Efficient Estimation for Random Dot Product Graphs via a One-Step Procedure

Fangzheng Xie
Department of Statistics, Indiana University
Department of Applied Mathematics and Statistics, Johns Hopkins University
and Yanxun Xu*
Department of Applied Mathematics and Statistics, Johns Hopkins University

Abstract

We propose a one-step procedure to estimate the latent positions in random dot product graphs efficiently. Unlike the classical spectral-based methods, the proposed one-step procedure takes advantage of both the low-rank structure of the expected adjacency matrix and the Bernoulli likelihood information of the sampling model simultaneously. We show that for each vertex, the corresponding row of the one-step estimator converges to a multivariate normal distribution after proper scaling and centering up to an orthogonal transformation, with an efficient covariance matrix. The initial estimator for the one-step procedure needs to satisfy the so-called approximate linearization property. The one-step estimator improves the commonly-adopted spectral embedding methods in the following sense: Globally for all vertices, it yields an asymptotic sum of squares error no greater than those of the spectral methods, and locally for each vertex, the asymptotic covariance matrix of the corresponding row of the one-step estimator dominates those of the spectral embeddings in spectra. The usefulness of the proposed one-step procedure is demonstrated via numerical examples and the analysis of a real-world Wikipedia graph dataset.

Keywords: Approximate linearization property; Asymptotic normality; Bernoulli likelihood information; Latent position estimation; Normalized Laplacian.

*Correspondence should be addressed to Yanxun Xu (yanxun.xu@jhu.edu)

1 Introduction

Statistical inference on graph data, an important topic in statistics and machine learning, has been pervasive in a variety of application domains, such as social networks (Young and Scheinerman, 2007; Girvan and Newman, 2002; Wasserman and Faust, 1994), brain connectomics (Priebe et al., 2017; Tang et al., 2019), political science (Ward et al., 2011), computer networks (Neil et al., 2013; Rubin-Delanchy et al., 2016), etc. Due to the high dimensional nature and the complex structure of graph data, classical statistical methods typically begin with finding a low-dimensional representation for the vertices in a graph using a collection of points in some Euclidean space, referred to as *latent positions* of the vertices. These latent positions are further used as features for subsequent inference tasks, such as vertex clustering (Sussman et al., 2012) and classification (Sussman et al., 2014; Tang et al., 2013), regression (Mele et al., 2019), and nonparametric graph testing (Tang et al., 2017b).

Hoff et al. (2002) proposed the latent position graphs to formalize the idea of latent positions: Each vertex i in the graph is assigned a Euclidean vector $\mathbf{x}_i \in \mathbb{R}^d$, and the occurrence of an edge linking vertices i and j is a Bernoulli random variable with success probability $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, where $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ is a symmetric link function. In this work, we study the random dot product graphs (Young and Scheinerman, 2007), a particular class of latent position graphs taking the link function to be the dot product of latent positions: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$. Random dot product graphs are of special interest due to the following two reasons: Firstly, the adjacency matrix of a random dot product graph can be viewed as the sum of a low-rank matrix and a mean-zero noise matrix, which facilitates the use of low-rank matrix factorization techniques for statistical inference; Secondly, random dot product graphs are sufficiently flexible as they can approximate general latent position graphs with symmetric positive definite link functions when the dimension d of the latent positions grows with the number of vertices at a certain rate (Tang et al., 2013). The readers are referred to the survey paper Athreya et al. (2018a) for a thorough review on the recent development of random dot product graphs.

Low-rank matrix factorization methods, or more precisely, spectral-based methods, have

been broadly used for estimating latent positions for random dot product graphs due to the low expected rank of the observed adjacency matrix. Sussman et al. (2014) proposed to estimate latent positions using the eigenvectors associated with the top d eigenvalues of the adjacency matrix. The resulting estimator is referred to as the *adjacency spectral embedding* (ASE). Asymptotic characterization of the global behavior of the ASE for all vertices have been established, including the consistency (Sussman et al., 2014) and the limit of the sum of squares error (Tang et al., 2017a) as the number of vertices goes to infinity. Locally, for each vertex, Athreya et al. (2016) proved that the distribution of the corresponding row of the adjacency spectral embedding converges to a mean-zero multivariate normal mixture distribution after proper scaling and centering, up to an orthogonal transformation, as the number of vertices goes to infinity. Another popular spectral-based method is the *Laplacian spectral embedding* (LSE), which computes the eigenvectors of the normalized Laplacian matrix of the adjacency matrix associated with the top d eigenvalues (Rohe et al., 2011). The asymptotic theory of the LSE has also been established (Sarkar and Bickel, 2015; Tang and Priebe, 2018). Notably, Tang and Priebe (2018) showed that each row of the LSE converges to a mean-zero multivariate normal mixture distribution after proper scaling and centering, up to an orthogonal transformation. These theoretical studies of the spectral-based methods lay a solid foundation for the development of subsequent inference tasks, such as vertex clustering (Sussman et al., 2012; Rohe et al., 2011; Sarkar and Bickel, 2015), vertex classification (Sussman et al., 2014; Tang et al., 2013), testing between graphs (Tang et al., 2017a,b), and parameter estimation in latent structure random graphs (Athreya et al., 2018b).

Despite the great success of the spectral-based methods for random dot product graphs, it has been pointed out in Xie and Xu (2019) that they are formulated in a low-rank matrix factorization fashion, whereas the Bernoulli likelihood information contained in the sampling model has been neglected. A fundamental question remains open: whether or not the adjacency/Laplacian spectral embedding is optimal for estimating latent positions (or the transformation of them) due to the negligence of the the likelihood information? In this paper, we prove the sub-optimality of the ASE by showing that the asymptotic covariance

matrix of each row of the ASE is sub-optimal. We propose a novel one-step procedure for estimating latent positions, and show that for each vertex, the corresponding row of the proposed one-step estimator converges to a multivariate normal distribution after $\sqrt{n}$-scaling and centering at the underlying true latent position, up to an orthogonal transformation. More importantly, the corresponding asymptotic covariance matrix is the same as the maximum likelihood estimator as if the rest of the latent positions are known, provided that the procedure is initialized at an estimator satisfying the approximate linearization property, which will be defined later. This phenomenon of the one-step estimator is referred to as the local efficiency, the formal definition of which is provided in Section 3. In particular, we show that the efficient covariance matrix is no greater than the asymptotic covariance matrix of the corresponding row of the ASE in spectra. We also provide an example where the difference between the efficient covariance matrix and the asymptotic covariance matrix of the ASE has at least one negative eigenvalue. Besides the local efficiency for each vertex, the proposed one-step estimator for latent positions has a smaller sum of squares error than that of the ASE globally for all vertices as well.

The general one-step procedure, which finds a new estimator via a single iteration of the Newton-Raphson update given a $\sqrt{n}$-consistent initial estimator, has been applied to M-estimation theory in classical parametric models to produce an efficient estimator (Van der Vaart, 2000). Even when the maximum likelihood estimator does not exist (*e.g.*, Gaussian mixture models), the one-step estimator could still be efficient. This motivates us to extend the one-step procedure from classical parametric models to efficient estimation in high-dimensional random graphs, because neither the existence nor the uniqueness of the maximum likelihood estimator for random dot product graphs has been established. Unlike the ASE, the proposed one-step procedure takes both the low-rank structure of the mean matrix and the likelihood information of the sampling model into account simultaneously. This work represents, to the best of our knowledge, the first effort in the literature addressing the efficient estimation problem for random dot product graphs.

Moreover, we prove the asymptotic sub-optimality of the widely adopted LSE by applying the one-step procedure to construct an estimator for the population version of the LSE

and showing that it dominates the LSE in the following sense: Locally for each vertex, the corresponding row of the new estimator converges to a mean-zero multivariate normal distribution after proper scaling and centering, up to an orthogonal transformation, and the asymptotic covariance matrix is no greater than that of the corresponding row of the LSE in spectra; Globally for all vertices, it yields a sum of squares error no greater than that of the LSE.

Recently, there has been substantial progress on generalized random dot product graphs (Rubin-Delanchy et al., 2017), which fall into the category of general latent position graphs as well but allow for a more general link function than random dot product graphs. The link function of a generalized random dot product graph is of the form $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\mathrm{T}} \mathbf{I}_{p,q} \mathbf{x}_j$, where $\mathbf{I}_{p,q}$ is a diagonal matrix with p ones and q minus ones on its diagonals and p, q are non-negative integers such that $p+q = d$. This class of random graphs include a broad class of popular network models (*e.g.*, mixed-membership stochastic block models). We remark that the theory and method established in this work can be extended to generalized random dot product graphs as long as p, q are either provided or can be estimated consistently.

The remaining part of the paper is structured as follows. We review the background on random dot product graphs and present the limit theorem for the ASE (modified theorem from Athreya et al., 2016) in Section 2.1. The theory for the maximum likelihood estimation of a single latent position with the rest of the latent positions being known, which motivates us to pursue the efficient estimation task, is established in Section 2.2. Section 3 elaborates on the proposed one-step procedure for estimating the entire latent position matrix, establishes its asymptotic theory, and shows that it dominates the ASE as the number of vertices goes to infinity. In Section 4, we apply the proposed one-step procedure to construct an estimator for the population version of the LSE, and show that it dominates the LSE asymptotically. Section 5 demonstrates the usefulness of the proposed one-step procedure via numerical examples and the analysis of a real-world Wikipedia graph data. We conclude the paper with a discussion in Section 6.

Notations: The $d \times d$ identity matrix is denoted by $\mathbf{I}_d$ and the vector with all entries being 1 is denoted by the boldface $\mathbf{1}$. We define the notation $[n]$ to be the set of all con-

secutive positive integers from 1 to n: $[n] := \{1, 2, \ldots, n\}$. The symbols $\lesssim$ and $\gtrsim$ mean the corresponding inequality up to a constant, *i.e.*, $a \lesssim b$ ($a \gtrsim b$) if $a \leq Cb$ ($a \geq Cb$) for some constant $C > 0$, and we denote $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. The shorthand notation $a \vee b$ denotes the maximum value between a and b, namely, $a \vee b = \max(a, b)$ for any $a, b \in \mathbb{R}$. We use the notation $\mathbb{O}(n, d)$ to denote the set of all orthonormal d-frames in $\mathbb{R}^n$, *i.e.*, $\mathbb{O}(n, d) = \{\mathbf{U} \in \mathbb{R}^{n \times d} : \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_d\}$, where $n \geq d$, and write $\mathbb{O}(d) = \mathbb{O}(d, d)$. The notation $\|\mathbf{x}\|$ is used to denote the Euclidean norm of a vector $\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{T}} \in \mathbb{R}^d$, *i.e.*, $\|\mathbf{x}\| = (\sum_{k=1}^d x_k^2)^{1/2}$. For any two vectors $\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{T}}$ and $\mathbf{y} = [y_1, \ldots, y_d]^{\mathrm{T}}$ in $\mathbb{R}^d$, the inequality $\mathbf{x} \leq \mathbf{y}$ means that $x_k \leq y_k$ for all $k = 1, 2, \ldots, d$. For any two positive semidefinite matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ of the same dimension, the notation $\mathbf{\Sigma}_1 \preceq \mathbf{\Sigma}_2$ ($\mathbf{\Sigma}_1 \succeq \mathbf{\Sigma}_2$) means that $\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1$ ($\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$) is positive semidefinite, and we say that $\mathbf{\Sigma}_1$ is no greater (no less) than $\mathbf{\Sigma}_2$ in spectra. For any rectangular matrix $\mathbf{X}$, we use $\sigma_k(\mathbf{X})$ to denote its kth largest singular value. For a matrix $\mathbf{X} = [x_{ik}]_{n \times d}$, we use $\|\mathbf{X}\|_2$ to denote the spectral norm $\|\mathbf{X}\|_2 = \sigma_1(\mathbf{X})$, $\|\mathbf{X}\|_{\mathrm{F}}$ to denote the Frobenius norm $\|\mathbf{X}\|_{\mathrm{F}} = (\sum_{i=1}^n \sum_{k=1}^d x_{ik}^2)^{1/2}$, and $\|\mathbf{X}\|_{2\to\infty}$ to denote the two-to-infinity norm $\|\mathbf{X}\|_{2\to\infty} = \max_{i \in [n]} (\sum_{k=1}^d x_{ik}^2)^{1/2}$.

2 Preliminaries

2.1 Background on random dot product graphs

Denote $\mathcal{X} = \{\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{T}} \in \mathbb{R}^d : x_1, \ldots, x_d > 0, \|\mathbf{x}\| < 1\}$ the space of latent positions, and $\mathcal{X}^n = \{\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d} : \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}\}$. For any $\delta \in (0, 1/2)$, denote $\mathcal{X}(\delta)$ the set of all $\mathbf{x} \in \mathcal{X}$ such that $\mathbf{x}^{\mathrm{T}}\mathbf{u} \in [\delta, 1 - \delta]$ for all $\mathbf{u} \in \mathcal{X}(\delta)$. Given an $n \times d$ matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathcal{X}^n$ and a sparsity factor $\rho_n \in (0, 1]$, a symmetric and hollow (*i.e.*, the diagonal entries are zeros) random matrix $\mathbf{A} = [A_{ij}]_{n \times n} \in \{0, 1\}^{n \times n}$ is said to be the adjacency matrix of a random dot product graph on n vertices $[n] = \{1, 2, \ldots, n\}$ with latent positions $\mathbf{x}_1, \ldots, \mathbf{x}_n$, denoted by $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X})$, if $A_{ij} \sim \mathrm{Bernoulli}(\rho_n \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)$ independently, $1 \leq i < j \leq n$. We refer to the matrix $\mathbf{X}$ as the latent position matrix. Namely, the distribution of $\mathbf{A}$ can be written as $p_{\mathbf{X}}(\mathbf{A}) = \prod_{i<j} (\rho_n \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^{A_{ij}} (1 - \rho_n \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^{1 - A_{ij}}$. When $\rho_n \equiv 1$ for all n, the resulting graph is dense, in the sense that the expected number of edges

$\mathbb{E}(\sum_{i<j} A_{ij})$ grows quadratically in n, and when $\rho_n \to 0$ as $n \to \infty$, the corresponding graph is sparse, namely, the expected number of edges is sub-quadratic in n ($\mathbb{E}(\sum_{i<j} A_{ij}) = o(n^2)$).

The goal of this work is to estimate the latent positions $\mathbf{x}_1, \ldots, \mathbf{x}_n$, which are treated as deterministic parameters. In some cases, the latent positions $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are considered as latent random variables that are independently sampled from some underlying distribution F on $\mathcal{X}$ (see, for example, Athreya et al., 2016; Sussman et al., 2014; Tang et al., 2017b; Tang and Priebe, 2018). For deterministic latent positions, we require that there exists some cumulative distribution function F on $\mathcal{X}$, such that

$$\sup_{\mathbf{x} \in \mathcal{X}} |F_n(\mathbf{x}) - F(\mathbf{x})| \to 0 \quad \text{as } n \to \infty, \tag{2.1}$$

where $F_n(\mathbf{x}) = (1/n) \sum_{i=1}^{n} \mathbb{1}\{\mathbf{x}_i \leq \mathbf{x}\}$. Condition (2.1) is similar to the case where $\mathbf{x}_i$'s are random in the following sense: When $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent random variables sampled from F, the Glivenko-Cantelli theorem asserts that (2.1) holds with probability one with respect to the randomness of the infinite i.i.d. sequence $(\mathbf{x}_i)_{i=1}^{\infty}$.

Remark 1. *The latent position matrix* $\mathbf{X}$ *can only be identified up to an orthogonal transformation since for any orthogonal matrix* $\mathbf{W} \in \mathbb{O}(d)$ *and* $i, j \in [n]$, $\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j = (\mathbf{W}\mathbf{x}_i)^{\mathrm{T}}(\mathbf{W}\mathbf{x}_j)$. *Furthermore, for any* $d' > d$ *and any latent position matrix* $\mathbf{X} \in \mathbb{R}^{n \times d}$, *there exists another matrix* $\mathbf{X}' \in \mathbb{R}^{n \times d'}$, *such that* RDPG($\mathbf{X}$) *and* RDPG($\mathbf{X}'$) *yield the same distribution of* $\mathbf{A}$. *The latter source of non-identifiability can be avoided for large* n *by requiring the second moment matrix* $\mathbf{\Delta} = \int_{\mathcal{X}} \mathbf{x}\mathbf{x}^{\mathrm{T}} F(\mathrm{d}\mathbf{x})$ *to be non-singular (Tang and Priebe, 2018).*

Random dot product graphs have connections with the simplest Erdős-Rényi models and the popular stochastic block models. When $F(\mathrm{d}x) = \delta_p(\mathrm{d}x)$, the resulting random dot product graph coincides with an Erdős-Rényi graph, with $(A_{ij})_{i<j}$ being independent Bernoulli(p^2) random variables. When $F(\mathrm{d}\mathbf{x}) = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\nu}_k}(\mathrm{d}\mathbf{x})$ for $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_k \in \mathcal{X}$ and $\sum_{k=1}^{K} \pi_k = 1$, there exists a function $\tau : [n] \to [K]$ such that $(1/n) \sum_{i=1}^{n} \mathbb{1}\{\tau(i) = k\} \to \pi_k$ for all $k = 1, 2, \ldots, K$ as $n \to \infty$. Denoting $\mathbf{B} = [B_{kl}]_{K \times K} := [\boldsymbol{\nu}_k^{\mathrm{T}} \boldsymbol{\nu}_l]_{K \times K}$ and $\mathbf{x}_i = \boldsymbol{\nu}_{\tau(i)}$, $i \in [n]$, we see that A_{ij} follows Bernoulli($B_{\tau(i)\tau(j)}$) = Bernoulli($\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j$) for $i < j$ independently, where $i, j \in [n]$. In this case, the random dot product graph RDPG($\mathbf{X}$) with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}}$ becomes a stochastic block model with a positive semidefinite block probability matrix $\mathbf{B}$

and a cluster assignment function τ.

To estimate the latent positions, Sussman et al. (2014) proposed to solve the least squares problem

$$\widehat{\mathbf{X}}^{(\mathrm{ASE})} = \underset{\mathbf{X}\in\mathbb{R}^{n\times d}}{\arg\min} \|\mathbf{A} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\|_{\mathrm{F}}^2. \tag{2.2}$$

The resulting solution $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$ to (2.2) is referred to as the *adjacency spectral embedding* (ASE) of $\mathbf{A}$ into $\mathbb{R}^d$. Note that $\mathbb{E}(\mathbf{A})$ is a positive semidefinite low-rank matrix modulus the diagonal entries and $\|\mathbf{A} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\|_{\mathrm{F}}^2 = \sum_{i=1}^n \sum_{j=1}^n (A_{ij} - \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^2$ is exactly the empirical squared-error loss. Hence the problem (2.2) becomes a naive empirical risk minimization problem if we regard $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$ as an estimator for $\rho_n^{1/2}\mathbf{X}$, and the solution to (2.2) can be conveniently computed (Eckart and Young, 1936): $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$ is the matrix of eigenvectors associated with the top d eigenvalues of $\mathbf{A}$, scaled by the square roots of these eigenvalues.

Sussman et al. (2014) proved that $\widehat{\mathbf{X}}^{(\mathrm{ASE})} = [\widehat{\mathbf{x}}_1^{(\mathrm{ASE})}, \ldots, \widehat{\mathbf{x}}_n^{(\mathrm{ASE})}]^{\mathrm{T}}$ is a consistent estimator for $\rho_n^{1/2}\mathbf{X}$ globally for all vertices: $(1/n)\|\widehat{\mathbf{X}}^{(\mathrm{ASE})}\mathbf{W}_n - \mathbf{X}\|_{\mathrm{F}}^2$ converges to 0 in probability as $n \to \infty$ for a sequence of orthogonal $(\mathbf{W}_n)_{n=1}^{\infty} \subset \mathbb{O}(d)$. Furthermore, for each fixed vertex $i \in [n]$, the asymptotic distribution of $\widehat{\mathbf{x}}_i^{(\mathrm{ASE})}$ after proper scaling and centering has been established (Athreya et al., 2016; Tang and Priebe, 2018) in the case where $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{i.i.d.}}{\sim} F$. The setup in this work is slightly different since we posit that the latent positions are deterministic. To distinguish between an arbitrary element $\mathbf{X} \in \mathcal{X}^n$ and the ground truth, we denote $\mathbf{X}_0$ the true latent position matrix that generates the observed adjacency matrix $\mathbf{A}$.

We modify the limit theorem of the ASE originally presented in Athreya et al. (2016) to accommodate the deterministic setup for $\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}$ and summarize the results in the following theorem. In the current framework, the proof technique for the asymptotic normality of the rows of the ASE is very different from that presented in Athreya et al. (2016) and Tang and Priebe (2018). The proof of Theorem 1 is deferred to Supplementary Material.

Theorem 1. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *with a sparsity factor* ρ_n *and condition* (2.1) *hold for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$. *Suppose either* $\rho_n \equiv 1$ *for all* n *or* $\rho_n \to 0$ *but* $(\log n)^4/(n\rho_n) \to 0$ *as* $n \to \infty$, *and denote* $\rho = \lim_{n\to\infty} \rho_n$. *Let* $\widehat{\mathbf{X}}^{(\mathrm{ASE})} = [\widehat{\mathbf{x}}_1^{(\mathrm{ASE})}, \ldots, \widehat{\mathbf{x}}_n^{(\mathrm{ASE})}]^{\mathrm{T}}$ *be the ASE defined*

by (2.2). *Denote*

$$\mathbf{\Delta} = \int_{\mathcal{X}} \mathbf{x}\mathbf{x}^{\mathrm{T}} F(\mathrm{d}\mathbf{x}), \quad \mathbf{\Sigma}(\mathbf{x}) = \mathbf{\Delta}^{-1} \left[\int_{\mathcal{X}} \left\{ \mathbf{x}_1^{\mathrm{T}}\mathbf{x} \left(1 - \rho \mathbf{x}_1^{\mathrm{T}}\mathbf{x}\right) \right\} \mathbf{x}_1\mathbf{x}_1^{\mathrm{T}} F(\mathrm{d}\mathbf{x}_1) \right] \mathbf{\Delta}^{-1},$$

and assume that $\mathbf{\Delta}$ *and* $\mathbf{\Sigma}(\mathbf{x})$ *are strictly positive definite for all* $\mathbf{x} \in \mathcal{X}$. *Then there exists a sequence of orthogonal matrices* $(\mathbf{W})_{n=1}^{\infty} = (\mathbf{W}_n)_{n=1}^{\infty} \subset \mathbb{O}(d)$, *such that*

$$\|\widehat{\mathbf{X}}^{(\mathrm{ASE})}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 \xrightarrow{a.s.} \int_{\mathcal{X}} \mathrm{tr}\{\mathbf{\Sigma}(\mathbf{x})\} F(\mathrm{d}\mathbf{x}), \tag{2.3}$$

and for any fixed index $i \in [n]$,

$$\sqrt{n}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i^{(\mathrm{ASE})} - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}(\mathbf{x}_{0i})). \tag{2.4}$$

In the rest of the paper, we drop the subscript n in $\mathbf{W}_n$ for notational simplicity and make the convention that the orthogonal alignment matrix $\mathbf{W}$ implicitly depends on n.

2.2 Motivation: Efficiency in estimating a single latent position

Theorem 1 suggests the following two properties of the ASE: Globally for all vertices, $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$ is a consistent estimator for $\rho_n^{1/2}\mathbf{X}_0$; Locally, for each fixed vertex $i \in [n]$, the distribution of the ith row $\widehat{\mathbf{x}}_i^{(\mathrm{ASE})}$ of $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$ after $\sqrt{n}$-scaling and centering at $\rho_n^{1/2}\mathbf{x}_{0i}$, converges to a mean-zero multivariate normal distribution with covariance matrix $\mathbf{\Sigma}(\mathbf{x}_{0i})$, up to a sequence of orthogonal transformations. Nevertheless, it remains open whether the results of Theorem 1 are optimal. In this work, we will propose an estimator $\widehat{\mathbf{X}}$ for $\mathbf{X}_0$ that dominates the ASE asymptotically in the following sense: Globally for all vertices, it yields a smaller asymptotic sum of squares error $\|\widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2$ than (2.3); Locally for each fixed vertex $i \in [n]$, the corresponding row of $\widehat{\mathbf{X}}$, after $\sqrt{n}$-scaling and centering at $\rho_n^{1/2}\mathbf{x}_{0i}$, also converges to a mean-zero multivariate normal distribution, up to a sequence of orthogonal transformations, but the asymptotic covariance matrix is no greater than $\mathbf{\Sigma}(\mathbf{x}_{0i})$ in spectra.

Before elaborating on the estimator for the entire latent position matrix $\mathbf{X}_0$, we begin with the problem of estimating a single latent position $\mathbf{x}_{0i}$ when the rest of the latent positions are known. The theory established herein motivates the development of the proposed efficient estimation procedure. Specifically, for a fixed $i \in [n]$, we estimate $\mathbf{x}_{0i}$ via the maximum likelihood estimator, assuming that the rest of the latent positions $\{\mathbf{x}_{0j} : j \in [n], j \neq i\}$ are

known. For simplicity, we assume that the sparsity factor $\rho_n \equiv 1$ for all n in this subsection. The result is summarized in the following theorem.

Theorem 2. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$ *with* $\rho_n \equiv 1$ *for all* n*, and condition* (2.1) *hold. Suppose that there exists some constant* $\delta > 0$ *such that* $(\mathbf{x}_{0j})_{j=1}^n \subset \mathcal{X}(\delta)$*. Let* $i \in [n]$ *be fixed and consider the problem of estimating* $\mathbf{x}_{0i}$ *where* $\{\mathbf{x}_{0j} : j \in [n], j \neq i\}$ *are known. Further assume that* $\mathbf{x}_{0i}$ *is in the interior of* $\mathcal{X}(\delta)$*, and for any* $\mathbf{x} \in \mathcal{X}(\delta)$*, define* $\mathbf{G}(\mathbf{x}) = \int_{\mathcal{X}} \mathbf{x}_1\mathbf{x}_1^{\mathrm{T}}\{\mathbf{x}^{\mathrm{T}}\mathbf{x}_1(1-\mathbf{x}^{\mathrm{T}}\mathbf{x}_1)\}^{-1}F(\mathrm{d}\mathbf{x}_1)$*. Then the maximum likelihood estimator* $\widehat{\mathbf{x}}_i^{(\mathrm{MLE})} = \arg\max_{\mathbf{x}\in\mathcal{X}(\delta)} \ell_{\mathbf{A}}(\mathbf{x})$ *is consistent for* $\mathbf{x}_{0i}$*, where* $\ell_{\mathbf{A}}(\mathbf{x})$ *is the log-likelihood function:* $\ell_{\mathbf{A}}(\mathbf{x}) = \sum_{j\neq i}\{A_{ij}\log(\mathbf{x}^{\mathrm{T}}\mathbf{x}_{0j}) + (1-A_{ij})\log(1-\mathbf{x}^{\mathrm{T}}\mathbf{x}_{0j})\}$*. Furthermore, the following asymptotic normality holds:*

$$\sqrt{n}(\widehat{\mathbf{x}}_i^{(\mathrm{MLE})} - \mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \mathbf{G}(\mathbf{x}_{0i})^{-1}). \tag{2.5}$$

Furthermore, $\mathbf{\Sigma}(\mathbf{x}) - \mathbf{G}(\mathbf{x})^{-1}$ *is always positive semidefinite for all* $\mathbf{x} \in \mathcal{X}(\delta)$*.*

Remark 2. *Recall that the cumulative distribution function* F *is defined on* $\mathcal{X}$*. Note that under the conditions of Theorem 2,* $(\mathbf{x}_{0j})_{j=1}^n \subset \mathcal{X}(\delta)$ *for a constant* δ *that does not depend on* n*. Therefore the cumulative distribution function* F *can be further restricted to the compact subset* $\mathcal{X}(\delta)$ *of* $\mathcal{X}$*, and* $\mathbf{G}(\mathbf{x})$ *can be written as* $\int_{\mathcal{X}(\delta)} \mathbf{x}_1\mathbf{x}_1^{\mathrm{T}}\{\mathbf{x}^{\mathrm{T}}\mathbf{x}_1(1-\mathbf{x}^{\mathrm{T}}\mathbf{x}_1)\}^{-1}F(\mathrm{d}\mathbf{x}_1)$ *alternatively.*

Remark 3. *Although the definition of* $\mathbf{G}(\mathbf{x})$ *given in Theorem 2 is with regard to the case where* $\rho_n \equiv 1$ *for all* n*, we remark that it can also be generalized to the case where the sparsity factor* $\rho_n \to 0$ *as* $n \to \infty$ *(see equation* (3.4) *in Section 3).*

Although the inequality $\mathbf{\Sigma}(\mathbf{x}) \succeq \mathbf{G}(\mathbf{x})^{-1}$ is not strict, we will present an example where there exists at least one negative eigenvalue of $\mathbf{G}(\mathbf{x}_{0i})^{-1} - \mathbf{\Sigma}_n(\mathbf{x}_{0i})$ in Section 3. The conclusion of this example is that the ASE is *inefficient* for estimating the latent position $\mathbf{x}_{0i}$ for vertex i when the rest of the latent positions are known, in contrast to the efficiency of the maximum likelihood estimator. The notion of efficiency in estimating a single latent position of a random dot product graph model is slightly subtle, as this special case does not belong to the classical (i.i.d.) parametric models. Here we make the convention that the

notion of efficiency is taken in analogy to the case of parametric models. Namely, we say an estimator $\widehat{\mathbf{x}}_i^{(\mathrm{Eff})}$ is asymptotically efficient for estimating a single latent position vector $\mathbf{x}_{0i}$, if $\sqrt{n}(\widehat{\mathbf{x}}_i^{(\mathrm{Eff})} - \mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, \mathbf{G}(\mathbf{x}_{0i})^{-1})$. We will see in Section 3 that when all the latent positions are unknown, we can still construct an estimator $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$, such that for each vertex i, $\sqrt{n}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, \mathbf{G}(\mathbf{x}_{0i})^{-1})$ still holds up to a sequence of orthogonal alignment matrices $(\mathbf{W})_{n=1}^{\infty} = (\mathbf{W}_n)_{n=1}^{\infty} \subset \mathbb{O}(d)$.

3 Efficient Estimation via a One-step Procedure

The inefficiency of the ASE, indicated by $\mathbf{\Sigma}(\mathbf{x}_{0i}) \succeq \mathbf{G}(\mathbf{x}_{0i})^{-1}$, is due to the fact that the ASE is a least squares estimator not depending on the likelihood function of the sampling model. In contrast, the maximum likelihood estimator $\widehat{\mathbf{x}}_i^{(\mathrm{MLE})}$ utilizes the Bernoulli likelihood function, and this is a main factor for the asymptotic efficiency. For estimating the entire latent position matrix $\mathbf{X}$, one strategy that takes advantage of the likelihood information is the maximum likelihood method. Unfortunately, when all latent positions are unknown, random dot product graphs belong to a curved exponential family rather than a canonical exponential family, and neither the existence nor the uniqueness of the maximum likelihood estimator of random dot product graphs has been established. As pointed out in Bickel and Doksum (2015), properties of the maximum likelihood estimator in curved exponential families are harder to develop than the canonical ones. Therefore, we seek another approach to find an estimator that is asymptotically equivalent to the maximum likelihood estimator. Recall that when $\{\mathbf{x}_{0j} : j \in [n], j \neq i\}$ are known, the maximum likelihood estimator for $\mathbf{x}_{0i}$ is a solution to the estimating equation

$$\mathbf{\Psi}_n(\mathbf{x}) := \frac{1}{n}\sum_{j\neq i}^{n} \frac{(A_{ij} - \mathbf{x}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{x}_{0j}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \mathbf{x}^{\mathrm{T}}\mathbf{x}_{0j})} = \mathbf{0}.$$

Then, given an "appropriate" initial guess of the solution $\widetilde{\mathbf{x}}_i$, we can perform a one-step Newton-Raphson update to obtain another estimator $\widehat{\mathbf{x}}_i^{(\mathrm{OS})}$ that is closer to the zero of the

estimating equation $\mathbf{\Psi}_n$ (see, for example, Section 5.7 of Van der Vaart, 2000):

$$\widehat{\mathbf{x}}_i^{(\mathrm{OS})} = \widetilde{\mathbf{x}}_i + \left\{ \frac{1}{n} \sum_{j \neq i}^{n} \frac{\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}}{\widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{x}_{0j}(1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{x}_{0j})} \right\}^{-1} \left\{ \frac{1}{n} \sum_{j \neq i}^{n} \frac{(A_{ij} - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{x}_{0j}}{\widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{x}_{0j}(1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{x}_{0j})} \right\}. \tag{3.1}$$

In the case of estimating $\mathbf{x}_{0i}$ with the rest of the latent positions being known, the requirement for $\widetilde{\mathbf{x}}_i$ is that it is $\sqrt{n}$-consistent for $\mathbf{x}_{0i}$, and the resulting one-step estimator $\widehat{\mathbf{x}}_i^{(\mathrm{OS})}$ is as efficient as the maximum likelihood estimator $\widehat{\mathbf{x}}_i^{(\mathrm{MLE})}$. This result is summarized in the following theorem, which is a variation of Theorem 5.45 of Van der Vaart (2000).

Theorem 3. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$ *with* $\rho_n \equiv 1$ *for all* n*, and assume that the conditions of Theorem 2 hold. Consider the problem of estimating* $\mathbf{x}_{0i}$ *with* $\{\mathbf{x}_{0j} : j \in [n], j \neq i\}$ *being known. Let* $\widetilde{\mathbf{x}}_i$ *be a* $\sqrt{n}$*-consistent estimator of* $\mathbf{x}_{0i}$*,* i.e.*,* $\sqrt{n}(\widetilde{\mathbf{x}}_i - \mathbf{x}_{0i}) = O_{\mathbb{P}_0}(1)$*. Then* $\sqrt{n}(\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}\left(0, \mathbf{G}(\mathbf{x}_{0i})^{-1}\right)$.

The above result motivates us to generalize the one-step estimator (3.1) to the case where the latent positions $\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}$ are all unknown. Let $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d}$ be an initial estimator $\widetilde{\mathbf{X}}$ for $\mathbf{X}_0$. An intuitive choice for generalizing the one-step updating scheme (3.1) to the case of unknown $(\mathbf{x}_{0j})_{j \neq i}$ is to substitute the unknown $\mathbf{x}_{0j}$ by the initial estimator $\widetilde{\mathbf{x}}_j$ for all $j \neq i$ in (3.1). We define the following one-step estimator $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$ for $\mathbf{X}_0$:

$$\widehat{\mathbf{x}}_i = \widetilde{\mathbf{x}}_i + \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}}{\widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j(1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)} \right\}^{-1} \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{(A_{ij} - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)\widetilde{\mathbf{x}}_j}{\widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j(1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)} \right\}, \quad i = 1, 2, \ldots, n. \tag{3.2}$$

In this case, we require that the initial estimator $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]$ satisfies a finer condition than the $\sqrt{n}$-consistency requirement, referred to as the *approximate linearization property*.

Definition 1 (Approximate linearization property)**.** Given $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ with a sparsity factor ρ_n, where $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$, an estimator $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}}$ is said to satisfy the approximate linearization property, if for all n, there exists an orthogonal matrix $\mathbf{W} = \mathbf{W}_n \in \mathbb{O}(d)$ and an $n \times d$ matrix $\widetilde{\mathbf{R}} = [\widetilde{\mathbf{R}}_1, \ldots, \widetilde{\mathbf{R}}_n]^{\mathrm{T}}$ with $\|\widetilde{\mathbf{R}}\|_{\mathrm{F}}^2 = O_{\mathbb{P}_0}((n\rho_n)^{-1}(\log n)^{\omega})$ for some $\omega \geq 0$, such that

$$\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i} = \rho_n^{-1/2} \sum_{j=1}^{n} (A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\boldsymbol{\zeta}_{ij} + \widetilde{\mathbf{R}}_i, \quad i = 1, 2, \ldots, n, \tag{3.3}$$

where $\{\boldsymbol{\zeta}_{ij} : i, j \in [n]\}$ is a collection of vectors in $\mathbb{R}^d$ with $\sup_{i,j \in [n]} \|\boldsymbol{\zeta}_{ij}\| \lesssim 1/n$.

The approximate linearization property describes that the deviation of the estimator $\widetilde{\mathbf{X}}$ from $\mathbf{X}_0$ can be approximately controlled by a linear combination of the centered Bernoulli random variables $(A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j})_{i<j}$. It has been shown in Athreya et al. (2016),Tang and Priebe (2018), and Tang et al. (2017a) that the ASE satisfies the approximate linearization property (3.3) with $\omega = 0$ and $\boldsymbol{\zeta}_{ij}$ being the jth row of $\mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}$, and hence, $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$ can be chosen to be an initial estimator for the one-step procedure in practice. Another initial estimator satisfying the approximate linearization property will be given in Theorem 7 using the Laplacian spectral embedding.

We present the complete procedure for obtaining the one-step estimator (3.2) initialized at the ASE in Algorithm 1.

Algorithm 1 One-step procedure initialized with the ASE

1: **Input:** The adjacency matrix $\mathbf{A} = [A_{ij}]_{n\times n}$ and the embedding dimension d.

2: **Step 1:** Compute the eigen-decomposition of the adjacency matrix:

$$\mathbf{A} = \sum_{i=1}^{n} \widehat{\lambda}_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i^{\mathrm{T}},$$

where $|\widehat{\lambda}_1| \geq |\widehat{\lambda}_2| \geq \ldots \geq |\widehat{\lambda}_n|$, and $\widehat{\mathbf{u}}_i^{\mathrm{T}}\widehat{\mathbf{u}}_j = \mathbb{1}(i=j)$ for all $i, j \in [n]$.

3: **Step 2:** Compute the ASE

$$\widetilde{\mathbf{X}} = \widehat{\mathbf{X}}^{(\mathrm{ASE})} = \sum_{k=1}^{d} |\widehat{\lambda}_k|^{1/2} \widehat{\mathbf{u}}_k$$

and write $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times d}$.

4: **Step 3:** For $i = 1, 2, \ldots, n$, compute

$$\widehat{\mathbf{x}}_i = \widetilde{\mathbf{x}}_i + \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{\widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}}}{\widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j (1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)} \right\}^{-1} \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{(A_{ij} - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)\widetilde{\mathbf{x}}_j}{\widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j (1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)} \right\}.$$

5: **Output:** The one-step estimator $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$.

The notion of efficiency for random dot product graphs becomes less clear when the number of unknown latent positions grows with the number of vertices. This is because in random dot product graphs, the dimension of the parameter space $\mathcal{X}^n$ grows with the number of vertices, and the definition of the efficiency for classical i.i.d. parametric models does not apply. To this end, we introduce the notion of local efficiency for random dot product graphs. The idea is that any row of the estimator $\widehat{\mathbf{X}}$ has the same asymptotic

covariance matrix with that of the maximum likelihood estimator as if the rest of the latent positions are known.

Definition 2 (Local efficiency)**.** Let $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ with a sparsity factor ρ_n for some $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$, $\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n} \in \mathcal{X}(\delta)$ for some $\delta > 0$ that does not depend on n, and either $\rho_n \equiv 1$ or $\rho_n \to 0$. Denote $\rho = \lim_{n\to\infty} \rho_n$. Assume the condition (2.1) holds. An estimator $\widehat{\mathbf{X}}^{(\mathrm{Eff})} = [\widehat{\mathbf{x}}_1^{(\mathrm{Eff})}, \ldots, \widehat{\mathbf{x}}_n^{(\mathrm{Eff})}]^{\mathrm{T}}$ is said to be a locally efficient estimator for $\mathbf{X}_0$, if there exists a sequence of orthogonal alignment matrices $(\mathbf{W})_{n=1}^{\infty} = (\mathbf{W}_n)_{n=1}^{\infty}$, such that for all $i \in [n]$, $\sqrt{n}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i^{(\mathrm{Eff})} - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \mathbf{G}(\mathbf{x}_{0i})^{-1})$, where $\mathbf{G}$ is a matrix-valued function $\mathbf{G} : \mathcal{X}(\delta) \to \mathbb{R}^{d\times d}$ defined by

$$\mathbf{G}(\mathbf{x}) = \int_{\mathcal{X}} \frac{\mathbf{x}_1\mathbf{x}_1^{\mathrm{T}}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}_1(1 - \rho\mathbf{x}^{\mathrm{T}}\mathbf{x}_1)} F(\mathrm{d}\mathbf{x}_1). \tag{3.4}$$

Theorem 4 and Theorem 5 below, which are the main technical results of this paper, establish the asymptotic behavior of the one-step estimator (3.2). In particular, Theorem 5 shows that the one-step estimator $\widehat{\mathbf{X}}$ is locally efficient.

Theorem 4. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *with a sparsity factor* ρ_n *for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$*. Assume that condition* (2.1) *holds, and there exists some constant* $\delta > 0$ *that is independent of* n *such that* $(\mathbf{x}_{0i})_{i=1}^n \subset \mathcal{X}(\delta)$*. Denote* $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$ *the one-step estimator defined by* (3.2) *initialized at an estimator* $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}}$ *that satisfies the approximate linearization property* (3.3)*. Denote* $\mathbf{G}_n(\mathbf{x}) = (1/n)\sum_{j=1}^n \mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\{\mathbf{x}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}^{\mathrm{T}}\mathbf{x}_{0j})\}^{-1}$ *for any* $\mathbf{x} \in \mathcal{X}(\delta)$*. If either* $\rho_n \equiv 1$ *for all* n *or* $\rho_n \to 0$ *but* $(\log n)^{2(1\vee\omega)}/(n\rho_n^5) \to 0$ *as* $n \to \infty$*, then there exists a sequence of orthogonal matrices* $(\mathbf{W})_{n=1}^{\infty} = (\mathbf{W}_n)_{n=1}^{\infty} \subset \mathbb{O}(d)$ *such that*

$$\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i} = \frac{1}{n\sqrt{\rho_n}} \sum_{j=1}^n \frac{(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} \mathbf{G}_n(\mathbf{x}_{0i})^{-1}\mathbf{x}_{0j} + \widehat{\mathbf{R}}_i, \quad i = 1, \ldots, n, \tag{3.5}$$

where $\|\widehat{\mathbf{R}}_i\| = O_{\mathbb{P}_0}(n^{-1}\rho_n^{-5/2}(\log n)^{(1\vee\omega)})$ *and* $\sum_{i=1}^n \|\widehat{\mathbf{R}}_i\|^2 = O_{\mathbb{P}_0}((n\rho_n^5)^{-1}(\log n)^{2(1\vee\omega)})$*.*

Theorem 5. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *with a sparsity factor* ρ_n *for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$*. Assume that the conditions of Theorem 4 hold, and denote* $\rho = \lim_{n\to\infty} \rho_n$*. Let* $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$ *be the one-step estimator* (3.2) *based on an initial estimator* $\widetilde{\mathbf{X}}$ *that satisfies the approximate linearization property. Then there exists a sequence of orthogonal matrices*

$(\mathbf{W})_{n=1}^{\infty} = (\mathbf{W}_n)_{n=1}^{\infty} \subset \mathbb{O}(d)$ *such that as* $n \to \infty$,

$$\left\| \widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0 \right\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}_0} \int_{\mathcal{X}} \mathrm{tr}\left\{ \mathbf{G}(\mathbf{x})^{-1} \right\} F(\mathrm{d}\mathbf{x}), \tag{3.6}$$

and for each fixed $i \in [n]$,

$$\sqrt{n}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \mathbf{G}(\mathbf{x}_{0i})^{-1}), \tag{3.7}$$

where $\mathbf{G}(\mathbf{x})$ *is given by equation* (3.4).

Since we have already shown that $\mathbf{\Sigma}(\mathbf{x}_{0i}) \succeq \mathbf{G}(\mathbf{x}_{0i})^{-1}$ for all $i \in [n]$, it follows that

$$\|\widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 - \|\widehat{\mathbf{X}}^{(\mathrm{ASE})}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}_0} \int_{\mathcal{X}} \mathrm{tr}\{\mathbf{\Sigma}(\mathbf{x}) - \mathbf{G}(\mathbf{x})^{-1}\} F(\mathrm{d}\mathbf{x}) \geq 0,$$

and hence we conclude that the one-step estimator $\widehat{\mathbf{X}}$ improves the ASE $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$ globally for all vertices asymptotically. Furthermore, for every fixed vertex $i \in [n]$, the ith row of the one-step estimator $\widehat{\mathbf{x}}_i$ is locally efficient by definition, and the corresponding asymptotic covariance matrix is no greater than that of the corresponding row of the ASE in spectra.

Remark 4. *Theorem 4 has the following implication: when the graph is dense* ($\rho_n \equiv 1$ *for all* n)*, one can apply the one-step procedure multiple times, and the resulting estimator still satisfies the approximate linearization property and has the same asymptotic behavior as given by Theorem 4. This multi-step updating strategy is of practical interest for more accurate estimation when the sample size is insufficient for asymptotic approximation.*

Proofs sketch for Theorem 4 and Theorem 5. The key to the proofs of Theorem 4 and Theorem 5 is formula (3.5). From here, we can apply the logarithmic Sobolev concentration inequality to (3.5) (see, for example, Section 6.4 in Boucheron et al., 2013) to show that $\|\widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2$ converges in probability to its expectation, which is exactly the quantity on the right-hand side of (3.6). The asymptotic normality (3.7) of $\widehat{\mathbf{x}}_i$ can be obtained by directly applying the Lyapunov's central limit theorem to

$$\frac{1}{\sqrt{n\rho_n}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} \mathbf{G}_n(\mathbf{x}_{0i})^{-1}\mathbf{x}_{0j},$$

which is a sum of independent random variables. For (3.5), by construction of the one-step

estimator (3.2) and a Taylor expansion device, we have,

$$\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i} = \frac{1}{n\sqrt{\rho_n}}\sum_{j=1}^{n}\frac{(A_{ij}-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}\mathbf{G}_n(\mathbf{x}_{0i})^{-1}\mathbf{x}_{0j} + (\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}) + \mathbf{G}_n(\mathbf{x}_{0i})^{-1}\mathbf{R}_{i1} + o_{\mathbb{P}_0}(n^{-1/2}),$$

where $\mathbf{R}_{i1} = -\mathbf{G}_n(\mathbf{x}_{0i})(\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}) + o_{\mathbb{P}_0}(n^{-1/2})$. Thus we obtain that

$$\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i} = \frac{1}{n\sqrt{\rho_n}}\sum_{j=1}^{n}\frac{(A_{ij}-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}\mathbf{G}_n(\mathbf{x}_{0i})^{-1}\mathbf{x}_{0j} + o_{\mathbb{P}_0}\left(n^{-1/2}\right).$$

The detailed technical derivation of(3.5) is deferred to Supplementary Material.

Remark 5. *Theorem 5 asserts that the one-step estimator* $\widehat{\mathbf{X}}$ *dominates the ASE under the density condition* $(n\rho_n^5)^{-1}(\log n)^{2(1\vee\omega)} \to 0$ *as* $n \to \infty$. *When the graph is dense,* i.e., $\rho_n \equiv 1$ *for all* n, *it is easy to show that this condition holds. When* ρ_n^{-1} *is a polynomial of* $\log n$, *indicating that the graph is moderately sparse, this condition still holds. This condition starts to fail when the graph becomes very sparse,* e.g., $\rho_n^{-1} \asymp n^t$ *for some* $t \geq 1/5$, *in which case a broad range of statistical inference tasks become challenging due to the weak signal.*

Remark 6. *Theorem 4 requires that the sparsity factor* ρ_n *is lower bounded by* $n^{-1/5}$ *times a polynomial factor of* $\log n$. *This causes the average expected degree to grow at a polynomial rate of* n, *and the resulting graph is considered as moderately sparse. In contrast, Theorem 1 only requires* ρ_n *to be lower bounded by* n^{-1} *times a polynomial factor of* $\log n$, *and this results in the average expected degree to grow at a polynomial rate of* $\log n$, *which is a sparser regime than that required by Theorem 4. The stronger density assumption that the average expected degree is a polynomial factor of* n *is essential for the proof strategy employed in this work. Nevertheless, we remark that the proof strategy is standard (see, for example, Section 5.7 of Van der Vaart, 2000). In fact, the stronger density assumption stems from the Lipschitz continuity of the Hessian of the average log-likelihood function, which is guaranteed by the continuity of the third derivatives. This is referred to as the* classical conditions *for M-estimators (see, for example, Section 5.6 of Van der Vaart, 2000). Further discussion of the sparsity condition for the one-step estimator (3.2) is provided in Supplementary Material.*

Theorem 5 claims that the asymptotic covariance matrix of any fixed row of the one-step

estimator (3.2) is no greater than that of the ASE in spectra. The following example shows that there exist situations where $\mathbf{G}(\mathbf{x}_{0i})^{-1} - \mathbf{\Sigma}(\mathbf{x}_{0i})$ contains at least one strictly negative eigenvalue. This implies that the one-step estimator dominates the ASE asymptotically.

Example 1. (Two-block stochastic block model) Consider the following two-block stochastic block model, which has also been considered in Tang and Priebe (2018). Let $F = \pi_1\delta_p + \pi_2\delta_q$ be the distribution on $(0,1)$ giving rise to the latent positions $x_{01}, \ldots, x_{0n}$ via (2.1), where $p, q \in (0,1)$ and $p \neq q$. This results in an $n \times n$ adjacency matrix $\mathbf{A}$ drawn from RDPG($\mathbf{X}_0$) with $\mathbf{X}_0 = [x_{01}, \ldots, x_{0n}]^{\mathrm{T}} \in \mathbb{R}^{n\times 1}$. Let $\tau : [n] \to \{1,2\}$ be a cluster assignment function such that $\tau(i) = 1$ if $x_{0i} = p$, $\tau(i) = 2$ if $x_{0i} = q$, and denote

$$\mathbf{B} = \begin{bmatrix} p^2 & pq \\ pq & q^2 \end{bmatrix}.$$

Then the distribution of $\mathbf{A}$ can be also regarded as a stochastic block model with a block probability matrix $\mathbf{B}$ and a cluster assignment function τ. Let $\widehat{\mathbf{X}}^{(\mathrm{ASE})} = [\widehat{\mathbf{x}}_1^{(\mathrm{ASE})}, \ldots, \widehat{\mathbf{x}}_n^{(\mathrm{ASE})}]^{\mathrm{T}}$ be the ASE and $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$ be the one-step estimator satisfying the conditions of Theorem 4. Using formulas (2.4) and (3.7), we obtain:

$$\sqrt{n}(\widehat{\mathbf{x}}_i^{(\mathrm{ASE})} - p) \xrightarrow{\mathcal{L}} \mathrm{N}\left(0, \Sigma(p)\right) \text{ if } x_{0i} = p, \quad \sqrt{n}(\widehat{\mathbf{x}}_i^{(\mathrm{ASE})} - q) \xrightarrow{\mathcal{L}} \mathrm{N}\left(0, \Sigma(q)\right) \text{ if } x_{0i} = q,$$

where $\Sigma(p) = \frac{\pi_1 p^4(1-p^2) + \pi_2 pq^3(1-pq)}{(\pi_1 p^2 + \pi_2 q^2)^2}$, $\Sigma(q) = \frac{\pi_1 p^3 q(1-pq) + \pi_2 q^4(1-q^2)}{(\pi_1 p^2 + \pi_2 q^2)^2}$, and

$$\sqrt{n}(\widehat{\mathbf{x}}_i - p) \xrightarrow{\mathcal{L}} \mathrm{N}\left(0, G(p)^{-1}\right) \text{ if } x_{0i} = p, \quad \sqrt{n}(\widehat{\mathbf{x}}_i - q) \xrightarrow{\mathcal{L}} \mathrm{N}\left(0, G(q)^{-1}\right) \text{ if } x_{0i} = q,$$

where $G(p) = \frac{\pi_1 p^2}{p^2(1-p^2)} + \frac{\pi_2 q^2}{pq(1-pq)}$, $G(q) = \frac{\pi_1 p^2}{pq(1-pq)} + \frac{\pi_2 q^2}{q^2(1-q^2)}$. By Cauchy-Schwartz inequality, we see that $G(p)^{-1} \leq \Sigma(p)$ and $G(q)^{-1} \leq \Sigma(q)$ for all $p, q \in (0,1)$, and in particular, $G(p)^{-1} = \Sigma(p)$ if and only if $q = (1-p^2)/p$, and $G(q)^{-1} = \Sigma(q)$ if and only if $q = (1/2)(\sqrt{p^2+4} - p)$ (recall that $p \neq q$). Namely, the asymptotic variance of the one-step estimator is strictly smaller than that of the ASE for almost every (p,q) pair in $(0,1)^2 \backslash \{(p,q) : p = q\}$. The comparison of variances between the ASE and the one-step estimator is further visualized in Figure 1 through the relative improvements of the variances $\{\Sigma(p) - G(p)^{-1}\}/G(p)^{-1}$ and $\{\Sigma(q) - G(q)^{-1}\}/G(q)^{-1}$ for different values of p and q.

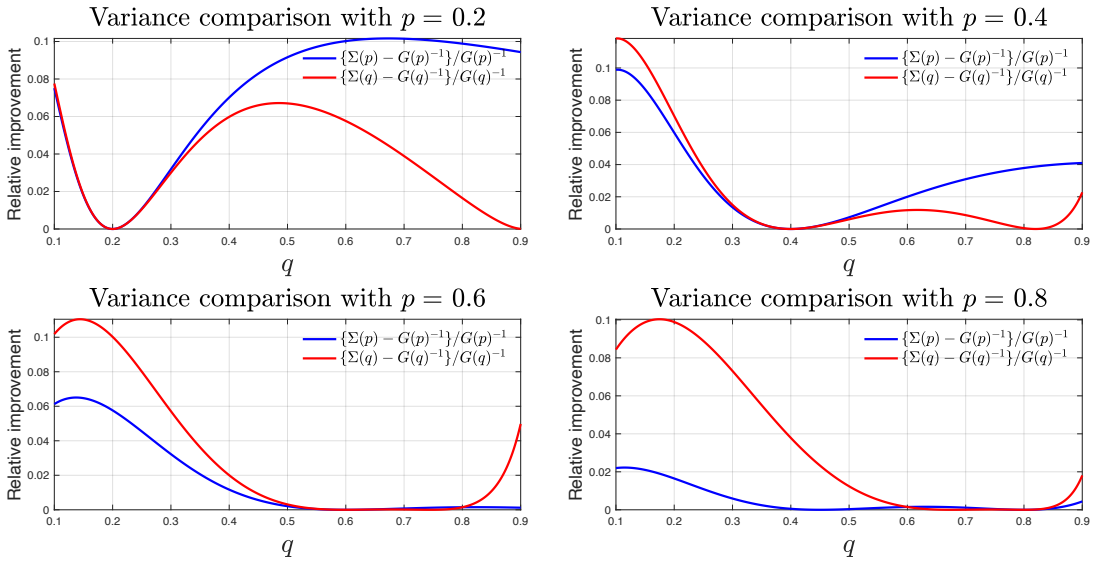

Figure 1: Relative improvements of the one-step estimator variances $\{\Sigma(p)-G(p)^{-1}\}/G(p)^{-1}$ and $\{\Sigma(q) - G(q)^{-1}\}/G(q)^{-1}$ for different values of $p, q \in (0, 1)$ in Example 1. The cluster assignment probabilities are set to $\pi_1 = 0.6$ and $\pi_2 = 0.4$. Note that all the variances $G(p)^{-1}, G(q)^{-1}, \Sigma(p), \Sigma(q)$ depend on both p and q.

4 Application to Estimating the Laplacian Matrix

Instead of directly analyzing the adjacency matrix $\mathbf{A}$, another broadly adopted technique for statistical analysis on random graphs is based on the normalized Laplacian of $\mathbf{A}$ (Rohe et al., 2011; Sarkar and Bickel, 2015). Formally, given a matrix $\mathbf{M}$ with non-negative entries and positive row sums, the normalized Laplacian of $\mathbf{M}$, denoted by $\mathcal{L}(\mathbf{M})$, is defined by $(\mathrm{diag}(\mathbf{M}\mathbf{1}))^{-1/2}\mathbf{M}(\mathrm{diag}(\mathbf{M}\mathbf{1}))^{-1/2}$, Here, for a vector $\mathbf{z} = [z_1, \ldots, z_n]^{\mathrm{T}} \in \mathbb{R}^n$, $\mathrm{diag}(\mathbf{z})$ is the $n \times n$ diagonal matrix with $z_1, \ldots, z_n$ being its diagonal entries. We follow the definition of the normalized Laplacian adopted in Tang and Priebe (2018) in contrast to the combinatorial Laplacian $\mathrm{diag}(\mathbf{M}\mathbf{1}) - \mathbf{M}$ that has been applied to graph theory (Merris, 1994). The (i, j) entry of the normalized Laplacian matrix $\mathcal{L}(\mathbf{A})$ can be interpreted as the connection between vertices i and j normalized by the square roots of the degrees of the two vertices.

Recall that the edge probability matrix $\rho_n\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is positive semidefinite low-rank when $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X})$ with a sparsity factor ρ_n. Similarly, the normalized Laplacian of $\rho_n\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is also a positive semidefinite low-rank matrix: $\mathcal{L}(\rho_n\mathbf{X}\mathbf{X}^{\mathrm{T}}) = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}$, where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^{\mathrm{T}} \in$

$\mathbb{R}^{n\times d}$, and $\mathbf{y}_i = \mathbf{x}_i(\sum_{j=1}^n \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^{-1/2}$. Following the same spirit of the formulation of the ASE through (2.2), one can analogously define the Laplacian spectral embedding (LSE) $\breve{\mathbf{X}}$ of $\mathbf{A}$ into $\mathbb{R}^d$ by solving the least squares problem (Rohe et al., 2011)

$$\breve{\mathbf{X}} = \underset{\mathbf{Y}\in\mathbb{R}^{n\times d}}{\arg\min} \|\mathcal{L}(\mathbf{A}) - \mathbf{Y}\mathbf{Y}^{\mathrm{T}}\|_{\mathrm{F}}^2. \tag{4.1}$$

Since the LSE $\breve{\mathbf{X}}$ is an estimator for $\mathbf{Y}$, we refer to the $n \times d$ matrix $\mathbf{Y}$ as the population LSE. The estimator $\breve{\mathbf{X}}$, which is the LSE of $\mathbf{A}$ into $\mathbb{R}^d$, is also referred to as the sample LSE as opposed to the population LSE $\mathbf{Y}$. Alternatively, the population LSE can be viewed as a transformation $\mathbf{Y} = \mathbf{Y}(\mathbf{X})$ of the latent position matrix $\mathbf{X}$ defined by

$$\mathbf{Y}(\mathbf{X}) = [\mathbf{y}_1(\mathbf{X}), \ldots, \mathbf{y}_n(\mathbf{X})]^{\mathrm{T}}, \quad \mathbf{y}_i = \frac{\mathbf{x}_i}{\sqrt{\sum_{j=1}^n \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j}}, \quad i = 1, \ldots, n. \tag{4.2}$$

The asymptotic results for the (sample) LSE in random dot product graphs with independent and identically distributed latent positions $\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}$ have been established in Tang and Priebe (2018). In the context of the deterministic latent positions framework adopted in this work, we provide the analogous results for the LSE in Theorem 6. The proof is deferred to Supplementary Material.

Theorem 6. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *with a sparsity factor* ρ_n *for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n \subset \mathbb{R}^{n\times d}$, *where* $\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}$ *satisfy* (2.1). *Suppose either* $\rho_n \equiv 1$ *for all* n *or* $\rho_n \to 0$ *but* $(\log n)^4/(n\rho_n) \to 0$ *as* $n \to \infty$, *and denote* $\rho = \lim_{n\to\infty} \rho_n$. *Let* $\breve{\mathbf{X}} = [\breve{\mathbf{x}}_1, \ldots, \breve{\mathbf{x}}_n]^{\mathrm{T}}$ *be the LSE of* $\mathbf{A}$ *into* $\mathbb{R}^d$ *defined by* (4.1). *Define the following quantities:*

$$\mathbf{Y}_0 = \mathbf{Y}(\mathbf{X}_0), \quad \boldsymbol{\mu} = \int_{\mathcal{X}} \mathbf{x}F(\mathrm{d}\mathbf{x}), \quad \widetilde{\boldsymbol{\Delta}} = \int_{\mathcal{X}} \frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{\mathbf{x}^{\mathrm{T}}\boldsymbol{\mu}} F(\mathrm{d}\mathbf{x}),$$
$$\widetilde{\boldsymbol{\Sigma}}(\mathbf{x}) = \left(\widetilde{\boldsymbol{\Delta}}^{-1} - \frac{\mathbf{x}\boldsymbol{\mu}^{\mathrm{T}}}{2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}}\right)\left[\int_{\mathcal{X}} \left\{\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}_1(1-\rho\mathbf{x}^{\mathrm{T}}\mathbf{x}_1)}{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}_1)^2}\mathbf{x}_1\mathbf{x}_1^{\mathrm{T}}\right\} F(\mathrm{d}\mathbf{x}_1)\right]\left(\widetilde{\boldsymbol{\Delta}}^{-1} - \frac{\mathbf{x}\boldsymbol{\mu}^{\mathrm{T}}}{2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}}\right)^{\mathrm{T}}.$$

Then there exists a sequence of orthogonal $(\mathbf{W})_{n=1}^{\infty} = (\mathbf{W}_n)_{n=1}^{\infty} \subset \mathbb{R}^{d\times d}$ *such that as* $n \to \infty$,

$$n\rho_n\|\breve{\mathbf{X}}\mathbf{W} - \mathbf{Y}_0\|_{\mathrm{F}}^2 \xrightarrow{a.s.} \int \mathrm{tr}\{\widetilde{\boldsymbol{\Sigma}}(\mathbf{x})\}F(\mathrm{d}\mathbf{x}). \tag{4.3}$$

Furthermore, assume the graph model falls into one of the following two regimes:

(i) ***Dense regime:*** $\rho_n \equiv 1$ *for all* n;

(ii) ***Sparse stochastic block model regime:*** $\rho_n \to 0$ *with* $(\log n)^4/(n\rho_n) \to 0$ *as* $n \to$

∞, *and there exists* $K \geq d$ *linearly independent* $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K \in \mathcal{X}$ *and a probability vector* $[\pi_1, \ldots, \pi_K]$ *with* $\sum_{k=1}^K \pi_k = 1$, *such that* $F(\mathrm{d}\mathbf{x}) = \sum_{k=1}^K \pi_k \delta_{\boldsymbol{\nu}_k}(\mathrm{d}\mathbf{x})$. *Namely, the random dot product graph coincides with a stochastic block model.*

Then for any fixed $i \in [n]$,

$$n\rho_n^{1/2}(\mathbf{W}^{\mathrm{T}}\breve{\mathbf{x}}_i - \mathbf{y}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}(\mathbf{x}_{0i})). \tag{4.4}$$

The LSE can be applied to construct another initial estimator that satisfies the approximate linearization property. This is given in the following theorem.

Theorem 7. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *with a sparsity factor* ρ_n *for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n \subset \mathbb{R}^{n\times d}$, *where* $\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}$ *satisfy* (2.1). *Suppose either* $\rho_n \equiv 1$ *for all* n *or* $\rho_n \to 0$ *but* $(\log n)^4/(n\rho_n) \to 0$ *as* $n \to \infty$, *and denote* $\rho = \lim_{n\to\infty} \rho_n$. *Let* $\breve{\mathbf{X}}$ *be the LSE of* $\mathbf{A}$ *into* $\mathbb{R}^d$ *defined by* (4.1). *Then the estimator* $\widetilde{\mathbf{X}} = \mathrm{diag}(\sum_{j=1}^n A_{1j}, \ldots, \sum_{j=1}^n A_{nj})^{1/2}\breve{\mathbf{X}}$ *satisfies the approximate linearization property.*

Similar to the ASE, the LSE is also a least squares type estimator and does not involve the likelihood function. Therefore, to estimate the population LSE $\mathbf{Y}_0 = \mathbf{Y}(\mathbf{X}_0)$ using the Bernoulli likelihood information, we propose the following one-step estimator $\widehat{\mathbf{Y}}$ for $\mathbf{Y}_0$ based on the one-step estimator $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$ defined in (3.2) and an initial estimator $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}}$ that satisfies the approximate linearization property (3.3):

$$\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_n]^{\mathrm{T}}, \quad \widehat{\mathbf{y}}_i = \frac{\widehat{\mathbf{x}}_i}{\sqrt{\sum_{j=1}^n \widehat{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j}}, \quad i = 1, 2, \ldots, n. \tag{4.5}$$

In matrix form, we can write $\widehat{\mathbf{Y}} = \{\mathrm{diag}(\widehat{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{1})\}^{-1/2}\widehat{\mathbf{X}}$. The likelihood information is thus absorbed into $\widehat{\mathbf{Y}}$ through the one-step estimator $\widehat{\mathbf{X}}$. We characterize the global and local behavior of the one-step estimator $\widehat{\mathbf{Y}}$ for the population LSE via the following two theorems.

Theorem 8. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *with a sparsity factor* ρ_n *for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$. *Assume that the conditions of Theorem 4 hold. Denote* $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_n]^{\mathrm{T}}$ *the one-step estimator for the population LSE defined by* (4.5), *and* $\boldsymbol{\mu}_n = (1/n)\sum_{i=1}^n \mathbf{x}_{0i}$. *Then there exists a sequence of orthogonal matrices* $(\mathbf{W})_{n=1}^{\infty} = (\mathbf{W}_n)_{n=1}^{\infty} \subset \mathbb{O}(d)$ *such that*

$$\sqrt{n}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{y}}_i - \mathbf{y}_{0i}) = \rho_n^{-1/2}\frac{1}{\sqrt{\boldsymbol{\mu}_n^{\mathrm{T}}\mathbf{x}_{0i}}}\left(\mathbf{I}_d - \frac{\mathbf{x}_{0i}\boldsymbol{\mu}_n^{\mathrm{T}}}{2\boldsymbol{\mu}_n^{\mathrm{T}}\mathbf{x}_{0i}}\right)(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}) + \mathbf{R}_i^{(\mathrm{L})}, \quad i = 1, 2, \ldots, n,$$

where $\|\mathbf{R}_i^{(\mathrm{L})}\| = O_{\mathbb{P}_0}((n\rho_n^2)^{-1}(\log n)^{1\vee\omega})$ *and* $\sum_{i=1}^n \|\mathbf{R}_i^{(\mathrm{L})}\|^2 = O_{\mathbb{P}_0}\left((n\rho_n^4)^{-1}(\log n)^{2(1\vee\omega)}\right)$.

Theorem 9. *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ *with a sparsity factor* ρ_n *for some* $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}} \in \mathcal{X}^n$. *Assume the conditions of Theorem 8 hold. Denote* $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_n]^{\mathrm{T}}$ *the one-step estimator for the population LSE defined by (4.5), and*

$$\widetilde{\mathbf{G}}(\mathbf{x}) = \frac{1}{(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x})}\left(\mathbf{I}_d - \frac{\mathbf{x}\boldsymbol{\mu}^{\mathrm{T}}}{2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}}\right)\mathbf{G}(\mathbf{x})^{-1}\left(\mathbf{I}_d - \frac{\mathbf{x}\boldsymbol{\mu}^{\mathrm{T}}}{2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}}\right)^{\mathrm{T}}$$

for any $\mathbf{x} \in \mathcal{X}(\delta)$, *where* $\boldsymbol{\mu} = \int_{\mathcal{X}} \mathbf{x}F(\mathrm{d}\mathbf{x})$ *and* $\mathbf{G}(\cdot)$ *is defined in equation (3.4). Then there exists a sequence of orthogonal matrices* $(\mathbf{W})_{n=1}^{\infty} = (\mathbf{W}_n)_{n=1}^{\infty} \subset \mathbb{O}(d)$ *such that*

$$n\rho_n \left\|\widehat{\mathbf{Y}}\mathbf{W} - \mathbf{Y}_0\right\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}_0} \int_{\mathcal{X}} \mathrm{tr}\left\{\widetilde{\mathbf{G}}(\mathbf{x})\right\} F(\mathrm{d}\mathbf{x}), \tag{4.6}$$

and for each fixed $i \in [n]$,

$$n\rho_n^{1/2}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{y}}_i - \mathbf{y}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \widetilde{\mathbf{G}}(\mathbf{x}_{0i})). \tag{4.7}$$

Furthermore, for any $\mathbf{x} \in \mathcal{X}(\delta)$, $\widetilde{\boldsymbol{\Sigma}}(\mathbf{x}) - \widetilde{\mathbf{G}}(\mathbf{x})$ *is always positive semidefinite, where the formula for* $\widetilde{\boldsymbol{\Sigma}}(\cdot)$ *is given in Theorem 6.*

Remark 7. *The key difference between the assumption of Theorem 8 for the one-step estimator for the population LSE and that of Theorem 6 for the (sample) LSE is that, under the sparse regime (ii), we drop the requirement that* F *is a finite mixture of point masses and* F *is allowed to be a general distribution function on* $\mathcal{X}^n$, *at the cost of a stronger density assumption* $(\log n)^{2(1\vee\omega)}/(n\rho_n^4) \to 0$.

In Section 3, it is shown that the one-step estimator $\widehat{\mathbf{X}}$ dominates the ASE $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$ for estimating $\mathbf{X}_0$ asymptotically. Similarly, since $\widetilde{\boldsymbol{\Sigma}}(\mathbf{x}) \succeq \widetilde{\mathbf{G}}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}(\delta)$, it follows that locally for a fixed vertex i, the one-step estimator $\widehat{\mathbf{Y}}$ improves the LSE $\breve{\mathbf{X}}$ asymptotically in terms of a smaller asymptotic covariance matrix in spectra. In addition,

$$n\rho_n\|\breve{\mathbf{X}}\mathbf{W} - \mathbf{Y}_0\|_{\mathrm{F}}^2 - n\rho_n\|\widehat{\mathbf{Y}}\mathbf{W} - \mathbf{Y}_0\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}_0} \int_{\mathcal{X}} \mathrm{tr}\{\widetilde{\boldsymbol{\Sigma}}(\mathbf{x}) - \widetilde{\mathbf{G}}(\mathbf{x})\}F(\mathrm{d}\mathbf{x}) \geq 0.$$

Namely, the one-step estimator $\widehat{\mathbf{Y}}$ also improves the LSE $\breve{\mathbf{X}}$ globally for all vertices in terms of the sum of squares error $\|\widehat{\mathbf{Y}}\mathbf{W} - \mathbf{Y}_0\|_{\mathrm{F}}^2$.

5 Numerical Examples

5.1 A latent curve random graph example

In this subsection, we consider a random dot product graph whose latent positions are generated from a curve. Consider a graph with n vertices and latent dimension $d = 1$. The latent position x_{0i} for the ith vertex is set to $x_{0i} = 0.8 \sin\{\pi(i-1)/(n-1)\} + 0.1$, where $i \in [n]$. Let $\mathbf{X}_0 = [x_{01}, \ldots, x_{0n}]^{\mathrm{T}}$ and suppose an adjacency matrix $\mathbf{A}$ is generated from RDPG($\mathbf{X}_0$). The four estimators involved are the ASE $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$, the one-step estimator $\widehat{\mathbf{X}}$ initialized at the ASE (OSE-A), the LSE $\breve{\mathbf{X}}$, and the one-step estimator $\widehat{\mathbf{Y}}$ for the population LSE (OSE-L). We focus on the following objectives:

(i) Comparison between the ASE and the OSE-A, and the comparison between the LSE and the OSE-L. We evaluate the performance of these estimates by computing their sum of squares errors (SSEs):

$$SSE_{\mathrm{ASE}} = \inf_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{X}}^{(\mathrm{ASE})}\mathbf{W} - \mathbf{X}_0\|_2^2, \quad SSE_{\mathrm{OSE-A}} = \inf_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_2^2,$$
$$SSE_{\mathrm{LSE}} = \inf_{\mathbf{W}\in\{\pm1\}} \|\breve{\mathbf{X}}\mathbf{W} - \mathbf{Y}_0\|_2^2, \quad SSE_{\mathrm{OSE-L}} = \inf_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{Y}}\mathbf{W} - \mathbf{Y}_0\|_2^2.$$

(ii) Performance of the vertex-wise confidence intervals (CIs) for the latent positions and the population LSE. The vertex-wise CIs can be derived from Theorem 5 and Theorem 9. Let $\widehat{\mathbf{X}} = [\widehat{x}_1, \ldots, \widehat{x}_n]^{\mathrm{T}}$ be the OSE-A. By Theorem 5, $\sqrt{n}(|\widehat{x}_i| - x_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, G(x_{0i})^{-1})$, where $G(x_{0i}) = \int x_1\{x_{0i}(1 - x_{0i}x_1)\}^{-1}F(\mathrm{d}x_1)$. To compute a $1-\alpha$ confidence interval for x_{0i}, we need to estimate $G(x_{0i})$ using $\widehat{\mathbf{X}}$ because neither x_{0i} nor the function form of G is accessible from the data. Specifically, let $\widehat{G}(\widehat{x}_i) = (1/n)\sum_{j=1}^n \widehat{x}_j\{\widehat{x}_i(1 - \widehat{x}_i\widehat{x}_j)\}^{-1}$. Then a $1-\alpha$ confidence interval for x_{0i} is given by

$$\left(|\widehat{x}_i| - \frac{q_z(1-\alpha/2)}{\sqrt{\widehat{G}(\widehat{x}_i)n}}, |\widehat{x}_i| + \frac{q_z(1-\alpha/2)}{\sqrt{\widehat{G}(\widehat{x}_i)n}}\right), \tag{5.1}$$

where $q_z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of the standard normal distribution. Similarly, the asymptotic normality $n(|\widehat{y}_i| - y_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, \widetilde{G}(x_{0i}))$ from Theorem 9 can be employed

to construct a $1-\alpha$ confidence interval for the coordinate y_{0i} of the population LSE $\mathbf{Y}_0$, where $\widehat{y}_i$ is the ith coordinate of $\widehat{\mathbf{Y}}$. The corresponding asymptotic variance can be estimated by $\{4\widehat{\mu}\widehat{x}_i\widehat{G}(\widehat{x}_i)\}^{-1}$, where $\widehat{\mu} = (1/n)\sum_{j=1}^{n}\widehat{x}_j$. Therefore, a $1-\alpha$ confidence interval for y_{0i} is given by

$$\left(|\widehat{y}_i| - \frac{q_z(1-\alpha/2)}{\sqrt{4n^2\widehat{\mu}\widehat{x}_i\widehat{G}(\widehat{x}_i)}}, |\widehat{y}_i| + \frac{q_z(1-\alpha/2)}{\sqrt{4n^2\widehat{\mu}\widehat{x}_i\widehat{G}(\widehat{x}_i)}}\right). \tag{5.2}$$

(iii) Performance of the hypothesis testing as a subsequent inference task. We consider testing hypothese $H_0 : \mathbf{A} \sim \text{RDPG}(\mathbf{X}_0)$ against $H_A : \mathbf{A} \sim \text{RDPG}(\mathbf{X}_\epsilon)$, where $\mathbf{X}_\epsilon = [\mathbf{X}_0, \epsilon\mathbf{1}]$, $\mathbf{1}$ is the n-dimensional vector of all ones, and $\epsilon \in \{0.001, 0.002, \ldots, 0.01\}$. To compare the impact of the ASE and OSE-A on hypothesis testing, we let $T_{\text{ASE}} = \inf_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{X}}^{(\text{ASE})}\mathbf{W} - \mathbf{X}_0\|_{\text{F}}^2$ and $T_{\text{OSE-A}} = \inf_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{\text{F}}^2$ be the test statistics associated with them. The goal is to explore the powers of T_{ASE} and $T_{\text{OSE-A}}$ as functions of ϵ.

For objectives (i) and (ii), we draw 1000 independent adjacency matrices from $\text{RDPG}(\mathbf{X}_0)$ as Monte Carlo replicates. Regarding objective (i), we compute the sum of squares errors across the 1000 Monte Carlo replicates and present the boxplots of $SSE_{\text{ASE}} - SSE_{\text{OSE-A}}$ and $nSSE_{\text{LSE}} - nSSE_{\text{OSE-L}}$ in Figures 2 (a) and (b), respectively. We can see clearly that, for each realization, $SSE_{\text{OSE-A}} < SSE_{\text{ASE}}$ and $SSE_{\text{OSE-L}} < SSE_{\text{LSE}}$ with large probability. The difference between SSE_{ASE} and $\text{SSE}_{\text{OSE-A}}$ and that between SSE_{LSE} and $\text{SSE}_{\text{OSE-L}}$ are both statistically significant at level $\alpha = 0.01$. These results support the theory developed in Sections 3 and 4. In terms of objective (ii), we construct the vertex-wise 95% confidence intervals for both $\mathbf{X}_0$ and $\mathbf{Y}_0$ based on each realization of the adjacency matrix, and compute the corresponding empirical coverage probabilities for each vertex $i \in [n]$. The results are visualized in Figures 2 (c) and (d), respectively. The empirical coverage probabilities concentrate near the nominal coverage probability. We also randomly select one realization of the adjacency matrix and visualize the vertex-wise CIs for $[x_{01}, \ldots, x_{0n}]^{\text{T}}$ and $[y_{01}, \ldots, y_{0n}]^{\text{T}}$ in Figures 2 (e) and (f), respectively, which further consolidate the asymptotic normality of the rows of the OSE-A and the OSE-L developed in Sections 3 and 4.

For objective (iii), we need to determine the null distributions of the test statistics T_{ASE}

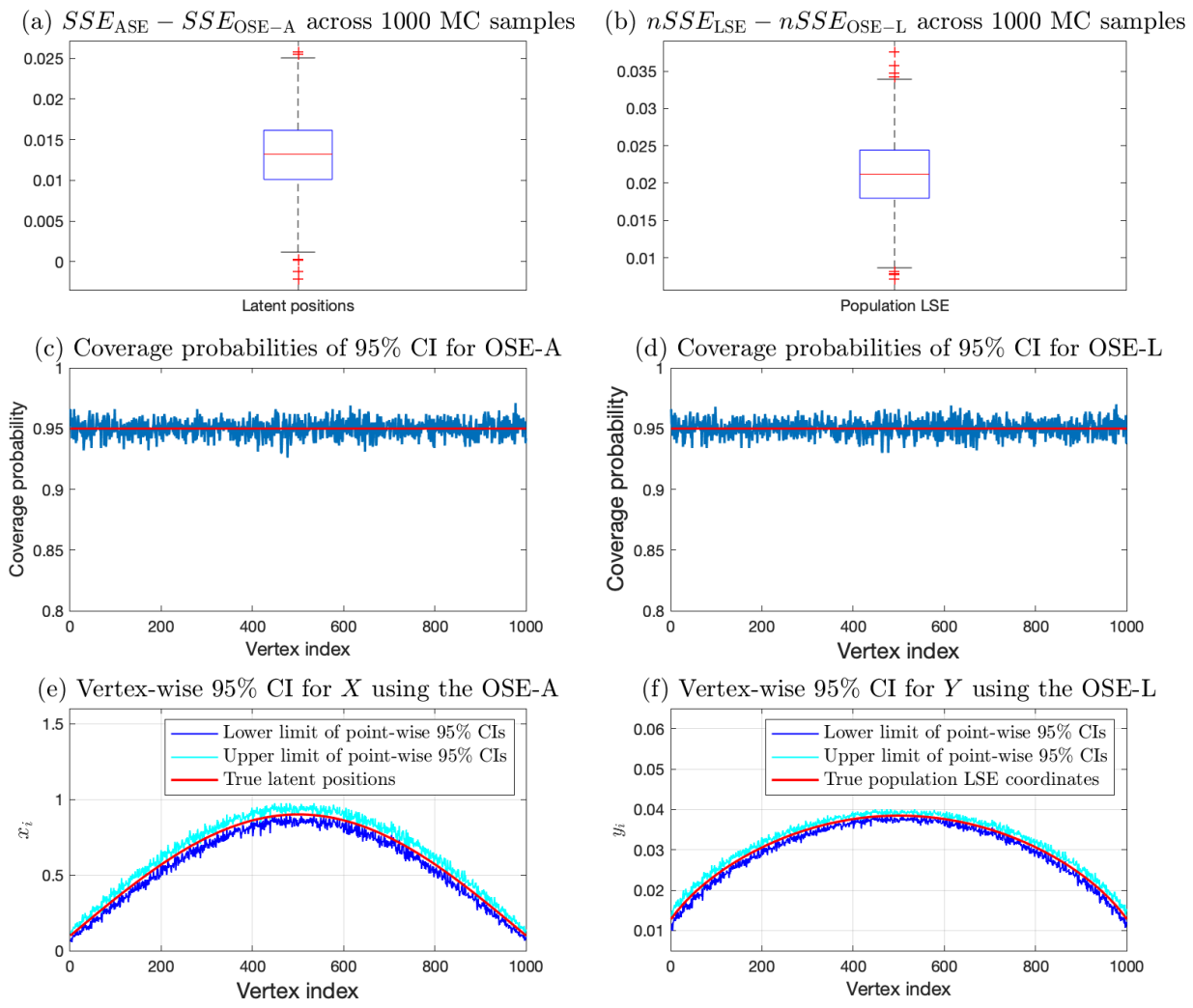

Figure 2: Numerical results for Subsection 5.1: Panels (a) and (b) are the boxplots of $SSE_{\mathrm{ASE}} - SSE_{\mathrm{OSE-A}}$ and $nSSE_{\mathrm{LSE}} - nSSE_{\mathrm{OSE-L}}$ across 1000 Monte Carlo replicates, respectively; Panels (c) and (d) are the coverage probabilities of the vertex-wise confidence intervals for the latent positions $\mathbf{X}_0 = [x_{01}, \ldots, x_{0n}]^{\mathrm{T}}$ and the population LSE $\mathbf{Y}_0 = [y_{01}, \ldots, y_{0n}]^{\mathrm{T}}$, respectively, and the red horizontal lines correspond to the nominal 95% coverage probability; Panels (e) and (f) are the realizations of the vertex-wise 95% confidence intervals for $\mathbf{X}_0 = [x_{01}, \ldots, x_{0n}]^{\mathrm{T}}$ and $\mathbf{Y}_0 = [y_{01}, \ldots, y_{0n}]^{\mathrm{T}}$ from a single draw of $\mathbf{A}$, respectively.

and T_{OSE}. We compute these null distributions using a Monte Carlo simulation with 1000 independent replicates. The rejection regions for level α tests based on T_{ASE} and $T_{\mathrm{OSE-A}}$ are $R_{\mathrm{ASE}} := \{\mathbf{A} : T_{\mathrm{ASE}} > q_\alpha(\mathrm{ASE})\}$ and $R_{\mathrm{OSE-A}} := \{\mathbf{A} : T_{\mathrm{OSE-A}} > q_\alpha(\mathrm{OSE\text{-}A})\}$, where $q_\alpha(\mathrm{ASE})$ and $q_\alpha(\mathrm{OSE\text{-}A})$ are the $(1-\alpha)$-quantiles of the distributions of T_{ASE} and $T_{\mathrm{OSE-A}}$

under the null hypothesis, respectively. We then compute the powers of the two test statistics under different values of $\epsilon \in \{0.001, 0.002, \ldots, 0.01\}$ using a Monte Carlo simulation with 1000 independent replicates and report the results in Table 1. Due to the improvement of the OSE-A over the ASE, we see clearly that the test based on $T_{\mathrm{OSE-A}}$ is more powerful than that based on T_{ASE}, which shows the usefulness of the proposed OSE-A for hypotheses testing as a subsequent inference task.

Table 1: Power comparison of T_{ASE} and T_{OSE} for Subsection 5.1

ϵ	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.010
Power of T_{ASE}	0.111	0.097	0.109	0.096	0.122	0.156	0.181	0.288	0.428	0.613
Power of T_{OSE}	0.154	0.137	0.156	0.157	0.208	0.261	0.317	0.437	0.582	0.744

5.2 Comparison with the method of maximum likelihood

This subsection aims at comparing the proposed one-step procedure with the ASE, and a local maximum likelihood estimator for the random dot product graph. Although neither the existence nor the uniqueness of the maximum likelihood estimator for the random dot product graph has been established, it is always possible to compute a local maximizer of the log-likelihood function using optimization algorithms. We first provide a simple block-coordinate descent method for finding a local maximizer of the log-likelihood function and then implement the algorithm in two concrete simulated examples. The goal is to compare the performance of the resulting estimate with the ASE and the one-step estimator (OSE) in terms of both the sum-of-squares errors and the computation time.

Let $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X})$ with sparsity factor $\rho_n = 1$ and let $\ell_{\mathbf{A}}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote the log-likelihood function of $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}}$. A potential local maximizer of $\ell_{\mathbf{A}}$ can be found using the block-coordinate ascent algorithm in Algorithm 2. Note that within each iteration, Algorithm 2 requires an exact line search along each $\mathbf{x}_i$ direction for all $i = 1, \ldots, n$. This step can be implemented using the Matlab function `fmincon` conveniently.

We next implement Algorithm 2 to Example **??** with $p = 0.6$, $q = 0.4$, $F(\mathrm{d}x) = 0.6\delta_p(\mathrm{d}x) + 0.4\delta_q(\mathrm{d}x)$, and $n = 300$. The same experiment is repeated for 1000 independent Monte Carlo replicates. We report the computation times for the ASE $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$, the

Algorithm 2 Block-coordinate ascent maximum likelihood

Input: The adjacency matrix $\mathbf{A} = [A_{ij}]_{n\times n}$ and the embedding dimension d.
Step 1: Compute the ASE $\widehat{\mathbf{X}}^{(\mathrm{ASE})}$
Step 2: Initialize $\widehat{\mathbf{X}}^{(0)} = \widehat{\mathbf{X}}^{(\mathrm{ASE})}$ and set $t = 0$.
Step 3: **While** not converged
    For $i = 1, 2, \ldots, n$

$$\widehat{\mathbf{x}}_i^{(t+1)} \longleftarrow \arg\max_{\mathbf{x}_i} \ell_{\mathbf{A}}\left(\widehat{\mathbf{x}}_1^{(t+1)}, \ldots, \widehat{\mathbf{x}}_{i-1}^{(t+1)}, \mathbf{x}_i, \widehat{\mathbf{x}}_{i+1}^{(t)}, , \ldots, \widehat{\mathbf{x}}_n^{(t)}\right).$$

    End For
    Set $t \longleftarrow t + 1$.
End While
Output: $\widehat{\mathbf{X}}^{(t)} = [\widehat{\mathbf{x}}_1^{(t)}, \ldots, \widehat{\mathbf{x}}_n^{(t)}]^{\mathrm{T}}$.

one-step estimator (OSE) initialized at the ASE $\widehat{\mathbf{X}}$, and the local maximum likelihood estimate (MLE) $\widehat{\mathbf{X}}^{(\mathrm{MLE})}$ for a single realization in Table 2 below. We also compute the sum-of-squares errors (SSEs) of the three estimates: $SSE_{\mathrm{ASE}} = \min_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{X}}^{(\mathrm{ASE})}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$, $SSE_{\mathrm{OSE}} = \min_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$, and $SSE_{\mathrm{MLE}} = \min_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{X}}^{(\mathrm{MLE})}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$. The average SSEs of these estimates across 1000 Monte Carlo replicates are tabulated in Table 2, together with the standard errors. Figure 3 visualizes $SSE_{\mathrm{ASE}} - SSE_{\mathrm{OSE}}$, $SSE_{\mathrm{ASE}} - SSE_{\mathrm{MLE}}$, and $SSE_{\mathrm{OSE}} - SSE_{\mathrm{MLE}}$ in the three panels, respectively. In particular, $SSE_{\mathrm{OSE}} - SSE_{\mathrm{MLE}}$ is mild in this example in contrast to $SSE_{\mathrm{ASE}} - SSE_{\mathrm{OSE}}$. These numerical results suggest that the improvement from the ASE to the OSE is more significant than the improvement from the OSE to the MLE in terms of the SSEs, whereas the computation cost of the MLE is much higher than that of the OSE and that of the ASE.

Table 2: Computation time and error comparison for Subsection 5.2: The latent positions are set as in Example 1 with $p = 0.6, q = 0.4$, and the number of vertices is $n = 300$.

Method	ASE	OSE-A	MLE
Computation time in `Matlab`	0.008057s	0.01882s	26.1634s
$SSE = \inf_{\mathbf{W}\in\{\pm1\}} \|\widehat{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$	0.7178	0.7040	0.7030
Standard error for SSE	0.0019	0.0018	0.0018

We finally consider a simulated example with a comparatively small sample size, which sheds some light to future research direction concerning the method of maximum likelihood in finite-sample problems. The setup is similar to the example in Subsection 5.1. Namely,

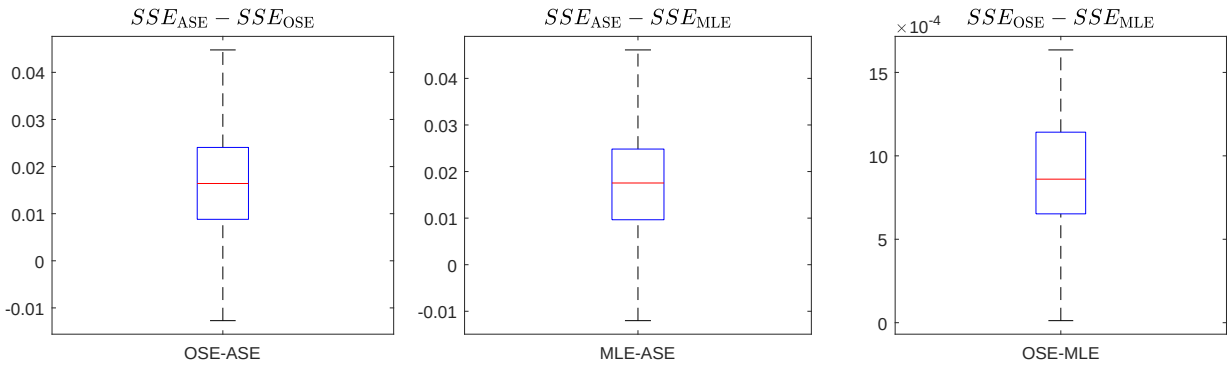

Figure 3: Numerical results for Subsection 5.2: The boxplots of $SSE_{\text{ASE}} - SSE_{\text{OSE}}$, $SSE_{\text{ASE}} - SSE_{\text{MLE}}$, and $SSE_{\text{OSE}} - SSE_{\text{MLE}}$ across 1000 Monte Carlo replicates; The latent positions are set as as in Example 1 with $p = 0.6, q = 0.4$, and the number of vertices is $n = 300$.

we consider a 1-dimensional random dot product graph whose latent positions are given by $x_{0i} = 0.8 \sin\{\pi(i-1)/(n-1)\} + 0.1$, $i \in [n]$. The number of vertices n is set to 30, and we generate an adjacency matrix $\mathbf{A} \sim \text{RDPG}(\mathbf{X}_0)$, where $\mathbf{X}_0 = [x_{01}, \ldots, x_{0n}]^{\text{T}}$. We compute the ASE, the proposed OSE, and a local MLE using Algorithm 2. We repeat the experiment for 1000 independent Monte Carlo replicates. The computation times for obtaining the ASE, the OSE, and the MLE for a single realization are reported in Table 3, together with the average SSEs and the corresponding standard errors across 1000 Monte Carlo replicates. We also visualize $SSE_{\text{ASE}} - SSE_{\text{OSE}}$, $SSE_{\text{ASE}} - SSE_{\text{MLE}}$, and $SSE_{\text{OSE}} - SSE_{\text{MLE}}$ in the three panels of Figure 4, respectively. Observe that in this example, with a relatively small number of vertices $n = 30$, the OSE does not provide improvement over the ASE, whereas the MLE shows significant improvement over the ASE as well as the OSE. The practical performance of the MLE for finite-sample problems is also inspiring for designing a multiple-step procedure that repeats the one-step update multiple times for finding a local MLE. This interesting direction is deferred to future work. Another implication of this experiment is that the practitioners are not recommended to apply the one-step procedure for network data with comparatively small vertices. Instead, it is recommended that a local MLE is used over the one-step estimate and the ASE.

Table 3: Computation time and error comparison for Subsection 5.2: The latent positions are set to $x_{0i} = 0.8\sin\{\pi(i-1)/(n-1)\} + 0.1$, $i = 1, \ldots, 30$.

Method	ASE	OSE-A	MLE
Computation time in `Matlab`	0.004213s	0.001644s	0.209708s
$SSE = \|\widehat{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$	0.4449	0.6260	0.4062
Standard error for SSE	0.0041	0.0079	0.0039

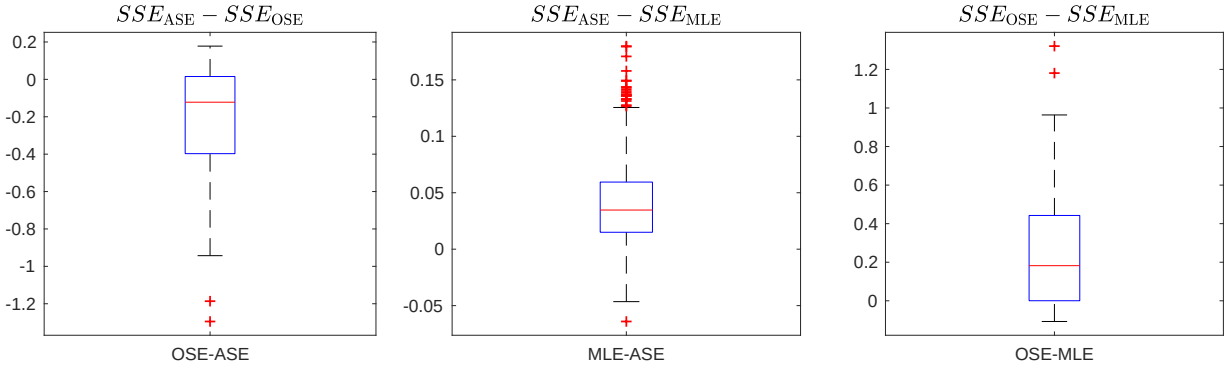

Figure 4: Numerical results for Subsection 5.2: The boxplots of $SSE_{\mathrm{ASE}} - SSE_{\mathrm{OSE}}$, $SSE_{\mathrm{ASE}} - SSE_{\mathrm{MLE}}$, and $SSE_{\mathrm{OSE}} - SSE_{\mathrm{MLE}}$ across 1000 Monte Carlo replicates; The latent positions are set as $x_{0i} = 0.8\sin\{\pi(i-1)/(n-1)\} + 0.1$, $i = 1, \ldots, 30$.

5.3 Wikipedia Graph data

We finally apply the proposed one-step procedure to a real-world Wikipedia graph dataset, which is available at `http://www.cis.jhu.edu/~parky/Data/data.html`. The Wikipedia graph dataset consists of an adjacency matrix among $n = 1382$ Wikipedia articles that are within two hyperlinks of the article "Algebraic Geometry", and these articles are further manually labeled according to one of the following 6 descriptions: People, Places, Dates, Things, Math, and Category. To determine a suitable embedding dimension d for the random dot product graph model, we follow the ad-hoc approach of Zhu and Ghodsi (2006) and computes

$$\widehat{d} = \underset{d=1,2,\ldots,q}{\arg\max} \left\{ \sum_{k=1}^{d} \log f(\sigma_k(\mathbf{A}); \widehat{\mu}_1, \widehat{\sigma}^2) + \sum_{k=d+1}^{q} \log f(\sigma_k(\mathbf{A}); \widehat{\mu}_2, \widehat{\sigma}^2) \right\},$$

where $f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-(x-\mu)^2/(2\sigma^2)\}$ is the normal density with mean μ and variance σ^2, $\mu_1 = \frac{1}{d}\sum_{k=1}^{d} \sigma_k(\mathbf{A})$, $\mu_2 = \frac{1}{p-d}\sum_{k=d+1}^{p} \sigma_k(\mathbf{A})$, $\widehat{\sigma}^2 = \frac{(d-1)s_1^2 + (p-d-1)s_2^2}{p-2}$, s_1^2, s_2^2

are the sample variances of $\{\sigma_k(\mathbf{A})\}_{k=1}^{d}$ and $\{\sigma_k(\mathbf{A})\}_{k=d+1}^{q}$, respectively, and q is an upper bound for the embedding dimension. Here we select $q = 50$ as a conservative upper bound, resulting in $\widehat{d} = 11$.

We next compute the ASE, the LSE, the OSE-A, and the OSE-L, with the embedding dimension $d = 11$, and then apply the GMM-based clustering algorithm to these estimates, with the number of clusters being 6. We next compare the similarity between the manually assigned 6 class labels and these clustering results by computing the respective Rand indices, which are tabulated in Table 4. The results show that the one-step procedure for the population LSE outperforms the rest of the competitors, as it provides the clustering result that is most similar to the original class label assignment among the four methods.

Table 4: Wikipedia Graph Data: Rand indices of the GMM-based clustering algorithm applied to the ASE, the LSE, the OSE-A, and the OSE-L, respectively, with the number of clusters being 6, in comparison with the corresponding manual labels.

Method	ASE	LSE	OSE-A	OSE-L
Rand Index	0.7429	0.7350	0.7413	**0.7538**

Table 5: Wikipedia Graph Data: Rand indices of the GMM-based clustering algorithm applied to the ASE, the LSE, the OSE-A, and the OSE-L, respectively, with the number of clusters being 2, in comparison with the corresponding one-versus-all manual labels for the class "Dates".

Method	ASE	LSE	OSE-A	OSE-L
Rand Index	0.5289	0.5097	**0.5432**	0.5313

Besides evaluating the performance of the overall clustering for the 6 manually-assigned labels, we also focus on the comparison of the article class "Dates" against the rest of the articles specifically. We apply the GMM-based clustering algorithm to the aforementioned four estimates again, but with the number of clusters being 2, and tabulate the Rand indices in Table 5. We can see that the proposed one-step procedure improves the clustering accuracy as well when we focus on the comparison between the article class "Dates" against the rest of the labels. The scatter plots of the first-versus-second dimension of the four estimates are visualized in Figure 5, along with the cluster-specific 95% empirical confidence ellipses in dashed lines.

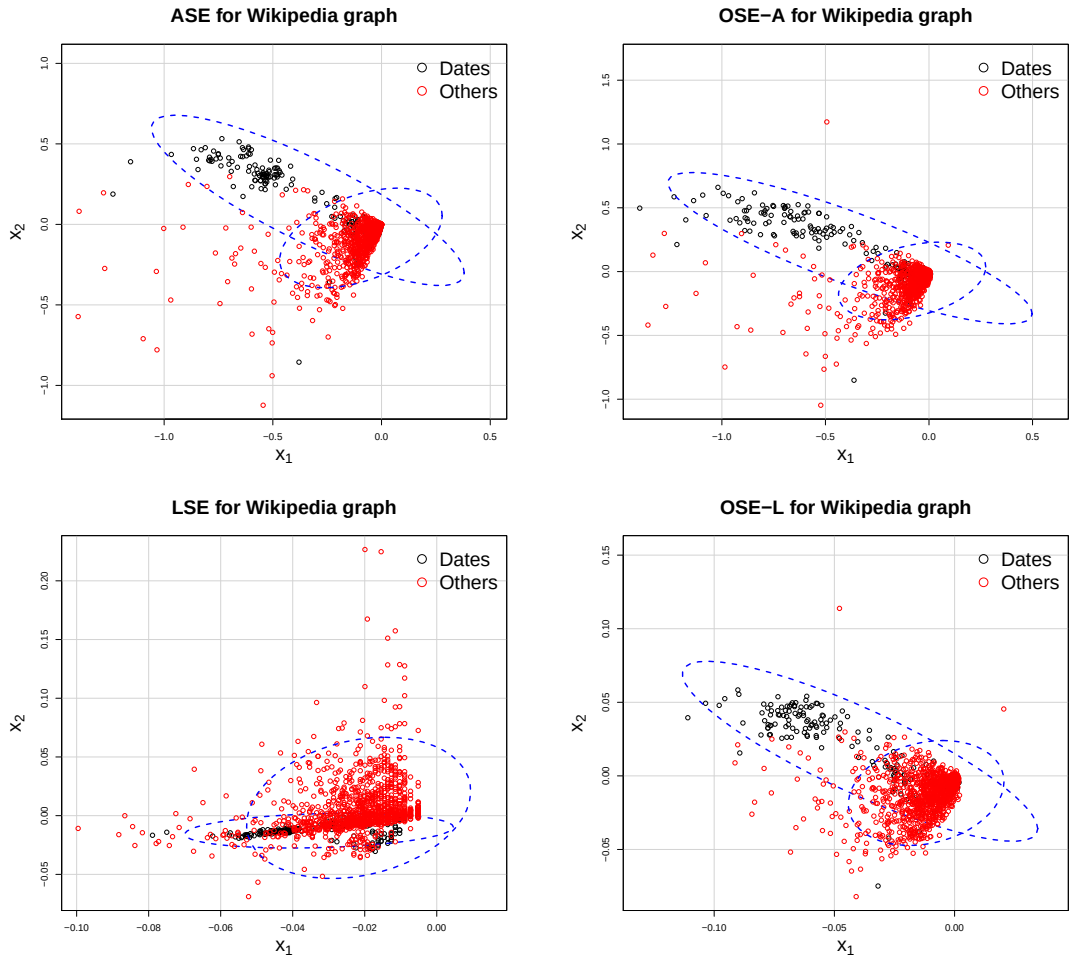

Figure 5: Wikipedia graph data: The scatter plots of the first-versus-second dimension of the four estimates. The scatter points are colored according to whether the articles are in the class "Dates" or the others. The 95% empirical cluster-specific confidence ellipses are displayed by the dashed lines.

6 Discussion

In the context of stochastic block models, Gao et al. (2017) proposed a vertex clustering approach that improves the solution provided by the ASE and/or the LSE. The algorithm in Gao et al. (2017) starts from the clustering solution of the ASE/LSE and then refines the cluster assignment of each vertex through the maximization of a penalized Bernoulli likelihood function, where the cluster memberships of the rest of the vertices are fixed at

their most recent values. This approach is similar to our one-step procedure for estimating the latent positions in spirit, as both methods are implemented in a vertex-by-vertex optimization fashion with a warm start solution (*i.e.*, the ASE/LSE or the cluster assignment given by them). Our method differs from the method of Gao et al. (2017) in that the proposed one-step procedure aims at maximizing the Bernoulli likelihood function with regard to the continuous-valued latent positions and takes the gradient information of the likelihood function into account, whereas Gao et al. (2017) focus on estimating cluster memberships of vertices, and no gradient information is available due to the discrete nature of the variables of interest.

We assume that the embedding dimension d for the random dot product graph is known throughout the paper. The proposed one-step procedure is also valid when the true dimension d for the underlying sampling model is unknown. In this case, the method proceeds by first finding the ASE into $\mathbb{R}^{d'}$ for some $d' \geq 1$ and $d' < d$ (*i.e.*, when the dimension is under-estimated) and then computing the one-step estimator based on d'. Our Theorem 5 and Theorem 9 still hold and can be easily proved as suggested by Tang and Priebe (2018). On the other hand, leveraging Bayesian methods when the dimension d is unknown is a promising future direction in light of the recent progress in Bayesian theory and methods for low-rank matrix models with undetermined rank (Bhattacharya and Dunson, 2011; Rocková and George, 2016) and network models (Caron and Fox, 2017; Xie and Xu, 2019; Geng et al., 2019).

We have shown that the one-step procedure produces an estimator enjoying fascinating asymptotic properties both globally for all vertices and locally for each vertex. Nevertheless, for problems with comparatively small sample sizes, we have also shown in a simulation example that the one-step estimators do not necessarily provide us with better numerical results compared to the classical adjacency/Laplacian spectral embedding. Instead, we have also observed that the method of maximum likelihood, which is implemented in a block-coordinate ascent algorithm, provides practical improvement over the adjacency spectral embedding. Since the one-step procedure only implements a single iteration of the Newton-Raphson algorithm with the observed Hessian matrix replaced by the negative Fisher information

matrix, we hope to develop an iterative algorithm for finding a local maximum likelihood estimator by repeating the one-step procedure multiple times until convergence. Such an iterative algorithm can be implemented in conjunction with the regularization of the Fisher information matrix and backtracking procedure for finding suitable step sizes to achieve faster convergence (Nocedal and Wright, 2006). Furthermore, developing a scalable version of such an algorithm will be highly desirable in the presence of big data and extremely large networks. It will also be useful to explore the statistical properties of the estimator obtained by the iterative algorithm, and to establish its theoretical guarantee. We defer these research topics to future work.

Supplementary Material

The supplementary material contains a comprehensive list of notations, the proofs of the technical results in Section 2, Section 3, and Section 4, the behavior of the one-step estimator for positive definite stochastic block models, further discussion regarding sparse graph models, and additional simulated examples.

Acknowledgements

The work of Xu was supported by NSF 1918854 and NSF 1940107.

References

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., Qin, Y., and Sussman, D. L. (2018a). Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92.

Athreya, A., Priebe, C. E., Tang, M., Lyzinski, V., Marchette, D. J., and Sussman, D. L. (2016). A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1):1–18.

Athreya, A., Tang, M., Park, Y., and Priebe, C. E. (2018b). On estimation and inference in latent structure random graphs. *arXiv preprint arXiv:1806.01401*.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, pages 291–306.

Bickel, P. J. and Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics*, volume 2. CRC Press.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Caron, F. and Fox, E. B. (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.*, 18(1):1980–2024.

Geng, J., Bhattacharya, A., and Pati, D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526):893–905.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Mele, A., Hao, L., Cape, J., and Priebe, C. E. (2019). Spectral inference for large stochastic blockmodels with nodal covariates. *arXiv preprint arXiv:1908.06438*.

Merris, R. (1994). Laplacian matrices of graphs: a survey. *Linear algebra and its applications*, 197:143–176.

Neil, J., Uphoff, B., Hash, C., and Storlie, C. (2013). Towards improved detection of attackers in computer networks: New edges, fast updating, and host agents. In *2013 6th International Symposium on Resilient Control Systems (ISRCS)*, pages 218–224.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Priebe, C. E., Park, Y., Tang, M., Athreya, A., Lyzinski, V., Vogelstein, J. T., Qin, Y., Cocanougher, B., Eichler, K., Zlatic, M., et al. (2017). Semiparametric spectral modeling of the drosophila connectome. *arXiv preprint arXiv:1705.03297*.

Rocková, V. and George, E. I. (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915.

Rubin-Delanchy, P., Adams, N. M., and Heard, N. A. (2016). Disassortativity of computer networks. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 243–247.

Rubin-Delanchy, P., Cape, J., Tang, M., and Priebe, C. E. (2017). A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*.

Sarkar, P. and Bickel, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist.*, 43(3):962–990.

Sussman, D. L., Tang, M., Fishkind, D. E., and Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128.

Sussman, D. L., Tang, M., and Priebe, C. E. (2014). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):48–57.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., Park, Y., and Priebe, C. E. (2017a). A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., and Priebe, C. E. (2017b). A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23(3):1599–1630.

Tang, M. and Priebe, C. E. (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann. Statist.*, 46(5):2360–2415.

Tang, M., Sussman, D. L., and Priebe, C. E. (2013). Universally consistent vertex classification for latent positions graphs. *Ann. Statist.*, 41(3):1406–1430.

Tang, R., Ketcha, M., Badea, A., Calabrese, E. D., Margulies, D. S., Vogelstein, J. T., Priebe, C. E., and Sussman, D. L. (2019). Connectome smoothing via low-rank approximations. *IEEE Transactions on Medical Imaging*, 38(6):1446–1456.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Ward, M. D., Stovel, K., and Sacks, A. (2011). Network analysis and political science. *Annual Review of Political Science*, 14(1):245–264.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Xie, F. and Xu, Y. (2019). Optimal bayesian estimation for random dot product graphs. *arXiv preprint arXiv:1904.12070*.

Young, S. J. and Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer.

Zhu, M. and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918 – 930.