castro et al., 2018), and common neighbor counts (Sarkar et al., 2010) have been used. Recently (Parthasarathy et al., 2017) 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 (Arias-Castro et al., 2018).

Other related problems: In **matrix completion**, we try to fill in matrix entries given a partially observed noisy matrix. (Song et al., 2016; Li et al., 2019c) 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 (Chan and Airoldi, 2014), Stochastic Blockmodel approximation (Airoldi et al., 2013), and neighborhood smoothing (Zhang et al., 2017); see also (Gao et al., 2015; Borgs and Chayes, 2017; Xu, 2018). Our problem is different; we need to predict node covariates, and not necessarily the latents.

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 (Györfi et al., 2006; Wied and Weißbach, 2012; Duchi, 2019):

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 (Vaart, 1998).

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:

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

where $\mathbf{W}_{ij}$ is a measure of similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$, and $topK(i)$ 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 (Györfi et al., 2006; Wied and Weißbach, 2012). But in our case, we do not know the latents.

Our approach is to use the network to find $topK(i)$, 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 $topK(i)$, 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) (Holland et al., 1983). 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 $topK(i)$ would be those in the same cluster as i. So they would have the same latent as i. Thus, averaging over the covariates of $topK(i)$ 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 $O(\sqrt{n})$ (Rohe et al., 2011; Sarkar and Chakrabarti, 2015). For the stochastic blockmodel under appropriate conditions, the nodes selected in $topK(i)$ 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 (Jeh and Widom, 2003) 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., $\mathbf{P}^2_{i\ell} \approx \mathbf{P}^2_{j\ell}$), then $\mathbf{C}_{i\ell} \approx \mathbf{P}^2_{i\ell} \approx \mathbf{P}^2_{j\ell} \approx \mathbf{C}_{j\ell}$ 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

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

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 (Airoldi et al., 2013; Zhang et al., 2017; Chan and Airoldi, 2014; Gao et al., 2018; Xu, 2018). 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 Graphs (GRDPG) (Young and Scheinerman, 2007; Rubin-Delanchy et al., 2017) which include Stochastic and Mixed Membership Blockmodels. The details are in the supplement.

By Assumption 3.2, the similarity between i and j can

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: $dist(j) \leftarrow \mathbf{K}_{ij}$ (by Eq. (4))
3: $topK(i) \leftarrow k$ nodes with the smallest values of $dist(j)$
4: $\hat{\mathbf{X}}_i \leftarrow \dfrac{\sum_{j \in topK(i)} \mathbf{X}_j}{|topK(i)|}$
5: **end for**

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 $\sum_{k \neq i,j} (\mathbf{C}_{ik} - \mathbf{C}_{jk})^2$ 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 $\sum_{k \neq i,j} (\mathbf{C}_{ik} - \mathbf{C}_{jk})^2$ 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:

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

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

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 (Sarkar and Chakrabarti, 2015). Our analysis is also completely different from (Sarkar and Chakrabarti, 2015), and requires finer analysis.

Thus, given a node i, ordering the nodes $j \in [n] \setminus \{i\}$ 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 $topK(i)$), 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 $j \in topK(i)$. 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 $\|(\mathbf{e}_i - \mathbf{e}_j)^T\mathbf{P}^2\left(I - \mathbf{e}_i\mathbf{e}_i^T - \mathbf{e}_j\mathbf{e}_j^T\right)\|$ as the distance metric yields weakly consistent estimates for node covariates when ties occur with probability 0:

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

3.2. Algorithm for Low-Rank Models

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**

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, $f(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T \mathbf{\Theta} \mathbf{z}_j$ for some model parameters $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$. For example, in the Stochastic Blockmodel (SBM) (Holland et al., 1983), 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 $\mathbf{P}_{ij} = \rho f(\mathbf{z}_i, \mathbf{z}_j) = \rho \mathbf{z}_i^T \mathbf{\Theta} \mathbf{z}_j = \rho \mathbf{\Theta}_{ij}$. That is, link probabilities are solely dependent on the community memberships of nodes, and $\mathbf{\Theta}$ represents the community interconnections. The Mixed Membership Stochastic Blockmodel (MMSB) (Airoldi et al., 2008) 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) (Young and Scheinerman, 2007; Rubin-Delanchy et al., 2017) model allows for more general $\mathbf{z}_i$ and sets $\mathbf{P}_{ij} = \rho \mathbf{z}_i^T \mathbf{I}_{p,d-p} \mathbf{z}_j$, where $\mathbf{I}_{p,d-p}$ is a diagonal matrix with first p elements on the diagonal as 1 and the rest as -1. Let $\mathbf{A} = \hat{\mathbf{U}}\hat{\mathbf{E}}\hat{\mathbf{U}}^T$ 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, $d \ll n$). 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 $\hat{\mathbf{v}}_i$) (Rubin-Delanchy et al., 2017). 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 (Chatterjee, 2015). 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 GRDPG. 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 $\theta = \tilde{\Theta}(n^{-\frac{1}{2d}})$, 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 $\max_{i\in[n]} \|\mathbf{O}_n \hat{\mathbf{v}}_i - \mathbf{z}_i\| = O_P\left(\frac{(\log n)^\xi}{n^{1/2}}\right)$ by (Rubin-Delanchy et al., 2017), for some matrix $\mathbf{O}_n \in \mathbb{R}^{d\times d}$. The details are presented in the supplementary material. □

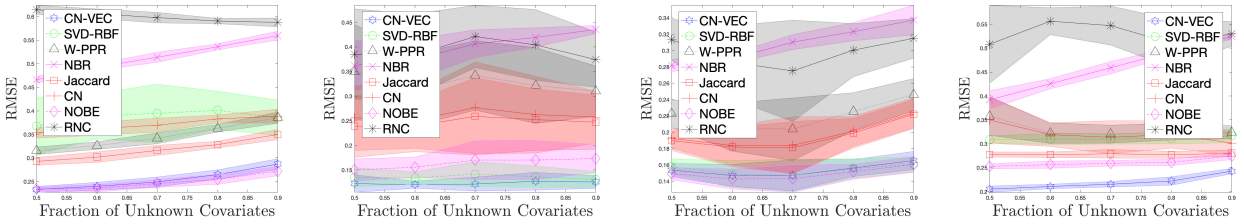
Proposition 3.1 gives a guidance on choosing the bandwidth θ for Algorithm 2, e.g., setting $\theta = \Theta\left((\log n)^{\frac{3}{d}}/n^{\frac{1}{2d}}\right)$, while the constant can be fine-tuned by cross-validation.

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 sparse complexity of $O(n^2)$ to store the pairwise node similarities.

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 (Jeh and Widom, 2003; Kloumann et al., 2017; Li et al., 2019b). 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 (Li et al., 2019b).



(a) LSM (b) SBM (c) MMSB (d) RDPG

Figure 1. RMSE on recovering hidden node covariates.

- 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 $\mathbf{W}$.
- NOBE: This constructs a node embedding $\mathbf{u}_i$ for each node i in the graph (Jiang et al., 2018). 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.

We also compare with a method that is not based on similarity measure: regression with network cohesion (RNC) (Li et al., 2019a; Le and Li, 2020). 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.

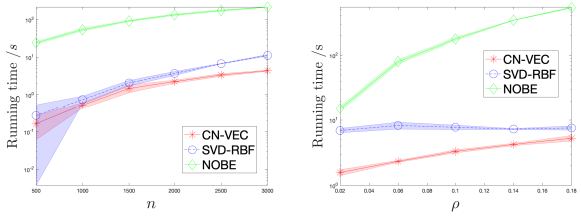
We do not compare against other popular node embeddings such as node2vec (Grover and Leskovec, 2016), DeepWalk (Perozzi et al., 2014), and LINE (Tang et al., 2015), since NOBE outperforms them (Jiang et al., 2018). All experiments are performed with Matlab R2018b on servers with 24-core Intel Xeon X5675 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 $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 100 \cdot \mathbf{I})$ and $\mathbf{z}_i \in \mathbb{R}^d$ is the latent vector of node i. The high variance allows us to have varied coefficients on the $\boldsymbol{\beta}$ vector.

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.

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



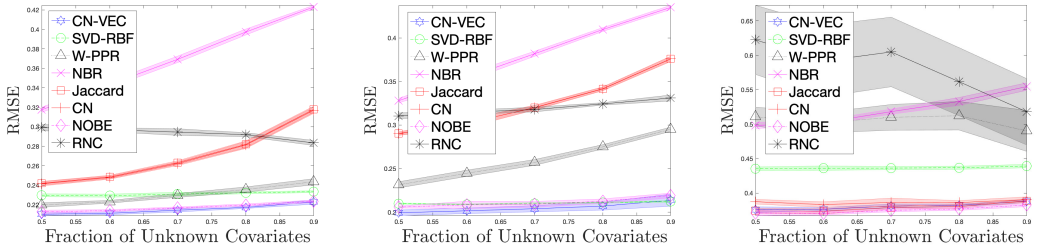
(a) Increasing n with ρn fixed. (b) Increasing ρ with n fixed.

Figure 2. Running time (log scale).

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 1(a) shows that CN-VEC 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 1(b) shows that CN-VEC perform best, followed by SVD-RBF 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



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

Figure 3. RMSE on recovering hidden topic distributions for each node.

do well. Figure 1(c) shows that CN-VEC, SVD-RBF 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 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 1(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.

To summarize, we find that CN-VEC, 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, and outperforms NOBE for the SBM and RDPG models. Note that NOBE has no convergence guarantees and takes 10x longer time than CN-VEC, as can be seen from the wall-clock timing results in Figure 2. The timing results are for the SBM graph using Matlab implementations of all algorithms. 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 NOBE.

4.2. Real networks

We evaluated our method on two citation networks, namely Cora (McCallum et al., 2000) and CiteSeer (Giles et al., 1998)[2], and one social network, namely Sinanet (Jia et al., 2017)[3]. 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[4] (Jia et al., 2017). The nodes are users of the website, and the node covariates are the topic distributions published by Jia et al. (2017). It has roughly 3500 nodes, 10 topics and average degree 16. A table with more details can be found in the Supplement.

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 $\text{RMSE} = \sqrt{\frac{1}{|U|}\sum_{i \in U} \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|^2}$, where U is the set of unlabeled nodes.

In Figure 3, we see that CN-VEC 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 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. Conclusions

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 method while being 10x-100x faster.

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

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

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

REFERENCES

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

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

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.

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.

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.

Ery Arias-Castro, Antoine Channarond, Bruno Pelletier, and Nicolas Verzelen. On the estimation of latent distances using graph distances. *arXiv preprint arXiv:1804.10611*, 2018.

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.

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

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

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

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.

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

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

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

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

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

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.

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

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.

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.

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.

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

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.

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

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.

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

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.

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.

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

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

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

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

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

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.

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.

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

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

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

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

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*, 2019c.

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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

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*, 2017.

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.

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

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.

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.

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.

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.

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.

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.

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

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

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

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.

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.

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

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.

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.

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.

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

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

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

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.

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.

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

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.