Iterative Collaborative Filtering for Sparse Matrix Estimation

Iterative Collaborative Filtering for Sparse Matrix Estimation

Christian Borgs * Jennifer T. Chayes † Devavrat Shah ‡ Christina Lee Yu §

Abstract

We consider sparse matrix estimation where the goal is to estimate an $n \times n$ matrix from noisy observations of a small subset of its entries. We analyze the estimation error of the popularly utilized collaborative filtering algorithm for the sparse regime. Specifically, we propose a novel iterative variant of the algorithm, adapted to handle the setting of sparse observations. We establish that as long as the fraction of entries observed at random scale as $\frac{\log^{1+\kappa}(n)}{n}$ for any fixed $\kappa > 0$, the estimation error with respect to the max-norm decays to 0 as $n \to \infty$ assuming the underlying matrix of interest has constant rank r. Our result is robust to model mis-specification in that if the underlying matrix is approximately rank r, then the estimation error decays to the approximate error with respect to the max-norm. In the process, we establish algorithm's ability to handle arbitrary bounded noise in the observations.

1 Introduction

We consider the task of sparse matrix estimation given noisy observations. Let F be an $n \times n$ matrix which we would like to estimate, and let Z be a noisy signal of matrix F such that $\mathbb{E}[Z] = F$. Let $\mathcal{E} \subset [n] \times [n]$ denote the subset of indices that are observed. In particular, we observe matrix M where $M(u, v) = Z(u, v)$ for $(u, v) \in \mathcal{E}$, and $M(u, v) = 0$ for $(u, v) \notin \mathcal{E}$. We assume that the entries of Z are independent random variables, and we assume a Bernoulli sampling model; each $(u, v) \in [n] \times [n]$ is in $\mathcal{E}$ with probability $p \in (0, 1]$ independently. The goal is to estimate F.

As a prototype for such a problem, consider a noisy observation of a social network where observed interactions are signals of true underlying connections. We might want to predict the probability that two users would choose to connect if recommended by the platform, e.g. LinkedIn. As a second example, consider a recommendation system where we observe movie ratings provided by users, and we may want to predict the probability distribution over ratings for specific movie-user pairs. A popular collaborative filtering approach suggests using "similarities" between pairs of users to estimate the probability that a connection is formed or the probability a user likes a particular movie. Traditionally, the similarities between pair of users in a social network is computed by comparing the set of their friends, or in the context of movie recommendation, by comparing commonly rated movies. In the sparse setting, most pairs of users have no common friends, or most pairs of users have no commonly rated movies; thus there is insufficient data to compute the traditional similarity metrics.

In this work, the primary interest is to provide a principled way to extend the simple, intuitive approach of computing similarities between pair of users or items in order to perform sparse matrix estimation via nearest neighbor collaborative filtering. We propose to do so by incorporating

*UC Berkeley, Berkeley, CA; borgs@berkeley.edu
†UC Berkeley, Berkeley, CA; jchayes@berkeley.edu
‡Massachusetts Institute of Technology, Cambridge, MA; devavrat@mit.edu
§Cornell University, Ithaca NY; cleeyu@cornell.edu

information within a larger radius neighborhood of the data graph rather than restricting only to immediate neighbors. This variation of collaborative filtering and its analysis in this work can be viewed as a natural extension of the work by [2, 4] in the context of stochastic block model and [51, 35] for traditional collaborative filtering.

1.1 Summary of Contributions

The primary contribution of this work is an analysis of an iterative collaborative filtering algorithm in the sparse regime. We consider the setting of a latent variable model where the matrix $F = [F(u, v)]$ can be described by a latent function f evaluated over latent variables associated to the coordinates. In particular, we assume that $F(u, v) = f(\theta_u, \theta_v)$ where f is a piece-wise Lipschitz function, and $\theta_u, \theta_v \in [0, 1]$ are coordinate latent variables sampled uniformly at random. Details of the model are described in Section 2.

As the main result of this work, we establish that with high probability the max entry-wise error associated with the resulting estimate converges to 0 as long as the latent function f when regarded as an integral operator has finite spectrum with constant rank r and $p = \Omega(n^{-1+\kappa})$ for $\kappa > 0$. In addition, if we have knowledge of the spectrum, the algorithm can be improved so that the max entry-wise error of the estimate converges to zero as long as $p = \Omega(n^{-1} \ln^{1+\kappa} n)$ for any $\kappa > 0$. We also establish robustness of our result with respect to the low rank requirement of f. In particular, we provide a robust version of our result that holds when f has ε-approximate rank r, i.e. there exists a rank r function that approximates f within ε with respect to the ℓ_∞ norm. We establish it by arguing that if all the observed entries are perturbed arbitrarily or adversarially within ε, then the algorithm estimates for each entry are perturbed by at most $O(\max(\sqrt{\varepsilon}, \varepsilon))$. The efficacy of the proposed algorithm with respect to arbitrary noise is an interesting result on its own.

Algorithmically and methodologically, our work builds on [2, 3, 4], which estimates clusters of the stochastic block model by computing distances from local neighborhoods around vertices. We improve upon their algorithm and analysis to provide bounds on the maximum entrywise estimation error for the general latent variable model with finite spectrum. This includes a larger class of generative models such as mixed membership stochastic block models, in contrast to their work which focuses on the stochastic block model with non-overlapping communities. We note that the algorithm considered in this work, uses the knowledge of which entries are observed and which are not, in line with the literature on matrix estimation. In the setting of clustering cf. [2, 3, 4], such a knowledge is absent from the purview of the algorithm.

Withthe exception of a few recent results, by and large the literature on matrix estimation has focused on providing estimation error bounds with respect to the normalized Frobenius norm. In contrast, we provide bounds on the max entry-wise estimation error which is a lot more challenging. Our bounds are restricted to the latent variable model, while the traditional matrix estimation literature considers the underlying matrix to be an arbitrary instance from the family of (approximately) low-rank matrices with 'incoherence'-like conditions. Indeed, understanding the relationship between these two seemingly different model classes remains an important direction for future work.

A weaker version of this result was published in the NeurIPS conference as [6]. In contrast, this paper provides sharper bounds for both the MSE and max-norm error that improves the exponent in the convergence rates. We have also included a perturbation analysis of the algorithm that shows under "adversarial" bounded noise, the error scales gracefully with the bound on the noise. This enables analysis of our work for the approximately low-rank setting. We have also included a modified algorithm that achieves the same rates with a reduced computational complexity, and we

have shown extensions of our results to relaxed modeling assumptions on the latent variable model. We have added empirical evaluation of our method compared with state-of-art methods.

1.2 Related Work

The related work includes that of matrix estimation or completion, collaborative filtering, and graphon estimation arising from the asymptotic theory of graphs. We provide a brief overview of prior works for each of these topics.

In the context of matrix estimation or completion, there has been much progress under the low-rank assumption and additive noise model. Most theoretically founded methods are based on spectral decompositions or minimizing a loss function with respect to spectral constraints, c.f. [29, 30, 14, 16, 44, 42, 20, 18, 17, 49]. In a nutshell, this collection of works establishes that if the underlying matrix has rank r, then it can be estimated so that the estimator has normalized Mean Squared Error (MSE) going to 0 as $n \to \infty$ as long as $p = \Omega(rn^{-1} \log n)$. Furthermore, [30, 15] showed that $\omega(rn^{-1})$ samples are necessarily required for such a guarantee. These near optimal sample complexity results hold when the noise in each entry of the matrix is independent and identically distributed. For the setting of generic noise and the general latent variable model where the latent function is analytic, [17, 49] provide an estimator for which the MSE decays to 0 as $n \to \infty$ as long as $p = \Omega(n^{-1}\mathsf{poly}(\log n))$.

The guarantee with respect to MSE does not necessarily guarantee recovery of *all* entries accurately. Indeed, bounding max entrywise error provides such a guarantee as established by our result. In parallel with our work, there has been recent progress on developing matrix estimation methods that provide max entrywise bounds for matrices with rank r. In particular, for sufficiently 'nice' rank r matrices, [1] establish that a simple spectral algorithm can recover the matrix with max entrywise error decaying to 0 as long as $p = \Omega(\log n/n)$. Indeed, improving such max entrywise guarantee has been actively pursued over the past few years witnessed in the growing body of works, cf. [19], [52], [13], [39] and [23].

The collaborative filtering method has been successfully employed across industry applications (Netflix, Amazon, Youtube) due to its simplicity and scalability, c.f. [27, 37, 33, 43]; however the theoretical results have been relatively sparse. We call special attention to the recent works by [51, 35, 36] which provide a non-parametric statistical perspective for the traditional collaborative filtering method. In particular, they suggest that the practical success of these methods across a variety of applications may be due to its ability to capture local structure like the classical nearest neighbor or kernel regression method. They establish that as long as the latent function f is Lipschitz, the MSE of the resulting estimator decays to 0 as $n \to \infty$ as long as $p = \omega(n^{-\frac{1}{2}})$. A key limitation of this approach is that it requires a dense dataset with sufficient entries in order to compute similarity metrics, requiring that each pair of rows or columns has a growing number of overlapped observed entries, which does not hold when $p = o(n^{-1/2})$.

Graphons emerged as the limiting object of a sequence of large dense graphs, c.f. [12, 22, 38], with recent work extending the theory to sparse graphs, c.f. [10, 11, 9, 47]. In the graphon estimation problem, one observes a single instance of a random graph sampled from an underlying latent variable model, and the goal is to estimate the function that governs the edge probabilities of the graph. [24, 31] provide minimax optimal rates for graphon estimation; however a majority of the proposed estimators are not computable in polynomial time, since they require optimizing over an exponentially large space (e.g. least squares or maximum likelihood), c.f. [48, 8, 7, 24, 31]. [8] provides a polynomial time method based on degree sorting in the special case when the expected degree function is monotonic. [49] analyzes universal singular value thresholding (USVT) for graphon estimation in settings that the spectrum decays quickly, showing convergence rates

which matches the minimax optimal rate for low dimensional smooth functions.

Stochastic block model (SBM) parameter estimation is an instance of graphon estimation, where the underlying function has a specific structure. Under the SBM, each vertex is associated to one of r community types, and the probability of an edge is a function of the community types of both endpoints. This implies that the edge probability function is block constant. Estimating the $n \times n$ parameter matrix becomes an instance of matrix estimation with a technical distinction – all entries are fully observed, i.e. each edge is present (1) or absent (0). In SBM, the expected matrix is at most rank r due to its block structure. Precise thresholds for cluster detection (better than random) and estimation have been established by [2, 3, 4]. As mentioned before, our work, both algorithmically and methodically is closely related to their work. The mixed membership stochastic block model (MMSBM) allows each vertex to be associated to a length r vector, which represents its weighted membership in each of the r communities. The probability of an edge is a function of the weighted community memberships vectors of both endpoints, resulting in an expected matrix with rank at most r. Recent work by [45] provides an algorithm for weak detection for MMSBM with sample complexity $r^2 n$, when the community membership vectors are sparse and evenly weighted. They provide partial results to support a conjecture that $r^2 n$ is a computational lower bound, separated by a gap of r from the information theoretic lower bound of rn. This gap was first shown in the simpler context of the stochastic block model [21]. [50] proposed a spectral clustering method for inferring the edge label distribution for a network sampled from a generalized stochastic block model. When the expected function has a finite spectrum decomposition, i.e. low rank, then they provide a consistent estimator for the sparse data regime, with $\Omega(n \log n)$ samples.

In the above discussion, we have focused primarily on the sample complexity required for consistent estimation, i.e. the scaling of the number of samples required (pn) such that the normalized estimation error such as the MSE or max-norm goes to 0. When consistent estimation is feasible, we can further consider the rate of decay of the error guarantees. To that end, we provide a brief overview of the minimax scaling with respect to boudns on the MSE. [17] identifies a minimax lower bound on the scaling of the MSE for a generic matrix estimation task characterized by the nuclear norm of the target matrix. In particular, for symmetric matrices with nuclear norm bounded by δ, the minimax MSE scaling is lower bounded by $\min\left(\frac{\delta}{\sqrt{n^3 p}}, \frac{\delta^2}{n^2}, 1\right)$; furthermore [17] argues that the universal singular value thresholding achieves this scaling. This bound holds even in the scenario where observed entries are noiseless. This characterization however is loose for the setting of low-rank matrices. Observe that for rank r symmetric matrices with entries bounded in $[-1, 1]$, the nuclear norm can scale as $n\sqrt{r}$; resulting in a bound of $\sqrt{\frac{r}{np}}$ (for small enough p) [17]. For the setting of rank r matrices with noiseless observations, [29, 30] provide an estimator with MSE scaling as $\frac{r}{np}$ for $p = \Omega(1/n)$. This points to the fact that the class of matrices with bounded nuclear norm is more complex than the class of rank r matrices with bounded entries. In the setting of low rank graphon estimation (i.e. binary observations), [25, 32] show a minimax lower bound on the MSE scaling as $\frac{\log r}{pn}$ for small enough $p = \Omega(\log r/n)$; however the existence of a computationally efficient estimator that achieves this lower bound under the more general noise setting of graphon estimation is still an open research direction.

2 Setup

2.1 Model and Assumptions

Recall that our goal is to estimate the $n \times n$ matrix F; Z is a noisy signal of matrix F such that $\mathbb{E}[Z] = F$. The available data is denoted by $(\mathcal{E}, M)$, where $\mathcal{E} \subset [n] \times [n]$ denotes the subset of indices

for which data is observed, and M is the $n \times n$ data matrix where $M(u,v) = Z(u,v)$ for $(u,v) \in \mathcal{E}$, and $M(u,v) = 0$ for $(u,v) \notin \mathcal{E}$. The observations can be equivalently represented by an directed weighted graph $\mathcal{G}$ with vertex set $[n]$, edge set $\mathcal{E}$, and edge weights given by M. We assume that $\{Z(u,v)\}_{(u,v)\in[n]^2}$ are independent random variables across all indices with $\mathbb{E}[Z(u,v)] = F(u,v)$, and that the underlying matrix and observations are bounded, i.e. $F(u,v), Z(u,v) \in [0,1]$. We assume a uniform Bernoulli sampling model, where each entry is observed independently with probability p, i.e. $\{\mathbb{I}((u,v) \in \mathcal{E})\}_{(u,v)\in[n]^2}$ are independent Bernoulli(p) random variables.

Latent Variable Model. Assume that each $u \in [n]$ is associated to a latent feature variable $\theta_u \sim U[0,1]$, which is drawn independently across indices $[n]$ uniformly on the unit interval. We assume that the expected data matrix can be described by the latent function f, i.e. $F(u,v) = f(\theta_u, \theta_v)$, where $f : [0,1]^2 \to [0,1]$ is a symmetric bounded function. The symmetry assumption can be easily relaxed but is assumed for ease of notation in the analysis. The latent function f is assumed to be fixed and independent of the dimension n. We additionally impose local neighborhood properties that are primarily used in the nearest neighbor portion of the analysis. We will assume that f is Lipschitz, but this assumption can be relaxed as discussed in Section 2.2.

Low Rank. We assume that the latent function f has finite spectrum with rank r when regarded as an integral operator, i.e. for any $\theta_u, \theta_v \in [0,1]$,

$$f(\theta_u, \theta_v) = \sum_{k=1}^{r} \lambda_k q_k(\theta_u) q_k(\theta_v),$$

where $\lambda_k \in \mathbb{R}$ for $1 \le k \le r$, and q_k are orthonormal ℓ_2 functions for $1 \le k \le r$ such that

$$\int_0^1 q_k(y)^2 dy = 1 \text{ and } \int_0^1 q_k(y) q_h(y) dy = 0 \text{ for } k \neq h \in [r].$$

We assume there exists some B such that $\sup_{y\in[0,1]} |q_k(y)| \le B$ for all $k \in [r]$. Let Λ denote the $r \times r$ diagonal matrix with $\{\lambda_k\}_{k\in[r]}$ as the diagonal entries, and let Q denote the $r \times n$ matrix where $Q(k,u) = q_k(\theta_u)$. Since Q is a random matrix depending on the sampled θ, it is not guaranteed to be an orthonormal matrix (even though q_k are orthonormal functions). By definition, it follows that $F = Q^T \Lambda Q$. Let $r' \le r$ be the number of distinct valued eigenvalues amongst $\{\lambda_k\}_{k\in[r]}$. Let $\tilde{\Lambda}$ denote the $r \times r'$ matrix where $\tilde{\Lambda}(a,b) = \lambda_a^{b-1}$.

The finite spectrum assumption also implies that the model can be represented by latent variables in the r dimensional Euclidean space, where the latent variable for node i would be the vector $(q_1(\theta_i), \dots q_r(\theta_i))$, and the latent function would be bilinear, having the form

$$f(\vec{q}, \vec{q}') = \sum_k \lambda_k q_k q_k' = q^T \Lambda q'.$$

This condition also implies that the expected matrix F is low rank, which includes scenarios such as the mixed membership stochastic block model and finite degree polynomials. The function f is fixed with respect to n, the rank r is assumed to be finite in the low rank setting.

The mixed membership model for network data can be represented with a finite spectrum latent variable model. Each coordinate is associated to a vector $\pi \in \Delta_r$, sampled iid from a distribution P. For two nodes with respective types π and π', the observed interaction is $f(\pi, \pi') = \sum_{ij} \pi_i \pi_j' B_{ij} = \pi^T B \pi'$, where $B \in [0,1]^{r\times r}$ and assumed to be symmetric. Since B is symmetric, there exists a diagonal decomposition $B = U \tilde{\Lambda} U^T$ with u_k denoting the eigenvectors, such that

$f(\pi, \pi') = \sum_{k=1}^{r} \tilde{\lambda}_k u_k^T \pi u_k^T \pi'$. It follows from this decomposition that the Hilbert-Schmidt integral operator associated to function $f : \Delta_r \times \Delta_r \to [0,1]$ has finite spectrum with rank at most r.

Interaction data arising from symmetric finite degree polynomials also leads to finite spectrum latent variable models. Let $f(x,y)$ be a finite degree symmetric polynomial, represented by $f(x,y) = \sum_{i=0}^{r} \sum_{j=0}^{r} c_{ij} x^i y^j$, where $c_{ij} = c_{ji}$ for all ij. Let $\mathbf{x} = (1, x, x^2, \dots x^r)$ and $\mathbf{y} = (1, y, y^2, \dots y^r)$, and let C denote the $(r+1)\times(r+1)$ matrix with entries $[c_{ij}]$, so that $f(x,y) = \mathbf{x}^T C \mathbf{y}$. Since C is symmetric, there exists a diagonal decomposition $B = U\tilde{\Lambda}U^T$ with u_k denoting the eigenvectors, such that $f(x,y) = \sum_{k=1}^{r} \tilde{\lambda}_k u_k^T \mathbf{x} u_k^T \mathbf{y}$. It follows from this decomposition that the Hilbert-Schmidt integral operator associated to function f has finite spectrum with rank at most r.

Approximately Low Rank. More generally, we shall consider approximately low-rank f cf. [46]. Specifically, for a given $\varepsilon > 0$, a symmetric function f is said to have ε-approximate rank r if

$$\sup_{\theta_u, \theta_v \in [0,1]} \left| f(\theta_u, \theta_v) - \sum_{k=1}^{r} \lambda_k q_k(\theta_u) q_k(\theta_v) \right| \leq \varepsilon, \tag{1}$$

where $\lambda_k \in \mathbb{R}$ for $1 \leq k \leq r$, and q_k are orthonormal ℓ_2 functions for $1 \leq k \leq r$. In this case, it follows that $F = Q^T \Lambda Q + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} = [\varepsilon_{ij}] \in \mathbb{R}^{n \times n}$ is such that $\max_{ij} |\varepsilon_{ij}| \leq \varepsilon$. That is, the matrix F is approximately rank r. Functions f which do not have finite spectrum, but for which the eigenvalues decay quickly can be shown to have approximately low rank. [17, 49] use this observation to analyze the USVT algorithm for latent variable model estimation with Lipschitz, Holder, and Sobolev functions. [46] also show that any analytic function with bounded derivatives has approximately low rank. Recall again that we assume the function f is fixed with respect to n, but we can consider the choice of ε to be dependent on n, so that the ε approximate rank r would grow with respect to n.

2.2 Discussion on Latent Variable Model

The latent variable model assumes a random generative model on the underlying matrix F, as opposed to the typical deterministic incoherence style conditions found in the literature. The generative model assuming i.i.d. sampled latent variables and boundedness of the eigenfunctions of f guarantee similar properties as incoherence with high probability, as any single row or column will not dominate the signal in a way that deviates too much from the typical values of f. The i.i.d. sampling assumption on the latent variables is used in analyzing the local neighborhoods of the observation graph, however this assumption can likely be replaced by regularity assumptions over the empirical distribution of the latent factors for large n, e.g. if the latent factors are close to a typical sample set from a well-behaved underlying distribution.

The Lipschitzness assumption of f together with the assumption that $\theta_u \sim U[0,1]$, guarantees that for any given $u \in [n]$ there are sufficiently many other coordinates $v \in [n]$ such that the observed entries are similar across both rows or columns. These assumptions can be relaxed as long as the key property of "sufficiently many similarly behaving coordinates" is maintained. As examples, a piecewise Lipschitz function f or a setting with finite latent types would also satisfy the needed local neighborhood properties. Similarly, the scalar assumption on the latent variables and the uniform distribution $U[0,1]$ are not crucial and can be relaxed to i.i.d. sampled random latent vectors from a larger class of distributions. The critical conditions to maintain are the finite spectrum of f, boundedness of eigenfunctions, and local neighborhood properties. The local measure needs to be concentrated enough relative to the rate of change in the function f so that

when n points are sampled from the space, there are sufficently many "nearby neighbors" for whom the function behaves similarly for any given point we would want to estimate. This primarily affects the nearest neighbor portion of the algorithm and analysis. [36] also provides a formal discussion and results for extending the nearest neighbor analysis to accommodate settings beyond scalar Lipschitz functions. Our model can also be extended to asymmetric matrix settings and categorical data. Section 6 discuss how our theorem extend to some of these model variations.

2.3 Goal

The goal is to produce $\hat{F}$, an estimate of F, using observation matrix M and knowledge of $\mathcal{E}$. We measure the estimation error through the maximum entry-wise error and the mean squared error. The maximum entry-wise error or ∞-norm of the error matrix $\hat{F} - F$ is defined as

$$\|\hat{F} - F\|_{\max} = \max_{u,v} |\hat{F}(u,v) - F(u,v)|. \tag{2}$$

We will provide bounds on this that hold with high probability, that is, with probability converging to 1 as $n \to \infty$. The mean squared error (MSE) is defined as

$$\text{MSE}(\hat{F}) = \frac{1}{n^2}\mathbb{E}\Big[\sum_{u,v}(\hat{F}(u,v) - F(u,v))^2\Big]. \tag{3}$$

In measuring error either with high probability or in expectation, the randomness is considered over the data generation process.

3 Algorithm

We propose and analyze a variation of the similarity based collaborative filtering algorithm. At its core, the collaborative filtering algorithm attempts to produce the estimate $\hat{F}(u,v)$ by averaging over observed entries $F(u',v')$ for a subset of tuples (u',v') such that u' is "similar" to u and v' is "similar" to v.

Sample Splitting. To state the precise algorithm, for technical reasons, we shall use sample splitting. Recall that $\mathcal{E} \subset [n]^2$ denotes the set of indices for which we observe noisy signals of $F(u,v)$, i.e. for each $(u,v) \in \mathcal{E}$, $M(u,v) = Z(u,v)$ where $\mathbb{E}[Z(u,v)] = F(u,v)$. We assumed that $\mathcal{E}$ is generated according to a Bernoulli(p) sampling model, i.e. for each $(u,v) \in [n]^2$, it belongs to $\mathcal{E}$ with probability p independently. We split the samples $\mathcal{E}$ into three subsets as follows: for each tuple or edge $(u,v) \in \mathcal{E}$, with probability 1/4 it is placed in $\mathcal{E}'$, with probability 1/4 it is placed in $\mathcal{E}''$, and with the remaining 1/2 probability it is placed in $\mathcal{E}''' = \mathcal{E} \backslash (\mathcal{E}' \cup \mathcal{E}'')$.

We will use additional "virtual" edges that will aid in estimating the distance as part of the algorithm. To that end, note that conditioned on the edge set $\mathcal{E}'$, for some $(u,v) \notin \mathcal{E}'$, $\mathbb{P}\left((u,v) \in \mathcal{E}'' | \mathcal{E}'\right) = \frac{p}{4-p} = p'$. Furthermore, conditioned on $\mathcal{E}'$, $\mathbb{I}((u,v) \in \mathcal{E}'')$ are independent random variables. Conditioned on $\mathcal{E}'$, we generate a random subset $\mathcal{E}'_{\mathsf{ind}} \subseteq \mathcal{E}'$ such that each $(u,v) \in \mathcal{E}'$ is included in $\mathcal{E}'_{\mathsf{ind}}$ independently with probability $p' = \frac{p}{4-p}$. Therefore, conditioned on $\mathcal{E}'$, the set $\mathcal{E}'_{\mathsf{ind}} \cup \mathcal{E}''$ is distributed according to a Bernoulli(p') sampling model, where each $(u,v) \in [n]^2$ are included in $\mathcal{E}'_{\mathsf{ind}} \cup \mathcal{E}''$ independently with probability p'.

For each $u, v \in [n]$, define $M'(u,v) = \mathbb{I}((u,v) \in \mathcal{E}')M(u,v)$, $M'_{\mathsf{ind}}(u,v) = \mathbb{I}((u,v) \in \mathcal{E}'_{\mathsf{ind}})M(u,v)$, $M''(u,v) = \mathbb{I}((u,v) \in \mathcal{E}'')M(u,v)$, and $M'''(u,v) = \mathbb{I}((u,v) \in \mathcal{E}''')M(u,v)$; let $M' = [M'(u,v)]$, $M'_{\mathsf{ind}} = [M'_{\mathsf{ind}}(u,v)]$, $M'' = [M''(u,v)]$ and $M''' = [M'''(u,v)]$ denote the associated $n \times n$ matrices. Note that M'_{ind} is strictly contained within M' as $\mathcal{E}'_{\mathsf{ind}} \subseteq \mathcal{E}'$. The algorithm will use observations

M' and M'' to producing distance estimates $\hat{d}$, and it uses observations M''' to produce the final estimate $\hat{F}$ given $\hat{d}$.

Noisy Nearest Neighbor Algorithm. We consider the following noisy nearest neighbor algorithm described below, followed by three different subroutines to compute distances depending on the sparsity regime of the dataset.
(1) Compute distances $\hat{d}(u,v)$ between pairs of coordinates $u, v \in [n]^2$ using M' and M''.
(2) For each $u, v \in [n]^2$, produce an estimate

$$\hat{F}(u,v) = \frac{1}{|\mathcal{E}'''_{uv}|} \sum_{(a,b)\in\mathcal{E}'''_{uv}} M'''(a,b), \tag{4}$$

where $\mathcal{E}'''_{uv} = \{(a,b) \in \mathcal{E}''' \ : \ \hat{d}(u,a) < \eta, \hat{d}(v,b) < \eta\}$ for some small enough $\eta > 0$.

We will choose the threshold $\eta = \eta(n)$ depending on the local geometry of the latent feature space with respect to $\hat{d}(u,v)$, in order to guarantee that $\eta(n)$ is small enough to drive the bias to zero, yet large enough to ensure $|\mathcal{E}'''_{uv}|$ diverges so that the variance due to observation noise is small. The key part of the algorithm is determining how to estimate the distances $\hat{d}(u,v)$. In what follows, we describe three variations depending upon the observation density, p.

3.1 Estimating Distance $\hat{d}$

Dense Regime. When $p = \omega(n^{-\frac{1}{2}})$, it is feasible to compute distances by simply looking at the overlapping entries; this is popularly done in practice [27] as well as analyzed theoretically in the recent works [51, 35]. For any $(u,a) \in [n]^2$,

$$\hat{d}(u,a) = \frac{1}{|\mathcal{O}_{ua}|} \sum_{y\in\mathcal{O}_{ua}} (M(u,y) - M(a,y))^2, \tag{5}$$

where $\mathcal{O}_{ua} = \{y \in [n] : (u,y), (a,y) \in \mathcal{E}'\}$. This is a finite sample approximation of $\int_0^1 (f(\theta_u, y) - f(\theta_v, y))^2 \, dy$. When $p = \omega(n^{-\frac{1}{2}})$, it follows that $|\mathcal{O}_{ua}| = \omega(1)$ for all $u, a \in [n]^2$ with high probability, so that $\hat{d}(u,a) \approx \int_0^1 (f(\theta_u, y) - f(\theta_v, y))^2 \, dy$. [35] subsequently prove that for any Lipschitz latent function f the MSE decays to 0 as $n \to \infty$ as long as $p = \omega(n^{-\frac{1}{2}})$. The arguments of [35] can be adapted to show that the maximum entry-wise error decays to 0 with high probability as well. However, for $p = o(n^{-\frac{1}{2}})$, for most $u, a \in [n]^2$, $\mathcal{O}_{ua} = \emptyset$ with high probability and hence a different approach is needed – overcoming the sparse regime is the primary interest of this work.

Sparse Regime. Consider the sparse regime where $p = n^{-1+\kappa}$ for any $\kappa \in (0, \frac{1}{2})$; in this regime the overlap is small and thus new distance estimates are required. Recall that the function f has finite spectrum, i.e. $f(\theta_u, \theta_v) = \sum_k \lambda_{k=1}^r q_k(\theta_u) q_k(\theta_v)$. We propose an estimator which approximates $d(u,v) = \|\Lambda^t Q(e_u - e_v)\|_2^2$ by comparing depth t neighborhoods of u and v in the data graph $\mathcal{G} = ([n], \mathcal{E}')$. Specifically, let the weight of an edge $(a,b) \in \mathcal{E}'$ in graph $\mathcal{G}$ be the observed value $M(a,b)$ $(= M'(a,b))$. By assumption, in expectation this weight equals $F(a,b) = f(\theta_a, \theta_b)$. Therefore, the product of weights along a path from u to y, of length t, denoted as $(u, x_1, \ldots, x_{t-1}, y)$ with $(u,x_1), (x_1,x_2), \ldots, (x_{t-1}, y) \in \mathcal{E}'$, in expectation equals

$$\begin{aligned}
&\mathbb{E}_{X_1,\ldots,X_{t-1}}\Big[f(\theta_u, X_1) \times \prod_{s=1}^{t-2} f(X_s, X_{s+1}) \times f(X_{t-1}, \theta_y) | \theta_u, \theta_y\Big] \\
&\quad = \textstyle\sum_{k=1}^r \lambda_k^t q_k(\theta_u) q_k(\theta_y) \\
&\quad = e_u^T Q^T \Lambda^t Q e_y.
\end{aligned} \tag{6}$$

Therefore, the product of weights along the path connecting u to y is a good proxy of quantity $e_u^T Q^T \Lambda^t Q e_y$. Recall that each entry is observed independently with probability p due to our assumed Bernoulli sampling model. Therefore, for any $u \in [n]$, the number of neighbors of u in $\mathcal{G}$ scale as $pn = n^\kappa$. More generally, for $1 \le t \le 1/\kappa$, the number of nodes at distance t from u scale as $n^{\kappa t}$. We choose t large enough to guarantee that for any two nodes u and v, there is a sufficient overlap between the two subset of nodes at distance y from nodes u and v respectively. This suggests that we choose t so that $n^{\kappa t} \approx n^{\frac{1}{2}}$, which in effect aggregates enough data in the sparse regime to match the expected number of observations per row in the dense regime. We formalize this intuition in the following construction of the distance estimates.

Let $\mathcal{S}_{u,s}$ denote the set of vertices which are at distance s from vertex u in the graph defined by edge set $\mathcal{E}'$. Specifically, $i \in \mathcal{S}_{u,s}$ if the shortest path in $\mathcal{G} = ([n], \mathcal{E}')$ from u to i has a length of s. Let $\mathcal{T}_u$ denote a breadth-first tree in $\mathcal{G}$ rooted at vertex u. The breadth-first property ensures that the length of the path from u to i within $\mathcal{T}_u$ is equal to the length of the shortest path from u to i in $\mathcal{G}$. Let $\mathcal{T}_u^t \subset \mathcal{T}_u$ denote the sub-tree containing all nodes and edges in $\mathcal{T}_u$ up to and including depth t. If there is more than one valid breadth-first tree rooted at u, choose one uniformly at random. Let $N_{u,t} \in [0,1]^n$ denote the following vector with support on the boundary of the depth-t neighborhood of vertex u (we also call $N_{u,t}$ the neighborhood boundary):

$$N_{u,t}(i) = \begin{cases} \prod_{(a,b) \in \text{path}_{\mathcal{T}_u}(u,i)} M'(a,b) & \text{if } i \in \mathcal{S}_{u,t}, \\ 0 & \text{if } i \notin S_{u,t}, \end{cases}$$

where $\text{path}_{\mathcal{T}_u}(u,i)$ denotes the set of edges along the path from u to i in the tree $\mathcal{T}_u$. The sparsity of $N_{u,t}(i)$ is equal to $|\mathcal{S}_{u,t}|$, and the value of the coordinate $N_{u,t}(i)$ is equal to the product of weights along the path from u to i. Let $\tilde{N}_{u,t}$ denote the normalized neighborhood boundary such that $\tilde{N}_{u,t} = N_{u,t}/|\mathcal{S}_{u,t}|$. For each tuple $(u,v) \in [n]^2$, compute $\hat{d}(u,v)$ according to

$$\hat{d}(u,v) = \left(\tfrac{1}{p'}\right)\left(\tilde{N}_{u,t} - \tilde{N}_{v,t}\right)^T (M'' + M'_{\mathsf{ind}})\left(\tilde{N}_{u,t+1} - \tilde{N}_{v,t+1}\right). \tag{7}$$

Sparser Regime. Consider the even sparser regime where $p = n^{-1} \ln^{1+\kappa} n$ for some $\kappa > 0$. Let us assume that the algorithm knows the eigenvalues $\{\lambda_k\}_{k \in [r]}$. Recall that $r' \le r$ denotes the number of distinct valued eigenvalues amongst $\{\lambda_k\}_{k \in [r]}$. Recall that Λ is the diagonal matrix with $\Lambda_{kk} = \lambda_k$, and $\tilde{\Lambda}$ is the $r \times r'$ Vandermonde matrix where $\tilde{\Lambda}(a,b) = \lambda_a^{b-1}$. Let $z \in \mathbb{R}^{r'}$ be the vector that satisfies $\Lambda^{2t+2}\tilde{\Lambda} z = \Lambda^2 \mathbf{1}$; z always exists and is unique because $\tilde{\Lambda}$ is a Vandermonde matrix, and $\Lambda^{-2t}\mathbf{1}$ lies within the span of its columns. For every $(u,v) \in [n]^2$, compute distance according to

$$\hat{d}(u,v) = \left(\tfrac{1}{p'}\right) \textstyle\sum_{\ell \in [r']} z_\ell \left(\tilde{N}_{u,t} - \tilde{N}_{v,t}\right)^T (M'' + M'_{\mathsf{ind}})\left(\tilde{N}_{u,t+\ell} - \tilde{N}_{v,t+\ell}\right). \tag{8}$$

3.2 Reducing computation by subsampling vertices

The pairwise distances can only be estimated up to a limited precision depending on the sparsity of the data and amount of noise in the observations, and furthermore we tune the nearest neighbor threshold to tradeoff between bias and variance. As a result, the performance of the algorithm can be maintained with reduced computation by clustering the coordinates so that not all n^2 pairwise distances need to be computed. This would involve adding an extra step at the beginning of the algorithm that samples sufficiently many "anchor" vertices $\mathcal{K} \subset [n]$ that cover the space well. $|\mathcal{K}|$ should be chosen large enough such that for any vertex $u \in [n]$, there exists some anchor vertex $i \in \mathcal{K}$ which is "close" to u in the sense that $\|\Lambda Q(e_u - e_i)\|_2^2$ is small. For all n vertices, we only

compute the distances to each of the $|\mathcal{K}|$ anchor vertices, and we let $\pi : [n] \to \mathcal{K}$ be a mapping from each vertex to the anchor vertex that minimizes the estimated distance $\hat{d}$ as computed in the original algorithm statement, $\pi(u) = \arg\min_{i \in \mathcal{K}} \hat{d}(u, i)$. The final estimate then is given by

$$\hat{F}(u, v) = \hat{F}(\pi(u), \pi(v)) = \frac{1}{|\mathcal{E}_{\pi(u)\pi(v)}|} \sum_{(a,b) \in \mathcal{E}_{\pi(u)\pi(v)}} M'''(a, b),$$

where $\mathcal{E}_{\pi(u)\pi(v)}$ denotes the set of undirected edges (a, b) such that $(a, b) \in \mathcal{E}_3$ and both $\hat{d}(\pi(u), a)$ and $\hat{d}(\pi(v), b)$ are less than some threshold η. We can compute $\mathcal{E}_{\pi(u)\pi(v)}$ by the clustering assignments and distances of all vertices to the anchor vertices.

3.3 Computational Complexity

To analyze the computational complexity of the algorithm, we consider each step. Growing local neighborhoods around each vertex costs at most $n|\mathcal{E}|$, since there are n vertices and the BFS trees visit each edge at most once. Computing the inner product for all pairs of vertices given the local neighborhood vectors costs at most $n^2|\mathcal{E}|$, since there are n^2 vertex pairs and $|\mathcal{E}|$ entries in the data matrix M. The final nearest neighbor estimator involves a (weighted) average of the datapoints, which costs at most $n^2|\mathcal{E}|$, as there are n^2 entries in the matrix to estimate, and at worst the estimate would involve averaging over $|\mathcal{E}|$ datapoints. This extremely crude bound leads to a computational complexity of $O(pn^4)$. The bottleneck of the algorithm is the final nearest neighbor estimate, which may be reduced by using approximate nearest neighbor methods.

If we instead used the modified algorithm that subsamples $|\mathcal{K}|$ anchor vertices at random and treats them as "cluster centers", there are only $(|\mathcal{K}|^2 + n|\mathcal{K}|)$ pairwise distances computed, for a computational cost of $(|\mathcal{K}|^2 + n|\mathcal{K}|)|\mathcal{E}|$ instead of $n^2|\mathcal{E}|$. Once we cluster the vertices, the final estimate is only computed for the pairwise cluster blocks, as the final estimate is a block constant matrix with only $|\mathcal{K}|^2$ distinct valued estimates. This results in $|\mathcal{K}|^2|\mathcal{E}|$ computation for the final step of the estimation. The computational complexity reduces from $O(n^2|\mathcal{E}|)$ to $O((|\mathcal{K}|^2+n|\mathcal{K}|)|\mathcal{E}|)$. The choice of $|\mathcal{K}|$ depends on the distribution of latent variables, the shape of the latent function, and the error tolerance. In a setting with finitely many latent types, then $|\mathcal{K}|$ would be roughly linear in the number of latent types.

A practical benefit of our algorithm is that it is amenable to a distributed and parallelized implementation. The key computational step of our algorithm involves comparing the expanded local neighborhoods of pairs of vertices to find the "nearest neighbors". As the algorithm is inherently local with respect to the data graph, it can be easily implemented for large scale datasets where the data may be stored in a distributed fashion optimized for local graph computations. The local neighborhoods can be computed in parallel, as they are independent computations. Using approximate nearest neighbor techniques and subsampling vertices to cluster will additionally reduce the computation.

3.4 Discussion

In practice, we may not know the model parameters, and we would use cross validation to tune the BFS tree depth t and nearest neighbor threshold η. If the depth t is either too small or too large, then the vector $N_{u,t}$ will be too sparse, and will not optimally aggregate the datapoints. The threshold η trades off between bias and variance of the final estimate. When the sampled observations are not uniform across entries, the algorithm may require more modifications to properly normalize for high degree hub vertices, as the optimal choice of depth t may differ depending on the local sparsity.

In our algorithm, we assumed that we observed the edge set $\mathcal{E}$. Specifically, this means that we are able to distinguish between entries of the matrix that have value zero because they are not observed, i.e. $(i,j) \notin \mathcal{E}$, or if the entry was observed to be value zero, i.e. $(i,j) \in \mathcal{E}$ and $M(i,j) = Z(i,j) = 0$. This fits well for applications such as recommendations, where the system does know the information of which entries are observed or not. Some social network applications contain this information (e.g. facebook would know if they have recommended a link which was then ignored) but other network information may lack this information, e.g. we do not know if link does not exist because observations are sparse, or because observations are dense but the probability of an edge is small. The absence of this knowledge would primarily affect the normalization of the neighborhood vectors as well as the normalization in the final averaging step.

The idea of comparing vertices by looking at larger radius neighborhoods was introduced in [2], and has connections to belief propagation [21, 4] and the non-backtracking operator [34, 28, 41, 40, 5]. The non-backtracking operator was introduced to overcome the issue of sparsity. For sparse graphs, vertices with high-degree dominate the spectrum, such that the informative components of the spectrum get hidden behind the high degree vertices. The non-backtracking operator avoids paths that immediately return to the previously visited vertex in a similar manner as belief propagation, and its spectrum has been shown to be more well-behaved, perhaps adjusting for the high degree vertices, which get visited very often by paths in the graph. In our algorithm, the neighborhood paths are defined by first selecting a rooted tree at each vertex, thus enforcing that each vertex along a path in the tree is unique. This is important in our analysis, as it guarantees that the distribution of vertices at the boundary of each subsequent depth of the neighborhood is unbiased, since the sampled vertices are freshly visited.

4 Results

In all of the results below, we assume the latent variable model assumptions laid out in Section 2. As a reminder, we assume uniform Bernoulli sampling with density p, independent bounded observation noise, and a generative latent variable model where coordinates are associated to i.i.d. sampled latent variables and the underlying matrix behaves according to a bounded latent function f that is Lipschitz and low rank (or approximately low rank) with bounded eigenfunctions.

4.1 f has rank r

We first provide theoretical bounds for the estimation error in both sparse regimes mentioned above when f has finite spectrum with rank r.

Sparse Regime. Theorem 4.1 shows that the maximum entrywise error of the collaborative filtering algorithm using distance function (7) converges to zero in the sparse regime when $p = n^{-1+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$.

Theorem 4.1. *Let f have rank r, $p = n^{-1+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$ so that $1/\kappa$ is not an integer. Consider the estimates produced by the nearest neighbor algorithm using the distance defined in (7) for $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$ and selecting the nearest neighbor distance threshold to satisfy $\eta = \Theta(n^{-\frac{1}{2}(\kappa-\rho)})$ for any $\rho \in (0, \kappa)$. Let $C_f = |\lambda_1|/|\lambda_r|$ denote the condition number of the latent function f. With probability $1 - o(1)$,*

$$\|\hat{F} - F\|_{\max} = O\Big(rC_f^{1/\kappa} n^{-\frac{1}{4}(\kappa-\rho)}\Big). \tag{9}$$

Furthermore,

$$\mathrm{MSE}(\hat{F}) = \frac{1}{n^2}\|\hat{F} - F\|_{Fr}^2 = O\Big(r^2 C_f^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}\Big). \tag{10}$$

Sparser Regime. Theorem 4.2 shows that the maximum entrywise error of the collaborative filtering algorithm using distance function (8) converges to zero in the sparser regime when $p = n^{-1}\ln^{1+\kappa} n$ for some $\kappa > 0$.

Theorem 4.2. *Let f have rank r, $p = n^{-1}\ln^{1+\kappa} n$ for some $\kappa > 0$. Consider the estimates produced by the nearest neighbor algorithm using the distance defined in (8) for $t = \lceil \frac{\ln(0.08/p)}{\ln(0.275np)} - r' \rceil$ and selecting the nearest neighbor distance threshold to satisfy $\eta = \Theta\Big((\ln n)^{-\frac{1}{2}(\kappa-\rho)}\Big)$ for any $\rho \in (0, \kappa)$. With probability $1 - o(1)$,*

$$\|\hat{F} - F\|_{\max} = O\Big((\ln n)^{-\frac{1}{4}(\kappa-\rho)}\Big). \tag{11}$$

Furthermore,

$$\mathrm{MSE}(\hat{F}) = O\Big((\ln n)^{-\frac{1}{2}(\kappa-\rho)} n\Big). \tag{12}$$

Theorems 4.1 and 4.2 show that for symmetric sparse matrix estimation, as long as the fraction of entries observed at random scale as $\frac{\log^{1+\kappa}(n)}{n}$ for any fixed $\kappa > 0$, the estimation error of our proposed iterative variant of the classical collaborative filtering algorithm with respect to the max-norm decays to 0 as $n \to \infty$ assuming the underlying matrix of interest has constant rank r.

4.2 f has ε-approximate rank r

We extend the above stated result to the setting when the latent function f has ε-approximate rank r; this captures settings where f may have infinite but quickly decaying spectrum. We formally state the extension in the sparse regime ($p = n^{-1+\kappa}$), but we believe that a similar result is likely to hold for the sparser regime ($p = n^{-1}\log^{1+\kappa}(n)$) as well, which we omit for simplicity of presentation.

Theorem 4.3. *Let f have ε-approximate rank r for some $\varepsilon > 0$, $p = n^{-1+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$ so that $1/\kappa$ is not an integer. Consider the estimates produced by the nearest neighbor algorithm using the distance defined in (7) for $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor < \frac{1}{\kappa} - 1$ and selecting the nearest neighbor distance threshold to satisfy $\eta = \Theta(n^{-\frac{1}{2}(\kappa-\rho)})$ for any $\rho \in (0, \kappa)$. Let $C_{f,r} = |\lambda_1|/|\lambda_r|$ denote the condition number of the rank r approximation to the latent function f. With probability $1 - o(1)$,*

$$\|\hat{F} - F\|_{\max} = O\Big(r C_{f,r}^{1/\kappa} n^{-\frac{1}{4}(\kappa-\rho)}\Big) + O\Big(|\lambda_r|^{-\frac{1}{\kappa}}\sqrt{r}\left(\sqrt{\frac{\varepsilon}{\kappa}}(1+\varepsilon)^{\frac{1}{2\kappa}-\frac{1}{2}} + \frac{\varepsilon}{\kappa}(1+\varepsilon)^{\frac{1}{\kappa}-\frac{3}{2}}\right)\Big) \tag{13}$$

Furthermore,

$$\mathrm{MSE}(\hat{F}) = O\Big(r^2 C_{f,r}^{\frac{2}{\kappa}} n^{-\frac{1}{2}(\kappa-\rho)}\Big) + O\Big(|\lambda_r|^{-\frac{2}{\kappa}} r\left(\frac{\varepsilon}{\kappa}(1+\varepsilon)^{\frac{1}{\kappa}-1} + \frac{\varepsilon^2}{\kappa^2}(1+\varepsilon)^{\frac{2}{\kappa}-3}\right)\Big). \tag{14}$$

As we assume the function values are bounded in $[0, 1]$, we can assume that $\varepsilon \in [0, 1]$, such that the dominating terms in (13) are $O\Big(r C_{f,r}^{1/\kappa} n^{-\frac{1}{4}(\kappa-\rho)}\Big) + O\Big(\sqrt{\varepsilon r |\lambda_r|^{-\frac{2}{\kappa}} \kappa^{-1}}\Big)$, and the dominating

terms in (14) are $O\Big(r^2 C_{f,r}^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}\Big) + O\Big(\varepsilon r |\lambda_r|^{-\frac{2}{\kappa}} \kappa^{-1}\Big)$. While we assume the function f is fixed with respect to n, when the function f has infinite spectrum, we can choose ε to decrease with n in order to tradeoff between the two terms in the error bound. Note that the approximate rank r and the approximate condition number $C_{f,r}$ also depend on the choice of ε. In particular the relationship between ε, r, and $C_{f,r}$ will depend on the spectrum of f and how quickly the tail decays to zero. Choosing a larger value of r will increase the condition number as $|\lambda_r|$ will be smaller, and it will decrease the approximation error ε. Below we present a specific example as a consequence of Theorem 4.3.

Corollary 4.4. *Let $p = n^{-1+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$ so that $1/\kappa$ is not an integer. Consider f such that for any $r \geq 1$, it has ε_r-approximate rank r with $|\lambda_r|$ corresponding to rank r approximation with $C_{f,r} = |\lambda_1|/|\lambda_r|$ being the condition number such that $|\lambda_1| = O(1)$ and*

$$\lim_{r \to \infty} \varepsilon_r |\lambda_r|^{-2/\kappa} r = 0. \tag{15}$$

Then, for any $\delta > 0$, for all n large enough, with probability $1 - o(1)$, $\|\hat{F} - F\|_{\max} = O\Big(\sqrt{\delta}\Big)$. Further, $\mathrm{MSE}(\hat{F}) = O\Big(\delta\Big)$.

Proof of Corollary 4.4. For any $\delta > 0$, by (15), there exists large enough $r = r(\delta)$ such that $\varepsilon_r |\lambda_r|^{-2/\kappa} r \leq \delta$. Due to $|\lambda_1| = O(1)$, $C_{f,r} = O(|\lambda_r|^{-1})$. Given choice of $r = r(\delta)$, for n large enough we have $r C_{f,r}^{1/\kappa} n^{-\frac{1}{4}(\kappa-\rho)} \leq \sqrt{\delta}$. By (13) of Theorem 4.3 it follows that $\|\hat{F} - F\|_{\max} = O(\sqrt{\delta})$ with probability at least $1 - o(1)$. By (14) of Theorem 4.3, it follows that $\mathrm{MSE}(\hat{F}) = O(\delta)$. □

From Corollary 4.4, it follows that $\|\hat{F} - F\|_{\max} = o(1)$ with probability $1 - o(1)$ and $\mathrm{MSE}(\hat{F}) = o(1)$ when the spectrum decays in such a way that $\lim_{r \to \infty} \varepsilon_r |\lambda_r|^{-2/\kappa} r = 0$.

4.3 Discussion

In our latent variable model, the latent function f is fixed with respect to n, so the max norm of the truth matrix is constant $\|F\|_{\max} = \Theta(1)$, and the Frobenius norm of the truth matrix scales linear with the matrix dimension so that $\frac{1}{n^2}\|F\|_{Fr}^2 = \Theta(1)$. As a result the above stated results also show the convergence rates with respect to the relative errors of the max norm and normalized Frobenius norm.

The overall proof sketch can be split into two parts. First we prove that the estimated pairwise distances concentrate to a metric computed with respect to the true latent function f. Second we prove that given well behaved estimated distances, the nearest neighbor estimate with properly chosen thresholds to balance mean and variance will converge at the above stated rate. This second part of the proof is straightforward and follows the standard proof for any nearest neighbor style algorithm. The crux of the proof is arguing that in sparse settings the computed distances concentrate well. This relies on the uniform sampling assumption, independence of the observation noise, regularity of the latent feature variables, and the finite spectrum assumption of the latent function. The assumptions on the specific distribution of the latent variables and the Lipschitzness of the latent function are in fact primarily used for the second nearest neighbor portion of the proof, and thus can be relaxed. The key property needed is that there are sufficiently many "nearest neighbor" coordinates; the precise distribution of the latent variables and shape of the latent function will affect the tuning of the threshold parameter to tradeoff between bias and variance. We provide formal statements for a few variations of the model in Section 6.

In addition to providing bounds on the MSE, our theorem also provides bounds on the maximum entrywise error of the estimate. The rate of our maximum entrywise error is the square root of the MSE rate, which suggests that the error is uniformly spread across all entries. This is a stronger guarantee that the typical MSE bounds found in the literature, and it can be useful for downstream results that use the estimates for decision making such as ranking and recommendations.

Thus far, we have focused on finding conditions on p that allow for consistent estimation with respect to both the MSE and max entrywise error. Our results also provide the rate at which the error decays. Specifically, our bound for the mean squared error (MSE) scales as $O((pn)^{-1/2+\rho})$ for any arbitrarily small constant $\rho > 0$, and our bound for the max entrywise error is $O((pn)^{-1/4+\rho})$ for any small ρ.

5 Proof Sketch for Analyzing Noisy Nearest Neighbors

As the algorithm uses a fixed radius nearest neighbor estimate, the analysis boils down to arguing that the distance functions as defined in (7) and (8) have certain desired properties that enable the classical nearest neighbor algorithm to be effective. In this section we characterize the needed properties for the convergence of noisy nearest neighbors.

Our algorithm estimates $F(u, v)$, i.e. $f(\theta_u, \theta_v)$, according to (4), which simply averages over datapoints $M(u', v')$ corresponding to tuples (u', v') for which u' is close to u and v' is close to v according to the estimated distance function $\hat{d}$. This simple nearest neighbor averaging estimator suggests that the last step of the analysis involves choosing the threshold η to tradeoff between bias and variance.

The primary desired property is that the data-driven distance estimates $\hat{d}(u, v)$ concentrate around some ideal data-independent distance $d(\theta_u, \theta_v)$ for $d : [0, 1]^2 \to \mathbb{R}_+$. We can then subsequently argue that the nearest neighbor estimate produced by (4) using $d(\theta_u, \theta_v)$ in place of $\hat{d}(u, v)$ will yield a good estimate by properly choosing the threshold η to tradeoff between bias and variance. The bias will depend on the local geometry of the function f relative to the distances defined by d. The variance depends on the measure of the latent variables $\{\theta_u\}_{u\in[n]}$ relative to the distances defined by d, i.e. the number of observed tuples $(u', v') \in \mathcal{E}'''$ such that $d(\theta_u, \theta_{u'}) \le \eta$ and $d(\theta_v, \theta_{v'}) \le \eta$ needs to be sufficiently large. We formalize the above stated desired properties.

Property 5.1 (Good Distance)**.** *We call an ideal distance function $d : [0, 1]^2 \to \mathbb{R}_+$ to be a* `bias`*-good distance function for some* `bias` $: \mathbb{R}_+ \to \mathbb{R}_+$ *if for any given $\eta > 0$ it follows that $|f(\theta_a, \theta_b) - f(\theta_u, \theta_v)| \le$* `bias`$(\eta)$ *for all $(\theta_a, \theta_b, \theta_u, \theta_v) \in [0, 1]^4$ such that $d(\theta_u, \theta_a) \le \eta$ and $d(\theta_v, \theta_b) \le \eta$.*

Property 5.1 follows from choosing an appropriate ideal distance function d. In particular we will choose d with respect to the spectral representation of f, and the desired property and the expression for `bias`(η) will follow from the low rank assumption as well as the boundedness of the eigenfunctions.

Property 5.2 (Good Distance Estimation)**.** *For some $\Delta > 0$, we call distance $\hat{d} : [n]^2 \to \mathbb{R}_+$ a Δ-good estimate for ideal distance $d : [0, 1]^2 \to \mathbb{R}_+$, if $|d(\theta_u, \theta_a) - \hat{d}(u, a)| \le \Delta$ for all $(u, a) \in [n]^2$.*

Showing property 5.2 is the crux of the proof and follows from the design of the algorithm along with the assumptions of uniform sampling and the latent variable model. It essentially uses all the model assumptions except for Lipschitzness of f.

Property 5.3 (Sufficient Representation)**.** *The collection of coordinate latent variables* $\{\theta_u\}_{u\in[n]}$ *is called* `meas`*-represented for some* $\texttt{meas} : \mathbb{R}_+ \to \mathbb{R}_+$ *if for any* $u \in [n]$ *and* $\eta' > 0$, $\frac{1}{n}\sum_{a\in[n]} \mathbb{I}(d(u,a) \leq \eta') \geq \texttt{meas}(\eta')$.

Property 5.3 is only used for the final step of the nearest neighbor analysis. In particular, as the estimate averages datapoints within an estimated nearby region of the target coordinates, there is a bias variance tradeoff that depends on how the datapoints are locally distributed. In particular, we need to guarantee that for any $(a, b) \in [n]^2$, there exists sufficiently many observed pairs $(u, v) \in [n]^2$ such that the function behaves similarly, i.e. $f(a, b)$ is close to $f(u, v)$. This property follows from our assumption that the latent variables are sampled i.i.d. from $U[0, 1]$, and that the function f is L-Lipschitz. As discussed in section 2, these assumptions can be relaxed, but alternative assumptions would need to guarantee property 5.3 for some reasonable local measure function $\texttt{meas}(\eta)$.

Given the above three properties, we can then prove Lemma 5.1, which characterizes the error of the noisy nearest neighbor algorithm as a function of the `bias` function, `meas` function, and estimation error Δ. Section 8 uses Lemma 5.1 to establish Theorems 4.1, 4.2, and 4.3 by simply showing the three properties for suitable choices of $\texttt{bias}, \texttt{meas}$, and Δ, and tuning η accordingly to balance between different terms of the error. Proving that the distance estimates concentrate well, i.e. property 5.2, is the most involved part of the analysis, which we defer to sections 9 and 10. Property 5.1 follows from the low rank assumption and property 5.3 arises from the latent variable model assumptions, in particular the distribution of the latent variables and shape of the latent function.

Lemma 5.1. *Assume that properties 5.1-5.3 hold with probability* $1 - \alpha$ *for some* η, Δ, *and* $\eta' = \eta - \Delta$*; in particular* d *is a* `bias`*-good distance function,* $\hat{d}$ *as estimated from* M' *and* M'' *is a* Δ*-good distance estimate for* d*, and* $\{\theta_u\}_{u\in[n]}$ *is* `meas`*-represented. The noisy nearest neighbor estimate* $\hat{F}$ *computed according to* (4) *satisfies*

$$\mathrm{MSE}(\hat{F}) \leq \texttt{bias}^2(\eta + \Delta) + \frac{2}{(1-\delta)p\left(\texttt{meas}(\eta-\Delta)n\right)^2} + \exp\left(-\frac{\delta^2 p\left(\texttt{meas}(\eta-\Delta)n\right)^2}{4}\right) + \alpha,$$

for any $\delta \in (0, 1)$*. Furthermore, for any* $\delta' \in (0, 1)$*,*

$$\max_{(u,v)\in[n]^2} |\hat{F}(u, v) - f(\theta_u, \theta_v)| \leq \texttt{bias}(\eta + \Delta) + \delta',$$

with probability at least

$$1 - n^2 \exp\left(-\tfrac{1}{4}\delta^2 p\left(\texttt{meas}(\eta-\Delta)n\right)^2\right) - n^2 \exp\left(-\delta'^2(1-\delta)p\left(\texttt{meas}(\eta-\Delta)n\right)^2\right) - \alpha.$$

Proof of Lemma 5.1. Recall that the algorithm uses sample splitting, where $\hat{d}$ is computed using M' and M'', and the final estimate $\hat{F}$ is computed using M'''. Therefore, for some $(a, b) \in \mathcal{E}'''$, the observation $M(a, b) = Z(a, b)$ is independent of $\hat{d}$, and $\mathbb{E}[M(a, b)] = f(\theta_a, \theta_b)$. Conditioned on $\mathcal{E}'''$,

by definition of $\hat{F}$ and by assuming properties 5.1 and 5.2, it follows that

$$\begin{aligned}\mathbb{E}[(\hat{F}(u,v) - f(\theta_u,\theta_v))^2] &= \left(\frac{1}{|\mathcal{E}'''_{uv}|}\sum_{(a,b)\in\mathcal{E}'''_{uv}} f(\theta_a,\theta_b) - f(\theta_u,\theta_v)\right)^2 \\ &\quad + \frac{1}{|\mathcal{E}'''_{uv}|^2}\sum_{(a,b)\in\mathcal{E}'''_{uv}} \mathrm{Var}[M(a,b)] \\ &\overset{(a)}{\leq} \mathrm{bias}^2(\eta+\Delta) + \frac{1}{|\mathcal{E}'''_{uv}|}.\end{aligned}$$

Inequality (a) follows from Properties 5.1-5.2: $|d(u,a)-\hat{d}(u,a)| \leq \Delta$ and $\hat{d}(u,a) \leq \eta \implies d(u,a) \leq \eta+\Delta$. By definition $M(a,b) \in [0,1]$ for all (a,b), which implies $\mathrm{Var}[M(a,b)] \leq 1$ for all $(a,b) \in \mathcal{E}'''$. Define $\mathcal{V}_{uv} = \{(a,b) \in [n]^2 \ : d(u,a) < \eta-\Delta,\ d(v,b) < \eta-\Delta\}$. Assuming property 5.3,

$$\begin{aligned}|\mathcal{V}_{uv}| &= |\{a \in [n] : d(u,a) < \eta-\Delta\}|\ |\{b \in [n] : d(v,b) < \eta-\Delta\}| \\ &\geq (\mathtt{meas}(\eta-\Delta)n)^2\,.\end{aligned}$$

By the Bernoulli sampling model and sample splitting process, each tuple $(a,b) \in [n]^2$ belongs to $\mathcal{E}'''$ with probability $p/2$ independently. By a straightforward application of Chernoff's bound, it follows that for any $\delta \in (0,1)$,

$$\mathbb{P}\left(|\mathcal{E}''' \cap \mathcal{V}_{uv}| \leq \frac{(1-\delta)p}{2}\left(\mathtt{meas}(\eta-\Delta)n\right)^2\right) \leq \exp\left(-\frac{\delta^2 p\left(\mathtt{meas}(\eta-\Delta)n\right)^2}{4}\right). \tag{16}$$

Therefore, by assuming property 5.2, it follows that with probability at least $1-\exp\left(-\frac{\delta^2 p(\mathtt{meas}(\eta-\Delta)n)^2}{4}\right)$,

$$\begin{aligned}|\mathcal{E}'''_{uv}| &= |\{(a,b) \in \mathcal{E}''' \ : \hat{d}(u,a) < \eta,\ \hat{d}(v,b) < \eta\}| \\ &\geq |\{(a,b) \in \mathcal{E}''' \ : d(u,a) < \eta-\Delta,\ d(v,b) < \eta-\Delta\}| \\ &= |\mathcal{E}''' \cap \mathcal{V}_{uv}| \\ &\geq \frac{(1-\delta)p}{2}\left(\mathtt{meas}(\eta-\Delta)n\right)^2.\end{aligned}$$

Define the event $\mathcal{H} = \{|\mathcal{E}'''_{uv}| \geq \frac{(1-\delta)p}{2}\left(\mathtt{meas}(\eta-\Delta)n\right)^2|\}$. It follows that $\mathbb{P}\left(\mathcal{H}^c\right) \leq \exp\left(-\frac{1}{4}\delta^2 p\left(\mathtt{meas}(\eta-\Delta)n\right)^2\right)$. By definition, $F(u,v) = f(\theta_u,\theta_v) \in [0,1]$ for all $u,v \in [n]$. Therefore, assuming properties 5.1-5.3 hold,

$$\begin{aligned}&\mathbb{E}[(\hat{F}(u,v) - f(\theta_u,\theta_v))^2] \\ &\leq \mathbb{E}[(\hat{F}(u,v) - f(\theta_u,\theta_v))^2 \ \Big|\ \mathcal{H}] + \mathbb{P}\left(\mathcal{H}^c\right) \\ &\leq \mathtt{bias}^2(\eta+\Delta) + \frac{2}{(1-\delta)p\left(\mathtt{meas}(\eta-\Delta)n\right)^2} + \exp\left(-\frac{1}{4}\delta^2 p\left(\mathtt{meas}(\eta-\Delta)n\right)^2\right).\end{aligned}$$

We add an additional α in the final MSE bound to account for the probability that properties 5.1-5.3 are violated.

To obtain the high-probability bound on the maximum entry-wise error, note that $M(a,b)$ are independent across indices $(a,b) \in \mathcal{E}'''$ as well as independent of observations in $\mathcal{E}' \cup \mathcal{E}''$. Additionally, the model assumes that $M(a,b), F(a,b) \in [0,1]$, and $\mathbb{E}[M(a,b)] = F(a,b)$ for observed tuples (a,b).

By an application of Hoeffding's inequality for bounded, zero-mean independent variables, for any $\delta' \in (0,1)$ it follows that assuming properties 5.1-5.3 hold,

$$\mathbb{P}\left(\frac{\left| \sum_{(a,b)\in\mathcal{E}'''_{uv}} (M(a,b) - F(a,b)) \right|}{|\mathcal{E}'''_{uv}|} \geq \delta' \;\middle|\; \mathcal{H} \right) \leq \exp\left(-\delta'^2 (1-\delta) p \left(\texttt{meas}(\eta - \Delta) n \right)^2 \right).$$

By union bound it follows that

$$\max_{(u,v)\in[n]^2} |\hat{F}_{uv} - f(\theta_u, \theta_v)| \leq \texttt{bias}(\eta + \Delta) + \delta',$$

with probability at least

$$1 - n^2 \exp\left(-\frac{1}{4}\delta^2 p \left(\texttt{meas}(\eta - \Delta) n \right)^2 \right) - n^2 \exp\left(-\delta'^2 (1-\delta) p \left(\texttt{meas}(\eta - \Delta) n \right)^2 \right) - \alpha.$$

This completes the proof of Lemma 5.1. □

6 Extensions

6.1 Subsampled Anchor Vertices

As mentioned in Section 3.2, we can reduce the computational complexity of the algorithm by subsampling a set of anchor vertices $\mathcal{K}$ and only computing pairwise distances relative to the anchor vertices, equivalent to computing a clustering amongst vertices and using that to estimate. For pairs of anchor vertices $(a,b) \in \mathcal{K}^2$ which we also refer to as cluster centers, the algorithm estimates $\hat{F}(a,b)$ according to the original stated algorithm with no modifications. For $u \notin \mathcal{K}$, we denote $\pi(u) = \arg\min_{i\in\mathcal{K}} \hat{d}(u,i)$ to be a clustering that maps from u to the closest anchor vertex in $\mathcal{K}$. The final estimate for $(u,v) \notin \mathcal{K}^2$ is then given by the estimate of the associated anchor vertices, which act as cluster centers, $\hat{F}(u,v) = \hat{F}(\pi(u), \pi(v))$.

The original argument provides high probability bounds on $|\hat{F}(u,v) - F(u,v)|$ for cluster centers $(u,b) \in \mathcal{K}^2$, as nothing changed in the algorithm for the cluster centers. The only additional part of the proof is to bound the additional bias for non cluster centers, as $|\hat{F}(u,v) - F(u,v)| \leq |\hat{F}(\pi(u),\pi(v)) - F(\pi(u),\pi(v))| + |F(\pi(u),\pi(v)) - F(u,v)|$. The first term is directly bounded by the current analysis, and the bias from the second term will depend on the size of $|\mathcal{K}|$. Recall our latent variable model assumption that each vertex u is associated to a latent variable $\theta_u \sum U[0,1]$ such that $F(u,v) = f(\theta_u, \theta_v)$ and f is L-Lipschitz with respect to the latent variables. For $|\mathcal{K}| = \frac{2}{\delta}\log(\frac{1}{\delta})$, with probability at least $1-\delta$, each interval $[(i-1)\delta, i\delta]$ for $i \in [1/\delta]$ contains at least one anchor point in $\mathcal{K}$, as the latent variables of these anchor points are chosen at random. Under this good event, then $\max_{u\in[n]} \min_{i\in\mathcal{K}} |\theta_u - \theta_i| \leq \delta$.

We discuss the results and analysis for the sparse setting when $p = n^{-1+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$, however a similar argument applies for the sparser setting of $p = n^{-1}\ln^{1+\kappa} n$ as well. Equation (20) will show that $d(\theta_u, \theta_v) \leq |\lambda_1|^{2t} L^2 |\theta_u - \theta_v|^2$, so that for some $u \in [n]$, the closest anchor point $a \in \mathcal{K}$ with respect to the latent representation will also satisfy $d(\theta_u, \theta_a) \leq |\lambda_1|^{2t} L^2 \delta^2$. As Property 5.2 guarantees $|d(\theta_u,\theta_a) - \hat{d}(u,a)| \leq \Delta$ for all estimated distances, it follows that $d(\theta_u, \theta_{\pi(u)}) \leq |\lambda_1|^{2t} L^2 \delta^2 + 2\Delta$ for all $u \in [n]$. By Property 5.1, $|F(\pi(u),\pi(v)) - F(u,v)| \leq \texttt{bias}(|\lambda_1|^{2t} L^2 \delta^2 + 2\Delta)$. We choose $|\mathcal{K}|$ so that $\delta = \frac{\sqrt{\Delta}}{L|\lambda_1|^t}$, and we plug in the choice of Δ and t from Theorem 4.1, resulting in $\delta = Br|\lambda_1|^{(\kappa+1)/\kappa} L^{-1} n^{-\frac{1}{4}(\kappa-\rho)} = o(1)$ so that $|\mathcal{K}| = \Theta(n^{\frac{1}{4}(\kappa-\rho)})$. This choice of $|\mathcal{K}|$ will guarantee that the extra added bias does not change the existing guarantees in Theorem 4.1 by more than a constant.

6.2 Local Geometry

We can generalize the latent variable model beyond scalar valued latent variables and Lipschitz latent functions. These assumptions only affect the function `meas` in Property 5.3, and thus it only changes the last portion of the nearest neighbor proof in which we tune the threshold η to tradeoff between the bias and variance terms. We present two examples of extending our results to a different local geometry, illustrating the modifications for the sparse setting when $p = n^{-1+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$.

If there were only m distinct latent types such that $\theta_u \in [m]$ and $p_{\min} = \min_{i \in [m]} \mathbb{P}(\theta_u = i) > 0$, then $\texttt{meas}(\eta')$ could be chosen to be a constant slightly less than $p_{\min}$ for every value of $\eta' > 0$. If the minimum distance measured by $d(\theta_u, \theta_v)$ between any two distinct types $\theta_u \neq \theta_v$ is larger than 2Δ, then we can choose η to be Δ so that by Property 5.2, the algorithm will achieve perfect clustering. In particular, if Property 5.2 holds then no vertex of a different type will have estimated distance less than η and $\texttt{bias}(\eta + \Delta) = 0$. Given this, each type has at least $p_{\min} n$ instances realized, on average. Therefore, for a given $u, v \in [n]$, there are roughly $(p_{\min} n)^2$ entries $(u', v') \in [n] \times [n]$ such that u, u' and v, v' are of the same type. Each of these $(p_{\min} n)^2$ is observed with probability p. Therefore, by taking average over these observed entries, the Mean Squared Error should scale as $1/(p(p_{\min} n)^2)$ and the max entry-wise error would scale as $(p(p_{\min} n)^2)^{-1/2}$. In the case that the minimum distance between any two distinct types is less than Δ, then the bias term will still be there and the limiting term is still $\texttt{bias}(\Delta)$, and thus the convergence rate would be limited by the same rate as stated in Theorem 4.1.

Next we discuss a higher dimensional setting. Assume the latent variables are sampled uniformly over a m-dimensional hypercube such that $\theta_u \sim U([0,1]^m)$ and the latent function f is L-Lipschitz with respect to an underlying metric d_m, such that the measure of a ball with radius δ is $\Theta(\delta^m)$. Property 5.3 would instead hold for $\texttt{meas}(\eta') = \Theta((\frac{\sqrt{\eta'}}{\lambda^t L})^m)$, resulting in a different choice of threshold η to balance between bias and variance. If $m \leq (\kappa + 2)/\kappa$, then the current $\texttt{bias}(\Delta)$ term dominates such that we would choose $\eta = \Theta(\Delta)$, and the error convergence rate will be the same as that stated in Theorem 4.1. For high dimension $m > (\kappa + 2)/\kappa$, we choose the threshold $\eta = \Theta((pn^2)^{-1/(m+1)})$ such that the MSE bound will scale as $\Theta((pn^2)^{-1/(m+1)}) = \Theta(n^{-(1+\kappa)/(m+1)})$ and the max entrywise error bound will scale as $\Theta(n^{-(1+\kappa)/2(m+1)})$.

6.3 Asymmetric Matrix

Even though our stated results are for symmetric models, we can transform an asymmetric latent variable model to a symmetric model as long as the row and column dimensions grow proportionally to one another. Consider an $n \times m$ matrix F which we would like to learn, where $F(u, v) = f(\alpha_u, \beta_v) \in [0, 1]$, and f has finite spectrum. We can construct a $(n + m) \times (n + m)$ matrix where F is placed on the off-diagonal blocks and the diagonal $n \times n$ and $m \times m$ blocks are set to zero. We can argue that this constructed matrix is sampled form a symmetric latent model, so that we can apply our algorithm and analysis directly.

6.4 Categorical Valued Data

If the edge labels are categorical instead of real-valued, then the goal is instead to estimate the distribution over the different categories or labels. This is particularly suitable for a setting in which there is no obvious metric between the categories such that an aggregate statistic such as the expected label would not be meaningful. If the edge labels take values within m category types, we can split the data is split into m different matrices, each containing the information for a separate category (or edge label). For each category or label $\ell \in [m]$, the associated matrix F_ℓ represents the

probability that each datapoint is labeled with ℓ, such that $\mathbb{P}(Z(u,v)=\ell) = F_\ell(u,v) = f_\ell(\alpha_u, \alpha_v)$, where f is a symmetric function having finite spectrum. The algorithm can then be applied to each matrix separately to estimate the probability of each category across the different entries. Since we need the estimates across different categories for the same entry to sum to 1, we can simply let the estimate for the m-th category one minus the sum of the estimates for the first $m-1$ categories. To obtain an error bound, we can simply use union bound across the $m-1$ applications of the algorithm, which simply multiplies the error probability by $m-1$.

6.5 Non-Uniform Sampling

We assumed a uniform sampling model, where each entry is observed independently with probability p. However, in reality the probability that entries are observed may not be uniform across all pairs (i,j). Our results can be extended to a setting where the sampling probability is instead a function of the latent variable, i.e. entry (i,j) is observed with probability $c_n g(\theta_i, \theta_j)$ where g is a Lipschitz low rank function independent of n and c_n is a scaling factor governing the density. The observed data $M(i,j)$ would then be sampled according to

$$M(i,j) = \begin{cases} 0 & \text{with probability } 1 - c_n g(\theta_i, \theta_j) \\ Z(i,j) & \text{with probability } c_n g(\theta_i, \theta_j). \end{cases}$$

for $\mathbb{E}[Z(i,j)] = f(\theta_i, \theta_j)$. Whereas previously we had $\mathbb{E}[M(i,j)] = pf(\theta_i, \theta_j)$, in this modified model, $\mathbb{E}[M(i,j)] = c_n g(\theta_i, \theta_j) f(\theta_i, \theta_j)$. A limitation of this model is that we need the sampling probabilities to all scale at the same order with respect to n. The model is not fully identifiable as we could multiply c_n by a constant and divide g by the same constant and obtain the same data distribution, and thus we can only estimate up to a constant scaling factor.

We can essentially then apply our algorithm twice, first using data matrix M to estimate the product $g(\theta_i, \theta_j) f(\theta_i, \theta_j)$ up to a scaling factor. Second we apply our algorithm to the binary adjacency matrix representing the sparsity of the observation set Ω in order to estimate $g(\theta_i, \theta_j)$ up to scaling factor. The one nuance one would have to handle is that since the set of observed entries is not uniformly sampled, the constructed BFS trees will grow non-uniformly, which will affect the normalization and scaling terms. As the model is only recoverable up to scaling, this is the best we can do. If we had data from a two-step sampling process in which we first observe binary edges sampled uniformly with probability c_n, and then subsequently observed datapoints sampled with an additional probability $g(\theta_i, \theta_j)$, then the model would exactly fall into our assumptions and the results could directly be applied to estimating $g(\theta_i, \theta_j)$ and the product $g(\theta_i, \theta_j) f(\theta_i, \theta_j)$.

7 Experiments

We show results on synthetic data to illustrate the performance of our algorithm. We did not do sample splitting as it is primarily introduced for the purpose of the analysis. We computed distances according to equation (7) (but again without sample splitting) for fixed radius parameters of $t \in \{0,1,2,3,4\}$. Note that the depth for expanding the BFS tree is until $t+1$. We did not specifically tune the nearest neighbor threshold η, but simply chose it to be the 70th percentile amongst all estimated distances. As a result, the expected number of entries used to compute the final weighted average estimate is $0.49pn^2$. We compare against a naive baseline which predicts using the column-wise mean. And we compare against the softimpute implementation in python's fancyimpute package and alternating least squares with rank 2 from parafac algorithm in the python tensorly package (higher rank performed more poorly in the sparse setting as it overfit to noise).

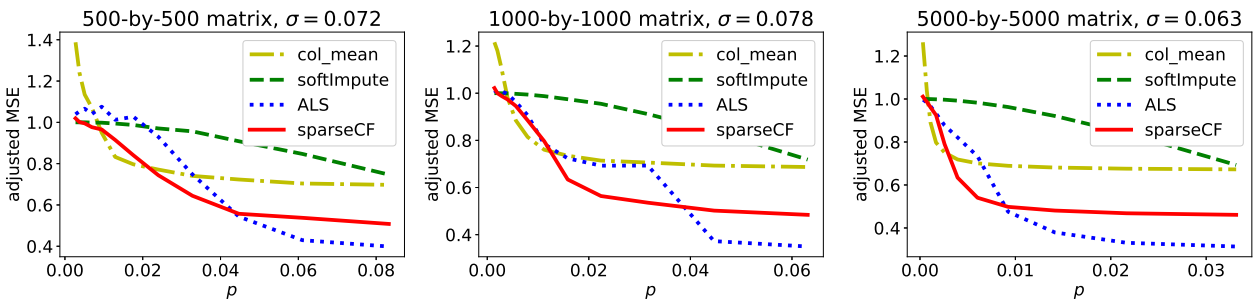

Figure 1: Adjusted MSE of missing entries vs. sampling probability p. The rank of the ground truth matrix is 10, and the observations are perturbed with mean zero additive Gaussian noise with variance σ^2. Results are shown left to right for matrices of sizes 500-by-500, 1000-by-1000, and 5000-by-5000.

Nuclear norm minimization was too slow for the size of instances that we show and thus was omitted.

The matrix F is generated as follows. For rank $r = 10$, we first sample two Gaussian $n \times r$ latent factor matrices $U \in \mathbb{R}^{n \times r}$ and $V \in \mathbb{R}^{n \times r}$. Each entry of the latent factor matrices is sampled from an independent Gaussian distribution with mean 10 and standard deviation 10. Next we compute F according to

$$F = \frac{(UV^T - \text{mean}(UV^T))}{\max(\text{abs}(UV^T))}.$$

For a $\kappa \in (0, 1]$, the density is chosen to be $p = n^{-1+\kappa}$, and each entry is observed (and thus included in sample set Ω) with probability p independently of all other entries. For each observed entry $(u, v) \in \Omega$, there is an added independent Gaussian noise $M(u, v) = F(u, v) + \varepsilon(u, v)$, where $\varepsilon(u, v) \sim N(0, \sigma^2)$ where σ is chosen to be the 40th percentile of the magnitude of entries in F. We show results for $n = 500, 1000$, and 5000.

We compute an adjusted mean squared error (MSE), limited to the error in predicting missing entries, and we normalize by the squared error of predicting with zeros. When the adjusted MSE is larger than 1, it means the estimate is worse than predicting all zeros.

$$\text{adjusted MSE} = \frac{\sum_{(u,v) \notin \Omega} (\hat{F}(u, v) - F(u, v))^2}{\sum_{(u,v) \notin \Omega} F(u, v)^2}$$

Figure 1 shows the adjusted MSE of the algorithms with respect to the sampling probability p. When p is very small, then our algorithm with the optimal choice of the depth parameter t performs better than ALS and SoftImpute, however when it is too sparse than either the simple mean estimate or predicting with all zeros is best. Note that we did not do any tuning of the nearest neighbor parameter η, and thus there may be additional gains possible for our algorithm. If we consider the minimum density for which the algorithm performs better than the simple mean, SoftImpute requires the most dense observation. The minimum density required for our algorithm depends on optimally choosing the depth parameter t, but for an optimal choice of t, our algorithm requires less data than ALS before it performance better than the simple mean.

Figure 2 shows the adjusted MSE of the algorithms with respect to the exponent of the density parameter κ where $p = n^{-1+\kappa}$. This rescales the x-axis so that the small values of p are more

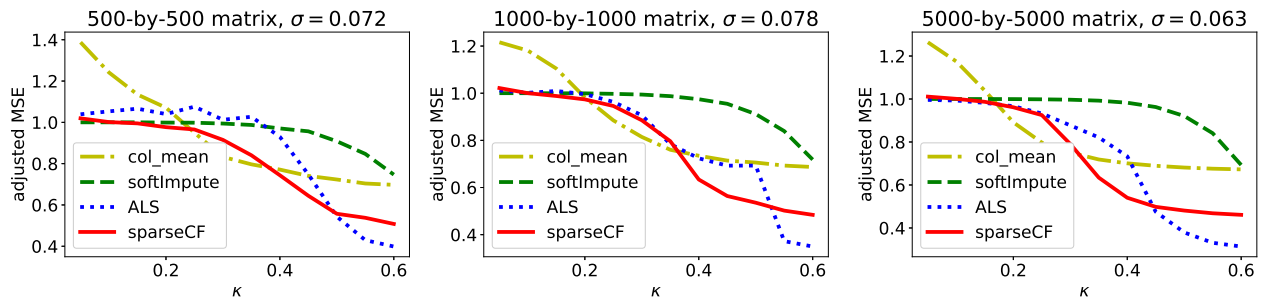

Figure 2: Adjusted MSE of missing entries vs. sampling probability exponent $\kappa = \ln(pn)/\ln(n)$. The rank of the ground truth matrix is 10, and observations are perturbed with mean zero additive Gaussian noise with variance σ^2. Results shown left to right for matrices of sizes 500-by-500, 1000-by-1000, and 5000-by-5000.

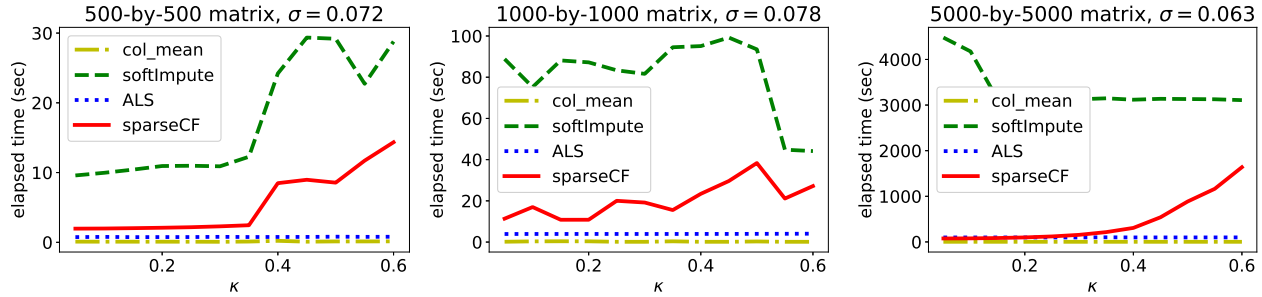

Figure 3: Computation time vs. sampling probability exponent $\kappa = \ln(pn)/\ln(n)$. The rank of the ground truth matrix is 10, and the observations are perturbed with mean zero additive Gaussian noise with variance σ^2. Results are shown left to right for matrices of sizes 500-by-500, 1000-by-1000, and 5000-by-5000.

visible. We plot only up to $\kappa = 0.6$ as we are focusing on the sparse regime with little overlaps in entries between pairs of rows and columns. Notice the dependence on the performance of our algorithm with respect to the radius parameter t as illustrated best in Figure **??**. For too small values of t the alg is suboptimal as it does not aggregate data sufficiently, but for too large values of t the algorithm again is suboptimal as it simply estimates zeros due to the BFS trees running out of vertices.

Figure 3 shows the time each of the algorithms took to run. We can see that our proposed algorithm is faster than SoftImpute, and this gap in speed is amplified with large n. Alternating Least Squares (ALS) is very fast, nearly as fast as the simple mean. Nuclear norm minimization was too slow to run on the size of instances in our example and thus was not included.

8 Proofs for Theorems 4.1, 4.2, and 4.3

In this section, we use the noisy nearest neighbor lemma 5.1 along with to establish Theorems 4.1, 4.2, and 4.3. Proofs of the concentration of distance estimates is deffered to sections 9 and 10.

8.1 Analyzing Sparse Regime: Proofs of Theorem 4.1 and 4.3

We prove that as long as $p = n^{-1+\kappa}$ for any $\kappa \in (0, \frac{1}{2})$, with high probability, properties 5.1-5.3 hold for an appropriately chosen function d, and for distance estimates $\hat{d}$ computed according to (7) with $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$. We subsequently use Lemma 5.1 to conclude Theorem 4.1. The most involved part in the proof is establishing that property 5.2 holds with high probability for an appropriately chosen Δ, which is delegated to Lemma 8.1.

Good distance d and Property 5.1. We start by defining the ideal distance d as follows. For all $(u, v) \in [n]^2$, let

$$d(\theta_u, \theta_v) = \|\Lambda^{t+1} Q(e_u - e_v)\|_2^2 = \sum_{k=1}^{r} \lambda_k^{2(t+1)} (q_k(\theta_u) - q_k(\theta_v))^2. \tag{17}$$

Recall that $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$. Assuming $p = n^{-1+\kappa}$, $\kappa \in (0, \frac{1}{2})$

$$t = \left\lfloor \frac{\ln(1/p)}{\ln(np)} \right\rfloor \;=\; \left\lfloor \frac{1}{\kappa} - 1 \right\rfloor. \tag{18}$$

We want to show that there exists $\texttt{bias} : \mathbb{R}_+ \to \mathbb{R}_+$ so that $|f(\theta_a, \theta_b) - f(\theta_u, \theta_v)| \le \texttt{bias}(\eta)$for any $\eta > 0$ and $(u, a, v, b) \in [n]^4$ such that $d(\theta_u, \theta_a) \le \eta$ and $d(\theta_v, \theta_b) \le \eta$. By the finite spectrum characterization of the function f, it follows that

$$\begin{aligned}
|f(\theta_u, \theta_v) - f(\theta_a, \theta_b)| &= |e_u^T Q^T \Lambda Q e_v - e_a^T Q^T \Lambda Q e_b| \\
&= |e_u^T Q^T \Lambda Q (e_v - e_b) - (e_a - e_u)^T Q^T \Lambda Q e_b| \\
&\overset{(a)}{\le} B\sqrt{r} \|\Lambda Q (e_v - e_b)\|_2 + B\sqrt{r} \|\Lambda Q (e_u - e_a)\|_2 \\
&\le B\sqrt{r} |\lambda_r|^{-t} \|\Lambda^{t+1} Q (e_v - e_b)\|_2 + B\sqrt{r} |\lambda_r|^{-t} \|\Lambda^{t+1} Q (e_u - e_a)\|_2 \\
&= B|\lambda_r|^{-t} \sqrt{r} \Big(\sqrt{d(\theta_v, \theta_b)} + \sqrt{d(\theta_u, \theta_a)} \Big) \\
&\le 2B|\lambda_r|^{-t} \sqrt{r\eta} \;\equiv \texttt{bias}(\eta),
\end{aligned} \tag{19}$$

where (a) follows from assuming that $|q_k(\theta)| \le B$ for all $k \in [r]$ and $\theta \in [0, 1]$. In summary, property 5.1 is satisfied for distance function d defined according to (17) and $\texttt{bias}(\eta) = 2B|\lambda_r|^{-t}\sqrt{r\eta}$.

Good distance estimate $\hat{d}$ and Property 5.2. We state the following Lemma when f has rank r, whose proof is delegated to Section 9.

Lemma 8.1. *Let f has rank r, $p = n^{-1+\kappa}$ for $\kappa \in (0, \frac{1}{2})$ such that $1/\kappa$ is not an integer. Consider $\hat{d}$ as computed in (7) with $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$. For any $\rho \in (0, \kappa)$*

$$\max_{u,a \in [n]^2} |d(\theta_u, \theta_a) - \hat{d}(u, a)| = O(Br|\lambda_1|^{2/\kappa} n^{-\frac{1}{2}(\kappa - \rho)}),$$

with probability at least $1 - O\Big(n^2 \exp\big(-\Theta(n^{\min(\rho, \kappa(t-\frac{1}{2}))})\big)\Big)$.

Lemma 8.1 implies that property 5.2 holds with probability $1 - o(1)$ for some $\Delta = \Theta(Br|\lambda_1|^{2/\kappa} n^{-(\kappa-\rho)/2})$ and any $\rho \in (0, \kappa)$. The distance error bound Δ is minimized by choosing ρ arbitrarily close to 0 so that Δ can be arbitrarily close to $\Theta(Br|\lambda_1|^{2/\kappa} n^{-\kappa/2}) = \Theta(Br|\lambda_1|^{2/\kappa}(pn)^{-1/2})$.

The corresponding statement for f that has ε-approximate rank r is stated below.

Lemma 8.2. *Let f have ε-approximate rank r, $p = n^{-1+\kappa}$ for $\kappa \in (0, \frac{1}{2})$ such that $1/\kappa$ is not an integer. Consider $\hat{d}$ as computed in (7) with $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$. For any $\rho \in (0, \kappa)$*

$$\max_{u,a \in [n]^2} |d(\theta_u, \theta_a) - \hat{d}(u,a)| = O(Br|\lambda_1|^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}) + O\Big(t\varepsilon(1+\varepsilon)^t + t^2\varepsilon^2(1+\varepsilon)^{2t-1}\Big),$$

with probability at least $1 - O\Big(n^2 \exp\big(-\Theta(n^{\min(\rho, \kappa(t-\frac{1}{2}))})\big)\Big)$.

Sufficient representation and Property 5.3. Since f is L-Lipschitz, the distance d as defined in (17) is bounded above by the squared ℓ_2 distance:

$$\begin{aligned} d(\theta_u, \theta_v) &= \|\Lambda^{t+1} Q(e_u - e_v)\|_2^2 \\ &\le |\lambda_1|^{2t} \|\Lambda Q(e_u - e_v)\|_2^2 \\ &= |\lambda_1|^{2t} \int_0^1 (f(\theta_u, y) - f(\theta_v, y))^2 dy \\ &\le |\lambda_1|^{2t} L^2 |\theta_u - \theta_v|^2. \end{aligned} \tag{20}$$

We assumed that the latent parameters $\{\theta_u\}_{u \in [n]}$ are sampled i.i.d. uniformly over $[0,1]$. Therefore, for any $\theta_u \in [0,1]$, for any $v \in [n]$ and $\eta' > 0$,

$$\begin{aligned} \mathbb{P}\left(d(\theta_u, \theta_v) \le \eta' \;\middle|\; \theta_u\right) &\ge \mathbb{P}\left(|\lambda_1|^{2t} L^2 |\theta_u - \theta_v|^2 \le \eta' \;\middle|\; \theta_u\right) \\ &= \mathbb{P}\left(|\theta_u - \theta_v| \le \frac{\sqrt{\eta'}}{|\lambda_1|^t L} \;\middle|\; \theta_u\right) \\ &\ge \min\left(1, \frac{\sqrt{\eta'}}{|\lambda_1|^t L}\right). \end{aligned}$$

Let us define

$$\texttt{meas}(\eta') = \frac{(1-\delta)\sqrt{\eta'}}{|\lambda_1|^t L} \tag{21}$$

for all $\eta' \in (0, |\lambda_1|^{2t} L^2)$. By an application of Chernoff's bound and a simple majorization argument, it follows that for all $\eta' \in (0, |\lambda_1|^{2t} L^2)$ and $\delta \in (0,1)$,

$$\mathbb{P}\left(\frac{1}{n-1} \sum_{a \in [n] \setminus u} \mathbb{I}\left(d(u,a) \le \eta'\right) \le \texttt{meas}(\eta') \;\middle|\; \theta_u\right) \le \exp\left(-\frac{\delta^2 (n-1)\sqrt{\eta'}}{2|\lambda_1|^t L}\right).$$

By using union bound over all n indices, it follows that for any $\eta' \in (0, |\lambda_1|^{2t} L^2)$, with probability at least $1 - n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2|\lambda_1|^t L}\right)$, property 5.3 is satisfied with `meas` as defined in (21).

Concluding Proof of Theorem 4.1. In summary, with probability at least $1 - \alpha$ for

$$\alpha = O(n^2 \exp\left(-\Theta(n^{\min(\rho, \kappa(t-\frac{1}{2}))})\right)) + n \exp\left(-\frac{\delta^2(n-1)\sqrt{\eta - \Delta}}{2|\lambda_1|^t L}\right),$$

properties 5.1-5.3 are satisfied for the estimate $\hat{d}$ computed from (7) with $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$, and the choices of

$$\begin{aligned} d(\theta_u, \theta_v) &= \|\Lambda^{t+1} Q(e_u - e_v)\|_2^2, \\ \texttt{bias}(\eta) &= 2B|\lambda_r|^{-t}\sqrt{r\eta}, \\ \Delta &= \Theta(Br|\lambda_1|^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}), \\ \texttt{meas}(\eta') &= \frac{(1-\delta)\sqrt{\eta'}}{|\lambda_1|^t L}, \end{aligned} \tag{22}$$

for any $\eta > 0$, $\rho \in (0, \kappa)$, $\delta \in (0,1)$ and $\eta' = \eta - \Delta \in (0, |\lambda_1|^{2t} L^2)$. By substituting the expressions for `bias`, `meas`, and α into Lemma 5.1, it follows that

$$\begin{aligned} \mathrm{MSE}(\hat{F}) \le 4B^2|\lambda_r|^{-2t} r(\eta+\Delta) + \frac{2|\lambda_1|^{2t}L^2}{(1-\delta)^3 pn^2(\eta-\Delta)} + \exp\left(-\frac{\delta^2 pn^2(1-\delta)^2(\eta-\Delta)}{4L^2|\lambda_1|^{2t}}\right) \\ + O(n^2 \exp\left(-\Theta(n^{\min(\rho,\kappa(t-\frac{1}{2}))})\right)) + n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2|\lambda_1|^t L}\right). \end{aligned} \tag{23}$$

Additionally, for any $\delta' \in (0,1)$,

$$\max_{(u,v)\in[n]^2} |\hat{F}(u,v) - f(\theta_u, \theta_v)| \le 2B|\lambda_r|^{-t}\sqrt{r(\eta+\Delta)} + \delta' \tag{24}$$

with probability at least

$$\begin{aligned} 1 - n^2\exp\left(-\tfrac{\delta^2(1-\delta)^2 pn^2(\eta-\Delta)}{4|\lambda_1|^{2t}L^2}\right) - n^2\exp\left(-\tfrac{\delta'^2(1-\delta)^3 pn^2(\eta-\Delta)}{|\lambda_1|^{2t}L^2}\right) \\ - O(n^2\exp\left(-\Theta(n^{\min(\rho,\kappa(t-\frac{1}{2}))})\right)) - n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2|\lambda_1|^t L}\right). \end{aligned}$$

By selecting $\eta = \Theta\big(Br|\lambda_1|^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}\big)$ with a large enough constant, it follows that

$$\begin{aligned} \eta \pm \Delta &= \Theta(\eta) = \Theta(\Delta), \\ pn^2\eta &= \Theta(Br|\lambda_1|^{2/\kappa} n^{1+\kappa-\frac{1}{2}(\kappa-\rho)}) = \Omega(Br|\lambda_1|^{2/\kappa} n^{1+\kappa/2}), \\ n\sqrt{\eta} &= \omega(\sqrt{Br}|\lambda_1|^{1/\kappa} n^{\frac{7}{8}}). \end{aligned}$$

By substituting this choice of η and $\delta = \frac{1}{2}$ into (23), it follows that

$$\mathrm{MSE}(\hat{F}) = O\Big(r^2 B^3 |\lambda_r|^2 (|\lambda_1|/|\lambda_r|)^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}\Big). \tag{25}$$

By choosing $\delta' = 2B|\lambda_r|^{-t}\sqrt{r(\eta+\Delta)}$, it follows that $\delta'^2 pn^2\eta = \Omega(n)$. Therefore, by substituting into (24), it follows that with probability $1 - o(1)$,

$$\max_{(u,v)\in[n]^2} |\hat{F}(u,v) - f(\theta_u,\theta_v)| = O\Big(rB^{3/2}|\lambda_r|(|\lambda_1|/|\lambda_r|)^{1/\kappa} n^{-\frac{1}{4}(\kappa-\rho)}\Big). \tag{26}$$

This completes the proof of Theorem 4.1. □

Concluding Proof of Theorem 4.3. Like Proof of Theorem 4.1, with probability at least $1 - \alpha$ for

$$\alpha = O(n^2\exp\left(-\Theta(n^{\min(\rho,\kappa(t-\frac{1}{2}))})\right)) + n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2|\lambda_1|^t L}\right),$$

properties 5.1-5.3 are satisfied for the estimate $\hat{d}$ computed from (7) with $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$, and the choices of

$$\begin{aligned} d(\theta_u, \theta_v) &= \|\Lambda^{t+1} Q(e_u - e_v)\|_2^2, \\ \texttt{bias}(\eta) &= 2B|\lambda_r|^{-t}\sqrt{r\eta}, \\ \Delta &= \Theta(Br|\lambda_1|^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}) + \Theta\Big(t\varepsilon(1+\varepsilon)^t + t^2\varepsilon^2(1+\varepsilon)^{2t-1}\Big), \\ \texttt{meas}(\eta') &= \frac{(1-\delta)\sqrt{\eta'}}{|\lambda_1|^t L}, \end{aligned} \tag{27}$$

for any $\eta > 0$, $\rho \in (0, \kappa)$, $\delta \in (0,1)$ and $\eta' = \eta - \Delta \in (0, |\lambda_1|^{2t} L^2)$. Note that the only difference is in choice of Δ due to Lemma 8.2 for f that has ε-approximate rank r. By substituting the expressions for `bias`, `meas`, and α into Lemma 5.1, it follows that

$$\begin{aligned} \mathrm{MSE}(\hat{F}) \le 4B^2|\lambda_r|^{-2t} r(\eta+\Delta) + \frac{2|\lambda_1|^{2t}L^2}{(1-\delta)^3 pn^2(\eta-\Delta)} + \exp\left(-\frac{\delta^2 pn^2(1-\delta)^2(\eta-\Delta)}{4L^2|\lambda_1|^{2t}}\right) \\ + O(n^2 \exp\left(-\Theta(n^{\min(\rho,\kappa(t-\frac{1}{2}))})\right)) + n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2|\lambda_1|^t L}\right). \end{aligned} \tag{28}$$

Additionally, for any $\delta' \in (0,1)$,

$$\max_{(u,v)\in[n]^2} |\hat{F}(u,v) - f(\theta_u, \theta_v)| \le 2B|\lambda_r|^{-t}\sqrt{r(\eta+\Delta)} + \delta' \tag{29}$$

with probability at least

$$\begin{aligned} 1 - n^2\exp\left(-\tfrac{\delta^2(1-\delta)^2 pn^2(\eta-\Delta)}{4|\lambda_1|^{2t}L^2}\right) - n^2\exp\left(-\tfrac{\delta'^2(1-\delta)^3 pn^2(\eta-\Delta)}{|\lambda_1|^{2t}L^2}\right) \\ - O(n^2\exp\left(-\Theta(n^{\min(\rho,\kappa(t-\frac{1}{2}))})\right)) - n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2|\lambda_1|^t L}\right). \end{aligned}$$

By selecting $\eta = \Theta\big(Br|\lambda_1|^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}\big) + \Theta(t\varepsilon(1+\varepsilon)^t + t^2\varepsilon^2(1+\varepsilon)^{2t-1}\Big)$ with appropriately large enough constants, it follows that

$$\begin{aligned} \eta \pm \Delta &= \Theta(\eta) = \Theta(\Delta), \\ pn^2\eta &= \Omega(n^{1+\kappa/2}), \\ n\sqrt{\eta} &= \omega(n^{\frac{7}{8}}). \end{aligned}$$

By substituting this choice of η and $\delta = \frac{1}{2}$ into (23), and using $t < 1/\kappa - 1$, it follows that

$$\mathrm{MSE}(\hat{F}) = O\Big(r^2(|\lambda_1|/|\lambda_r|)^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}\Big) + O\Big(|\lambda_r|^{-2/\kappa} r\left(\frac{\varepsilon}{\kappa}(1+\varepsilon)^{1/\kappa-1} + \frac{\varepsilon^2}{\kappa^2}(1+\varepsilon)^{2/\kappa-3}\right)\Big). \tag{30}$$

By choosing $\delta' = \Theta(B|\lambda_r|^{-t}\sqrt{r(\eta+\Delta)})$, it follows that $\delta'^2 pn^2\eta = \Omega(n)$. Therefore, by substituting into (24), it follows that with probability $1 - o(1)$,

$$\begin{aligned} &\max_{(u,v)\in[n]^2} |\hat{F}(u,v) - f(\theta_u,\theta_v)| \\ &\quad = O\Big(r(|\lambda_1|/|\lambda_r|)^{1/\kappa} n^{-\frac{1}{4}(\kappa-\rho)}\Big) + O\Big(|\lambda_r|^{-\frac{1}{\kappa}}\sqrt{r}\big(\sqrt{\frac{\varepsilon}{\kappa}}(1+\varepsilon)^{\frac{1}{2\kappa}-\frac{1}{2}} + \frac{\varepsilon}{\kappa}(1+\varepsilon)^{\frac{1}{\kappa}-\frac{3}{2}}\big)\Big). \end{aligned} \tag{31}$$

This completes the proof of Theorem 4.3. □

8.2 Analyzing Sparser Regime: Proof of Theorem 4.2

Similar to the proof of Theorem 4.1, we prove that as long as $p = \frac{\log n^{1+\kappa}}{n}$ for any $\kappa > 0$, with high probability, properties 5.1-5.3 are satisfied for an appropriately chosen function d and for distance estimates $\hat{d}$ computed according to (8) with $t = \lceil \frac{\ln(0.08/p)}{\ln(0.275pn)} - r' \rceil$. We subsequently use Lemma 5.1 to conclude Theorem 4.2. The most involved part in the proof is establishing that property 5.2 holds with high probability for an appropriately chosen Δ, which is delegated to Lemma 8.3.

Good distance d and Property 5.1. We start by defining the ideal distance d as follows. For all $(u, v) \in [n]^2$,

$$d(\theta_u, \theta_v) = \|\Lambda Q(e_u - e_v)\|_2^2 = \int_0^1 (f(\theta_u, y) - f(\theta_v, y))^2 dy. \tag{32}$$

For any $u, v, a, b \in [n]$ with corresponding $\theta_u, \theta_v, \theta_a, \theta_b \in [0, 1]$,

$$\begin{aligned}
|f(\theta_u, \theta_v) - f(\theta_a, \theta_b)| &= |e_u^T Q^T \Lambda Q e_v - e_a^T Q^T \Lambda Q e_b| \\
&= |e_u^T Q^T \Lambda Q (e_v - e_b) - (e_a - e_u)^T Q^T \Lambda Q e_b| \\
&\overset{(a)}{\leq} B\sqrt{r}\|\Lambda Q(e_v - e_b)\|_2 + B\sqrt{r}\|\Lambda Q(e_u - e_a)\|_2, \\
&= B\sqrt{r}(\sqrt{d(\theta_v, \theta_b)} + \sqrt{d(\theta_u, \theta_a)}),
\end{aligned}$$

where (a) follows from assuming that $|q_k(\theta)| \leq B$ for all $k \in [r]$ and $\theta \in [0, 1]$. It follows that for any $\eta > 0$, if $d(\theta_u, \theta_a) \leq \eta$ and $d(\theta_v, \theta_b) \leq \eta$, then $|f(\theta_u, \theta_v) - f(\theta_a, \theta_b)| \leq 2B\sqrt{r\eta}$. In summary, property 5.1 is satisfied for distance d defined in (32) with $\texttt{bias} : \mathbb{R}_+ \to \mathbb{R}_+$ defined as $\texttt{bias}(\eta) = 2B\sqrt{r\eta}$.

Good distance estimation $\hat{d}$ and Property 5.2. We state the following Lemma whose proof is delegated to Section 9.

Lemma 8.3. *Assume that $p = n^{-1} \ln^{1+\kappa} n$ for some $\kappa > 0$. Consider $\hat{d}$ as computed in (8) with*

$$t = \left\lceil \frac{\ln(0.08/p)}{\ln(0.275np)} - r' \right\rceil.$$

For any $\rho \in (0, \kappa)$,

$$\max_{u,a \in [n]^2} |d(\theta_u, \theta_a) - \hat{d}(u, a)| \leq c(\ln n)^{-\frac{1}{2}(\kappa - \rho)}$$

with probability at least

$$1 - O\Big(n^2 \exp(-\Theta((\ln n)^{1+\rho}))\Big),$$

where $c = c(\lambda_1, \lambda_r, \lambda_{gap}, r, B)$ is independent of n and $\lambda_{gap} = \min_{1 \leq s < s' \leq r} |\lambda_s - \lambda_s'|$.

Therefore, property 5.2 is satisfied with probability $1 - o(1)$ for some $\Delta = \Theta\left((\ln n)^{-\frac{1}{2}(\kappa - \rho)}\right)$ for any $\rho \in (0, \kappa)$.

Sufficient representation and Property 5.3. Since f is L-Lipschitz, the distance d as defined in (17) is bounded above by squared ℓ_2 distance:

$$\begin{aligned}
d(\theta_u, \theta_v) = \|\Lambda Q(e_u - e_v)\|_2^2 &= \int_0^1 (f(\theta_u, y) - f(\theta_v, y))^2 dy &(33)\\
&\leq L^2 |\theta_u - \theta_v|^2. &(34)
\end{aligned}$$

Note that the only difference in (20) and (34) is the constant $L^2|\lambda_1|^{2t}$ versus L^2. It follows by a similar argument that with probability at least $1 - n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2L}\right)$, for any $\eta' \in (0, L^2)$, property 5.3 is satisfied with $\texttt{meas}(\eta') = \frac{(1-\delta)\sqrt{\eta'}}{L}$.

Concluding Proof of Theorem 4.2. In summary, with probability at least $1 - \alpha$ for

$$\alpha = O\Big(n^2\exp(-\Theta((\ln n)^{1+\rho}))\Big) + n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2L}\right),$$

properties 5.1-5.3 are satisfied for the estimate $\hat{d}$ computed from (8) with $t = \lceil \frac{\ln(0.08/p)}{\ln(0.275np)} - r' \rceil$, and the choices of

$$\begin{aligned} d(\theta_u, \theta_v) &= \|\Lambda Q(e_u - e_v)\|_2^2, \\ \texttt{bias}(\eta) &= 2B\sqrt{r\eta}, \\ \Delta &= \Theta\left((\ln n)^{-\frac{1}{2}(\kappa-\rho)}\right), \\ \texttt{meas}(\eta') &= \frac{(1-\delta)\sqrt{\eta'}}{L}, \end{aligned} \tag{35}$$

for any $\eta > 0$, $\rho \in (0, \kappa)$, $\delta \in (0, 1)$ and $\eta' = \eta - \Delta \in (0, L^2)$. By substituting the expressions for $\texttt{bias}$, $\texttt{meas}$, and α into Lemma 5.1, it follows that

$$\begin{aligned} \mathrm{MSE}(\hat{F}) \le 4B^2 r(\eta + \Delta) + \frac{2\sigma^2 L^2}{(1-\delta)^3 pn^2(\eta - \Delta)} + \exp\left(-\frac{\delta^2 pn^2(1-\delta)^2(\eta-\Delta)}{4L^2}\right) \\ + O\Big(n^2\exp(-\Theta((\ln n)^{1+\rho}))\Big) + n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2L}\right). \end{aligned} \tag{36}$$

Additionally, for any $\delta' \in (0, 1)$,

$$\max_{(u,v)\in[n]^2} |\hat{F}(u,v) - f(\theta_u, \theta_v)| \le 2B\sqrt{r(\eta+\Delta)} + \delta' \tag{37}$$

with probability at least

$$\begin{aligned} 1 - n^2\exp\left(-\tfrac{\delta^2(1-\delta)^2 pn^2(\eta-\Delta)}{4L^2}\right) - n^2\exp\left(-\tfrac{\delta'^2(1-\delta)^3 pn^2(\eta-\Delta)}{L^2}\right) \\ - O\Big(n^2\exp(-\Theta((\ln n)^{1+\rho}))\Big) - n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2L}\right). \end{aligned}$$

By selecting $\eta = \Theta\left(\left(\frac{\ln^{1+\rho} n}{np}\right)^{1/2}\right) = \Theta\Big((\ln n)^{-\frac{1}{2}(\kappa-\rho)}\Big)$ with a large enough constant, it follows that

$$\begin{aligned} \eta \pm \Delta &= \Theta(\eta) = \Theta(\Delta), \\ pn^2\eta &= \Omega(n), \\ n\sqrt{\eta} &= \omega(\sqrt{n}). \end{aligned}$$

By substituting this choice of η and $\delta = \frac{1}{2}$ into (36) it follows that

$$\mathrm{MSE}(\hat{F}) = O(\eta) \;=\; O\Big((\ln n)^{-\frac{1}{2}(\kappa-\rho)}\Big). \tag{38}$$

By choosing $\delta' = \Theta(\sqrt{\eta})$, it follows that $\delta'^2 pn^2\eta = \omega(\sqrt{n})$. Therefore, by substituting into (37), it follows that with probability $1 - o(1)$,

$$\max_{(u,v)\in[n]^2} |\hat{F}(u,v) - f(\theta_u,\theta_v)| = O(\sqrt{\eta}) \;=\; O\Big((\ln n)^{-\frac{1}{4}(\kappa-\rho)}\Big). \tag{39}$$

This completes the proof of Theorem 4.2. □

9 Proving distance estimates are close when f has rank r

This section is dedicated to establishing that the distance estimates (7) and (8) are good approximations of the desired ideal distances as claimed in the statements of Lemmas 8.1 and 8.3 when f has rank r. We start by establishing key auxiliary concentration results which will lead to their proofs.

9.1 Regular enough growth of bread-first-search tree

Recall that we grow the neighborhood of each $u \in [n]$ in $\mathcal{G} = ([n], \mathcal{E}')$ and use associated observations in M' as well as M'' to compute the distance estimates $\hat{d}$. By the assumed Bernoulli sampling model, any tuple $(a,b) \in [n]^2$ is independently included in $\mathcal{E}'$ with probability $p/4$. Therefore, the expected number of immediate neighbors of u (not including itself) is $(n-1)p/4 \approx np/4$. The expected number of nodes at distance $s \geq 1$ from a given u scales as $(np/4)^s$. We define some necessary notation before we present the formal statement of this event. Given $\delta \in (0,1)$, define

$$\phi(\delta) = 1 - \left(\frac{1-\delta}{1-\delta\sqrt{2/3}}\right)^{1/2} \;<\; 1. \tag{40}$$

For any $p = \omega\big(\frac{1}{n}\big)$ and $p = o(1)$,

$$s^*(\delta,p,n) = \sup\Big\{s \geq 1 : \frac{p}{8}\left(\frac{(1+\delta)np}{4}\right)^{s-1} \leq \phi(\delta)\Big\}. \tag{41}$$

For any given δ, $s^*(\delta,p,n)$ is well defined for n large enough since $p = o(1)$.

Lemma 9.1. *Let $\omega(\frac{1}{n}) \leq p \leq o(1)$, $\delta \in (0,1)$. For $1 \leq s \leq s^*(\delta,p,n)$,*

$$\mathbb{P}\left(\bigcup_{h=1}^{s}\left\{|\mathcal{S}_{u,h}| \notin \left[\left(\frac{(1-\delta)np}{4}\right)^h, \left(\frac{(1+\delta)np}{4}\right)^h\right]\right\}\right) \leq 4\exp\left(-\frac{\delta^2((1-\delta)np)}{12(1-\delta\sqrt{2/3})}\right)$$

The proof of Lemma 9.1 follows from standard argument using repeated application of Chernoff's bound and is well known in the literature in various forms. For completeness, we have included it in the Appendix. Lemma 9.1 suggests definition of events that will hold with high probability. Specifically, for any $u \in [n]$ and $h \geq 1$, define

$$\mathcal{A}^1_{u,h}(\delta) = \left\{|\mathcal{S}_{u,h}| \in \left[\left(\frac{(1-\delta)np}{4}\right)^h, \left(\frac{(1+\delta)np}{4}\right)^h\right]\right\}. \tag{42}$$

We note that by event $\mathcal{A}^1_{u,h}(\delta)$ we simply require that the number of nodes at distance h from a given node $u \in [n]$ is nearly $(np/4)^h$. However, it does not impose any restrictions on how the nodes are connected or the latent parameters associated with the nodes themselves.

9.2 Concentration of a Quadratic Form One

The event $\cap_{h=1}^{s+\ell}\mathcal{A}_{u,h}^1(\delta)$ implies that the size of $|\mathcal{S}_{u,h}|$ grows regularly as expected for $h \le s+\ell \le s^*(\delta,p,n)$. Conditioned on this event, we prove that a specific quadratic form concentrates around its mean. This will be used as the key property to eventually establish that the distance estimates are a good approximation to the ideal distances.

Lemma 9.2. *Let $\omega(\frac{1}{n}) \le p \le o(1)$, $\delta \in (0,1)$, $s \ge 0, \ell \ge 1$ for $s+\ell \le s^*(\delta,p,n)$. Then*

$$\mathbb{P}\left(|e_k^T Q\tilde{N}_{u,s+\ell} - e_k^T\Lambda^\ell Q\tilde{N}_{u,s}| \ge \lambda_k^\ell((1-\delta)np/4)^{-(s+1)/2}x \;\middle|\; \cap_{h=1}^{s+\ell}\mathcal{A}_{u,h}^1(\delta)\right) \le 2\exp\left(-\frac{x^2\lambda_k^2}{4}\right),$$

as long as $x < \frac{2((1-\delta)np/4)^{(s+1)/2}}{B|\lambda_k|(1+|\lambda_k|)}$.

Proof of Lemma 9.2. Recall that conditioning on event $\cap_{h=1}^{s+\ell}\mathcal{A}_{u,h}^1(\delta)$ simply imposes the restriction that the neighborhood of $u \in [n]$ grows at a specific rate, i.e. number of nodes at distances $h \le s+\ell$ is within $((1\pm\delta)np/4)^h$. However, this event is independent from latent parameters $\{\theta_i\}_{i\in[n]}$ and the realization of observations $M(i,j) = Z(i,j)$ for $(i,j) \in [n]\times[n]$. Consider any realization of the tree $\mathcal{T}_u^{s+\ell}$ satisfying $\cap_{h=1}^{s+\ell}\mathcal{A}_{u,h}^1(\delta)$; the tree contains information regarding the depth $s+\ell$ neighborhood of u. Given such a realization, let $\mathcal{F}_{u,h}$ for $0 \le h \le s+\ell$ denote the sigma-algebra containing information about the latent parameters, edges and the values associated with $\mathcal{T}_u^h$, i.e. the depth h BFS tree rooted at u. Specifically, $\mathcal{F}_{u,0}$ contains information about latent parameter θ_u associated with $u \in [n]$; $\mathcal{F}_{u,s}$ contains information about latent parameters $\cup_{h=1}^s\{\theta_i\}_{i\in\mathcal{S}_{u,h}}$ and all edges and observations involved in the depth h BFS tree, i.e. $\{M(i,j)\}_{(i,j)\in\mathcal{T}_u^h}$. This implies that $\mathcal{F}_{u,0} \subset \mathcal{F}_{u,1} \subset \mathcal{F}_{u,2}$, etc.

We shall consider a specific martingale sequence with respect to the filtration $\mathcal{F}_{u,h}$ that will help establish the desired concentration of $e_k^T Q\tilde{N}_{u,s+\ell} - e_k^T\Lambda^\ell Q\tilde{N}_{u,s}$. For $s+1 \le h \le s+\ell$, define

$$\begin{aligned} Y_{u,h} &= e_k^T\Lambda^{s+\ell-h}Q\tilde{N}_{u,h} \\ D_{u,h} &= Y_{u,h} - Y_{u,h-1} \\ Y_{u,s+\ell} - Y_{u,s} &= e_k^T Q\tilde{N}_{u,s+\ell} - e_k^T\Lambda^\ell Q\tilde{N}_{u,s} \\ &= \sum_{h=s+1}^{s+\ell} D_{u,h} \end{aligned}$$

Note that $Y_{u,h}$ is measurable with respect to $\mathcal{F}_{u,h}$ because $e_k^T\Lambda^{s+\ell-h}Q\tilde{N}_{u,h}$ only depends on observations in $\mathcal{T}_u^h$ and latent variables associated to vertices in $\mathcal{S}_{u,h}$. We will show that $Y_{u,h}$ is martingale with finite mean with respect to $\mathcal{F}_{u,h}$ for $s+1 \le h \le s+\ell$,

$$\mathbb{E}[Y_{u,h} - Y_{u,h-1} \mid \mathcal{F}_{u,h-1}] = 0 \text{ and } \mathbb{E}[|D_{u,h}|] < \infty. \tag{43}$$

For any $s+1 \le h \le s+\ell$,

$$\begin{aligned} D_{u,h} &= Y_{u,h} - Y_{u,h-1} \\ &= \lambda_k^{s+\ell-h}\left(e_k^T Q \tilde{N}_{u,h} - \lambda_k e_k^T Q \tilde{N}_{u,h-1}\right) \\ &= \lambda_k^{s+\ell-h}\left(\frac{1}{|\mathcal{S}_{u,h}|} e_k^T Q N_{u,h} - \lambda_k e_k^T Q \tilde{N}_{u,h-1}\right) \\ &= \lambda_k^{s+\ell-h}\left(\frac{1}{|\mathcal{S}_{u,h}|} \sum_{i \in \mathcal{S}_{u,h}} N_{u,h}(i) q_k(\theta_i) - \lambda_k e_k^T Q \tilde{N}_{u,h-1}\right) \\ &= \sum_{i \in \mathcal{S}_{u,h}} X_i, \end{aligned}$$

where for $i \in \mathcal{S}_{u,h}$, we define

$$X_i \triangleq \frac{\lambda_k^{s+\ell-h}}{|\mathcal{S}_{u,h}|}\left(N_{u,h}(i) q_k(\theta_i) - \lambda_k e_k^T Q \tilde{N}_{u,h-1}\right). \tag{44}$$

By definition,

$$N_{u,h}(i) = \sum_{j \in S_{u,h-1}} \mathbb{I}((i,j) \in \mathcal{E}') M(i,j) N_{u,h-1}(j). \tag{45}$$

Conditioned on $\mathcal{F}_{u,h-1}$, $N_{u,h-1}(j)$ for $j \in S_{u,h-1}$ is determined and so is θ_j. However, θ_i is conditionally independent random variable. Also, given the construction of the breadth-first-search tree, for any given $i \in S_{u,h}$ any of the $j \in S_{u,h-1}$ is equally likely to be its parent with probability $1/|S_{u,h-1}|$. Therefore, we have that $X_i,\ i \in \mathcal{S}_{u,h}$ are independent and

$$\begin{aligned} &\mathbb{E}\Big[X_i | \mathcal{F}_{u,h-1}\Big] \\ &= \frac{\lambda_k^{s+\ell-h}}{|\mathcal{S}_{u,h}|}\Big(\sum_{j \in S_{u,h-1}} \frac{1}{|S_{u,h-1}|} \mathbb{E}[f(\theta_i,\theta_j) q_k(\theta_i) | \theta_j] N_{u,h-1}(j) - \lambda_k e_k^T Q \tilde{N}_{u,h-1}\Big). \end{aligned} \tag{46}$$

Now $N_{u,h-1}(j)/|S_{u,h-1}| = \tilde{N}_{u,h-1}(j)$. And

$$\begin{aligned} \mathbb{E}[f(\theta_i,\theta_j) q_k(\theta_i) | \theta_j] &= \sum_{k'=1}^{r} \lambda_{k'} \mathbb{E}[q_{k'}(\theta_i) q_{k'}(\theta_j) q_k(\theta_i) | \theta_j] \\ &= \sum_{k'=1}^{r} \lambda_{k'} q_{k'}(\theta_j) \mathbb{E}[q_{k'}(\theta_i) q_k(\theta_i)] \\ &= \lambda_k q_k(\theta_j), \end{aligned}$$

where we use the orthonormality of $q_{k'},\ k' \in [r]$. Therefore,

$$\begin{aligned} \sum_{j \in S_{u,h-1}} \frac{1}{|S_{u,h-1}|} \mathbb{E}[f(\theta_i,\theta_j) q_k(\theta_i) | \theta_j] N_{u,h-1}(j) &= \sum_{j \in S_{u,h-1}} \lambda_k q_k(\theta_j) \tilde{N}_{u,h-1}(j) \\ &= \lambda_k e_k^T Q \tilde{N}_{u,h-1}. \end{aligned}$$

Therefore, we conclude that for $i \in S_{u,h}$

$$\mathbb{E}\Big[X_i | \mathcal{F}_{u,h-1}\Big] = 0. \tag{47}$$

That is, $\mathbb{E}[Y_{u,h} - Y_{u,h-1} | \mathcal{F}_{u,h-1}] = 0$. By definition, we have $N_{u,h}(i) \in [0,1]$ for any $i \in \mathcal{S}_{u,h}$ and $\|q_k\|_\infty \leq B$. Therefore, it follows that for any $i \in S_{u,h}$,

$$|X_i| \leq \frac{B(1+|\lambda_k|)|\lambda_k|^{s+\ell-h}}{|\mathcal{S}_{u,h}|}. \tag{48}$$

Therefore, it follows that

$$|D_{u,h}| \leq B(1+|\lambda_k|)|\lambda_k|^{s+\ell-h}. \tag{49}$$

Thus, we have $\{(D_{u,h}, \mathcal{F}_{u,h}) : s+1 \leq h \leq s+\ell\}$ as a martingale difference sequence with differences being uniformly bounded. Now we wish to establish its concentration. To that end, consider X_i for $i \in \mathcal{S}_{u,h}$ as defined in (44). Its variance is bounded as

$$\begin{aligned}
&\mathrm{Var}[X_i \mid \mathcal{F}_{u,h-1}] \\
&= \frac{\lambda_k^{2(s+\ell-h)}}{|\mathcal{S}_{u,h}|^2} \mathrm{Var}\Big[\sum_{j \in S_{u,h-1}} \mathbb{I}((i,j) \in \mathcal{E}') M(i,j) N_{u,h-1}(j) q_k(\theta_i) \mid \mathcal{F}_{u,h-1} \Big].
\end{aligned}$$

Since $\mathrm{Var}[Z] \leq \mathbb{E}[Z^2]$ for any Z, we can upper bound the variance expression by the second moment, additionally using the fact that $\mathbb{I}((i,j) \in \mathcal{E}')$ only takes value 1 for a single $j \in S_{u,h-1}$ and otherwise takes value 0,

$$\begin{aligned}
&\mathrm{Var}[X_i \mid \mathcal{F}_{u,h-1}] \\
&= \frac{\lambda_k^{2(s+\ell-h)}}{|\mathcal{S}_{u,h}|^2} \mathbb{E}\Big[\sum_{j \in S_{u,h-1}} \mathbb{I}((i,j) \in \mathcal{E}') M(i,j)^2 N_{u,h-1}^2(j) q_k^2(\theta_i) \mid \mathcal{F}_{u,h-1} \Big].
\end{aligned}$$

We use the fact that $M(i,j)^2 \leq 1$, $\mathbb{E}[q_k^2(\theta_i)] = 1$ due to orthonormality assumptions on q_k, for $i \in \mathcal{S}_{u,h}$ it holds that $\mathbb{E}[\mathbb{I}((i,j) \in \mathcal{E}' \mid \mathcal{F}_{u,h-1}] = \frac{1}{|\mathcal{S}_{u,h-1}|}$, so that

$$\mathrm{Var}[X_i \mid \mathcal{F}_{u,h-1}] \leq \frac{\lambda_k^{2(s+\ell-h)}}{|\mathcal{S}_{u,h}|^2} \frac{\|N_{u,h-1}\|_2^2}{|\mathcal{S}_{u,h-1}|} \overset{(a)}{\leq} \frac{\lambda_k^{2(s+\ell-h)}}{|\mathcal{S}_{u,h}|^2}$$

where (a) follows from the assumption that $N_{u,h-1}$ has sparsity $\mathcal{S}_{u,h-1}$ and has entries bounded in $[0,1]$. It follows that X_i conditioned on $\mathcal{F}_{u,h-1}$ is sub-exponential with parameters

$$\left(\frac{\lambda_k^{(s+\ell-h)}}{|\mathcal{S}_{u,h}|}, \frac{B(1+|\lambda_k|)|\lambda_k|^{s+\ell-h}}{|\mathcal{S}_{u,h}|} \right).$$

Now $D_{u,h}$ is sum of such X_i for $i \in \mathcal{S}_{u,h}$ which are independent of each other conditioned on $\mathcal{F}_{u,h-1}$. Therefore, it follows that conditioned on $\mathcal{F}_{u,h-1}$, $D_{u,h}$ is sub-exponential with parameters

$$\Big(\frac{\lambda_k^{(s+\ell-h)}}{\sqrt{|\mathcal{S}_{u,h}|}}, \frac{B(1+|\lambda_k|)|\lambda_k|^{s+\ell-h}}{|\mathcal{S}_{u,h}|} \Big).$$

Since $\{(D_{u,h}, \mathcal{F}_{u,h}) : s+1 \le h \le s+\ell\}$ is a martingale difference sequence, $\sum_{h=s+1}^{s+\ell} D_{u,h}$ conditioned on $\mathcal{F}_{u,s}$ is sub-exponential with parameters

$$\left(\sqrt{\sum_{h=s+1}^{s+\ell} \frac{\lambda_k^{2(s+\ell-h)}}{|\mathcal{S}_{u,h}|}}, \max_{h\in[s+1,s+\ell]} \frac{B(1+|\lambda_k|)|\lambda_k|^{s+\ell-h}}{|\mathcal{S}_{u,h}|} \right).$$

Under event $\cap_{h=1}^{s+\ell} \mathcal{A}_{u,h}^1(\delta)$, for any realization of the breadth-first-search tree of u, $|\mathcal{S}_{u,h}| \in [((1-\delta)np/4)^h, ((1+\delta)np/4)^h]$ for all $h \in [s+\ell]$. Therefore, we can bound the sub-exponential parameters of $\sum_{h=s+1}^{s+\ell} D_{u,h}$ conditioned on $\mathcal{F}_{u,s}$ using the property $p = \omega(1/n)$ or $np = \omega(1)$ as

$$\left(\lambda_k^{\ell-1}\sqrt{2}\left(\frac{(1-\delta)np}{4}\right)^{-(s+1)/2}, B(1+|\lambda_k|)|\lambda_k|^{\ell-1}\left(\frac{(1-\delta)np}{4}\right)^{-(s+1)} \right).$$

By Azuma's concentration inequality, for $0 < x < \frac{2((1-\delta)np/4)^{(s+1)/2}}{B|\lambda_k|(1+|\lambda_k|)}$,

$$\begin{aligned}
&\mathbb{P}\left(|e_k^T Q \tilde{N}_{u,s+\ell} - e_k^T \Lambda^\ell Q \tilde{N}_{u,s}| \ge \lambda_k^\ell ((1-\delta)np/4)^{-(s+1)/2} x \mid \cap_{h=1}^{s+\ell} \mathcal{A}_{u,h}^1(\delta), \mathcal{F}_{u,s} \right) \\
&\le 2\exp\left(-\min\left(\frac{x^2\lambda_k^2}{4}, \frac{x|\lambda_k|((1-\delta)np/4)^{(s+1)/2}}{2B(1+|\lambda_k|)} \right) \right) \\
&\le 2\exp\left(-\frac{x^2\lambda_k^2}{4} \right).
\end{aligned}$$

This completes the proof of Lemma 9.2. □

Lemma 9.2 suggests the following high probability events: for any $u \in [n], k \in [r], x > 0, s \ge 0, \ell \ge 1$, $\delta \in (0,1)$, define

$$\mathcal{A}_{u,k,s,\ell}^2(x,\delta) = \left\{ |e_k^T Q \tilde{N}_{u,s+\ell} - e_k^T \Lambda^\ell Q \tilde{N}_{u,s}| \le \lambda_k^\ell ((1-\delta)np/4)^{-(s+1)/2} x \right\}.$$

9.3 Concentration of a Quadratic Form Two

We state a useful concentration that builds on Lemma 9.2 towards establishing Lemma 8.1.

Lemma 9.3. *Let $\omega(\frac{1}{n}) \le p \le o(1)$, $\delta \in (0,1)$, $s \ge 0, \ell \ge 1$ with $s+\ell \le s^*(\delta,p,n)$, and $x \le B((1-\delta)np/4)^{1/2}$. Consider any $u, v \in [n]$. Then, conditioned on event $\cap_{k=1}^r (\mathcal{A}_{u,k,0,s}^2(x,\delta) \cap \mathcal{A}_{v,k,0,s+\ell}^2(x,\delta))$, we have*

$$\left| \tilde{N}_{u,s}^T F \tilde{N}_{v,s+\ell} - e_u^T Q^T \Lambda^{2s+\ell+1} Q e_v \right| \le \frac{3Bx}{((1-\delta)np/4)^{1/2}} \Big(\sum_{k=1}^r |\lambda_k|^{2s+\ell+1} \Big).$$

and

$$\left| \tilde{N}_{u,s}^T F \tilde{N}_{v,s+\ell} \right| \le 4B^2 \Big(\sum_{k=1}^r |\lambda_k|^{2s+\ell+1} \Big).$$

Proof of Lemma 9.3. Assuming event $\cap_{k=1}^{r}(\mathcal{A}^2_{u,k,0,s}(x,\delta)\cap\mathcal{A}^2_{v,k,0,s+\ell}(x,\delta))$ holds, and using the fact that $F=Q^T\Lambda Q$, it follows that

$$\begin{aligned}
&|\tilde{N}^T_{u,s}F\tilde{N}_{v,s+\ell}-e_u^TQ^T\Lambda^{2s+\ell+1}Qe_v|\\
&\leq|(\tilde{N}^T_{u,s}Q^T-e_u^TQ^T\Lambda^s)(\Lambda Q\tilde{N}_{v,s+\ell}-\Lambda^{s+\ell+1}Qe_v)|\\
&\quad+|(\tilde{N}^T_{u,s}Q^T-e_u^TQ^T\Lambda^s)\Lambda^{s+\ell+1}Qe_v|+|e_u^TQ^T\Lambda^{s+1}(Q\tilde{N}_{v,s+\ell}-\Lambda^{s+\ell}Qe_v)|\\
&\leq\Big|\sum_{k=1}^{r}(e_k^TQ\tilde{N}_{u,s}-e_k^T\Lambda^sQe_u)(e_k^T\Lambda Q\tilde{N}_{v,s+\ell}-e_k^T\Lambda^{s+\ell+1}Qe_v)\Big|\\
&\quad+\Big|\sum_{k=1}^{r}(e_k^TQ\tilde{N}_{u,s}-e_k^T\Lambda^sQe_u)e_k^T\Lambda^{s+\ell+1}Qe_v\Big|\\
&\quad+\Big|\sum_{k=1}^{r}(e_k^T\Lambda^{s+1}Qe_u)(e_k^TQ\tilde{N}_{v,s+\ell}-e_k^T\Lambda^{s+\ell}Qe_v)\Big|\\
&\leq\frac{x}{((1-\delta)np/4)^{1/2}}\left(\frac{x}{((1-\delta)np/4)^{1/2}}+2B\right)\left(\sum_{k=1}^{r}|\lambda_k|^{2s+\ell+1}\right)\\
&\leq\frac{3Bx}{((1-\delta)np/4)^{1/2}}\left(\sum_{k=1}^{r}|\lambda_k|^{2s+\ell+1}\right),
\end{aligned}\tag{50}$$

where we have used the conditioned event $\cap_{k=1}^{r}(\mathcal{A}^2_{u,k,0,s}(x)\cap\mathcal{A}^2_{v,k,0,s+\ell}(x))$, the model assumption that $\|Q\|_\infty\leq B$, and the fact that $x\leq B((1-\delta)np/4)^{1/2}$ for n sufficiently large. From (50), it follows that

$$\begin{aligned}
&|\tilde{N}^T_{u,t}F\tilde{N}_{v,t+\ell}|\\
&\leq|e_u^TQ^T\Lambda^{2t+\ell+1}Qe_v|+|\tilde{N}^T_{u,t}F\tilde{N}_{v,t+\ell}-e_u^TQ^T\Lambda^{2t+\ell+1}Qe_v|\\
&\leq(B^2+3B^2)\Big(\sum_{k=1}^{r}|\lambda_k|^{2t+\ell+1}\Big).
\end{aligned}$$

□

9.4 Concentration of a Quadratic Form Three

We establish a final concentration that will lead us to the proof of good distance function property.

Lemma 9.4. *Let $\omega(\frac{1}{n})\leq p\leq o(1)$, $\delta\in(0,1)$, $s\geq 0,\ell\geq 1$ with $s+\ell\leq s^*(\delta,p,n)$ and $0<x\leq B((1-\delta)np/4)^{1/2}$. Let $u,v\in[n]$. Define event*

$$A'(u,v,s,\ell)(x)=\cap_{k=1}^{r}(\mathcal{A}^2_{u,k,0,s}(x)\cap\mathcal{A}^2_{v,k,0,s+\ell}(x))\cap\mathcal{A}^1_{u,s}\cap\mathcal{A}^1_{v,s+\ell}.$$

For $0<z\leq 4B^2\sqrt{\Big(\sum_{k=1}^{r}|\lambda_k|^{2s+\ell+1}\Big)\times p'((1-\delta)np/4)^{2s+\ell}}$, conditioned on the event $A'(u,v,s,\ell)(x)$, with probability at least

$$1-2\exp\left(-\frac{z^2}{8B^2}\right)-\exp\left(-\Theta\left(p'\left(\frac{(1-\delta)np}{4|\lambda_r|^{-1}}\right)^{2s+\ell-\frac{1}{2}}\right)\right),$$

it holds that

$$|\frac{1}{p'}\tilde{N}_{u,s}\big(M''+M'_{\mathsf{ind}}\big)\tilde{N}_{v,s+\ell}-\tilde{N}_{u,s}F\tilde{N}_{v,s+\ell}|\leq\frac{|\lambda_r|^{2s}}{(pn)^{1/2}}+z\sqrt{\frac{\sum_{k=1}^{r}|\lambda_k|^{2s+\ell+1}}{p'((1-\delta)np/4)^{2s+\ell}}}.$$

Proof of Lemma 9.4. We establish this result by arguing that conditioned on the event $A'(u,v,s,\ell)(x)$, the matrix $M'' + M'_{\mathsf{ind}}$ is statistically very similar to a freshly sampled dataset with density p'. Recall that $\mathcal{E}'_{\mathsf{ind}}$ was constructed so that conditioned on $\mathcal{E}'$, the set $\mathcal{E}'_{\mathsf{ind}} \cup \mathcal{E}''$ is distributed according to a Bernoulli(p') sampling model, where each $(u,v) \in [n]^2$ are included in $\mathcal{E}'_{\mathsf{ind}} \cup \mathcal{E}''$ independently with probability p'. The event $A'(u,v,s,\ell)(x)$ depends on $\mathcal{E}'$ and the values $M(i,j)$ such that $(i,j) \in \mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}$. Therefore datapoints $M(i,j) = Z(i,j)$ for tuples $(i,j) \notin \mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}$ are independent from the event $A'(u,v,s,\ell)(x)$.

Let us define $M''_{\mathsf{ind}} = [M''_{\mathsf{ind}}(i,j)]$ where

$$M''_{\mathsf{ind}}(i,j) = \begin{cases} M(i,j) = Z(i,j) & \text{if } (i,j) \in (\mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}}) \text{ and } (i,j) \notin \mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell} \\ Z_{\mathsf{ind}}(i,j) & \text{if } (i,j) \in \mathcal{E}'_{\mathsf{ind}} \text{ and } (i,j) \in \mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell} \end{cases},$$

and $Z_{\mathsf{ind}}(i,j)$ is a freshly sampled observation for edge (i,j), distributed equivalently to $Z(i,j)$. Conditioned on $\mathcal{E}'$ and the event $A'(u,v,s,\ell)(x)$, M''_{ind} has sparsity pattern $\mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}}$, which is distributed according to a Bernoulli(p') sampling model where each $(i,j) \in [n]^2$ is included in $\mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}}$ with probability p'. Furthermore, conditioned on $A'(u,v,s,\ell)$, for each $(i,j) \in \mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}}$ with probability p', the datapoint $M''_{\mathsf{ind}}(i,j)$ is independent of all observations used to compute $\tilde{N}_{u,s}$ and $\tilde{N}_{v,s+\ell}$. As a result, $M''_{\mathsf{ind}}(i,j)$ is a fresh independent signal of $F(i,j)$, distributed according to $Z(i,j)$. First we will argue that

$$\big(\tfrac{1}{p'}\big)\tilde{N}_{u,s}^T(M'' + M'_{\mathsf{ind}})\tilde{N}_{v,s+1} \approx \big(\tfrac{1}{p'}\big)\tilde{N}_{u,s}^T M''_{\mathsf{ind}}\tilde{N}_{v,s+1}.$$

By construction, M''_{ind} differs from $M'' + M'_{\mathsf{ind}}$ only for indices $(i,j) \in \mathcal{E}'_{\mathsf{ind}} \cap (\mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell})$. Therefore, it follows that

$$\begin{aligned} &|N_{u,s}M''_{\mathsf{ind}}N_{v,s+\ell} - N_{u,s}\big(M'' + M'_{\mathsf{ind}}\big)N_{v,s+\ell}| \\ &\le \sum_{i,j} \mathbb{I}((i,j) \in \mathcal{E}'_{\mathsf{ind}} \cap (\mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}))|Z_{\mathsf{ind}}(i,j) - Z(i,j)|N_{u,s}(i)N_{v,s+\ell}(j). \end{aligned}$$

By the boundedness assumption, $|Z_{\mathsf{ind}}(i,j) - Z(i,j)| \le 1$. Furthermore, $N_{u,s}(i)N_{v,s+\ell}(j) \in [0,1]$ is only nonzero for $(i,j) \in \mathcal{S}_{u,s} \times \mathcal{S}_{v,s+\ell}$. Therefore,

$$\begin{aligned} &|N_{u,s}M''_{\mathsf{ind}}N_{v,s+\ell} - N_{u,s}\big(M'' + M'_{\mathsf{ind}}\big)N_{v,s+\ell}| \\ &\le \sum_{i,j} \mathbb{I}((i,j) \in \mathcal{E}'_{\mathsf{ind}} \cap (\mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}))\mathbb{I}((i,j) \in \mathcal{S}_{u,s} \times \mathcal{S}_{v,s+\ell}) \\ &= |\{(i,j) \in \mathcal{E}'_{\mathsf{ind}} \cap (\mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}) \cap (\mathcal{S}_{u,s} \times \mathcal{S}_{v,s+\ell})\}| =: X. \end{aligned}$$

Conditioned on $\mathcal{E}'$ and the event $A'(u,v,s,\ell)(x)$, the quantity above, denoted as X, is distributed as a Binomial random variable, where each pair $(i,j) \in (\mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}) \cap (\mathcal{S}_{u,s} \times \mathcal{S}_{v,s+\ell})$ is included in the set $\mathcal{E}'_{\mathsf{ind}}$ independently with probability p'. The number of tuples in $(\mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}) \cap (\mathcal{S}_{u,s} \times \mathcal{S}_{v,s+\ell})$ is bounded above by $|\mathcal{S}_{u,s}| + |\mathcal{S}_{v,s+\ell}|$, since the only edges in $\mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}$ that intersect with $\mathcal{S}_{u,s} \times \mathcal{S}_{v,s+\ell}$ must be at the last layer of $\mathcal{T}_u^s$ or $\mathcal{T}_v^{s+\ell}$. By construction, the number of edges in tree $\mathcal{T}_u^s$ at depth s is equal to $|\mathcal{S}_{u,s}|$. For sufficiently large n, by event $A'(u,v,s,\ell)(x)$, it follows that $|\mathcal{S}_{u,s}| \le |\mathcal{S}_{v,s+\ell}|$. Therefore the random variable X is stochastically dominated by a Binomial$(2|\mathcal{S}_{v,s+\ell}|, p')$ random variable. For sufficiently large n, conditioned on $\mathcal{E}'$ and the event $A'(u,v,s,\ell)(x)$, by Chernoff's

bound,

$$\begin{aligned}
&\mathbb{P}\left(X \geq \frac{p'|\mathcal{S}_{u,s}||\mathcal{S}_{v,s+\ell}|}{|\lambda_r|^{-2s}(pn)^{1/2}}\right) \\
&\leq \exp\left(-\frac{1}{3}\left(\frac{p'|\mathcal{S}_{u,s}||\mathcal{S}_{v,s+\ell}|}{|\lambda_r|^{-2s}(pn)^{1/2}} - 2p'|\mathcal{S}_{v,s+\ell}|\right)\right) \\
&= \exp\left(-\frac{2}{3}p'|\mathcal{S}_{v,s+\ell}|\left(\frac{|\mathcal{S}_{u,s}|}{2|\lambda_r|^{-2s}(pn)^{1/2}} - 1\right)\right) \\
&\leq \exp\left(-\frac{2}{3}p'\left(\frac{(1-\delta)np}{4}\right)^{s+\ell}\left(\frac{(1-\delta)^{1/2}}{4|\lambda_r|^{-1}}\left(\frac{(1-\delta)np}{4|\lambda_r|^{-2}}\right)^{s-\frac{1}{2}} - 1\right)\right) \\
&= \exp\left(-\Theta\left(p'\left(\frac{(1-\delta)np}{4|\lambda_r|^{-1}}\right)^{2s+\ell-\frac{1}{2}}\right)\right)
\end{aligned}$$

It follows that conditioned on event $A'(u,v,s,\ell)(x)$, with probability at least $1-\exp\left(-\Theta\left(p'\left(\frac{(1-\delta)np}{4|\lambda_r|^{-1}}\right)^{2s+\ell-\frac{1}{2}}\right)\right)$,

$$\begin{aligned}
|\frac{1}{p'}\tilde{N}_{u,s}M''_{\mathsf{ind}}\tilde{N}_{v,s+\ell} - \frac{1}{p'}\tilde{N}_{u,s}\big(M'' + M'_{\mathsf{ind}}\big)\tilde{N}_{v,s+\ell}| &\leq \frac{X}{p'|\mathcal{S}_{u,s}||\mathcal{S}_{v,s+\ell}|} \\
&\leq \frac{|\lambda_r|^{2s}}{(pn)^{1/2}}. \qquad (51)
\end{aligned}$$

Next, we prove that with high probability,

$$\big(\tfrac{1}{p'}\big)\big(\tilde{N}_{u,s} - \tilde{N}_{v,s}\big)^T M''_{\mathsf{ind}}\big(\tilde{N}_{u,s+1} - \tilde{N}_{v,s+1}\big) \approx \tilde{N}_{u,s}^T F \tilde{N}_{v,s+\ell}.$$

Let $\mathcal{F}(u,v,s,\ell,x)$ denote all the information related to $\mathcal{T}_u^s$ and $\mathcal{T}_v^{s+\ell}$, including the node latent parameters and observations in M' that are associated to edges in $\mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}$. Furthermore, let $\mathcal{F}(u,v,s,\ell,x)$ be conditioned on the event that $A'(u,v,s,\ell)(x)$ holds, which is fully determined by the realization of edges and weights in $\mathcal{T}_u^s$ and $\mathcal{T}_v^{s+\ell}$. We establish concentration of $N_{u,s}^T M''_{\mathsf{ind}} N_{v,s+\ell}$ by showing that the expression can be written as a sum of independent random variables conditioned on $\mathcal{F}(u,v,s,\ell,x)$,

$$N_{u,s}^T M''_{\mathsf{ind}} N_{v,s+\ell} = \sum_{i,j} \mathbb{I}((i,j) \in \mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}}) M''_{\mathsf{ind}}(i,j) N_{u,s}(i) N_{v,s+\ell}(j),$$

where each term of the summation is bounded in $[0,1]$ due to the fact that all observed entries are bounded in $[0,1]$. Let

$$\phi(i,j) = \mathbb{I}((i,j) \in \mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}}) M''_{\mathsf{ind}}(i,j) N_{u,s}(i) N_{v,s+\ell}(j).$$

By construction, $\{\phi(i,j)\}_{(i,j)\in[n]^2}$ are independent random variables conditioned on $\mathcal{F}(u,v,s,\ell,x)$, because $N_{u,s}$ and $N_{v,s+\ell}$ are measurable with respect to $\mathcal{F}(u,v,s,\ell,x)$, and conditioned on $\mathcal{E}'$, $\mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}}$ is distributed according to the Bernoulli(p') sampling model, and the corresponding observations in M''_{ind} are constructed to be independent due to resampling observations $Z_{\mathsf{ind}}(i,j)$ for

$(i,j) \in \mathcal{T}_u^s \cup \mathcal{T}_v^{s+\ell}$. We can verify that

$$\begin{aligned}
&\mathbb{E}[\phi(i,j)|\mathcal{F}(u,v,s,\ell,x)] = p'F(i,j)N_{u,s}(i)N_{v,s+\ell}(j), \quad \text{and} \\
&\mathrm{Var}[\phi(i,j)|\mathcal{F}(u,v,s,\ell,x)] \\
&\quad = (N_{u,s}(i)N_{v,s+\ell}(j))^2\mathbb{E}[\mathbb{I}((i,j) \in \mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}})M''_{\mathsf{ind}}(i,j)^2 \mid \mathcal{F}(u,v,s,\ell,x)] \\
&\quad \overset{(a)}{\leq} N_{u,s}(i)N_{v,s+\ell}(j)\mathbb{E}[\mathbb{I}((i,j) \in \mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}})M''_{\mathsf{ind}}(i,j) \mid \mathcal{F}(u,v,s,\ell,x)] \\
&\quad \leq p'N_{u,s}(i)N_{v,s+\ell}(j)F(i,j)
\end{aligned}$$

where inequality (a) follows from the assumption that observed entries are within $[0,1]$. Therefore,

$$\mathbb{E}[N_{u,s}^T M''_{\mathsf{ind}} N_{v,s+\ell}|\mathcal{F}(u,v,s,\ell,x)] = p'N_{u,s}^T F N_{v,s+\ell}, \tag{52}$$

and

$$\begin{aligned}
\mathrm{Var}[N_{u,s}^T M''_{\mathsf{ind}} N_{v,s+\ell}|\mathcal{F}(u,v,s,\ell,x)] &\leq p'N_{u,s}^T F N_{v,s+\ell} \\
&\leq 4p'|\mathcal{S}_{u,s}||\mathcal{S}_{v,s+\ell}|B^2\Big(\sum_{k=1}^r |\lambda_k|^{2s+\ell+1}\Big).
\end{aligned} \tag{53}$$

The last inequality follows from Lemma 9.3. By an application of Bernstein's inequality, for $z \leq 4B^2\Big(\sum_{k=1}^r |\lambda_k|^{2s+\ell+1}\Big)$,

$$\begin{aligned}
&\mathbb{P}\left(\Big|\frac{1}{p'}\tilde{N}_{u,s}M''_{\mathsf{ind}}\tilde{N}_{v,s+\ell} - \tilde{N}_{u,s}F\tilde{N}_{v,s+\ell}\Big| > z \mid \mathcal{F}(u,v,s,\ell,x)\right) \\
&= \mathbb{P}\left(\big|N_{u,s}^T M''_{\mathsf{ind}} N_{v,s+\ell} - p'N_{u,s}^T F N_{v,s+\ell}\big| > p'|\mathcal{S}_{u,s}||\mathcal{S}_{v,s+\ell}|z \mid \mathcal{F}(u,v,s,\ell,x)\right) \\
&\quad \leq 2\exp\left(-\min\left(\frac{z^2p'|\mathcal{S}_{u,s}||\mathcal{S}_{v,s+\ell}|}{8B^2\Big(\sum_{k=1}^r |\lambda_k|^{2s+\ell+1}\Big)}, \frac{zp'|\mathcal{S}_{u,s}||\mathcal{S}_{v,s+\ell}|}{2}\right)\right) \\
&\quad \leq 2\exp\left(-\frac{p'|\mathcal{S}_{u,s}||\mathcal{S}_{v,s+\ell}|z^2}{8B^2\Big(\sum_{k=1}^r |\lambda_k|^{2s+\ell+1}\Big)|}\right).
\end{aligned}$$

Conditioned on the event $A'(u,v,s,\ell)(x)$, $|\mathcal{S}_{u,s}|$ and $|\mathcal{S}_{v,s+\ell}|$ are lower bounded by $((1-\delta)np/4)^s$ and $((1-\delta)np/4)^{s+\ell}$. By reparametrizing $z \to z\sqrt{\frac{\sum_{k=1}^r |\lambda_k|^{2s+\ell+1}}{p'((1-\delta)np/4)^{2s+\ell}}}$, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\Big|\frac{1}{p'}\tilde{N}_{u,s}M''_{\mathsf{ind}}\tilde{N}_{v,s+\ell} - \tilde{N}_{u,s}F\tilde{N}_{v,s+\ell}\Big| > z\sqrt{\frac{\sum_{k=1}^r |\lambda_k|^{2s+\ell+1}}{p'((1-\delta)np/4)^{2s+\ell}}} \;\middle|\; \mathcal{F}(u,v,s,\ell,x)\right) \\
&\leq 2\exp\left(-\frac{z^2}{8B^2}\right),
\end{aligned}$$

for $0 < z \leq 4B^2\sqrt{\Big(\sum_{k=1}^r |\lambda_k|^{2s+\ell+1}\Big) \times p'((1-\delta)np/4)^{2s+\ell}}$. The final step in the proof is to combine the above probability bound with the inequality stated in (51). □

Define event

$$\begin{aligned}
\mathcal{A}^3_{u,v,s,\ell}(z,\delta) = \Bigg\{ &\Big|\frac{1}{p'}\tilde{N}_{u,s}\big(M'' + M'_{\mathsf{ind}}\big)\tilde{N}_{v,s+\ell} - \tilde{N}_{u,s}F\tilde{N}_{v,s+\ell}\Big| \leq \\
&\frac{|\lambda_r|^{2s}}{(pn)^{1/2}} + z\sqrt{\frac{\sum_{k=1}^r |\lambda_k|^{2s+\ell+1}}{p'((1-\delta)np/4)^{2s+\ell}}}\Bigg\}.
\end{aligned} \tag{54}$$

9.5 Proof of Lemma 8.1

By statement of Lemma 8.1, we have $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$ with $p = n^{-1+\kappa}$ where $1/\kappa$ is not an integer. We wish to establish that distance $\hat{d}$, as defined in (7) is a good proxy of distance d as defined in (17). We shall establish this result under event $\mathcal{A}$ where

$$\mathcal{A} = \mathcal{A}^1(0.1) \cap \mathcal{A}^2(n^{\rho/2}, 0.1) \cap \mathcal{A}^3(n^{\rho/2}, 0.1), \tag{55}$$

where

$$\begin{aligned}
\mathcal{A}^3(n^{\rho/2}, 0.1) &= \cap_{u,v\in[n]} \mathcal{A}^3_{u,v,t,1}(n^{\rho/2}, 0.1), \\
\mathcal{A}^2(n^{\rho/2}, 0.1) &= \cap_{u\in[n]} \cap_{k\in[r]} \left(\mathcal{A}^2_{u,k,0,t}(n^{\rho/2}, 0.1) \ \cap \ \mathcal{A}^2_{u,k,0,t+1}(n^{\rho/2}, 0.1)\right), \\
\mathcal{A}^1(0.1) &= \cap_{u\in[n]} \cap_{s=1}^{t+1} \mathcal{A}^1_{u,s}(0.1).
\end{aligned}$$

We shall use Lemmas 9.1, 9.2, 9.3 and 9.4 to conclude the desired result. To that end, we verify that appropriate conditions required in the statement of these Lemmas are satisfied.

A crucial condition is that $t + 1 \leq s^*(n, p, \delta)$ originally imposed by Lemma 9.1. By definition of $s^*(n, p, \delta)$, it is sufficient to establish that

$$\frac{p}{8}\left(\frac{(1+\delta)np}{4}\right)^t \leq \phi(\delta) \tag{56}$$

where recall $\phi(\delta) = 1 - \left(\frac{1-\delta}{1-\delta\sqrt{2/3}}\right)^{1/2}$. We shall fix $\delta = 0.1$ for the convenience through the remainder of the proof. To that end, it can be checked that $\phi(0.1) > 0.01$. Therefore, it is sufficient to have

$$t \leq \frac{\ln(0.08/p)}{\ln(0.275np)} \ < \ \frac{\ln(8\phi(0.1)/p)}{\ln(0.275np)}.$$

We have chosen $t = \lfloor \frac{\ln(1/p)}{\ln(np)} \rfloor$. That is,

$$t = \left\lfloor \frac{(1-\kappa)\ln n}{\kappa \ln n} \right\rfloor \ = \ \left\lfloor \frac{(1-\kappa)}{\kappa} \right\rfloor \ < \ \frac{1-\kappa}{\kappa},$$

since $1/\kappa$ is not an integer. And,

$$\begin{aligned}
\frac{\ln(8\phi(0.1)/p)}{\ln(0.275np)} \geq \frac{\ln 0.08 + (1-\kappa)\ln n}{\ln 0.275 + \kappa \ln n} \ &\to \ \frac{1-\kappa}{\kappa} \\
&> \left\lfloor \frac{(1-\kappa)}{\kappa} \right\rfloor \ = \ t.
\end{aligned}$$

for n large enough. That is, for all n large enough, $t+1 \leq s^*(n, p, 0.1)$. Since $1/\kappa$ is not an integer, for some $\gamma \in (0, 1)$

$$t = \left\lfloor \frac{(1-\kappa)}{\kappa} \right\rfloor \ = \ \frac{1-\kappa}{\kappa} - \gamma.$$

That is,

$$\kappa(t+2) - 1 = \kappa(1-\gamma) > 0. \tag{57}$$

For $\rho \in (0, \kappa)$, we use $x = n^{\rho/2}$ in statement of Lemmas 9.2, 9.3 and 9.4, and $z = n^{\rho/2}$ in statement of Lemma 9.4. We need to verify condition on x and z. Note that $\delta, B, |\lambda_k|, r, t$ are all constant with respect to n. Lemma 9.2 requires

$$x < \frac{2((1-\delta)np/4)^{1/2}}{B|\lambda_k|(1+|\lambda_k|)} = \Theta((np)^{1/2})$$

and Lemma 9.3 requires

$$x < B((1-\delta)np/4)^{1/2} = \Theta((np)^{1/2}).$$

Since $np = n^{\kappa}$ and $x = n^{\rho/2}$ with $\rho < \kappa$, both of the above conditions are satisfied for sufficiently large n. For Lemma 9.4, we require

$$z < 4B^2\big(p'((1-\delta)np/4)^{2t+1} \times (\sum_{k=1}^{r} |\lambda_k|^{2t+2})\big)^{1/2}\big) = \Theta((p'(np)^{2t+1})^{1/2}).$$

Now $p'(np/4)^{2t+1} = \Theta(n^{2\kappa(t+1)-1})$. By (57), $2\kappa(t+1) - 1 = \kappa(t+2) - 1 + \kappa t > \kappa t \geq \kappa$. By choice, $z = n^{\rho/2}$ for $\rho < \kappa \leq 2\kappa(t+1) - 1$. Therefore, for sufficiently large n, the above condition is also satisfied.

Now we are ready to bound the difference between $d(u, v)$ and $\hat{d}(u, v)$ for any $u, v \in [n]$. Recall,

$$\begin{aligned} d(\theta_u, \theta_v) = \|\Lambda^{t+1}Q(e_u - e_v)\|^2 \ &= \ (e_u - e_v)^T Q^T \Lambda^{2t+2} Q(e_u - e_v) \\ &= e_u^T Q^T \Lambda^{2t+2} Q e_u + e_v^T Q^T \Lambda^{2t+2} Q e_v - e_u^T Q^T \Lambda^{2t+2} Q e_v - e_v^T Q^T \Lambda^{2t+2} Q e_u. \end{aligned} \tag{58}$$

Recall, that according to (7),

$$\begin{aligned} \hat{d}(u, v) &= \big(\tfrac{1}{p'}\big)\big(\tilde{N}_{u,t} - \tilde{N}_{v,t}\big)^T (M'' + M'_{\mathsf{ind}})\big(\tilde{N}_{u,t+1} - \tilde{N}_{v,t+1}\big), \\ &= \frac{1}{p'}\tilde{N}_{u,t}^T (M'' + M'_{\mathsf{ind}})\tilde{N}_{u,t+1} + \frac{1}{p'}\tilde{N}_{v,t}^T (M'' + M'_{\mathsf{ind}})\tilde{N}_{v,t+1} \\ &\quad - \frac{1}{p'}\tilde{N}_{u,t}^T (M'' + M'_{\mathsf{ind}})\tilde{N}_{v,t+1} - \frac{1}{p'}\tilde{N}_{v,t}^T (M'' + M'_{\mathsf{ind}})\tilde{N}_{u,t+1}. \end{aligned} \tag{59}$$

Under event $\mathcal{A}$ as defined in (55), by Lemmas 9.3 and 9.4,

$$\begin{aligned} &\big|\frac{1}{p'}\tilde{N}_{u,t}^T (M'' + M'_{\mathsf{ind}})\tilde{N}_{u,t+1} - e_u^T Q^T \Lambda^{2t+2} Q e_u\big| \\ &\leq \frac{3Bx}{((1-\delta)np/4)^{1/2}}\Big(\sum_{k=1}^{r} |\lambda_k|^{2t+2}\Big) + \frac{|\lambda_r|^{2t}}{(pn)^{1/2}} + z\sqrt{\frac{\sum_{k=1}^{r} |\lambda_k|^{2t+2}}{p'((1-\delta)np/4)^{2t+1}}} \\ &\leq \frac{3Bn^{\rho/2}}{(0.225np)^{1/2}}\Big(\sum_{k=1}^{r} |\lambda_k|^{2t+2}\Big) + \frac{|\lambda_r|^{2t}}{(pn)^{1/2}} + n^{\rho/2}\sqrt{\frac{\sum_{k=1}^{r} |\lambda_k|^{2t+2}}{p'(0.225np)^{2t+1}}} \\ &= O(Br|\lambda_1|^{2t+2} n^{-(\kappa-\rho)/2}) + O(|\lambda_r|^{2t} n^{-\kappa/2}) + O((r|\lambda_1|^{2t+2})^{1/2} n^{-(2\kappa(t+1)-1-\rho)/2}) \\ &= O\Big(Br|\lambda_1|^{2t+2} n^{-\frac{1}{2}(\kappa-\rho)}\Big), \end{aligned}$$

where the last equality follows from observing that the first term asymptotically dominates with respect to n as $\rho < \kappa \leq 2\kappa(t+1) - 1$. Similarly, all other three terms on the right hand side in (58) and (59) can be bounded by same quantities. Therefore, we conclude that for any $u, v \in [n]$

$$\Big|d(\theta_u, \theta_v) - \hat{d}(u, v)\Big| = O\Big(Br|\lambda_1|^{2/\kappa} n^{-\frac{1}{2}(\kappa-\rho)}\Big), \tag{60}$$

where we used $t < \frac{1-\kappa}{\kappa}$.

To conclude the proof, we need to argue that event $\mathcal{A}$ holds with high enough probability. To that end, through union bound and Lemmas 9.1, 9.2, and 9.4, we have

$$\begin{aligned}\mathbb{P}\left(\neg\mathcal{A}\right) \leq \mathbb{P}\left(\neg\mathcal{A}^3(n^{\rho/2}, 0.1) \mid \mathcal{A}^1(0.1) \cap \mathcal{A}^2(n^{\rho/2}, 0.1)\right) + \\ \mathbb{P}\left(\neg\mathcal{A}^2(n^{\rho/2}, 0.1) \mid \mathcal{A}^1(0.1)\right) + \mathbb{P}\left(\neg\mathcal{A}^1(0.1)\right).\end{aligned}$$

By union bound and Lemma 9.4, we have that

$$\begin{aligned}&\mathbb{P}\left(\neg\mathcal{A}^3(n^{\rho/2}, 0.1) \mid \mathcal{A}^1(0.1) \cap \mathcal{A}^2(n^{\rho/2}, 0.1)\right) \\ &\leq O\left(n^2 \exp\big(-\Theta(n^{\rho})\big) + n^2 \exp\left(-\Theta\left(p'\left(\frac{(1-\delta)np}{4|\lambda_r|^{-1}}\right)^{2t+\frac{1}{2}}\right)\right)\right) \\ &\overset{(a)}{\leq} O\left(n^2 \exp\big(-\Theta(n^{\rho})\big) + n^2 \exp\left(-\Theta\left((np)^{t-\frac{1}{2}}\right)\right)\right) \\ &\leq O\left(n^2 \exp\big(-\Theta(n^{\rho})\big) + n^2 \exp\left(-\Theta\left(n^{\kappa/2}\right)\right)\right).\end{aligned}$$

where the inequality (a) follows from the choice of t, and the fact that δ and t are constant with respect to n. By union bound and Lemma 9.2, we have that

$$\mathbb{P}\left(\neg\mathcal{A}^2(n^{\rho/2}, 0.1) \mid \mathcal{A}^1(0.1)\right) \leq O\big(nr \exp\big(-\Theta(n^{\rho})\big)\big).$$

By union bound and Lemma 9.1, we have that

$$\mathbb{P}\left(\neg\mathcal{A}^1(0.1)\right) \leq O\big(n \exp\big(-\Theta(n^{\kappa})\big)\big).$$

In summary, (60) holds with probability $1 - O\big(n^2 \exp\big(-\Theta(n^{\min(\rho, \kappa(t-\frac{1}{2}))})\big)\big)$. This completes the proof of Lemma 8.1. □

9.6 Concentration in The Sparser Regime

We state consequence of earlier results that will help establish Lemma 8.3.

Lemma 9.5. *Fix $\delta = 0.1$, $p = n^{-1}\ln^{1+\kappa} n$ for some $\kappa > 0$. Let*

$$t = \left\lceil \frac{\ln(0.08/p)}{\ln(0.275np)} - r' \right\rceil.$$

Let $\rho \in (0, \kappa)$. Suppose the events, $\cap_{k=1}^{r}(\mathcal{A}^2_{u,k,0,t}(\ln^{(1+\rho)/2}(n), \delta) \cap \mathcal{A}^2_{v,k,0,t}(\ln^{(1+\rho)/2}(n), \delta))$, $\cap_{k\in[r]}\cap_{\ell=1}^{r'}$ $\mathcal{A}^2_{v,k,t,\ell}(\ln^{(1+\rho)/2}(n), \delta)$, $\cap_{\ell=1}^{r'}\mathcal{A}^3_{u,v,t,\ell}(\ln^{(1+\rho)/2}(n), \delta)$ and $\cap_{s=1}^{t+r'}\big(\mathcal{A}^1_{u,s}(\delta) \cap \mathcal{A}^1_{v,s}(\delta)\big)$ hold. Then,

$$\left|\sum_{k'\in[r']} z_{k'}\big(\tfrac{1}{p'}\big)\tilde{N}^T_{u,t}(M'' + M'_{\mathsf{ind}})\tilde{N}_{v,t+k'} - e_u^T Q^T \Lambda^2 Q e_v\right| \leq c\ln^{-\frac{(\kappa-\rho)}{2}} n$$

for some constant $c = c(\lambda_1, \lambda_r, \lambda_{gap}, r, B)$, independent of n with $\lambda_{gap} = \min_{1\leq s<s'\leq r}|\lambda_s - \lambda_{s'}|$.

Proof of Lemma 9.5. By choice of t, we have that

$$\frac{\ln(0.08/p)}{\ln(0.275np)} - r' \le t \ < \frac{\ln(0.08/p)}{\ln(0.275np)} - r' + 1. \tag{61}$$

We would like to verify that $t + r' \le s^*(\delta, p, n)$ for $\delta = 0.1$. By definition of $s^*(n, p, \delta)$, it is sufficient to establish that

$$\frac{1}{8}p\left(\frac{(1+\delta)np}{4}\right)^{t+r'-1} \le \phi(\delta)$$

where recall $\phi(\delta) = 1 - \left(\frac{1-\delta}{1-\delta\sqrt{2/3}}\right)^{1/2}$. For $\delta = 0.1$, it can be verified that $\phi(0.1) > 0.01$. Therefore, it is sufficient to have

$$t + r' - 1 \le \frac{\ln(0.08/p)}{\ln(0.275np)},$$

which is implied by (61).

For $p = n^{-1}\ln^{1+\kappa} n$, $\ln np = \ln\ln^{1+\kappa} n = (1+\kappa)\ln\ln n$. We choose $\rho \in (0, \kappa)$, which implies $\rho \in (0, \frac{\ln(np)}{\ln\ln n} - 1)$. Throughout the proof, we will denote $x = \ln^{(1+\rho)/2} n = \omega(1)$. It follows that for sufficiently large n,

$$x^2((1-\delta)np/4)^{-1} = 4(1-\delta)^{-1}(\ln n)^{-(\kappa-\rho)} = o(1). \tag{62}$$

Next, we verify properties of z. Recall that z is a vector that satisfies $\Lambda^{2t+2}\tilde{\Lambda}z = \Lambda^2\mathbf{1}$. That is, for any $k \in [r]$,

$$\sum_{k' \in [r']} z_{k'}\lambda_k^{k'-1} = \lambda_k^{-2t}. \tag{63}$$

Therefore,

$$\sum_{k' \in [r']} z_{k'} e_u^T Q^T \Lambda^{2t+k'+1} Q e_v = e_u^T Q^T \Lambda^2 Q e_v. \tag{64}$$

Let L be the $r' \times r'$ diagonal matrix containing only the distinct eigenvalues amongst $\{\lambda_k\}_{k\in[r]}$, such that L_{hh} denotes the h-th distinct eigenvalue. Let $\tilde{L}$ denote the associated $r' \times r'$ Vandermonde matrix containing only the distinct eigenvalues, i.e. if $\tilde{L}_{ab}$ takes the value of the a-th distinct eigenvalue raised to the $(b-1)$-th power. Note that $\Lambda^{2t+2}\tilde{\Lambda}z = \Lambda^2\mathbf{1}$ is satisfied whenever

$$L^{2t+2}\tilde{L}z = L\mathbf{1}$$

is satisfied. Let us define a diagonal matrix D with $D_{bb} = |\lambda_1|^{-(b-1)}$. Therefore the explicit expression for z is given by

$$z = D(\tilde{L}D)^{-1}L^{-2t}\mathbf{1},$$

such that for $\ell \in [r']$,

$$z_\ell = \sum_{h \in [r']} |\lambda_1|^{-(h-1)}(\tilde{L}D)^{-1}_{\ell h} L_{hh}^{-2t}. \tag{65}$$

Theorem 1 of [26] provides bounds on the sum of entries of the inverse of a Vandermonde matrix. It states that for a $N \times N$ Vandermonde matrix V such that $V_{ab} = \lambda_a^{b-1}$, if V^{-1} denotes the inverse of V, then

$$\max_{j\in[N]} \sum_{i\in[N]} |(V^{-1})_{ij}| \leq \max_{j\in[N]} \prod_{i\neq j} \frac{1+|\lambda_i|}{|\lambda_i - \lambda_j|}.$$

Using this result, we obtain

$$\begin{aligned}
\sum_{j\in[r']} \sum_{i\in[r']} |(\tilde{L}D)^{-1}_{ij}| &\leq \sum_{j\in[r']} \prod_{i\neq j} \left(\frac{1+|L_{ii}|/|\lambda_1|}{|L_{ii}-L_{jj}|/|\lambda_1|} \right) \\
&\leq r' \left(\frac{|\lambda_1|+|\lambda_1|}{\min_{i,j}|L_{ii}-L_{jj}|} \right)^{r'-1} \\
&= r' \left(\frac{2|\lambda_1|}{\lambda_{\text{gap}}} \right)^{r'-1}, \qquad (66)
\end{aligned}$$

where λ_{gap} is the minimum gap between eigenvalues only amongst the distinct eigenvalues,

$$\lambda_{\text{gap}} = \min_{i,j} |L_i - L_j| = \min_{i,j:\lambda_i\neq\lambda_j} |\lambda_i - \lambda_j|.$$

Our interest is in bounding

$$\begin{aligned}
&|\textstyle\sum_{k'\in[r']} z_{k'} \left(\frac{1}{p'}\right) \tilde{N}^T_{u,t}(M'' + M'_{\mathsf{ind}})\tilde{N}_{v,t+k'} - e_u^T Q^T \Lambda^2 Q e_v| \\
&\leq |\sum_{k'\in[r']} z_{k'} \left(\left(\tfrac{1}{p'}\right) \tilde{N}^T_{u,t}(M'' + M'_{\mathsf{ind}})\tilde{N}_{v,t+k'} - \tilde{N}^T_{u,t} F \tilde{N}_{v,t+k'} \right)| \qquad (67) \\
&+ |\sum_{k'\in[r']} z_{k'} \left(\tilde{N}^T_{u,t} Q^T \Lambda Q \tilde{N}_{v,t+k'} - \tilde{N}^T_{u,t} Q^T \Lambda^{k'+1} Q \tilde{N}_{v,t} \right)| \qquad (68) \\
&+ |\sum_{k'\in[r']} z_{k'} \left(\tilde{N}^T_{u,t} Q^T \Lambda^{k'+1} Q \tilde{N}_{v,t} - e_u^T Q^T \Lambda^{2t+k'+1} Q e_v \right)| \qquad (69)
\end{aligned}$$

Conditioned on events $\cap_{k=1}^r (\mathcal{A}^2_{u,k,0,t}(x,\delta) \cap \mathcal{A}^2_{v,k,0,t}(x,\delta))$ and given that all conditions of Lemma 9.3 are satisfied, it follows that

$$\begin{aligned}
|(69)| &= \left| \sum_{k\in[r]} \lambda_k^2 \left((e_k^T Q \tilde{N}_{u,t})(e_k^T Q \tilde{N}_{v,t}) - (e_k^T \Lambda^t Q e_u)(e_k^T \Lambda^t Q e_v) \right) \left(\sum_{k'\in[r']} z_{k'} \lambda_k^{k'-1} \right) \right| \\
&\overset{(a)}{=} \left| \sum_{k\in[r]} \lambda_k^{2-2t} \left(e_k^T Q \tilde{N}_{u,t} - e_k^T \Lambda^t Q e_u \right) \left(e_k^T Q \tilde{N}_{v,t} - e_k^T \Lambda^t Q e_v + e_k^T \Lambda^t Q e_v \right) \right| \\
&\quad + \sum_{k\in[r]} \lambda_k^{2-2t} e_k^T \Lambda^t Q e_u \left(e_k^T Q \tilde{N}_{v,t} - e_k^T \Lambda^t Q e_v \right) \\
&\leq \sum_{k\in[r]} |\lambda_k|^{2-2t} \left(|\lambda_k|^{2t} x^2 \left(\frac{(1-\delta)np}{4} \right)^{-1} + 2B|\lambda_k|^{2t} x \left(\frac{(1-\delta)np}{4} \right)^{-1/2} \right) \\
&\leq x \left(\frac{(1-\delta)np}{4} \right)^{-1/2} \left(x \left(\frac{(1-\delta)np}{4} \right)^{-1/2} + 2B \right) \sum_{k\in[r]} |\lambda_k|^2,
\end{aligned}$$

where (a) follows from (63).

Similarly, conditioned on events $\cap_{k=1}^{r}\cap_{\ell=1}^{r'}(\mathcal{A}^2_{u,k,t,\ell}(x,\delta)\cap\mathcal{A}^2_{v,k,t,\ell}(x,\delta))$ with $x=\ln^{(1+\rho)/2}n$ and $\delta=0.1$, we have

$$\begin{aligned}
|(68)| &\le \sum_{k'\in[r']} z_{k'}\left|\sum_k \lambda_k (e_k^T Q\tilde{N}_{u,t})\left(e_k^T Q\tilde{N}_{v,t+k'} - e_k^T\Lambda^{k'}Q\tilde{N}_{v,t}\right)\right| \\
&\overset{(a)}{\le} \sum_{k'\in[r']} z_{k'} Bx\left(\frac{(1-\delta)np}{4}\right)^{-(t+1)/2}\sum_{k\in[r]}|\lambda_k|^{k'+1} \\
&\overset{(b)}{=} \sum_{k'\in[r']}\sum_{h\in[r']}|\lambda_1|^{-k'+1}(\tilde{L}D)^{-1}_{k'h}L_{hh}^{-2t}Bx\left(\frac{(1-\delta)np}{4}\right)^{-(t+1)/2}\Big(\sum_{k\in[r]}|\lambda_k|^{k'+1}\Big) \\
&\overset{(c)}{\le} |\lambda_1|^2|\lambda_r|^2 Brx\left(\frac{(1-\delta)|\lambda_r|^4np}{4}\right)^{-(t+1)/2}\Big(\sum_{k'\in[r']}\sum_{h\in[r']}(\tilde{L}D)^{-1}_{k'h}\Big) \\
&\overset{(d)}{\le} |\lambda_1|^2|\lambda_r|^2 Brx\left(\frac{(1-\delta)|\lambda_r|^4np}{4}\right)^{-(t+1)/2} r'\left(\frac{2|\lambda_1|}{\lambda_{\rm gap}}\right)^{r'-1},
\end{aligned}$$

where (a) follows from events $\cap_{k=1}^{r}\cap_{\ell=1}^{r'}(\mathcal{A}^2_{u,k,t,\ell}(x,\delta)\cap\mathcal{A}^2_{v,k,t,\ell}(x,\delta))$ and showing that $e_k^TQ\tilde{N}_{u,t}\le B$ due to the boundedness of Q and $\|\tilde{N}_{u,t}\|_1\le 1$ by normalization; (b) follows from (65); (c) follows from $|\lambda_k|\le|\lambda_1|$ and $|L^{-1}_{hh}|\le|\lambda_r|^{-1}$;(d) follows from (66).

Conditioned on the event $\cap_{\ell=1}^{r'}\mathcal{A}^3_{u,v,t,\ell}(\ln^{(1+\rho)/2}(n),\delta)$ and Lemma 9.3, $x=\ln^{(1+\rho)/}2n$ and $\delta=0.1$ it follows that

$$\begin{aligned}
|(67)| &\le \sum_{k'\in[r']} z_{k'}\left(x\left(\frac{\sum_{k=1}^r|\lambda_k|^{2t+k'+1}}{p'((1-\delta)np/4)^{2t+k'}}\right)^{1/2}+\frac{|\lambda_r|^{2t}}{(pn)^{1/2}}\right) \\
&\overset{(a)}{\le} \sum_{k'\in[r']}\sum_{h\in[r']}L_{hh}^{-2t}(\tilde{L}D)^{-1}_{k'h}|\lambda_1|^{-k'+1}\left(x\left(\frac{\sum_{k=1}^r|\lambda_k|^{2t+k'+1}}{p'((1-\delta)np/4)^{2t+k'}}\right)^{1/2}+\frac{|\lambda_r|^{2t}}{(pn)^{1/2}}\right) \\
&\le |\lambda_r|^{-2t}\left(x\left(\frac{r|\lambda_1|^{2t+2}}{p'((1-\delta)np/4)^{2t+1}}\right)^{1/2}+\frac{\max(1,|\lambda_1|^{-r'+1})}{(pn)^{1/2}}\right)\Big(\sum_{k'\in[r']}\sum_{h\in[r']}(\tilde{L}D)^{-1}_{k'h}\Big) \\
&\overset{(b)}{\le} \left(\left(\frac{x^2r|\lambda_r|^2|\lambda_1|}{p'((1-\delta)|\lambda_r|^2|\lambda_1|^{-1}np/4)^{2t+1}}\right)^{1/2}+\frac{\max(1,|\lambda_1|^{-r'+1})}{(pn)^{1/2}}\right)r'\left(\frac{2|\lambda_1|}{\lambda_{\rm gap}}\right)^{r'-1}
\end{aligned}$$

where (a) follows using (65) as well as the fact that $np=\omega(1)$ and hence for n sufficiently large, $((1-\delta)np/4)^{-t}\ge((1-\delta)np/4)^{-t-k'}$ for any $k'\ge 0$; (b) follows using (66).

In summary, we conclude

$$
\begin{aligned}
&|\textstyle\sum_{k'\in[r']} z_{k'}\left(\frac{1}{p'}\right)\tilde{N}_{u,t}^T(M''+M'_{\mathsf{ind}})\tilde{N}_{v,t+k'} - e_u^TQ^T\Lambda^2Qe_v| \\
&\leq \left(\left(\frac{x^2r|\lambda_r|^2|\lambda_1|}{p'((1-\delta)|\lambda_r|^2|\lambda_1|^{-1}np/4)^{2t+1}}\right)^{1/2} + \frac{\max(1,|\lambda_1|^{-r'+1})}{(pn)^{1/2}}\right) r'\left(\frac{2|\lambda_1|}{\lambda_{\text{gap}}}\right)^{r'-1} \qquad (70)\\
&+ |\lambda_1|^2|\lambda_r|^2Brx\left(\frac{(1-\delta)|\lambda_r|^4np}{4}\right)^{-(t+1)/2} r'\left(\frac{2|\lambda_1|}{\lambda_{\text{gap}}}\right)^{r'-1} \qquad (71)\\
&+ x\left(\frac{(1-\delta)np}{4}\right)^{-1/2}\left(x\left(\frac{(1-\delta)np}{4}\right)^{-1/2}+2B\right)\sum_{k\in[r]}|\lambda_k|^2. \qquad (72)
\end{aligned}
$$

Observe that due to (62), $x((1-\delta)np/4)^{-1/2} = o(1)$ and $t = \Theta(\ln n/\ln\ln n) = \omega(1)$, hence there exists some constant $c_1 = c_1(\lambda_1,\lambda_r,\lambda_{\text{gap}},r,B)$, independent of n, such that

$$
\begin{aligned}
&|\mathsf{term}(71)+\mathsf{term}(72)| + \frac{\max(1,|\lambda_1|^{-r'+1})}{(pn)^{1/2}}r'\left(\frac{2|\lambda_1|}{\lambda_{\text{gap}}}\right)^{r'-1}\\
&\leq c_1x(np)^{-\frac{1}{2}}. \qquad (73)
\end{aligned}
$$

Recall that we chose t such that by (61),

$$
\begin{aligned}
\ln(p') &= \ln(p) - \ln(4-p)\\
&= \ln(0.08/(4-p)) - \ln(0.08/p)\\
&\geq \ln(0.08/(4-p)) - (t+r')\ln(0.275np).
\end{aligned}
$$

It follows by $t = \Theta(\frac{\ln(1/p)}{\ln(np)}) = \Theta(\frac{\ln(n)}{\ln\ln n}) = \omega(1)$ that,

$$
\begin{aligned}
&\ln(p'((1-\delta)|\lambda_r|^2|\lambda_1|^{-1}np/4)^{2t})\\
&\geq \ln(0.08/(4-p)) - (t+r')\ln(0.275np) + 2t(\ln(\frac{(1-\delta)|\lambda_r|^2}{4|\lambda_1|}) + \ln np)\\
&= t\ln(np) + \ln(0.08/(4-p)) - r'\ln(0.275np) + t\left(2\ln(\frac{(1-\delta)|\lambda_r|^2}{4|\lambda_1|}) - \ln(0.275)\right)\\
&= \Theta(t\ln(np)) = \Theta(\ln(n)) = \omega(1). \qquad (74)
\end{aligned}
$$

This implies that for some constant $c_2 = c_2(\lambda_1,\lambda_r,\lambda_{\text{gap}},r,B)$, the square of the first term in (70) satisfies

$$
\frac{x^2r|\lambda_r|^2|\lambda_1|}{p'((1-\delta)|\lambda_r|^2|\lambda_1|^{-1}np/4)^{2t+1}}(r')^2\left(\frac{2|\lambda_1|}{\lambda_{\text{gap}}}\right)^{2(r'-1)} \leq c_2x^2(np)^{-1}. \qquad (75)
$$

Putting everything together, we have that for some constant $c = c(\lambda_1,\lambda_r,\lambda_{\text{gap}},r,B)$

$$
|\textstyle\sum_{k'\in[r']} z_{k'}\left(\frac{1}{p'}\right)\tilde{N}_{u,t}^T(M''+M'_{u,v,t,k'}))\tilde{N}_{v,t+k'} - e_u^TQ^T\Lambda^2Qe_v| \leq cx(np)^{-1/2}. \qquad (76)
$$

Replacing $x = \ln^{(1+\rho)/2} n$, we obtain the desired result. □

9.7 Proof of Lemma 8.3

The proof of Lemma 8.3 would follow from Lemma 9.5 and once we verify the probability of events required to hold for Lemma 9.5 to be applicable. To that end, given $\kappa > 0$ so that $p = n^{-1}\ln^{1+\kappa} n$, let $\rho \in (0, \kappa)$ be parameter of choice. We set

$$t = \left\lceil \frac{\ln(0.08/p)}{\ln(0.275np)} - r' \right\rceil.$$

Define event $\mathcal{A}$ where

$$\mathcal{A} = \mathcal{A}^1(0.1) \cap \mathcal{A}^2(\ln^{(1+\rho)/2}(n), 0.1) \cap \mathcal{A}^3(\ln^{(1+\rho)/2}(n), 0.1), \tag{77}$$

where

$$\begin{aligned}
\mathcal{A}^3(\ln^{(1+\rho)/2}(n), 0.1) &= \cap_{u,v\in[n]} \cap_{\ell=1}^{r'} \mathcal{A}^3_{u,v,t,\ell}(\ln^{(1+\rho)/2}(n), 0.1),\\
\mathcal{A}^2(\ln^{(1+\rho)/2}(n), 0.1) &= \cap_{u\in[n]} \cap_{k\in[r]} \mathcal{A}^2_{u,k,0,t}(\ln^{(1+\rho)/2}(n), 0.1)\\
&\qquad \cap_{u\in[n]} \cap_{k\in[r]} \cap_{\ell=1}^{r'} \mathcal{A}^2_{u,k,t,\ell}(\ln^{(1+\rho)/2}(n), 0.1),\\
\mathcal{A}^1(0.1) &= \cap_{u\in[n]} \cap_{s=1}^{t+r'} \mathcal{A}^1_{u,s}(0.1).
\end{aligned}$$

We shall use Lemmas 9.1, 9.2, 9.3 and 9.4 to conclude the desired result. To that end, we verify that appropriate conditions required in the statement of these Lemmas are satisfied.

To argue that $\mathcal{A}^1(0.1)$ holds with high probability, we wish to apply Lemmas 9.1 which requires verifying $t+r' \le s^*(n, p, 0.1)$ which is done in proof of Lemma 9.5. To argue that $\mathcal{A}^2(\ln^{(1+\rho)/2}(n), 0.1)$ and $\mathcal{A}^3(\ln^{(1+\rho)/2}(n), 0.1)$ hold with high probability, we will utilize Lemmas 9.2, 9.3 and 9.4 with $x = \ln^{(1+\rho)/2}(n)$ as well as $z = \ln^{(1+\rho)/2}(n)$ in statement of Lemma 9.4. We need to verify condition on x and z. Lemma 9.2 requires

$$x \le \frac{2((1-\delta)np/4)^{1/2}}{B|\lambda_k|(1+|\lambda_k|)}$$

and Lemma 9.3 requires

$$x \le B((1-\delta)np/4)^{1/2}.$$

For sufficiently large n these conditions are satisfied by our choice of x due to $\rho < \kappa$. For Lemma 9.4, we require

$$z \le 4B^2\big(p'((1-\delta)np/4)^{2t+\ell} \times (\sum_{k=1}^{r} |\lambda_k|^{2t+\ell+1})\big)^{1/2}.$$

Now $z = \ln^{(1+\rho)/2} n$ and $np = \ln^{1+\kappa} n$ and since $\rho < \kappa$ we have that $z = o((np)^{1/2})$. By the same argument as (74) in the proof of Lemma 9.5, $p'((1-\delta)|\lambda_r|np/4)^{2t} = \omega(1)$. As a result, the right hand side of the inequality is $\omega((np)^{\ell/2})$, which implies that for sufficiently large n, the above condition on z is satisfied.

Conditioned on event $\mathcal{A}$, by Lemma 9.5 it follows immediately that for distances defined as per (32) and (8),

$$\max_{u,v\in[n]} |d(\theta_u, \theta_v) - \hat{d}(u, v)| = O\Big(\ln^{-\frac{\kappa-\rho}{2}} n\Big) \;=\; O\left(\sqrt{\frac{\ln^{1+\rho} n}{np}}\right). \tag{78}$$

To conclude the proof, we need to argue that event $\mathcal{A}$ holds with high enough probability. To that end, through union bound and Lemmas 9.1, 9.2, and 9.4, we have

$$\mathbb{P}\left(\neg\mathcal{A}\right) \leq \mathbb{P}\left(\neg\mathcal{A}^3(\ln^{(1+\rho)/2}(n), 0.1) \mid \mathcal{A}^1(0.1) \cap \mathcal{A}^2(\ln^{(1+\rho)/2}(n), 0.1)\right) + \mathbb{P}\left(\neg\mathcal{A}^2(\ln^{(1+\rho)/2}(n), 0.1) \mid \mathcal{A}^1(0.1)\right) + \mathbb{P}\left(\neg\mathcal{A}^1(0.1)\right).$$

By union bound and Lemma 9.4, we have that

$$\mathbb{P}\left(\neg\mathcal{A}^3(\ln^{(1+\rho)/2}(n), 0.1) \mid \mathcal{A}^1(0.1) \cap \mathcal{A}^2(\ln^{(1+\rho)/2}(n), 0.1)\right) \tag{79}$$

$$\leq O\Big(n^2 r' \exp\big(-\Theta(\ln^{1+\rho} n)\big)\Big) + O\Big(n^2 r' \exp\left(-\Theta\left(p'\left(\frac{(1-\delta)np}{4|\lambda_r|^{-1}}\right)^{2t+\frac{1}{2}}\right)\right)\Big). \tag{80}$$

By the choice of t to satisfy (61), it follows that $p(0.275np)^{t+r'} \geq 0.08$. Therefore,

$$\begin{aligned} p'\left(\frac{(1-\delta)np}{4|\lambda_r|^{-1}}\right)^{2t+\ell-\frac{1}{2}} &\geq \frac{p}{4-p}(0.275np)^{t+r'}\left(\frac{(1-\delta)}{1.1|\lambda_r|^{-1}}\right)^{t+r'}\left(\frac{(1-\delta)np}{4|\lambda_r|^{-1}}\right)^{t+\frac{1}{2}-r'} \\ &= \frac{0.08}{4-p}\left(\frac{(1-\delta)}{1.1|\lambda_r|^{-1}}\right)^{2r'-\frac{1}{2}}\left(\frac{(1-\delta)^2np}{4.4|\lambda_r|^{-2}}\right)^{t+\frac{1}{2}-r'} \\ &= \Theta\left(\left(\frac{(1-\delta)^2np}{4.4|\lambda_r|^{-2}}\right)^{t+\frac{1}{2}-r'}\right) \\ &= \Omega(np) = \Theta(\ln^{1+\kappa} n), \end{aligned}$$

where we used the fact that $\delta, |\lambda_r|, r'$ are all constants, while $t = \omega(1)$ and $np = \omega(1)$. By union bound and Lemma 9.2, we have that

$$\mathbb{P}\left(\neg\mathcal{A}^2(\ln^{(1+\rho)/2}(n), 0.1) \mid \mathcal{A}^1(0.1)\right) \leq O\Big(nrr' \exp\big(-\Theta(\ln^{1+\rho} n)\big)\Big). \tag{81}$$

By union bound and Lemma 9.1, we have that

$$\mathbb{P}\left(\neg\mathcal{A}^1(0.1)\right) \leq O\Big(n \exp\big(-\Theta(\ln^{1+\kappa} n)\big)\Big). \tag{82}$$

In summary, the desired claim holds with probability $1 - O\Big(n^2 \exp\big(-\Theta((\ln n)^{1+\rho})\big)\Big)$. This completes the proof of Lemma 8.3. □

10 Proving distance estimate is close when f has ε-approximate rank r

In this section, we extend the result that distance estimate (7) is good approximation of the desired ideal distance as claimed in the statement of Lemma 8.2 when f has ε-approximate rank r. We will primarily establish robustness of the distance estimate with respect to arbitrary, additional error of magnitude at most ε in each observed entry. This will help conclude Lemma 8.2 from Lemma 8.1.

10.1 Robustness of The Quadratic Form In (7)

When f has ε-approximate rank r, the $F = Q^T \Lambda Q + \boldsymbol{\varepsilon}$ with $\|\boldsymbol{\varepsilon}\|_{\max} \leq \varepsilon$. In contrast, when f has rank r, $\boldsymbol{\varepsilon} = 0$, i.e. $F = Q^T \Lambda Q$. That is, the setting of f has ε-approximate rank r can be viewed as a perturbation of the setting with f having rank r: each observation $M(i,j)$ is first generated as per rank r setting and then arbitrary perturbation or adversarial noise ε_{ij} is added to it where $|\varepsilon_{ij}| \leq \varepsilon$. Therefore, we shall analyze the distance estimate as defined in (7) for the setting of f that has ε-approximate rank r by bounding the perturbation (or change) induced in distance estimates for the setting of f that is rank r, due to the addition of such an arbitrary perturbation ε_{ij}.

Lemma 10.1. *Let f have rank r, $\omega(\frac{1}{n}) \leq p \leq o(1)$, $\delta \in (0,1)$, $t \geq 0$ with $t+1 \leq s^*(\delta, p, n)$ and $0 < x \leq B((1-\delta)np/4)^{1/2}$. Let $u, v \in [n]$. As before, define event*

$$A'(u,v,t,1)(x) = \cap_{k=1}^{r} (\mathcal{A}^2_{u,k,0,t}(x) \cap \mathcal{A}^2_{v,k,0,t+1}(x)) \cap \mathcal{A}^1_{u,t} \cap \mathcal{A}^1_{v,t+1}.$$

We condition on the event that $A'(u,v,t,1)(x)$ holds. Let $\hat{d}(u,v)$ be the distance estimate computed according to (7). Upon adding arbitrary $\varepsilon_{ij} \in [-\varepsilon, \varepsilon]$ to $M(i,j)$ for each $(i,j) \in \mathcal{E}$, with probability at least

$$1 - \exp\left(-\Theta\left(p'\left(\frac{(1-\delta)np}{4}\right)^{2t+1}\right)\right),$$

$\hat{d}(u,v)$ changes at most by $O(t\varepsilon(1+\varepsilon)^t + t^2\varepsilon^2(1+\varepsilon)^{2t-1})$.

Proof of Lemma 10.1. Recall that $\hat{d}(u,v)$ is the sum of four quadratic terms (see (59) for example). For each of these terms, we shall argue that it changes by $O(\varepsilon + t\varepsilon(1+\varepsilon)^t + t^2\varepsilon^2(1+\varepsilon)^{2t-1})$ with high probability as claimed. This will conclude the proof. To that end, let us start by considering $\frac{1}{p'}\tilde{N}^T_{u,t}\bar{M}\tilde{N}_{u,t+1}$ where $\bar{M} = M'' + M'_{\mathsf{ind}}$; others follow in a similar manner. Specifically, consider

$$N^T_{u,t}\bar{M}N_{v,t+1} = \sum_{i,j} \mathbb{I}((i,j) \in \mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}})\bar{M}(i,j)N_{u,t}(i)N_{v,t+1}(j).$$

Let $\mathcal{F}(u,v,t,1,x)$ denote all the information related to $\mathcal{T}^t_u$ and $\mathcal{T}^{t+1}_v$, including the node latent parameters and observations in $\bar{M}$ that are associated to edges in $\mathcal{T}^t_u \cup \mathcal{T}^{t+1}_v$. Furthermore, let $\mathcal{F}(u,v,t,1,x)$ be conditioned on the event that $A'(u,v,t,1)(x)$ holds, which is fully determined by the realization of edges and weights in $\mathcal{T}^t_u$ and $\mathcal{T}^{t+1}_v$. We wish to understand how $N^T_{u,t}\bar{M}N_{v,t+1}$ changes if we perturb each entry $M(i,j)$ by adding arbitrary ε_{ij} so that $|\varepsilon_{ij}| \leq \varepsilon$ for all $(i,j) \in \mathcal{E}$. To that end, define

$$\phi(i,j) = \mathbb{I}((i,j) \in \mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}})\bar{M}(i,j)N_{u,t}(i)N_{v,t+1}(j).$$

By construction, $\{\phi(i,j)\}_{(i,j)\in[n]^2}$ in non-zero only if all four terms in its product are. Given $\mathcal{F}(u,v,t,1,x)$ and conditioned on $\mathcal{E}'$, $\mathbb{I}((i,j) \in \mathcal{E}'' \cup \mathcal{E}'_{\mathsf{ind}})$ are i.i.d. Bernoulli(p'). Each $\bar{M}(i,j)$ is perturbed at most by ε. By definition $N_{u,t}(i)$ is a product t terms, each of which takes value in $[0,1]$ and is perturbed by at most ε (in absolute) value. Let $N_{u,t}(i) = \prod_{s=1}^{t} w_s$ with $|w_s| \leq 1$ for all $s \leq t$. Let ε_s be perturbation added to w_s with $|\varepsilon_s| \leq \varepsilon$ for all $s \leq t$. The change in $N_{u,t}(i)$ is

bounded as

$$
\begin{aligned}
\Big|\prod_{s=1}^{t} w_s - \prod_{s=1}^{t}(w_s+\varepsilon_s)\Big| &= \Big|\sum_{S\subset[t]:S\neq\emptyset}\prod_{s\in S}\varepsilon_s\prod_{s\in[t]\setminus S} w_s\Big| \;\le\; \sum_{S\subset[t]:S\neq\emptyset}\prod_{s\in S}|\varepsilon_s|\prod_{s\in[t]\setminus S}|w_s| \\
&\le \sum_{S\subset[t]:S\neq\emptyset}\varepsilon^{|S|} \;=\sum_{s=1}^{t}\binom{t}{s}\varepsilon^s \\
&= \;\varepsilon\Big(\sum_{s=0}^{t-1}\frac{t!}{(t-s-1)!(s+1)!}\varepsilon^s\Big) \\
&\le t\varepsilon\Big(\sum_{s=0}^{t-1}\frac{(t-1)!}{((t-1)-s)!s!}\varepsilon^s\Big) \;=\; t\varepsilon\Big(\sum_{s=0}^{t-1}\binom{t-1}{s}\varepsilon^s\Big) \\
&= t\varepsilon(1+\varepsilon)^{t-1}. \qquad (83)
\end{aligned}
$$

Similarly, the perturbation in $N_{v,t+1}(j)$ can be bounded above by $(t+1)\varepsilon(1+\varepsilon)^t$. That is, the overall perturbation in $\bar{M}(i,j)N_{u,t}(i)N_{v,t+1}(j)$ is bounded above as $O(t\varepsilon(1+\varepsilon)^t + t^2\varepsilon^2(1+\varepsilon)^{2t-1})$ since each of the $\bar{M}(i,j), N_{u,t}(i), N_{v,t+1}(j)$ are $O(1)$. Therefore, the overall perturbation in $N_{u,t}^T\bar{M}N_{v,t+1}$ is bounded above by $O(t\varepsilon(1+\varepsilon)^t + t^2\varepsilon^2(1+\varepsilon)^{2t-1})$ times the number of (i,j) such that $N_{u,t}(i), N_{v,t+1}(j)$ are non-zero and $\mathbb{I}((i,j)\in\mathcal{E}''\cup\mathcal{E}'_{\mathsf{ind}})$. Given $\mathcal{F}(u,v,t,1,x)$, this is precisely Binomial$(|\mathcal{S}_{u,t}||\mathcal{S}_{v,t+1}|,p')$. Therefore, by Chernoff's bound, it follows that $\sum_{i,j\in[n]}\mathbb{I}((i,j)\in\mathcal{E}''\cup\mathcal{E}'_{\mathsf{ind}})$ is at most $2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t+1}|p'$ with probability at least $1-\exp\big(-|\mathcal{S}_{u,t}||\mathcal{S}_{v,t+1}|p'/3\big)$. That is, perturbation in $N_{u,t}^T\bar{M}N_{v,t+1}$ is bounded above by $O(|\mathcal{S}_{u,t}||\mathcal{S}_{v,t+1}|p')\times O(t\varepsilon(1+\varepsilon)^t+t^2\varepsilon^2(1+\varepsilon)^{2t-1})$ with probability at least $1-\exp\big(-|\mathcal{S}_{u,t}||\mathcal{S}_{v,t+1}|p'/3\big)$. Conditioned on the event $A'(u,v,t,1)(x)$, $|\mathcal{S}_{v,s+\ell}|$ are lower bounded by $((1-\delta)np/4)^t$ and $((1-\delta)np/4)^{t+1}$. That is, the above claim holds with probability at least $1-\exp\big(-(1-\delta)np/4)^{2t+1}p'/3\big)$. Recall that $\tilde{N}_{u,t}=N_{u,t}/|\mathcal{S}_{u,t}|$. It follows that the perturbation in $\frac{1}{p'}\tilde{N}_{u,t}^T\bar{M}\tilde{N}_{u,t+1}$ is bounded above by $O(t\varepsilon(1+\varepsilon)^t+t^2\varepsilon^2(1+\varepsilon)^{2t-1})$ with probability at least $1-\exp\big(-(1-\delta)np/4)^{2t+1}p'/3\big)$. Using an identical argument, the same conclusion holds for perturbation induced in the other three terms in distance estimate (7). This completes the proof of Lemma 10.1. □

10.2 Proof of Lemma 8.2

Using Lemma 10.1 and Lemma 8.1, we establish the proof of Lemma 8.2. As argued in the proof of Lemma 8.1, for choice of $t=\lfloor\frac{\ln(1/p)}{\ln(np)}\rfloor$ with $p=n^{-1+\kappa}$ where $1/\kappa$ is not an integer and $\delta=0.1$, we have that $t+1\le s^*(n,p,0.1)$ for n large enough. Further, $np=n^\kappa$ and $p'(np/4)^{2t+1}=\Theta(n^{2\kappa(t+1)-1})$ with $\kappa\le 2\kappa(t+1)-1$. As in Lemma 8.1, we choose $x=n^{\rho/2}$ for $\rho\in(0,\kappa)$ in Lemma 10.1. By this selection, we have $x\le(np/4)^{\frac{1}{2}}$ for n large enough. As in Lemma 8.1, the event $\mathcal{A}$ (recall definition from (55)) holds with probability at least $1-O\big(n^2\exp\big(-\Theta(n^{\min(\rho,\kappa(t-\frac{1}{2}))})\big)\big)$. Indeed, $\mathcal{A}$ implies the condition required for Lemma 10.1 to hold with $x=n^{\rho/2}$ for all $u\neq v\in[n]$. Finally, given this, the conclusion of Lemma 10.1 holds for all $u\neq v\in[n]$ with probability at least $1-\exp\Big(n^2\exp\big(-\Theta(n^\kappa)\big)\Big)$. In summary, from Lemma 10.1 and Lemma 8.1, it follows that

$$
|d(u,v)-\hat{d}(u,v)|\le O\Big(Br|\lambda_1|^{2/\kappa}n^{-\frac{1}{2}(\kappa-\rho)}\Big)+O\Big(t\varepsilon(1+\varepsilon)^t+t^2\varepsilon^2(1+\varepsilon)^{2t-1}\Big), \qquad (84)
$$

holds with probability at least $1-O\big(n^2\exp\big(-\Theta(n^{\min(\rho,\kappa(t-\frac{1}{2}))})\big)\big)$. This completes the proof of Lemma 8.2. □

Acknowledgements

We gratefully acknowledge funding from the NSF under grants CCF-1948256, CNS-1955997, CMMI-1462158, CMMI-1634259, and a TRIPODS phase I project.

References

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, Yiqiao Zhong, et al. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3):1452–1474, 2020.

[2] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015.

[3] Emmanuel Abbe and Colin Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in neural information processing systems*, 2015.

[4] Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *Advances in neural information processing systems*, 2016.

[5] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.

[6] Christian Borgs, Jennifer Chayes, Christina E. Lee, and Devavrat Shah. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4715–4726. Curran Associates, Inc., 2017.

[7] Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*, pages 1369–1377, 2015.

[8] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Shirshendu Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675*, 2015.

[9] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Nina Holden. Sparse exchangeable graphs and their limits via graphon processes. *arXiv preprint arXiv:1601.07134*, 2016.

[10] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *arXiv preprint arXiv:1401.2906*, 2014.

[11] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An L^p theory of sparse graph convergence II: Ld convergence, quotients, and right convergence. *arXiv preprint arXiv:1408.0744*, 2014.

[12] Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008.

[13] Changxiao Cai, Gen Li, Yuejie Chi, H Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: l2,∞ statistical guarantees. *arXiv preprint arXiv:1910.04267*, 2019.

[14] Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2009.

[15] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[16] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

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

[18] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[19] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019.

[20] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[21] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011.

[22] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica*, VII(28):33–61, 2008.

[23] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 2020.

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

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

[26] Walter Gautschi. On inverses of vandermonde and confluent vandermonde matrices. *Numer. Math.*, 4(1):117–123, dec 1962.

[27] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 1992.

[28] Brian Karrer, M. E. J. Newman, and Lenka Zdeborová. Percolation on sparse networks. *Phys. Rev. Lett.*, 113:208702, Nov 2014.

[29] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[30] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

[31] Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *To appear in Annals of Statistics*, 2015.

[32] Olga Klopp, Alexandre B Tsybakov, Nicolas Verzelen, et al. Oracle inequalities for network models and sparse graphon estimation. *Annals of Statistics*, 45(1):316–354, 2017.

[33] Yehuda Koren and Robert Bell. Advances in collaborative filtering. In *Recommender Systems Handbook*, pages 145–186. Springer US, 2011.

[34] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[35] Christina E. Lee, Yihua Li, Devavrat Shah, and Dogyoon Song. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems 29*, pages 2155–2163, 2016.

[36] Y. Li, D. Shah, D. Song, and C. L. Yu. Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 66(3):1760–1784, 2020.

[37] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[38] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Society Providence, 2012.

[39] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354, 2018.

[40] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703, New York, NY, USA, 2014. ACM.

[41] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, Aug 2017.

[42] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[43] Xia Ning, Christian Desrosiers, and George Karypis. *Recommender Systems Handbook*, chapter A Comprehensive Survey of Neighborhood-Based Recommendation Methods, pages 37–76. Springer US, 2015.

[44] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[45] David Steurer and Sam Hopkins. Bayesian estimation from few samples: community detection and related problems. 2017.

[46] 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.

[47] Victor Veitch and Daniel M Roy. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*, 2015.

[48] Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

[49] Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5433–5442, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[50] Jiaming Xu, Laurent Massoulié, and Marc Lelarge. Edge label inference in generalized stochastic block models: from spectral theory to impossibility results. In *Conference on Learning Theory*, pages 903–920, 2014.

[51] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.

[52] Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

A Proof of Extra Lemmas

Lemma A.1. *We use two simple inequalities to argue when a summation is dominated by the single largest term. For any $\rho \geq 2$,*

$$\sum_{s=1}^{r} \rho^s \leq 2\rho^r$$

For any $\rho \geq r^{1/(r-1)}$, it holds that $\rho^s \geq s\rho$ for all $s \leq r$. If additionally $\exp(-a\rho) \leq \frac{1}{2}$,

$$\sum_{s=1}^{r} \exp(-a\rho^s) \leq 2\exp(-a\rho)$$

Recall the definitions of ϕ and s^*,

$$\phi(\delta) = 1 - \left(\frac{1-\delta}{1-\delta\sqrt{2/3}}\right)^{1/2} \; < \; 1. \tag{85}$$

For any $p = \omega\big(\frac{1}{n}\big)$ and $p = o(1)$,

$$s^*(\delta, p, n) = \sup\Big\{s \geq 1 : \frac{p}{8}\left(\frac{(1+\delta)np}{4}\right)^{s-1} \leq \phi(\delta)\Big\}. \tag{86}$$

For any given δ, $s^*(\delta, p, n)$ is well defined for n large enough since $p = o(1)$. Event $\mathcal{A}^1_{u,s}(\delta)$ is defined as

$$\mathcal{A}^1_{u,s}(\delta) := \left\{|\mathcal{S}_{u,s}| \in \left[\left(\frac{(1-\delta)np}{4}\right)^s, \left(\frac{(1+\delta)np}{4}\right)^s\right]\right\}.$$

Lemma A.2. *Let $\omega(\frac{1}{n}) \le p \le o(1)$, $\delta \in (0,1)$. For $1 \le s \le s^*(\delta,p,n)$,*

$$\mathbb{P}\left(\neg\mathcal{A}^1_{u,s}(\delta) \mid \cap_{h=1}^{s-1}\mathcal{A}^1_{u,h}(\delta)\right) \le 2\exp\left(-\frac{\delta^2}{3(1-\delta\sqrt{2/3})}\left(\frac{(1-\delta)np}{4}\right)^s\right).$$

It follows that for $t+\ell \le s^*(\delta,p,n)$,

$$\mathbb{P}\left(\cup_{s=1}^{t+\ell}\neg\mathcal{A}^1_{u,s}(\delta)\right) \le 4\exp\left(-\frac{\delta^2((1-\delta)np)}{12(1-\delta\sqrt{2/3})}\right).$$

Proof of A.2. By definition, $s \le s^*(\delta,p,n)$ implies that

$$\frac{1}{8}p\left(\frac{(1+\delta)np}{4}\right)^{s-1} \le 1-\left(\frac{1-\delta}{1-\delta\sqrt{2/3}}\right)^{1/2} =: \phi(\delta), \tag{87}$$

Let us denote $\mathcal{B}_{u,s-1} = \cup_{h=1}^{s-1}\mathcal{S}_{u,h}$. Conditioned on $\cap_{h=1}^{s-1}\mathcal{A}^1_{u,h}(\delta)$, we can upper bound $|\mathcal{B}_{u,s-1}|$ by

$$|\mathcal{B}_{u,s-1}| = 1+\sum_{h=1}^{s-1}|\mathcal{S}_{u,h}| \;\le\; 1+\sum_{h=1}^{s-1}(\frac{(1+\delta)np}{4})^h \;\le\; 1+2(\frac{(1+\delta)np}{4})^{s-1},$$

where the last step follows from Lemma A.1 showing that the summation is dominated by the largest term for sufficiently large n. By assuming $s \le s^*(\delta,p,n)$, it follows that for sufficiently large n, because $np=\omega(1)$,

$$|\mathcal{B}_{u,s-1}| \le 1+\frac{16\phi(\delta)n}{np} \le \phi(\delta)n.$$

Conditioned on the set $\mathcal{B}_{u,s-1}$ and the set $\mathcal{S}_{u,s-1}$, any vertex $i \in [n]\setminus\mathcal{B}_{u,s-1}$ is in $\mathcal{S}_{u,s}$ independently with probability $(1-(1-\frac{p}{4})^{|\mathcal{S}_{u,s-1}|})$. Thus the number of vertices in $\mathcal{S}_{u,s}$ is distributed as a binomial random variable. By Chernoff's bound,

$$\begin{aligned}
&\mathbb{P}\left(|\mathcal{S}_{u,s}| > (1+\delta)(n-|\mathcal{B}_{u,s-1}|)\left(1-\left(1-\frac{p}{4}\right)^{|\mathcal{S}_{u,s-1}|}\right) \;\middle|\; \mathcal{B}_{u,s-1},\mathcal{S}_{u,s-1},\mathcal{A}^1_{u,s-1}\right)\\
&\le \exp\left(-\frac{1}{3}\delta^2(n-|\mathcal{B}_{u,s-1}|)\left(1-\left(1-\frac{p}{4}\right)^{|\mathcal{S}_{u,s-1}|}\right)\right)\\
&\overset{(a)}{\le} \exp\left(-\frac{1}{3}\delta^2(n-|\mathcal{B}_{u,s-1}|)\left(\frac{p|\mathcal{S}_{u,s-1}|}{4}\right)\left(1-\frac{1}{8}p|\mathcal{S}_{u,s-1}|\right)\right)\\
&\overset{(b)}{\le} \exp\left(-\frac{1}{12}\delta^2 np(1-\phi(\delta))\left(\frac{(1-\delta)np}{4}\right)^{s-1}(1-\phi(\delta))\right)\\
&= \exp\left(-\frac{1}{3}\delta^2\frac{(1-\phi(\delta))^2}{1-\delta}(\frac{(1-\delta)np}{4})^s\right)\\
&\overset{(c)}{=} \exp\left(-\frac{\delta^2}{3(1-\delta\sqrt{2/3})}\left(\frac{(1-\delta)np}{4}\right)^s\right),
\end{aligned}$$

where inequality (a) follows from $(1-(1-x)^y) \ge xy(1-\frac{1}{2}xy)$ for $x \in (0,1)$ and $y \in \mathbb{Z}_+$, inequality (b) follows from the event $\mathcal{A}^1_{u,s-1}$ and the assumption $s \le s^*(\delta,p,n)$, and equality (c) follows from

the fact that we constructed ϕ such that $(1-\delta\sqrt{2/3})(1-\phi(\delta))^2 = (1-\delta)$. We obtain a lower bound on $|\mathcal{S}_{u,s}|$ by a similar argument using Chernoff's bound,

$$\begin{aligned}
&\mathbb{P}\left(|\mathcal{S}_{u,s}| < (1-\delta\sqrt{2/3})(n-|\mathcal{B}_{u,s-1}|)\left(1-\left(1-\frac{p}{4}\right)^{|\mathcal{S}_{u,s-1}|}\right) \;\middle|\; \mathcal{B}_{u,s-1}, \mathcal{S}_{u,s-1}, \mathcal{A}^1_{u,s-1}\right) \\
&\leq \exp\left(-\frac{1}{2}(\delta\sqrt{2/3})^2(n-|\mathcal{B}_{u,s-1}|)\left(1-\left(1-\frac{p}{4}\right)^{|\mathcal{S}_{u,s-1}|}\right)\right) \\
&\leq \exp\left(-\frac{1}{3}\delta^2(n-|\mathcal{B}_{u,s-1}|)\left(\frac{p|\mathcal{S}_{u,s-1}|}{4}\right)\left(1-\frac{1}{8}p|\mathcal{S}_{u,s-1}|\right)\right) \\
&\leq \exp\left(-\frac{\delta^2}{3(1-\delta\sqrt{2/3})}\left(\frac{(1-\delta)np}{4}\right)^s\right).
\end{aligned}$$

Conditioned on $\mathcal{A}^1_{u,s-1}$, the above two inequalities show that $\mathcal{A}^1_{u,s}$ holds with high probability. The upper bound follows from

$$\begin{aligned}
|\mathcal{S}_{u,s}| &\leq (1+\delta)(n-|\mathcal{B}_{u,s-1}|)\left(1-\left(1-\frac{p}{4}\right)^{|\mathcal{S}_{u,s-1}|}\right) \\
&\leq (1+\delta)\frac{np}{4}|\mathcal{S}_{u,s-1}| \;\leq\; \left(\frac{(1+\delta)np}{4}\right)^s
\end{aligned}$$

and the lower bound follows from

$$\begin{aligned}
|\mathcal{S}_{u,s}| &\geq (1-\delta\sqrt{2/3})(n-|\mathcal{B}_{u,s-1}|)\left(1-\left(1-\frac{p}{4}\right)^{|\mathcal{S}_{u,s-1}|}\right) \\
&\geq (1-\delta\sqrt{2/3})n(1-\phi(\delta))\frac{p|\mathcal{S}_{u,s-1}|}{4}\left(1-\frac{1}{8}p|\mathcal{S}_{u,s-1}|\right) \\
&\geq (1-\delta\sqrt{2/3})\frac{np}{4}(1-\phi(\delta))|\mathcal{S}_{u,s-1}|\left(1-\frac{1}{8}p\left(\frac{(1+\delta)np}{4}\right)^{s-1}\right) \\
&\geq (1-\delta\sqrt{2/3})\frac{np}{4}(1-\phi(\delta))|\mathcal{S}_{u,s-1}|(1-\phi(\delta)) \\
&= (1-\delta\sqrt{2/3})\frac{np}{4}|\mathcal{S}_{u,s-1}|(1-\phi(\delta))^2 \\
&\overset{(b)}{=} \frac{(1-\delta)np}{4}|\mathcal{S}_{u,s-1}| \;\geq\; \left(\frac{(1-\delta)np}{4}\right)^s.
\end{aligned}$$

where equality (b) follows from the fact that we constructed ϕ such that $(1-\delta\sqrt{2/3})(1-\phi(\delta))^2 = (1-\delta)$.

We finally lower bound the probability of event $\cap_{s=1}^{t+\ell}\mathcal{A}^1_{u,s}$, by a repeated application of Chernoff's bound for all $s \in [t+\ell]$,

$$\begin{aligned}
\mathbb{P}\left(\cup_{s=1}^{t+\ell}\neg\mathcal{A}^1_{u,s}(\delta)\right) &= \sum_{s=1}^{t+\ell}\mathbb{P}\left(\neg\mathcal{A}^1_{u,s}(\delta) \mid \cap_{h=1}^{s-1}\mathcal{A}^1_{u,h}(\delta)\right) \\
&\leq \sum_{s=1}^{t+\ell} 2\exp\left(-\frac{\delta^2}{3(1-\delta\sqrt{2/3})}\left(\frac{(1-\delta)np}{4}\right)^s\right) \\
&\overset{(a)}{\leq} 4\exp\left(-\frac{\delta^2((1-\delta)np)}{12(1-\delta\sqrt{2/3})}\right),
\end{aligned}$$

where inequality (a) follows from the assumption that $pn = \omega(1)$ such that the largest term in the summation dominates. □