Abstract

Random-walk based network embedding algorithms like node2vec and DeepWalk are widely used to obtain Euclidean representation of the nodes in a network prior to performing down-stream network inference tasks. Nevertheless, despite their impressive empirical performance, there is a lack of theoretical results explaining their behavior. In this paper we studied the node2vec and DeepWalk algorithms through the perspective of matrix factorization. We analyze these algorithms in the setting of community detection for stochastic blockmodel graphs; in particular we established large-sample error bounds and prove consistent community recovery of node2vec/DeepWalk embedding followed by k-means clustering. Our theoretical results indicate a subtle interplay between the sparsity of the observed networks, the window sizes of the random walks, and the convergence rates of the node2vec/DeepWalk embedding toward the embedding of the true but unknown edge probabilities matrix. More specifically, as the network becomes sparser, our results suggest using larger window sizes, or equivalently, taking longer random walks, in order to attain better convergence rate for the resulting embeddings. The paper includes numerical experiments corroborating these observations.

Keywords: stochastic blockmodel, network embedding, matrix factorization, random walk, node2vec, DeepWalk

1. Introduction

Given a network $G$, a popular approach for analyzing $G$ is to first map or embed the vertices of $G$ into some low dimensional latent space and then apply statistical learning and inference methods in the embedded space. Through this embedding process, multiple tasks could be conducted on the network such as community detection (e.g., von Luxburg (2007); Wang et al. (2017)), link prediction (e.g., Liben-Nowell and Kleinberg (2007)), node classification (e.g., Hamilton et al. (2017a); Perozzi et al. (2014)) and network visualization (e.g., Theocharidis et al. (2009)).

There has been a large and diverse collection of network embedding algorithms proposed in the literature, including those based on spectral embedding (Rohe et al., 2011; Sussman et al., 2012), multivariate statistical dimension reduction (Robinson and Bennett, 1995; Ye...
et al., 2005), random walks (Grover and Leskovec, 2016; Perozzi et al., 2014) and neural network (Kipf and Welling, 2017). See Hamilton et al. (2017b), Cui et al. (2018) and Chen et al. (2020) for detailed surveys of network embedding and graph representation learning.

In recent years there has been a significant rise in popularity of network embedding algorithms based on random-walks. The most well-known examples include DeepWalk (Perozzi et al., 2014) and node2vec (Grover and Leskovec, 2016). These algorithms are computationally efficient and furthermore yield impressive empirical performance in many different scientific applications including recommendation systems (Palumbo et al., 2018), biomedical natural language processing (Zhang et al., 2019b), human protein identification (Zhang et al., 2019a), traffic prediction (Zheng et al., 2020) and city road layout modeling (Chu et al., 2019). Nevertheless, despite their widespread use, there is still a lack of theoretical results for these algorithms. In particular it is unclear what the node embeddings represents as well as their large-sample behavior as the number of nodes increases.

Theoretical properties for DeepWalk, node2vec, and related algorithms had been studied previously in the computer science community. The focus here had been mostly on the convergence of the entries of the co-occurrence matrix as the length and/or number of random walks goes to infinity. For example Levy and Goldberg (2014) showed that node2vec using the skip-gram model with negative sampling is equivalent to factorizing a matrix whose entries are obtained by taking the entrywise logarithm of a co-occurrence matrix, provided that the embedding dimension is sufficiently large, i.e., the embedding dimension \( d \) could exceed the number of nodes \( n \). The transformed co-occurrence entries can be interpreted as yielding a point-wise mutual information (PMI) between the nodes. In a related vein, Qiu et al. (2018) also studied DeepWalk and node2vec as specific instances of a matrix factorization problem using a (transformed) co-occurrence matrix. They then derived the limiting form of the entries of this matrix as the length of the random walks goes to infinity. Qiu et al. (2020) strengthened these results by providing finite-sample concentration bounds for the co-occurrence entries as the length of the random walks increases. Note, however, that the above cited works focused exclusively on the case of a fixed graph and thus do not provide results on the large sample behavior of these random-walk based algorithms as the number of nodes increases.

The statistical community, in contrast, had extensively studied the large-sample theoretical properties of graph embedding algorithms based on matrix factorization. However, the embedding algorithms considered are based on singular value decomposition of either the adjacency matrix or the Laplacian matrix and its normalized and/or regularized variants. For example, in the setting of the popular stochastic blockmodel random graphs, Rohe et al. (2011) and Sussman et al. (2012) derive consistency results for a truncated singular value decomposition of the normalized Laplacian matrix and the adjacency matrix, respectively. Rubin-Delanchy et al. (2017); Tang and Priebe (2018) complement these results by providing central limit theorems for the components of the eigenvectors of either the adjacency matrix or the normalized Laplacian matrix under the more general random dot product graphs model. Since DeepWalk and node2vec are based on taking the entrywise logarithm of a random-walk co-occurrence matrix, the techniques used in deriving the above cited results do not readily translate to this random-walk based setting.
1.1. Contributions of the Current Paper

The current paper studied the large-sample theoretical properties of random-walk based embedding algorithms. We present convergence results for the embeddings of DeepWalk and node2vec in the case of stochastic blockmodel graphs. We then show that running k-means on the resulting embeddings is sufficient for weak recovery of the latent community assignments. Our theoretical results thus provide a bridge between the previous theoretical results in the computer science community and their statistics counterpart.

We emphasize that our focus on stochastic blockmodel graphs is done purely for ease of exposition. Indeed, most of our results will still hold for the more general inhomogeneous Erdős-Rényi (ER) random graphs model, provided that the edge probabilities are sufficiently homogeneous, i.e., the minimum and maximum values for the edge probabilities are of the same order as $n$ increases. We recall that the inhomogeneous ER model only assume that the edges are independent Bernoulli random variables and hence is the most general model for edge independent random graphs. However, we might no longer be able to show that inhomogeneous ER random graphs possess a low-dimensional approximate representation, even when $n$ increases. See Section 5 for further discussion and reference to specific results.

We now outline our approach. The original node2vec/DeepWalk algorithms are based on optimizing the non-convex skip-gram model using stochastic gradient descent; the resulting optimization problem can have multiple local minimima and the obtained embeddings can thus be numerically unstable. We instead consider, for each embedding dimension $d$, the optimal low-rank approximation of a transformed co-occurrence matrix similar to that used in Levy and Goldberg (2014) and Qiu et al. (2020). We then show that the entries of the co-occurrence matrix computed using the observed adjacency matrix is uniformly close to the entries of the co-occurrence matrix computed using the true but unknown edge probabilities matrix. This uniform bound then implies that the entrywise logarithm of the two co-occurrence matrices are also uniformly close. We emphasize that the uniform entrywise error bounds are essential for this step. Indeed, a priori the co-occurrence matrices are only guaranteed to be non-negative matrices and hence their entrywise logarithm is possibly unbounded. Now, by definition, the co-occurrence matrix constructed using the true edge probabilities will be a positive matrix. Our uniform entrywise error bounds therefore implies, with high probability, that the co-occurrence matrix constructed using the observed graph is also a positive matrix. In contrast, an error bound using the Frobenius norm is not sufficient to guarantee the positivity of this matrix.

In summary, the uniform entrywise error bounds for the co-occurrence matrices imply that the matrix we constructed using the observed graph is close, in Frobenius norm, to the matrix constructed using the true edge probabilities. In the case of stochastic blockmodel graphs, the true edge probabilities matrix yield an embedding in at most $K$-dimension, where $K$ is the number of blocks. Thus, for stochastic blockmodel graphs with $K \ll n$, an application of the celebrated Davis-Kahan theorem then shows that the truncated low-rank representation of both matrices are close, i.e., the embeddings of the observed graph is approximately the same, up to orthogonal transformation, as the embeddings of the true edge probabilities matrix. Therefore, by running $k$-means on the low-rank embeddings of the observed graph, we weakly recover the latent community structures.
Our paper is organized as follows. In Section 2, we give a brief introduction of the node2vec algorithm of Grover and Leskovec (2016). We then describe the matrix factorization perspective for node2vec. The DeepWalk algorithm can be treated as a special case of node2vec by setting the 2nd-order random-walk parameters \((p, q)\) to be \((1, 1)\), which will also be assumed in Section 3 for the simplicity of theoretical analysis. In Section 3 we provide uniform entrywise error bounds for the entries of the \(k\)-step random-walk transition matrix. This error bound then yields, as a corollary, a Frobenius norm error bound for the difference between the transformed co-occurrence matrix of the observed graph and the transformed co-occurrence matrix of the edge probabilities matrix. The theoretical results in Section 3 are provided under both the dense and sparse regimes where the average degree grows linearly and sublinearly in the number of nodes, respectively. We note that the proofs for the sparse regime are substantially more involved, both technically and conceptually, than the proofs for the dense regime. Our theoretical results also imply a subtle interplay between the sparsity of the observed networks, the window sizes of the random walks, and the convergence rates of node2vec/DeepWalk embedding of the observed graph toward the embedding of the true edge probabilities matrix. In Section 4 we present simulations to corroborate our theoretical results. We conclude the paper in Section 5 with a discussion of some open questions and potential improvements. Proofs of the stated results and the associated technical lemmas are provided in the appendix.

1.2. Notation

We first introduce some general notations that are used throughout this paper. For a given positive integer \(K\), we denote by \([K]\) the set \(\{1, 2, \ldots, K\}\). We denote a graph on \(n\) vertices by \(G = (V, E)\) where \(V = \{v_i\}_{i=1}^n\) and \(E = \{e_{ii'}\}_{i,i'=1}^n\) are the vertices and edge sets, respectively. Unless specified otherwise, all graphs in this paper are assumed to be undirected and unweighted. For each node \(v_i\) we denote by \(N(v_i)\) the set of nodes \(v_{i'}\) adjacent to \(v_i\). If \(G\) is a graph on \(n\) vertices then its \(n \times n\) adjacency matrix is denoted as \(A = [a_{ii'}]\). In the subsequent discussion, we often assume that the upper triangular entries of \(A\) are independent random variables with \(a_{ii'} \sim \text{Bernoulli}(p_{ii'})\) for all \(i < i', a_{ii'} = 0\) for \(i = i'\) and we denote by \(P = [p_{ii'}]\) the corresponding \(n \times n\) matrix of edge probabilities.

Given a graph \(G\) with adjacency matrix \(A\), let \(D_A = \text{diag}(d_1, \ldots, d_n)\) be a diagonal matrix with \(d_i = \sum_{i'=1}^n a_{ii'}\) as its \(i\)th diagonal element. Assuming \(G\) is connected, we define a random walk on \(G\) with a 1-step transition matrix \(\tilde{W} = AD_A^{-1}\). Correspondingly, when appropriate, we also define \(W = PD_P^{-1}\) where \(D_P = \text{diag}(p_1, \ldots, p_n)\) is the diagonal matrix with \(p_i = \sum_{i'=1}^n p_{ii'}\).

We use \(\|\cdot\|, \|\cdot\|_F\) and \(\|\cdot\|_{\max}\) to denote the spectral norm, Frobenius norm, and maximum entrywise value of a matrix, respectively. We also use \(\|\cdot\|_{\max, \text{off}}\) and \(\|\cdot\|_{\max, \text{diag}}\) to denote the maximum value for the off-diagonal and diagonal entries of a matrix, i.e., for any \(M = [m_{ii'}]_{m \times m} \in \mathbb{R}^{m \times m}\)

\[
\|M\|_{\max, \text{off}} \equiv \max_{i \neq i'} |m_{ii'}|, \quad \|M\|_{\max, \text{diag}} \equiv \max_{i} |m_{ii}|.
\]

We use \(|\cdot|\) to denote the absolute value of a real number as well as the cardinality of a finite set. The vectors \(0_d\) and \(1_d \in \mathbb{R}^d\) are \(d\) dimensional vectors with all elements equal to 0 and
respectively. The set of \( d \times d' \) matrices with orthonormal columns is denoted as \( \mathbb{O}_{d \times d'} \) while the set of \( d \times d \) orthogonal matrices is denoted as \( \mathbb{O}_d \).

For two given terms \( a \) and \( b \), let \( a \land b \equiv \min\{a, b\} \). We also write \( a \preceq b \) and \( a \succeq b \) if there exists a constant \( c \) not depending on \( a \) and \( b \) such that \( a \leq cb \) and \( a \geq cb \), respectively. If \( a \preceq b \) and \( a \succeq b \) then \( a \asymp b \). Finally, for random sequences \( A_n, B_n \), we write \( A_n = o_p(B_n) \) if \( A_n/B_n \to 0 \) in probability.

### 2. Summary of Node2vec and Stochastic Blockmodel Graphs

In this section we first provide a brief overview of the node2vec algorithm. We then discuss the popular stochastic blockmodel for random graphs. Finally we discuss a matrix factorization perspective to node2vec and show that, for a graph \( \mathcal{G} \) generated from a stochastic blockmodel, this matrix factorization approach leads to an approximate low-rank decomposition of a certain elementwise non-linear transformation of the random walk transition matrix for \( \mathcal{G} \).

#### 2.1. Node2vec with Negative Sampling

First introduced in Grover and Leskovec (2016), node2vec is a computationally efficient and widely-used algorithm for network embedding. Motivated by the ideas behind Word2vec embeddings in text documents (Mikolov et al., 2013b), node2vec generates nodes sequences as “textual corpuses of words” using random walks starting from every node. These random walks are then feed into a skip-gram model (Mikolov et al., 2013a) to yield the node embeddings. The original skip-gram model is quite computationally demanding for large networks and hence, in practice, usually replaced by a skip-gram model with negative sampling (SGNS). The resulting algorithm is briefly summarized below.

1. **(Sampling Random Paths):** The algorithm first generates \( r \) 2\textsuperscript{nd} order random walk on \( \mathcal{G} \) with each random walk having a fixed length \( L \). In this paper we assume that the starting vertex of each random walk is sampled according to a stationary distribution \( S = (S_i)_{n} \in \mathbb{R}^n \) on \( \mathcal{G} \) with

   \[
P(\text{Starting Vertex is } v_i) = S_i = \frac{d_i}{2|E|}, \forall v_i \in V.
\]

   We denote by \( r_i \) the number of random walks starting from vertex \( v_i \), \( l_j^{(i)} \) as the \( j \)th random walk starting from \( v_i \) and \( \mathcal{L}_i = \{l_j^{(i)} ; j \in [r_i]\} \) as the set of all random walks starting from \( v_i \).

   We now describe the notion of a 2\textsuperscript{nd} order random walk on \( \mathcal{G} \). Let \( p > 0 \) and \( q > 0 \). A 2\textsuperscript{nd} order random walk of length \( L \) starting at \( v_i \) with parameters \( p \) and \( q \) is generated as follows. Let \( v_1^{(i,j)} = v_i \). Next sample \( v_2^{(i,j)} \) from \( \mathcal{N}(v_1^{(i,j)}) \) uniformly at random. For
3 \leq l \leq L$, we sample $v_l^{(i,j)}$ from $\mathcal{N}(v_{l-1}^{(i,j)})$ with probability,

$$P(v_l^{(i,j)} = v_0) = \begin{cases} \frac{1}{p} \cdot J(v_0) & \text{if } v_0 = v_{l-2}^{(i,j)} \\ J(v_0) & \text{if } v_0 \in \mathcal{N}(v_{l-2}^{(i,j)}) \cap \mathcal{N}(v_{l-1}^{(i,j)}) \\ \frac{1}{q} \cdot J(v_0) & \text{if } v_0 \in \mathcal{N}(v_{l-2}^{(i,j)})^c \cap \mathcal{N}(v_{l-1}^{(i,j)}) \\
\end{cases}$$

$$J(v_0) = \left(\frac{1}{p} + |\mathcal{N}(v_{l-1}^{(i,j)}) \cap \mathcal{N}(v_{l-1}^{(i,j)})| + \frac{1}{q} \cdot |\mathcal{N}(v_{l-2}^{(i,j)})^c \cap \mathcal{N}(v_{l-1}^{(i,j)})|\right)^{-1}$$

We then let $l_j^{(i)} = (v_1^{(i,j)}, v_2^{(i,j)}, \ldots, v_L^{(i,j)})$. The form of $J(v_0)$ allows for $v_l^{(i,j)}$ to have possibly unbalanced probabilities of reaching three different types of nodes in the neighborhood of $v_{l-1}^{(i,j)}$, namely: (1) the previous node $v_{l-2}^{(i,j)}$; (2) nodes belonging to both the neighborhoods of $v_{l-2}^{(i,j)}$ and $v_{l-1}^{(i,j)}$; i.e., nodes in $\mathcal{N}(v_{l-2}^{(i,j)}) \cap \mathcal{N}(v_{l-1}^{(i,j)})$; (3) nodes belonging only to the neighborhood of $v_{l-1}^{(i,j)}$ but not the neighborhood of $v_{l-2}^{(i,j)}$, i.e., nodes belonging to $\mathcal{N}(v_{l-2}^{(i,j)})^c \cap \mathcal{N}(v_{l-1}^{(i,j)})$. The parameters $p$ and $q$ control the probability weights for these three different type of nodes and hence control the speed at which the random walk leaves the neighborhood of the original node $v_i$.

**Remark 1** In this paper we consider only the case of $(p, q) = (1, 1)$ for our theoretical analysis. The choice $(p, q) = (1, 1)$ is the default setting for node2vec as suggested in the original paper of Grover and Leskovec (2016). Furthermore, if $(p, q) = (1, 1)$ then the sampling strategy of node2vec is equivalent to that of DeepWalk (Perozzi et al., 2014) and so the analysis presented here will also apply to DeepWalk.

2. **(Calculating C)**: Borrowing the ideas from Word2vec in Mikolov et al. (2013b), node2vec treats nodes as “words” and “contexts” and creates $C = [C_{iv}]_{n \times n}$ as a node-context matrix. The $ii$th entry of $C$ records the number of times the pair of nodes $(v_i, v_j)$ appears among all random paths in $\bigcup_{i=1}^{n} \mathcal{L}_i$. More specifically, for a given window size $(t_L, t_U)$, $C_{iv}$ is the number of times that $(v_i, v_j)$ appears within a sequence

$$\ldots, v_i, \ldots, v_{i'}, \ldots \text{ or } \ldots, v_{i'}, \ldots, v_i, \ldots$$

for any $t_L \leq t \leq t_U \leq L - 1$ (2.1) among all random paths in $\bigcup_{i=1}^{n} \mathcal{L}_i$.

**Remark 2** The original node2vec algorithm fixed $t_L = 1$ while in this paper we allow for varying $t_L$. In Section 3 we show that different values for $(t_L, t_U)$ could lead to different convergence rates for the embedding and furthermore appropriate values for $(t_L, t_U)$ depend intrinsically on the sparsity of the network.

3. **(Skip-gram model with negative sampling)**: Given the $n \times n$ matrix $C$ and an embedding dimension $d$, node2vec uses the SGNS model to learn the node embedding matrix $F \in \mathbb{R}^{n \times d}$ and the context embedding matrix $F' \in \mathbb{R}^{n \times d}$. The $i$th row of $F$ is the $d$-dimensional embedding vector of node $v_i$. In slight contrasts to the original node2vec, in this paper we do not require the constraint $F = F'$. The objective
function of SGNS model for a given $C$ as a function of $(F, F')$ is defined as

$$g(F, F') = \sum_{i,i'} C_{ii'} \cdot \left( \log (\sigma(f_i^T \cdot f'_{i'})) + k \cdot \mathbb{E}_{f'_N \sim P_N} \left( \log \sigma(-f_i^T \cdot f'_N) \right) \right) \tag{2.2}$$

where $k$ is the ratio of negative samples to positive samples, $P_N(f'_N) = \frac{\sum_{i,i'} C_{ii'} f'_{i}}{\sum_{i,i'} C_{ii'}}$ is the empirical unigram distribution for the negative samples, and $\sigma$ is the logistic function. The original node2vec algorithm solves for $(\hat{F}, \hat{F}')$ by minimizing Eq. (2.2) over $(F, F')$ using Stochastic Gradient Descent (SGD). In this paper we use a matrix factorization approach, described in section 2.3, for obtaining $(\hat{F}, \hat{F}')$.

### 2.2. Stochastic Block Model

The stochastic blockmodel of Holland et al. 1983 is one of the most popular generative model for network data and our theoretical analysis of node2vec and DeepWalk is situated in the context of this model. A $K$-blocks stochastic block graph is defined by two parameters $(B, \Theta)$ where $B = [B_{ii'}] \in [0,1]^{K \times K}$ is a symmetric matrix of blocks connectivity and $\Theta \in \{0,1\}^{n \times K}$ is a matrix whose rows denote the block assignments for the nodes; we use $k(i) \in [K]$ to represent the community assignment for node $i$, i.e., the $i$th row of $\Theta$ contains a single 1 in the $k(i)$th element and 0 everywhere else. Given $B$ and $\Theta$, the edges $a_{ii'}$ of $G$ are independent Bernoulli random variables with $\mathbb{P}[a_{ii'} = 1] = B_{k(i), k(i')}$, i.e., the probability of connection between $i$ and $i'$ depends only on the communities assignment of $i$ and $i'$. Denote by $P = [p_{ii'}] = \Theta B \Theta^T$ the matrix of edge probabilities. We denote a graph with adjacency matrix $A$ sampled from a stochastic blockmodel as $A \sim \text{SBM}(B, \Theta)$, and, for any stochastic blockmodel graph, we denote by $n_k$ the number of vertices assigned to block $k$, i.e., $n_k = |\{i | k(i) = k\}|$.

We shall assume, for ease of exposition and without loss of generality, that $\Theta$ and $P$ satisfy

$$\Theta = \begin{pmatrix} 1_{n_1} & 0 & \ldots & 0 \\ 0 & 1_{n_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1_{n_K} \end{pmatrix}, \quad P = \begin{pmatrix} B_{11} 1_{n_1} 1_{n_1}^T \ldots & B_{1K} 1_{n_1} 1_{n_K}^T \\ \vdots & \vdots & \ddots & \vdots \\ B_{K1} 1_{n_K} 1_{n_1}^T \ldots & B_{KK} 1_{n_K} 1_{n_K}^T \end{pmatrix}. \tag{2.3}$$

In particular, $P$ has a $K \times K$ block structure and $\text{Rank}(P) \leq K$.

### 2.3. Node2vec and Matrix Factorization

For a fixed given embedding dimension $d < n$, minimization of the objective function in Eq. (2.2) leads to a non-convex optimization problem and the potential convergence of SGD into local minima makes the asymptotic analysis of $\hat{F}$ quite complicated. We thus desire a different approach for finding $\hat{F}$, e.g., one for which the form of $\hat{F}$ is more readily apparent. One such approach is the use of matrix factorization. For example, in the context of word2vec embedding, Levy and Goldberg (2014) showed that minimization of Eq. (2.2) when $C$ is a word-context matrix is equivalent to a matrix factorization problem on some elementwise non-linear transformation of $C$ and that this non-linear transformation can be related to the notion of pointwise mutual information between the words. Motivated
by this line of inquiry, we consider an alternative formulation of node2vec wherein \( \hat{\mathbf{F}} \hat{\mathbf{F}}^T \) is a low-rank approximate factorization of some elementwise non-linear transformation \( \hat{\mathbf{M}} \) of \( \mathbf{W} \); recall that \( \mathbf{W} \) is the 1-step transition matrix for the canonical random walk on \( \mathcal{G} \). We now describe the matrix \( \hat{\mathbf{M}} \) together with the corresponding low-rank approximation \( \hat{\mathbf{F}} \hat{\mathbf{F}}^T \). For ease of exposition, we first describe the matrix factorization in the case when \( d \geq n \). We emphasize that this matrix factorization perspective of node2vec and other random-walk based algorithms had been considered previously in Qiu et al. (2018). The main contribution of our paper is in showing that the resulting matrix factorization leads to consistent recovery of blocks or communities in stochastic blockmodel graphs.

**Case 1** \( (d \geq n) \): In the context of the word2vec algorithm Levy and Goldberg (2014) showed that there exists an embedding dimension \( d \geq n \) such that the minimizer of Eq. (2.2) over \( \mathbf{F} \in \mathbb{R}^{n \times d} \) and \( \mathbf{F}' \in \mathbb{R}^{n \times d} \) satisfy

\[
\hat{\mathbf{F}} \cdot \hat{\mathbf{F}}^T = \left[ \log \left( \frac{C_{ii'} \cdot (\sum_i C_{ii'})}{\sum_i C_{ii'} \cdot \sum_i' C_{ii'}} \right) \right]_{n \times n} - \log(k) \mathbf{1} \cdot \mathbf{1}^T \equiv \hat{\mathbf{M}}(C, k),
\]

and hence \( (\hat{\mathbf{F}}, \hat{\mathbf{F}}') \) could be viewed as a solution of a matrix factorization problem. Using the same idea for our analysis of node2vec, we first fixed \( n \) and show that as the number of sampled random paths starting at each node goes to \( \infty \), \( \hat{\mathbf{M}}(C, k) \) converges, elementwise, to a limiting matrix \( \hat{\mathbf{M}}_0 \) almost surely, provided that the following assumption is satisfied.

**Assumption 1** There exists \( (t_L, t_U) \) such that the entries of \( \sum_{t=t_L}^{t_U} \hat{\mathbf{W}}^t \) are all positive, i.e., that \( \mathcal{G} \) is connected.

**Theorem 3** Applying node2vec sampling strategy introduced in Section 2.1 on a fixed network \( \mathcal{G} = (V, E) \), if \( (t_L, t_U) \) satisfies Assumption 1, we have

\[
\hat{\mathbf{M}}(C, k) \xrightarrow{a.s.} \log \left[ \frac{2|A|}{\gamma(L, t_L, t_U)} \sum_{t=t_L}^{t_U} (L - t) \cdot \left( D_A^{-1} \mathbf{W}^t \right) \right] - \log(k) \mathbf{1} \cdot \mathbf{1}^T \equiv \hat{\mathbf{M}}_0(\mathcal{G}, t_L, t_U, k, L)
\]

(2.5)

as \( r \to \infty \) for any fixed but arbitrary \( n \). The convergence of the matrix \( \hat{\mathbf{M}}(C, k) \) to the matrix \( \hat{\mathbf{M}}_0 \) is element-wise and uniform over all the elements of \( \hat{\mathbf{M}}(C, k) \). Here \( |A| \) denote the sum of the entries in \( A \) and the constant \( \gamma \) is defined as \( \gamma(L, t_L, t_U) = \frac{(2L - t_L - t_U)(t_U - t_L + 1)}{2} \).

Combining Eq. (2.4) and Theorem 3 we have that, for any fixed \( n \), there exists an embedding dimension \( d \geq n \) such that for \( r \to \infty \), the matrices \( \hat{\mathbf{F}} \) and \( \hat{\mathbf{F}}' \) are approximate factors for factorizing \( \hat{\mathbf{M}}_0(\mathcal{G}, t_L, t_U, k, L) \). To reduce notation clutter, we will henceforth drop the dependency of \( \hat{\mathbf{M}}_0 \) on the parameters \( \mathcal{G}, t_L, t_U, k, L \). Note that \( D_A^{-1} \mathbf{W}^t \) is symmetric for any \( t \geq 1 \) and hence \( \hat{\mathbf{M}}_0 \) is symmetric. Next, since the value of \( r \) is chosen purely for ease of computation, i.e., smaller values of \( r \) require sampling fewer random walks, we will thus take the conceptual view that \( r \) is arbitrarily large so that \( (\hat{\mathbf{F}}, \hat{\mathbf{F}}') \) are exact factors for factorizing \( \hat{\mathbf{M}}_0 \) when \( d \geq n \).

The previous discussion assumes that the embedding dimension \( d \geq n \), which is useless in practice. In contrasts, if \( d < n \) then exact factors \( (\hat{\mathbf{F}}, \hat{\mathbf{F}}') \) for factorizing \( \hat{\mathbf{M}}_0 \) might no longer exists. However, the requirement that \( (\hat{\mathbf{F}}, \hat{\mathbf{F}}') \) is an exact factor is generally misleading. Indeed, taking the view that the observed graph is a single noisy sample generated according
to some true but unobserved edge probabilities matrix $P$, what we really want to recover is the factorization induced by $P$. More specifically, replacing $\hat{W}^t$ and $|A|$ with $W^t$ and $|P|$ in $\tilde{M}_0(G, t_L, t_U, k, L)$, we define

$$M_0 = M_0(P, t_L, t_U, k, L) \equiv \log \left[ \frac{2|P|}{\gamma(L, t_L, t_U)} \sum_{t=t_L}^{t_U} (L-t) \cdot \left( D_P^{-1} W^t \right) \right] - \log(k) 1 \cdot 1^T \quad (2.6)$$

as the counterpart of $\tilde{M}_0$ using the true but unknown edge probabilities matrix $P$.

Recall from Eq. (2.3) that for stochastic blockmodel graphs, the matrix $P$ has a $K \times K$ block structure. Thus both $W^t$ and $D_P^{-1} W^t$ also have $K \times K$ block structures. Eq. (2.6) then implies that $M_0$ also has a $K \times K$ block structure and hence $\text{rank}(M_0) \leq K$. Most importantly, the $K \times K$ block structure of $M_0$ is also sufficient for recovering the community structure in $G$. We will show in Section 3 that the relative error, in Frobenius norm, between $\tilde{M}_0$ and $M_0$ converges to 0 as $n \to \infty$. This convergence, together with the Davis-Kahan theorem for perturbation of eigenspaces, implies the existence of an embedding dimension $d \leq K$ for which the $n \times d$ matrices $\hat{F}$ and $\hat{F}'$ obtained by factorizing $\tilde{M}_0$ also lead to consistent recovery of the community structure in $G$. We emphasize, however, that if $P$ does not arise from a stochastic blockmodel graph then $M_0$ need not have a low-rank structure. Nevertheless we can still consider a rank-$d$ approximation to $M_0$ for some $d < \text{rk}(M_0)$. Furthermore, as we will clarify in Section 5, the bound for $||M_0 - \tilde{M}_0||_F$ in Section 3 also holds for general edge independent random graphs, provided that the entries of $P$ is reasonably homogeneous. Hence if $\tilde{M}_0$ has an approximate low-rank structure then $M_0$ also has an approximate low-rank structure, and vice versa.

In summary, motivated by the low-rank structure of the matrix $M_0$ in the case of stochastic blockmodel graphs, we view the matrix factorization approach for node2vec as equivalent to finding, given a $d < n$, the best rank $d$ approximation $\hat{F} : \hat{F}'^T$ of $M_0$ under Frobenius norm, i.e.,

$$(\hat{F}, \hat{F}') = \arg\min_{(F,F') \in \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d}} ||M_0 - F \cdot F'^T||_F. \quad (2.7)$$

The minimizer of Eq. (2.7) is obtained by truncating the SVD of $\tilde{M}_0$. More specifically, let

$$\hat{M}_0 = \hat{U} \Sigma \hat{V}^T$$

with a decreasing order of singular values in $\hat{\Sigma}$. Then for a given $d \leq \text{rk}(M_0)$, let

$$\hat{F} = \hat{U}_d, \hat{F}' = \hat{V}_d \hat{\Sigma}_d$$

where $U_d \in \mathbb{R}^{n \times d}, V_d \in \mathbb{R}^{n \times d}$ are the first $d$ columns of $U$ and $V$, respectively, and $\hat{\Sigma}_d \in \mathbb{R}^{d \times d}$ is the diagonal matrix containing the $d$ largest singular values in $\hat{\Sigma}$.

**Remark 4** We emphasize that the appropriate embedding dimension $d$ for factorizing $\tilde{M}_0$ depends on knowing the rank of the matrix $M_0$. In contrasts, the convergence of $\tilde{M}_0$ to that of $M_0$ does not require knowing $d$. For ease of exposition our subsequent analysis for the factorization of $\tilde{M}_0$ will assume that $d$ is known; in practice $d$ can be estimated consistently using an eigenvalue thresholding procedure provided that $M_0$ has a low-rank
structure. Finally, in the context of stochastic blockmodel graphs and their variants, the recovery of the community assignment using \( \hat{\Theta} \) also depends on knowing \( K \). For simplicity we will also assume that \( K \) is known; consistent estimates for \( K \) are provided in Lei (2016); Sarkar and Bickel (2016).

3. Theoretical Results

In this section we provide results for the convergence rate of the embedding \( \hat{\mathcal{F}}_0 \), obtained by factorizing the matrix \( \tilde{M}_0 \) defined in Eq. (2.5), to the ground truth as the number of vertices in \( \mathcal{G} \) increases. For ease of exposition we first consider, in Section 3.1, the dense regime where the entries of the edge probabilities matrix \( P \) are bounded away from 0. The bounds derived in this regime are simpler to state and, for the casual reader, still include all the main conceptual insights of the paper. The results for the sparse regime, while similar to that of the dense regime, require substantially more technical machinery and involved derivations and will be presented in Section 3.2.

3.1. Dense Regime

This regime is characterized by the following assumption on the entries of \( P \).

**Assumption 2 (Dense Regime)** \( \mathcal{G} \) is generated from \( \text{SBM}(\Theta, B) \) with \( K \) communities and there exists \( c_0 > 0 \) not depending on \( n \) such that \( B_{k,\ell} > c_0 \) for all \( 1 \leq k \leq \ell \leq K \).

Recall that \( \hat{\mathcal{F}}_0 \) is obtained from computing the eigendecomposition of \( \tilde{M}_0 \) while the true embedding is obtained from the eigendecomposition of the matrix \( M_0 \) defined in Eq. (2.6). We therefore first consider the convergence of \( \tilde{M}_0 \) to \( M_0 \). This convergence is facilitated by the following Theorem 5 which provides a uniform bound for the entrywise difference between the \( t \)-step transition matrix \( \hat{W}^t \) defined using the adjacency matrix \( A \) and the \( t \)-step transition matrix \( W^t \) defined using the unobserved edge probabilities matrix \( P \).

**Theorem 5** When \( \mathcal{G} \) satisfies Assumption 2, we have, for sufficiently large \( n \),

\[
\|W - \hat{W}\|_{\text{max}} = O_P\left(n^{-1}\right), \tag{3.1}
\]

\[
\|W^2 - \hat{W}^2\|_{\text{max, diag}} = \max_i |w_{ii}^{(2)} - \hat{w}_{ii}^{(2)}| = O_P\left(n^{-1}\right),
\]

\[
\|W^2 - \hat{W}^2\|_{\text{max, off}} = \max_{i \neq i'} |w_{ii'}^{(2)} - \hat{w}_{ii'}^{(2)}| = O_P\left(n^{-3/2} \sqrt{\log n}\right). \tag{3.2}
\]

Furthermore, for any fixed \( t \geq 3 \),

\[
\|W^t - \hat{W}^t\|_{\text{max}} = O_P\left(n^{-3/2} \sqrt{\log n}\right). \tag{3.3}
\]

We then have, from Theorem 5 together with an entrywise Taylor series expansion for \( \log x \), the following bound for \( \|\tilde{M}_0 - M_0\|_F \).

**Theorem 6** Suppose Assumption 2 holds and \( t_U \geq t_L \geq 2 \). Then \( \tilde{M}_0 \) is well-defined with high probability. Moreover, we have

\[
\|\tilde{M}_0 - M_0\|_F = O_P(n^{1/2} \sqrt{\log n}), \tag{3.4}
\]

10
than the entries of $\hat{\Theta}$ and $\tilde{\Theta}$ as implied by Theorem 6. One particular Eq. (3.1) and Eq. (3.2) show that the entries of $\hat{\Theta}$ and $\tilde{\Theta}$ might lead to a convergence rate of $\tilde{\Theta}$ that is slower than that given in Eq. (3.4). In particular, Eq. (3.1) and Eq. (3.2) show that the entries of $\tilde{\Theta} - \tilde{\Theta}$ are of larger magnitude than the entries of $\tilde{\Theta} - \tilde{\Theta}$ for $t \geq 2$. Theorem 6 implies that $\hat{\Theta}$ is close to $\Theta_0$, i.e., the relative Frobenius norm error satisfies $\|\hat{\Theta} - \Theta_0\|_F^2/\|\Theta_0\|_F^2 = o_p(1)$ for sufficiently large $n$. Now, by Eq. (2.3), $\Theta_0$ has a $K \times K$ block structure and hence $\text{rk}(\Theta_0) \leq K$ and furthermore the eigenvectors of $\Theta_0$ associated with the $d = \text{rk}(\Theta_0)$ largest eigenvalues of $\Theta_0$ are sufficient for recovering the community assignments induced by $\Theta$. Remark 7 Throughout this paper we assume that $t_L \geq 2$ as opposed to $t_L \geq 1$ as in the original node2vec formulation. Recall the definition of $\hat{\Theta}_0$ in Eq. (2.5). For the dense regime if we allow $t$ to start from 1 in the sum $\sum_{t=L}^{t}(L - t)\cdot (D^{-1}\tilde{\Theta})^t$ then the term $\tilde{\Theta}$ might lead to a convergence rate of $\tilde{\Theta}$ that is slower than that given in Eq. (3.4). In particular, Eq. (3.1) and Eq. (3.2) show that the entries of $\tilde{\Theta} - \tilde{\Theta}$ are of larger magnitude than the entries of $\tilde{\Theta} - \tilde{\Theta}$ for $t \geq 2$. Theorem 6 implies that $\hat{\Theta}_0$ is close to $\Theta_0$, i.e., the relative Frobenius norm error satisfies $\|\hat{\Theta}_0 - \Theta_0\|_F^2/\|\Theta_0\|_F^2 = o_p(1)$ for sufficiently large $n$. Now, by Eq. (2.3), $\Theta_0$ has a $K \times K$ block structure and hence $\text{rk}(\Theta_0) \leq K$ and furthermore the eigenvectors of $\Theta_0$ associated with the $d = \text{rk}(\Theta_0)$ largest eigenvalues of $\Theta_0$ are sufficient for recovering the community assignments induced by $\Theta$. The following result, which follows from Theorem 6 together with the Davis-Kahan theorem (Davis and Kahan, 1970; Yu et al., 2015), shows that the embedding $\hat{\Theta}$ given by the $d$ largest eigenvectors of $\Theta_0$ is approximately the same as that given by the $d$ largest eigenvectors of $\Theta_0$. 

**Corollary 8** Under the condition of Theorem 6, let $\hat{\Theta}_0\hat{\Phi}_0$ and $\Theta_0\Phi_0$ be the eigendecomposition of $\Theta_0$ and $\Theta_0$, respectively. Let $d = \text{rk}(\Theta_0)$ and note that $\Phi$ is a $n \times d$ matrix. Let $\hat{\Phi} = \hat{\Phi}_d$ be the matrix formed by the columns of $\hat{\Theta}_0$ corresponding to the $d$ largest eigenvalues of $\Theta_0$. We then have, for sufficiently large $n$,

$$
\min_{\mathbf{T} \in \mathbb{B}_d} \| \hat{\Phi} \cdot \mathbf{T} - \Phi \|_F = O_p \left( n^{-1/2} \sqrt{\log n} \right).
$$

We note that each column of $\Phi$ has unit norm and that there are only $K$ distinct rows in $\Phi$. Let $U_i$ denote the $i$th row of $\Phi$. Then $\|U_i - U_i\| = \Omega(n^{-1/2})$ whenever $k(i) \neq k(i')$; recall that $k(i) \in \{1, 2, \ldots, K\}$ denote the block assignment for node $i$. Eq. (3.5) implies that there exists an orthogonal $\mathbf{T}$ such that on average, $\| \hat{\Phi} \cdot \mathbf{T} - U_i \| = O(n^{-1/2}\sqrt{\log n})$. The number of vertices $i$ for which

$$
\| \hat{\Phi} \cdot \mathbf{T} - U_i \| \geq \min_{i' : k(i') \neq k(i)} \| \hat{\Phi} \cdot \mathbf{T} - U_{i'} \|
$$

is therefore of order $O(\log n)$. In other words, clustering the rows of $\hat{\Theta}$ will recover the block assignments $k$ for all but $O(\log n)$ vertices.

**3.2. Sparse Regime**

We now extend the results in Section 3.1 to the sparse regime where the edge probabilities could decrease to 0 as the number of vertices increases. For ease of exposition and to avoid tedious bookkeeping, we shall assume that the edges probabilities are all of the same order as in the following assumption.

**Assumption 3 (Sparse Regime)** $G$ is generated from SBM$(\Theta, B)$ with $K$ communities. There exists a $\beta \in [0, 1)$ such that for all sufficiently large $n$, we have $B_{uv} \asymp \rho_n \asymp n^{-\beta}$ for all entries $B_{uv}$ of $B$. 


It is well known that, for sufficiently large $n$, if $G$ satisfies Assumption 3 then $G$ is connected with high probability (see, e.g., Abbe (2017)). Analogous to Theorem 5, we have the following theorem to give convergence rate of $t$ step transition matrix $\hat{W}^t$ under sparse regime.

**Theorem 9** Let $G$ be a graph on $n$ vertices that is sampled from a model satisfying Assumption 3. Let $t \geq 4$ satisfies $\frac{t-3}{t-1} > \beta$, then we have, for sufficiently large $n$,

$$\|\hat{W}^t - W^t\|_{\max} = O_P(n^{-3/2} \rho_n^{-1/2} \sqrt{\log n}). \tag{3.6}$$

Furthermore, if $0 \leq \beta < \frac{1}{2}$, then we have, for sufficiently large $n$,

$$\|\hat{W}^3 - W^3\|_{\max} = O_P(n^{-3/2} \rho_n^{-1} \sqrt{\log n})$$

$$\|\hat{W}^2 - W^2\|_{\max, \text{off}} = O_P(n^{-3/2} \rho_n^{-1} \sqrt{\log n}). \tag{3.7}$$

**Remark 10** To keep the statement of Theorem 9 simple we have omitted the following more general but weaker bounds

$$\|\hat{W}^2 - W^2\|_{\max, \text{diag}} = \max_{i=1} \left| W^{(2)}_{ii} - \hat{W}^{(2)}_{ii} \right| = O_P((n \rho_n)^{-1}),$$

$$\|\hat{W}^2 - W^2\|_{\max, \text{off}} = O_P(\max\{n^{-3/2} \rho_n^{-1} \sqrt{\log n}, (n \rho_n)^{-2} \sqrt{\log n}\}),$$

$$\|\hat{W}^t - W^t\|_{\max} = O_P(\max\{n^{-3/2} \rho_n^{-1} \sqrt{\log n}, (n \rho_n)^{-2} \sqrt{\log n}\}),$$

when $t \geq 3$. While we do not use these bounds in our subsequent analysis, we note that they do not depend on the values of $\beta$ and might thus be of independent interest.

Before discussing the convergence rate of $\hat{M}_0$ to $M_0$ we should first find a value of $t_U$ such that, for sufficiently large $n$, $M_0$ is well defined with high probability. Looking at Eq. (3.6) we see that the entries of $W^t$ are bounded uniformly away from $0$ whenever $(t-3)/(t-1) \geq \beta$. Indeed, the entries of $\{W_t^t\}_{t \geq 1}$ are uniformly of order $\Theta(n^{-1})$. By Eq. (3.6), if $(t-3)/(t-1) > \beta$ then the entries of $\hat{W}^t$ are uniformly of order $\Omega(n^{-1} - n^{-3/2} \rho_n^{1/2} \sqrt{\log n}) = \Omega(n^{-1})$ with high probability. Now recall that the matrix $\hat{M}_0$ is of the form

$$\log \left\{ \frac{2|A|}{\gamma(L, t_L, t_U)} \sum_{t=t_L}^{t_U} (L - t) \cdot \left( D_A^{-1} \hat{W}^t \right) \right\}$$

We therefore have, for $(t_U - 3)/(t_U - 1) > \beta$, that the entries of the inner sum are bounded away from $0$ with high probability. Therefore, with high probability, the elementwise logarithm is well-defined for all entries of $\hat{M}_0$, i.e., $\hat{M}_0$ is well-defined.

Given the existence of $\hat{M}_0$, the following result shows the convergence rate of $\hat{M}_0$ to $M_0$.

**Theorem 11** For $G$ generated under the Assumption 3 we could choose $t_U \geq t_L \geq 2$. Then with high probability, $\hat{M}_0$ is well defined and

$$\|\hat{M}_0 - M_0\|_F = \begin{cases} O_P(n^{1/2} \rho_n^{-1} \sqrt{\log n}) & \text{when } t_L \geq 2 \text{ and } \beta < \frac{1}{2}, \\ O_P(n^{1/2} \rho_n^{-1/2} \sqrt{\log n}) & \text{when } t_L \geq 4 \text{ and } \beta < \frac{t_L-3}{t_L-1}. \end{cases} \tag{3.8}$$

Moreover, for sufficiently large $n$ we have $\|M_0\|_F \asymp n$ and hence, for sufficiently large $n$, $\|\hat{M}_0 - M_0\|_F \ll \|M_0\|_F$ with high probability.
Theorem 11 indicates that as $\beta$ increases (equivalently, as $\rho_n$ decreases) so that the graph $\mathcal{G}$ becomes sparser, we could (1) still guarantee the existence of $\check{M}_0$ by increasing $t_U$ and (2) control the convergence rate of $\|\check{M}_0 - M_0\|_F$ relative to $\|M_0\|_F$ by increasing $t_L$.

Analogous to Corollary 8, we now state the convergence rate of $\check{F}$ given by the $d$ largest eigenvectors of $\check{M}_0$ to the $d$ largest eigenvectors of $M_0$, under the sparse regime.

**Corollary 12** Assume the setting in Theorem 11. Let $\check{U}\Sigma\check{U}^\top$ and $U\Sigma U^\top$ be the eigendecomposition of $\check{M}_0$ and $M_0$, respectively. Let $d = \text{rk}(M_0)$ and note that $U$ is a $n \times d$ matrix. Let $\check{F} = \check{U}_d$ be the matrix formed by the columns of $\check{U}$ corresponding to the $d$ largest eigenvalues of $\check{M}_0$. We then have, for sufficiently large $n$,

$$
\min_{T \in \mathcal{O}_d} \|\check{F} \cdot T - U\|_F = \begin{cases} O_{\mathbb{P}}\left(n^{-1/2}\rho_n^{-1}\sqrt{\log n}\right) & \text{when } t_L \geq 2 \text{ and } \beta < \frac{1}{2}, \\
O_{\mathbb{P}}\left(n^{-1/2}\rho_n^{-1/2}\sqrt{\log n}\right) & \text{when } t_L \geq 4 \text{ and } \beta < \frac{t_U-3}{t_U-1}.\end{cases}
$$

**4. Simulation**

We now present a collection of simulations experiments for the matrix factorization approach to node2vec and DeepWalk embedding. These experiments complement our theoretical results in Section 3; in particular they illustrate the interplay between the sparsity of the graphs and the choice of window sizes and their effects on the resulting embedding.

**4.1. Error bound for $\|\check{M}_0 - M_0\|_F$**

In this section we perform numerical experiments to illustrate the large sample behavior of $\|\check{M}_0 - M_0\|_F$ as suggested in Theorem 6 and 9. We shall simulate random undirected graphs generated from a 2 block SBM with parameters

$$
B(\rho_n) \equiv \begin{pmatrix} 0.8\rho_n & 0.3\rho_n \\ 0.3\rho_n & 0.8\rho_n \end{pmatrix}, \quad \pi = (0.4, 0.6).
$$

where the sparsity parameter $\rho_n$ are set to $\rho_n \in \{1, 3n^{-1/3}, 3n^{-1/2}, 3n^{-2/3}\}$. We emphasize that while this 2 blocks setting is quite simple it nevertheless still displays the effect of the sparsity $\rho_n$ and the window size $(t_L, t_U)$ on the upper bound for $\|\check{M}_0 - M_0\|_F$.

For each value of $n$ and sparsity parameter $\rho_n$, we run 100 independent replications. For each replication we generate a simple graph $\mathcal{G} \sim \text{SBM}(B(\rho_n), \Theta_n)$ and calculate $\check{M}_n$ for different choices of window size $(t_L, t_U)$. In particular we consider two types of window size, namely $t_U - t_L = 1$ and $t_U - t_L = 3$. While window sizes with $t_U - t_L = 1$ are not commonly used in practice, for simulation purpose this choice clearly show the effects of increase in the random walks’ length $t$ on the error bound for $\|\check{M}_0 - M_0\|_F$. This is in contrast to the choice of $t_U - t_L = 3$ which is more realistic but also harder to see the effect of increasing $t$ on $\|\check{M}_0 - M_0\|_F$. Recall that, from the discussion in Section 3.2, sparser values of $\rho_n$ requires larger values of $t_U$ to guarantee that $M_0$ is well-defined. We now summarize the combinations of $(\rho_n, n, (t_L, t_U))$ in the simulation below.

- When $\rho_n \in \{1, 3n^{-1/3}, 3n^{-1/2}\}$, we choose $n \in \{100, 200, 300, \ldots 1500\}$. For each $\rho_n$ and $n$, we calculate $\check{M}_0$ with $2 \leq t_L \leq 7$ when $t_U - t_L = 1$ and $2 \leq t_L \leq 5$ when $t_U - t_L = 3$. 
When \( \rho_n = 3n^{-2/3} \), we choose \( n \in \{800, 900, 1000, \ldots, 4000\} \). For each \( \rho_n \) and \( n \), we calculate \( \bar{M}_0 \) with \( 4 \leq t_L \leq 7 \) when \( t_U - t_L = 1 \) and \( 3 \leq t_L \leq 5 \) when \( t_U - t_L = 3 \).

We calculate two relative error criteria for \( \| \bar{M}_0 - M_0 \|_F \), namely

\[
\text{Error}_1 = \frac{\| \bar{M}_0 - M_0 \|_F}{\| M_0 \|_F}, \quad \text{and} \quad \text{Error}_2 = \frac{\| \bar{M}_0 - M_0 \|_F}{n^{1/2} \rho_n^{1/2} \sqrt{\log(n)}}.
\]

We expect that the first criteria converges to 0 as \( n \) increases while the second criteria remains bounded as \( n \) increases.

**Relative Error 1:** We first confirm the convergence of \( \| \bar{M}_0 - M_0 \|_F/\| M_0 \|_F \) to 0. Figure 1 and 2 shows the means and 95% confidence intervals of simulated \( \| \bar{M}_0 - M_0 \|_F/\| M_0 \|_F \) over 100 Monte Carlo replicates for different values of \( \rho_n, (t_L, t_U) \). Figure 1 and 2 indicates the following general patterns as supported by the theoretical results in Theorem 6 and Theorem 11.

- The relative error \( \| \bar{M}_0 - M_0 \|_F/\| M_0 \|_F \) are smallest in the dense case and deteriorates as the sparsity factor \( \rho_n \) decreases.
- The relative error depends on \( (t_L, t_U) \) as \( \rho_n \) decreases for increasing \( n \). Furthermore bigger values of \( t_U - t_L \) usually leads to smaller relative error.
- For a fixed window size \( (t_L, t_U) \), the relative error usually increases as \( \rho_n \) decreases. In particular when the window size is small, e.g., \( (t_L, t_U) = (2, 3) \) or \( (t_L, t_U) = (2, 5) \), the matrix \( \bar{M}_0 \) is often times not well-defined.

**Relative Error 2** Figure 1 and Figure 2 corroborate our theoretical results in Section 3. Nevertheless there are two additional questions we should consider. The first question is whether or not the error bound \( \| \bar{M}_0 - M_0 \| = O_p(n^{1/2} \rho_n^{-1/2} \sqrt{\log n}) \) in Theorem 11 is tight and, if it is tight, the second question is whether or not the restriction that \( t_L \) satisfies \( \beta \leq (t_L - 2)/t_L \) is necessary to achieve this error rate.

Analogous to the previous two figures, Figure 3 and 4 show the means and 95% empirical confidence intervals for the relative error \( \| \bar{M}_0 - M_0 \|_F/(n^{1/2} \rho_n^{-1/2} \sqrt{\log(n)}) \) over 100 Monte Carlo replicates for different values of \( \rho_n \) and \( (t_L, t_U) \). Based on these simulation results, we address the above questions as follows.

- For simulation settings with \( (t_L, t_U) \) and \( \rho_n \) satisfying \( \beta \leq \frac{t_L - 3}{t_L - 1} \), the relative error \( \text{Error}_2 \) converges to a constant as \( n \) increases. There is thus some evidence that the rate \( n^{1/2} \rho_n^{-1/2} \sqrt{\log n} \) is close to being optimal. However, when \( t_L \) is large relative to \( \rho_n \), e.g., \( (t_L, t_U) \in \{(6, 7), (7, 8)\} \) when \( \rho_n \in \{3n^{-1/3}, 3n^{-1/2}\} \), it appears that \( \text{Error}_2 \) is converging to 0 which suggests that for a given \( \rho_n \) the error rate could be smaller than \( n^{1/2} \rho_n^{-1/2} \sqrt{\log n} \) for large values of \( t_L \); this might be due to convergence of a \( t \)-step random walk toward the stationary distribution as \( t \) increases.
- For cases such as \( (t_L, t_U) \in \{(3, 4), (3, 6)\} \) and \( \rho_n = 3n^{-1/2} \) or \( (t_L, t_U) \in \{(4, 5), (3, 6)\} \) and \( \rho_n = 15n^{-2/3} \), the \( t \)-step’s do not satisfy the condition \( \beta \leq \frac{t_L - 3}{t_L - 1} \). Nevertheless the quantities \( \text{Error}_2 \) still appear to converge to a constant as \( n \) increases. This suggests
Figure 1: Sample means and 95% empirical confidence intervals for $\frac{\|\hat{M}_0 - M_0\|_F}{\|M_0\|_F}$ based on 100 Monte Carlo replicates for different values of $n, \rho_n, (t_L, t_U)$. Here we set $t_U - t_L = 1$. 
Figure 2: Sample means and 95% empirical confidence intervals for \( \|\tilde{M}_0 - M_0\|_F \) based on 100 Monte Carlo replicates for different settings of \( n, \rho_n, (t_L, t_U) \). Here we set \( t_U - t_L = 3 \).
that the condition $\beta \leq \frac{t_L - 3}{t_U - 1}$ is not necessary for the bound in Eq. (3.8) to hold. On the other hand, for fixed $n$ and $\rho_n$, the error $\|\hat{M}_0 - M_0\|_F$ generally decreases as the window sizes become larger.

$\begin{align*}
\text{(a) } \rho_n &= 1 \\
\text{(b) } \rho_n &= 3n^{-1/3} \\
\text{(c) } \rho_n &= 3n^{-1/2} \\
\text{(d) } \rho_n &= 15n^{-2/3}
\end{align*}$

Figure 3: Sample means and 95% empirical confidence intervals for $\frac{\|\hat{M}_n - M_0\|_F}{n^{1/2} \rho_n^{-1/2} \sqrt{\log n}}$ based on 100 Monte Carlo replicates for different values of $n$, $\rho_n$, and $(t_L, t_U)$ with $t_L - t_U = 1$. 
Figure 4: Sample means and 95% empirical confidence intervals for \( \| \tilde{M}_0 - M_0 \|_F \) based on 100 Monte Carlo replicates for different values of of \( n, \rho_n, \) and \((t_L, t_U)\) with \( t_L - t_U = 3.\)
Finally we note that if \((t_L, t_U) \in \{(2, 3), (2, 5)\}\) and \(\rho_n \in \{3n^{-1/3}, 3n^{-1/2}\}\), Error increases when \(n\) increases. This implies that there is phase transition of \(||\hat{M}_0 - M_0||_F\)'s convergence rate when \(t_L\) increases as implied by Theorem 11.

In summary, the simulation results in Figure 1 through Figure 4 numerically confirm the conclusion of Theorem 11 and also suggests that the error rate in Theorem 11 is sharp and that the condition \(\beta \leq \frac{t_L - 3}{t_L - 1}\) is sufficient but not necessary.

4.2. Embedding Performance

We have shown in Corollary 8 and 12, together with the simulation results in Section 4.1, that the convergence of \(\hat{\mathcal{F}}\) and its rate depends on \(\rho_n\) and the choice of \((t_L, t_U)\). However it is not clear, \textit{a priori}, if the convergence rate of \(\hat{\mathcal{F}}\) has a direct influence on the performance of node2vec on the downstream inference task.

In this section we perform numerical experiments to compare the performance of node2vec embedding for downstream community detection. We consider, in addition to the standard SBM, the degree-corrected SBM (Karrer and Newman, 2011). We will vary the window sizes \((t_L, t_U)\) and sparsity factor \(\rho_n\) in these simulations and investigate the effect of \((t_L, t_U)\) and \(\rho_n\) on the community detection accuracy.

More specifically, for each simulation with a specified value of \(n\) and \(\rho_n\), we run \(N = 100\) Monte Carlo replications. In each replication we apply node2vec with different window sizes on the simulated random graph to obtain the embeddings. Community detection is then done using \(K\)-means on the embedding. With the true cluster labels denoted as \(\{k(i)\}_{i=1}^n\) and the cluster labels estimated by \(K\)-means denoted as \(\{\hat{k}(i)\}_{i=1}^n\), we calculate the accuracy as (here \(P(\cdot)\) denote all possible permutations of \(\{1, 2, \ldots, K\}\))

\[
\text{Accuracy} = \min_{P(\cdot)} \frac{\#\{i|P(\hat{k}(i)) \neq k(i)\}}{n}.
\]

We also fix \(\rho_n = 3n^{-1/2}\), \(n = 600\) and sample one random realization from both the SBM and the degree-corrected SBM to illustrate the node2vec embeddings for different choices of window sizes. These visualizations, which are depicted in Figure 9 and Figure 10, provide us with more intuition about when increasing the window size could help separate nodes from different communities and thereby improve the community detection accuracy. We now describe the detailed settings of the network generation models used in these simulations.

**Stochastic Block Model:** We consider three-blocks stochastic blockmodels with block probabilities being either

\[
B_1 = \begin{pmatrix} 0.8 & 0.5 & 0.3 \\ 0.5 & 0.8 & 0.6 \\ 0.3 & 0.6 & 0.8 \end{pmatrix}, \quad \text{or} \quad B_2 = \begin{pmatrix} 0.8 & 0.5 & 0.5 \\ 0.5 & 0.8 & 0.5 \\ 0.5 & 0.5 & 0.8 \end{pmatrix}.
\]

We set the block assignment probabilities to \(\pi = (0.3, 0.3, 0.4)\) for both \(B_1\) and \(B_2\).

**Remark 13** Our rationale for choosing these specific values for \(B_1\) and \(B_2\) are as follows. When \(t \to \infty\), \(W^t\) converges to \(\frac{dp}{p^T}1^T\) where \(dp = (d_1, \ldots, d_n)^T \in \mathbb{R}^n\) are the expected
Figure 5: Community detection accuracy of node2vec followed by $K$-means for SBM graphs. The boxplots of the accuracy for each value of $n$, $\rho_n$, and $(t_L, t_U)$ are based on 100 Monte Carlo replications. The first and second row plot the results when the block probabilities for the SBM is $B_1$ and $B_2$, respectively.
Figure 6: Community detection accuracy of node2vec followed by K-means for DCSBM graphs. The boxplots of the accuracy for each value of $n, \rho_n$ and $(t_L, t_U)$ are based on 100 Monte Carlo replications. The first and second row plot the results when the block probabilities for the DCSBM is $B_1$ and $B_2$, respectively.
nodes’ degrees. Thus, as $t_L$ and $t_U$ increase, $D_p^{-1}W^t$ will get closer to $\mathbf{1}\mathbf{1}^T/|P|$ and $M_0$ will get closer to $C\mathbf{1}\mathbf{1}^T$ for some constant $C$; we note that the matrix $C\mathbf{1}\mathbf{1}^T$ contains no useful information for community detection.

Therefore, if $\pi$ is relatively balanced, the speed at which $M_0$ converges to $C\mathbf{1}\mathbf{1}^T$ as $t_L$ increases is determined by the second largest eigenvalue of $BD_B^{-1}$. Here $D_B = \text{diag}(B1)$. For the SBM with block probabilities matrix $B_1$ and $B_2$, this second largest eigenvalue is 0.309 and 0.167, respectively. We thus expect the model with block probabilities matrix $B_2$ to have less accurate community detection when $t_L$ increases, compared to the model with block probabilities matrix $B_1$.

**Degree-Corrected Stochastic Block Model:** DCSBMs are direct generalization of SBMs with the only difference being that each node $i$ has a degree-correction parameter $\theta_i$ and that the probability of connection between nodes $i$ and $j$ is given by

$$p_{ij} = \theta_i \theta_j B_{k(i)k(j)}$$

instead of $p_{ij} = B_{k(i)k(i')}$ as in the case of SBMs. For more on DCSBMs and their inference, see Gao et al. (2018); Karrer and Newman (2011); Zhao et al. (2012). While we had not provide formal theoretical results on consistency of node2vec embedding for DCSBMs, we are still interested in assessing the empirical performance of node2vec via simulation.

For our DCSBMs simulation, we generate the degree correction parameters $\theta_i$ as

$$\theta_i = |Z_i| + 1 - (2\pi)^{-1/2}, \ Z_i \sim \mathcal{N}(0, 0.25), \ i = 1, \ldots, n. \quad (4.1)$$

This procedure for generating $\theta_i$ is the same as that in Gao et al. (2018).

The simulation results for the SBMs and the DCSMBs are presented in Figures 5 and 6. We now summarize the main trend in these figures.

- The box plots when $\rho_n = 1$ (dense regime) have large variability because there are a few replications where, due to sampling variability, the community detection algorithm has low accuracy. If we ignore these replications then generally the community detection accuracy are almost 1 for all values of $n$ and furthermore, increasing the window size does not yields noticeable improvement in the accuracy.

- When the block probabilities matrix is $B_1$ then the second largest eigenvalue of $B_1 \text{diag}(B_1)\mathbf{1}^{-1}$ is not too small. Then $M_0$ does not converge to $C\mathbf{1}\mathbf{1}^T$ too fast as $t_L$ increases and we see that increasing window size significantly improves the community detection accuracy under the sparse regime for both the SBM and DCSBM settings. In particular as we increase $t_L, t_U$ the median accuracy increases and the number of replications with low accuracy also decreases. This corroborates our theoretical results in Section 3.2, that is, having larger window size could improve the convergence rate of $\hat{\mathcal{F}}$ to $\mathcal{F}$ and thus improve the community detection accuracy.

- In contrast when the block probabilities matrix is $B_2$ then the second largest eigenvalue of $B_2 \text{diag}(B_2)\mathbf{1}^{-1}$ is relatively small and hence $M_0$ converges to $C\mathbf{1}\mathbf{1}^T$ much faster. We now see that increasing the window size did not lead to improvement in the community detection accuracy; panels $(f)$ in Figure 5 and Figure 6 even show slight
Figure 7: Visualizations of node2vec embeddings for a realization of a SBM graph on \( n = 600 \) vertices and \( \rho_n = 3n^{-1/2} \) as \((t_L, t_U)\) changes. The first and second row plot the results when the block probabilities for the SBM is \(B_1\) and \(B_2\), respectively. The community detection is done using \(K\)-means on the embeddings.
Figure 8: Visualizations of node2vec embeddings for a realization of a DCSBM graph on $n = 600$ vertices and $\rho_n = 3n^{-1/2}$ as $(t_L, t_U)$ changes. The first and second row plot the results when the block probabilities for the DCSBM is $B_1$ and $B_2$, respectively. The community detection is done using $K$-means on the embeddings.
decrease in accuracy. Roughly speaking, as \( t_L, t_U \) increase, the information in \( M_0 \) and \( \mathcal{F} \) became weak quite quickly and the improvement in the convergence rate of \( \mathcal{F} \) to \( \mathcal{F} \) for larger values of \((t_L, t_U)\) could not overcome this loss of information in \( \mathcal{F} \) and thus the community detection accuracy does not improve. To see the improvement of detection performance in these cases will require much larger values of \( n \).

- Continuing the above observations, the node2vec embeddings visualized in Figure 9 and Figure 10 are the three dimensional embeddings labeled according to their true communities for one random realization from each SBM and DCSBM model for \( \rho_n = 3n^{-1/2}, n = 600 \) and with different window sizes. We see that when the block probabilities matrix is \( B_1 \) then as the window size becomes larger the embedded points from different communities are more separated which implies higher accuracy. Meanwhile for \( B_2 \) the embedded points from different communities are not separated and as a result the accuracy remains low.

5. Discussion

In this paper, we prove results on the embedding consistency of the random-walk based network embedding algorithms DeepWalk and node2vec with \((p, q) = (1, 1)\), assuming that the observed graphs are instances of the stochastic blockmodel graphs. Our results are valid under both the dense and sparse regimes. Under the sparse regime, we further demonstrate that (a) under certain sparsity of the network, there is a phase transition phenomenon as the window sizes increase and (b) as the network becomes sparser there is a need for larger window sizes in order to guarantee a tight error bound between the embeddings of the observed but noisy adjacency matrix and the true but unknown edge probabilities matrix. The simulation results corroborate our theoretical findings; in particular our simulation shows that increasing window sizes can improve the community detection accuracy for both sparse SBM graphs and DCSBM graphs as long as the random walk using the true edge probability matrix does not converge to the stationary distribution too quickly, i.e., \( D^{-1}_p W^t \) does not converge to \( 11^T /|P| \) too quickly as \( t \) increases.

We emphasize that our paper does not contain any real data analysis as DeepWalk and node2vec are widely-used algorithms with numerous existing papers demonstrating their uses for analyzing real graphs in diverse applications. In contrast, our paper is one of a few that addresses the theory underpinning these algorithms and is the first paper to establish consistency of community detection for SBMs using these random-walk based embedding algorithms.

There are several open questions for future research:

1. In this paper we only consider the case of \( p = q = 1 \) for node2vec embedding. Recall that \( p = q = 1 \) is the default option for typical uses of node2vec. For general values of \( p \) and \( q \), the form of the factorization matrix \( M_0 \) can no longer be represented in terms of the adjacency matrix \( A \) or the transition matrix \( W^t \); this renders the theoretical analysis for general values of \( p \) and \( q \) substantially more involved. One potential approach to address this problem is to consider, similar to the notion of the non-backtracking matrix in community detection for sparse SBM (Bordenave et al.,
a transition matrix associated with the edges of $G$ as opposed to the transition matrix associated with the vertices in $G$. Indeed, when $p \neq q$, the node2vec random walk transition from a given vertex $v$ to a vertex $w$ depends also on the vertex, say $u$, preceding $v$, i.e., the transition probability for $(v, w)$ depends on the choice of $(u, v)$.

2. In this paper we focus on the relative error of node2vec/DeepWalk embedding and prove weak recovery of the community assignments in stochastic blockmodel graphs. An important question is whether or not stronger limit results are available for these algorithms. For example one can show that the spectral embeddings of stochastic blockmodel graphs obtained by factorizing either the adjacency or the normalized Laplacian matrix are approximated by mixture of multivariate Gaussians. See Rubin-Delanchy et al. (2017); Tang and Priebe (2018) for more precise statements of these results and their implications for statistical inference in networks. It is thus natural to inquire if a normal approximation also holds for the embeddings obtained from node2vec/DeepWalk.

We ran several simulations to visualize the large-sample embeddings of node2vec/DeepWalk. Let $B = \begin{pmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{pmatrix}$ and $\pi = (0.4, 0.6)$. Figure 9 plots the embeddings of several stochastic blockmodel graphs with parameters $B$ and $\pi$ as $n \in \{1000, 1500, 2000, 3000\}$ varies; the window size is set to $(t_L, t_U) = (2, 5)$. As another example, Figure 10 plots the embeddings of a DCSBM with the same parameters and degree-correction factors $\{\theta_i\}$ generated according to Eq. (4.1). Both Figure 9 and Figure 10 suggest that for large samples the node2vec/DeepWalk embeddings of SBM and DCSBM graphs are approximately mixtures of multivariate Gaussians. We leave the theoretical justification of this phenomenon for future work.

3. As we allude to in the introduction, for simplicity we only consider stochastic blockmodel graphs in this paper. If carefully checking the proofs, one could see that all of our theoretical results will still hold in the case of the degree-corrected SBMs. For the more general inhomogeneous Erdős-Rényi random graphs model, we expect that Theorem 5, Theorem 6, Theorem 9, and Theorem 11 to still hold, provided that the edge probabilities are sufficiently homogeneous, i.e., the minimum and maximum values for the edge probabilities values are of the same order as $n$ increases. However, the error bounds in Corollary 8 and Corollary 12 might no longer apply since the structure of the true but unobserved edge probabilities matrix together with the entrywise logarithmic transformation of the co-occurrence matrices can lead to the setting wherein the eigenvalues of $M_0$ exhibit small eigengaps for which (1) no approximate low-rank representations exists and (2) the Davis-Kahan theorem will not yield meaningful upper bounds.

4. Finally, in this paper we consider the node2vec and DeepWalk embedding through matrix factorization. The original node2vec algorithm uses stochastic gradient descent for optimizing Eq. (2.2) to obtain the target embedding. The objective function in Eq. (2.2) is non-convex and thus can have a large number of local-minima, thereby making the theoretical analysis non-trivial. There are a few recent papers that address the large sample properties of local minima in non-convex optimization problems in
Figure 9: Scatter plots of node2vec/DeepWalk embeddings for two-blocks SBM with $B = \begin{pmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{pmatrix}$ and $\pi = (0.4, 0.6)$ as $n$ varies. The points are colored according to their community membership. The dashed ellipses are the 95% level curves for the block-conditional empirical distributions. The two black points are the two distinct embedding vectors obtained by factorizing $M_0$; note that these points had been transformed by the appropriate orthogonal matrices so as to align them with the node2vec/DeepWalk embedding obtained from the observed graphs.

statistics. See Chi et al. (2019) for an overview. In particular the ideas in Chi et al. (2019) suggest that if we can relate a subset of the local minimas of Eq. (2.2) to our matrix factorization embedding $\hat{F}$, then the consistency results in this paper can be adapted to study the consistency of node2vec using stochastic gradient descent.
Figure 10: Paired scatter plots for the first four dimensions of the node2vec/DeepWalk embeddings for two-blocks DCSBM with $B = \begin{pmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{pmatrix}$ and $\pi = (0.4, 0.6)$ as $n$ varies. For each panel, the diagrams below the diagonal show the paired scatter plots for each pairs of dimensions. The diagrams on the diagonal show the block-conditional empirical distributions of the embedding for each dimension. The diagrams above the diagonal show the paired density estimation curves.
Appendix A. Proofs under Dense Regime

In this section, we give proofs of theorems and corollary in Section 3.1. We first list two basic lemmas that will be used repeatedly in the subsequent proofs. Proofs of these lemmas are deferred to Appendix C.

Lemma 14 For any \( t > 0 \) is an integer, \( W^t \) and \( W^t \) satisfy

\[
1^T_n \cdot W^t = 1^T_n, 1^T_n \cdot W^t = 1^T_n, W^t d = d, W^t p = p.
\]

Lemma 15 Under Assumption 2, we have

\[
\|D_A\| = \max_{1 \leq i \leq n} d_i = O_P(n), \quad \|D_A^{-1}\| = \max_{1 \leq i \leq n} 1/d_i = O_P(1/n),
\]

\[
\|D_A - D_P\| = \max_i |d_i - p_i| = O_P(\sqrt{n \log n}), \quad \text{(A.1)}
\]

\[
\|D_A^{-1} - D_P^{-1}\| = \max_i |d_i^{-1} - p_i^{-1}| = O_P(n^{-3/2} \sqrt{\log n}),
\]

\[
\|W^t\|_{\text{max}} = O_P(n^{-1}), \max_{i,i'} w_i^{(t)} \asymp \frac{1}{n}, \min_{i,i'} w_i^{(t)} \asymp \frac{1}{n},
\]

for any \( t \geq 1, \) as \( n \to +\infty. \)

A.1. Proof of Theorem 5

We use the following three steps to bound Eq. (3.1)-Eq. (3.3) in turn.

Step 1 (Bounding \( \|W - W\|_{\text{max}} \)) We start with the decomposition

\[
W - W = AD_A^{-1} - PD_P^{-1} = \underbrace{AD_A^{-1}D_A^{-1}(D_P - D_A)}_{\Delta_1^{(1)}} + \underbrace{(A - P)D_P^{-1}}_{\Delta_2^{(1)}}.
\]

For the first term, we have

\[
\Delta_1^{(1)} = AD_A^{-1}D_A^{-1}(D_P - D_A) = \left[ \frac{a_{ii'} d_{i'} - p_{i'}}{d_{i'} - p_{i'}} \right]_{n \times n}
\]

and hence

\[
\|\Delta_1^{(1)}\|_{\text{max}} = \max_{1 \leq i,i' \leq n} \left| \frac{a_{ii'} d_{i'} - p_{i'}}{d_{i'} - p_{i'}} \right| \gtrsim \max_{1 \leq i,i' \leq n} \frac{1}{n} \left| \frac{d_{i'} - p_{i'}}{d_{i'}} \right|
\]

by \( |a_{ii'}| \leq 1 \) and \( c_0 < p_{ii'} < c_1. \) Lemma 15 then implies

\[
\|\Delta_1^{(1)}\|_{\text{max}} \gtrsim \frac{1}{n} \max_{1 \leq i \leq n} |d_{i'} - p_{i'}| \cdot \max_{1 \leq j \leq n} \frac{1}{d_{i'}} = O_P(n^{-3/2} \sqrt{\log n}). \quad \text{(A.2)}
\]

For the second term we have

\[
\|\Delta_2^{(1)}\|_{\text{max}} = \|(A - P)D_P^{-1}\|_{\text{max}} = \max_{1 \leq i,i' \leq n} \left| \frac{a_{ii'} - p_{ii'}}{p_{i'}} \right| \leq \max_{1 \leq j \leq n} \frac{1}{p_{i'}} = O_P(n^{-1}) \quad \text{(A.3)}
\]

due to Assumption 2. Combining Eq. (A.2) and Eq. (A.3) yield

\[
\|W - W\|_{\text{max}} \leq \|\Delta_1^{(1)}\|_{\text{max}} + \|\Delta_2^{(1)}\|_{\text{max}} = O_P(n^{-3/2} \sqrt{\log n}) + O_P(n^{-1}) = O_P(n^{-1}).
\]
Step 2 (Bounding $\|\hat{W}^2 - W^2\|_{\text{max,diag}}, \|\hat{W}^2 - W^2\|_{\text{max,off}}$) We first rewrite $\hat{W}^2 - W^2$ as

$$\hat{W}^2 - W^2 = \Delta_1^{(2)} + \Delta_2^{(2)}$$

As with (Eq. A.1) we have,

$$\Delta_1^{(2)} = (AD_A^{-1} - PD_P^{-1})W = (AD_P^{-1}D_A^{-1}(D_P - D_A) + (A - P)D_P^{-1}W$$

$$= (AD_P^{-1}D_A^{-1}(D_P - D_A) - AD_P^{-2}(D_P - D_A))W + AD_P^{-2}(D_P - D_A)W + (A - P)D_P^{-1}W.$$ 

The $ii'$th element of $\Delta_1^{(2,1)}$ is

$$\sum_{i'=1}^{n} \left( \frac{a_{ii'} (p_{i'} - d_{i'})}{d_{i'} p_{i'}} (p_i - d_i) - \frac{a_{ii'} (p_{i'} - d_{i'})}{p_{i'}} (p_i - d_i) \right) \cdot w_{i'i'} = \sum_{i'=1}^{n} \frac{a_{ii'} (p_{i'} - d_{i'})}{p_{i'}} \cdot \frac{1}{p_{i'}} - \frac{1}{p_{i'}} \cdot w_{i'i'}$$

We therefore have, by Assumption 2 and Lemma 15,

$$\|\Delta_1^{(2,1)}\|_{\text{max}} \leq \max_{1 \leq i, i' \leq n} \sum_{i'=1}^{n} \frac{a_{ii'} (p_{i'} - d_{i'})^2}{p_{i'}^2 d_{i'}} \cdot w_{i'i'}$$

$$\leq n \cdot \left( \max_{1 \leq i \leq n} |p_i - d_i| \right)^2 \cdot \left( \max_{1 \leq i \leq n} \frac{1}{p_i} \right)^2 \cdot \left( \max_{1 \leq i \leq n} \frac{1}{d_i} \right) \cdot \left( \max_{i,i'} w_{i'i'} \right) = O_p(n^{-2} \log n).$$

For the term $\Delta_1^{(2,2)}$ we have

$$\|\Delta_1^{(2,2)}\|_{\text{max}} = \max_{1 \leq i, i' \leq n} \left| \sum_{i'=1}^{n} \frac{a_{ii'} (p_{i'} - d_{i'})}{p_{i'}^2} (p_i - d_i) w_{i'i'} \right|$$

$$\leq n \cdot \left( \max_{i} |p_i - d_i| \right) \cdot \max_{i,i'} w_{i'i'} \cdot \left( \max_{i} \frac{1}{p_i} \right)^2 = O_p(n^{-3/2} \sqrt{\log n}).$$

We now consider the term $\Delta_1^{(2,3)}$. We have

$$\|\Delta_1^{(2,3)}\|_{\text{max}} = \max_{i,i'} \left| \sum_{i'} a_{ii'} p_{ii'} - p_{ii'} \cdot w_{i'i'} \right| = (n^{-2}) \times \max_{i,i'} \left| \sum_{i'} (a_{ii'} - p_{ii'}) \cdot (n^2 \cdot w_{i'i'}/p_{ii'}) \right|.$$

Assumption 2 and Lemma 15 then imply

$$\max_{i,i'} n^2 \cdot w_{i'i'}/p_{ii'}, \min_{i,i'} n^2 \cdot w_{i'i'}/p_{ii'} \asymp 1.$$ 

Another application of Bernstein inequality similar to that in Eq. (A.1) yield

$$\max_{i,i'} \left| \sum_{i'} (a_{ii'} - p_{ii'}) \cdot (n^2 \cdot w_{i'i'}/p_{ii'}) \right| = O_p(\sqrt{n \log n})$$ (A.5)
and thus \( \| \Delta_1^{(2,3)} \|_{\max} = \mathcal{O}_p(n^{-3/2} \sqrt{\log n}) \). Combining the above bounds, we have
\[
\| \Delta_1^{(2)} \|_{\max} \leq \| \Delta_1^{(2,1)} \|_{\max} + \| \Delta_1^{(2,2)} \|_{\max} + \| \Delta_1^{(2,3)} \|_{\max} = \mathcal{O}_p(n^{-3/2} \sqrt{\log n}). \tag{A.6}
\]

We now consider the term \( \Delta_2^{(2)} = \hat{W}(W - \hat{W}). \) We have
\[
\| \Delta_2^{(2)} \|_{\max} = \| (W - \hat{W})(\hat{W} - W) - W(W - \hat{W}) \|_{\max}
\leq \| (W - \hat{W})^2 \|_{\max} + \| W(W - \hat{W}) \|_{\max}.
\]

The same argument for bounding \( \| \Delta_1^{(1)} \|_{\max} = \| (W - \hat{W})W \|_{\max} \) as given above also yields
\[
\| W(W - \hat{W}) \|_{\max} = \mathcal{O}_p(n^{-3/2} \sqrt{\log n}).
\]

We then bound \( \| (W - \hat{W})^2 \|_{\max} \) through the following expansion
\[
\| (W - \hat{W})^2 \|_{\max} = \max_{i,i'} \left\| \sum_{i^* = 1}^{n} \left( \frac{p_{ii^*}}{p_{i^*}} - \frac{a_{ii^*}}{d_{i^*}} \right) \left( \frac{p_{ii'}}{p_{i'}} - \frac{a_{ii'}}{d_{i'}} \right) \right\|
\leq \max_{i,i'} \left\| \sum_{i^* = 1}^{n} \left( \frac{a_{ii^*}}{p_{i^*}} - \frac{a_{ii'}}{p_{i'}} \right) \left( \frac{p_{ii'}}{p_{i'}} - \frac{a_{ii'}}{p_{i'}} \right) \right\|
+ \max_{i,i'} \left\| \sum_{i^* = 1}^{n} \left( \frac{p_{ii^*}}{p_{i^*}} - \frac{a_{ii^*}}{d_{i^*}} \right) \left( \frac{a_{ii'}}{p_{i'}} - \frac{a_{ii'}}{p_{i'}} \right) \right\|
+ \max_{i,i'} \left\| \sum_{i^* = 1}^{n} \left( \frac{p_{ii^*}}{p_{i^*}} - \frac{a_{ii^*}}{p_{i^*}} \right) \left( \frac{p_{ii'}}{p_{i'}} - \frac{a_{ii'}}{p_{i'}} \right) \right\|.
\]

Since \( |a_{ii'}| \leq 1 \) and \( |a_{ii'} - p_{ii'}| \leq 1 \), we have
\[
\delta_2^{(2,1)} \leq \max_{i'} \left( \frac{1}{p_{i^*}} - \frac{1}{d_{i^*}} \right) \cdot \frac{1}{p_{i'}} \cdot \frac{1}{p_{i'}} \cdot \frac{1}{p_{i'}} \cdot \frac{1}{p_{i'}} \leq n \cdot \max_{i^*} \left( \frac{1}{p_{i^*}} - \frac{1}{d_{i^*}} \right) = \mathcal{O}_p(n^{-3/2} \sqrt{\log n}).
\]

Similar reasoning also yield
\[
\delta_2^{(2,2)} \leq \max_{i,i'} \left( \frac{1}{p_{i^*}} - \frac{1}{d_{i^*}} \right) \cdot \frac{1}{p_{i'}} \cdot \frac{1}{d_{i'}} \leq n \cdot \left( \max_{i'} \frac{1}{p_{i'}} + \max_{i'} \frac{1}{d_{i'}} \right) \cdot \frac{1}{p_{i'}} \cdot \frac{1}{d_{i'}} = n \cdot \mathcal{O}_p(1/n).
\]

We now bound \( \delta_2^{(2,3)} \) by considering the diagonal and off-diagonal terms separately. For the diagonal terms with \( i = i' \), we have
\[
\max_{i} \left\| \sum_{i^* = 1}^{n} \left( \frac{a_{ii^*} - p_{ii^*}}{p_{ii^*} p_i} \right)^2 \right\| \leq \left( \max_{i} \frac{1}{p_i} \right) \frac{1}{p_i} \sum_{i^* = 1}^{n} \frac{1}{p_{i^*}} \leq \frac{1}{n} \cdot n \cdot \frac{1}{n} = \mathcal{O}(1/n). \tag{A.8}
\]
\[ \| \hat{W}^2 - W^2 \|_{\text{max}, \text{diag}} \leq \| \Delta_1^{(2)} \|_{\text{max}} + \| \Delta_2^{(2)} \|_{\text{max}, \text{diag}} \]
\[ \leq \| \Delta_1^{(2)} \|_{\text{max}} + \| W(W - \hat{W}) \|_{\text{max}} + \| (W - \hat{W})^2 \|_{\text{max}, \text{diag}} \]
\[ \leq \| \Delta_1^{(2)} \|_{\text{max}} + \| W(W - \hat{W}) \|_{\text{max}} + \delta_2^{(2,1)} + \delta_2^{(2,2)} + \max_i \sum_{i' = 1}^n \frac{(p_{i i^*} - a_{i i^*})^2}{p_{i^*} p_i} \]
\[ = O_p(1/n). \quad (A.9) \]

We now consider the off-diagonal terms with \( i \neq i' \) for \( \delta_2^{(2,3)} \). First define
\[ \zeta_i^{i^*} = \frac{1}{p_{i^*} p_i^*} (p_{i i^*} - a_{i i^*}) (p_{i^* i'} - a_{i^* i'}). \]

We now make the important observation that if \( i \neq i' \) then the collection of random variables \( \zeta_i^{i^*} \) for \( i^* = 1, 2, \ldots, n \) are independent mean 0 random variables. Indeed, when \( i \neq i' \) then \( a_{i i^*} \) and \( a_{i^* i'} \) are independent and hence
\[ \mathbb{E} \zeta_i^{i^*} = \frac{1}{p_{i^*} p_i^*} \mathbb{E} (p_{i i^*} - a_{i i^*}) \cdot \mathbb{E} (p_{i^* i'} - a_{i^* i'}) = 0. \]

We then have
\[ \max_{i \neq i'} \left| \sum_{i^* = 1}^n \left( \frac{p_{i i^*} - a_{i i^*}}{p_{i^*}} \right) \left( \frac{p_{i^* i'} - a_{i^* i'}}{p_i} \right) \right| = \max_{i \neq i'} \left| \sum_{i^* = 1}^n \zeta_i^{i^*} \right| = \max_{i \neq i'} \left| \sum_{i^* = 1}^n \zeta_i^{i^*} + \zeta_i^{i^*} + \zeta_i^{i^*} \right|. \]

Now fix a pair \( \{i, i'\} \) with \( i \neq i' \). Then by Bernstein inequality, we have
\[ \mathbb{P} \left( \left| \sum_{i^* = 1}^n \zeta_i^{i^*} \right| > \epsilon \right) \leq 2 \exp \left( - \frac{\epsilon^2}{2 \sigma_1^2 + \frac{M}{3} \epsilon} \right) \quad (A.10) \]

where the variance proxy \( \sigma_1^2 \) is bounded as
\[ \sigma_1^2 = \sum_{i^* = 1}^n \text{Var}(\zeta_i^{i^*}) \leq \frac{1}{p_{i^*} p_i^*} \sum_{i^* = 1}^n \text{Var} \left( \frac{p_{i i^*} - a_{i i^*}}{p_{i^*}} \right) \cdot \text{Var} \left( \frac{p_{i^* i'} - a_{i^* i'}}{p_i} \right) \leq \frac{n}{16 p_{i^*} p_i^*} \leq \frac{1}{16 n^3}, \]

and \( M \) is any constant bigger than \( \max_{i^*} |\zeta_i^{i^*}|. \) In particular we have
\[ |\zeta_i^{i^*}| = \left| \left( \frac{p_{i i^*} - a_{i i^*}}{p_{i^*}} \right) \left( \frac{p_{i^* i'} - a_{i^* i'}}{p_i} \right) \right| \leq \frac{2}{n c_0} = \frac{4}{c_0^2 n^2} \equiv M. \]

Plugging these bounds for \( \sigma_1^2 \) and \( M \) into Eq. \( (A.10) \), we have, for any \( C_1 > 0 \),
\[ \mathbb{P} \left( \left| \sum_{i^* = 1}^n \zeta_i^{i^*} \right| > C_1 n^{-3/2} \sqrt{\log n} \right) \leq 2 \exp \left( - \frac{\left( C_1 n^{-3/2} \sqrt{\log n} \right)^2}{2 \cdot \frac{1}{16 c_0^2 n^4} + \frac{4}{c_0^2 n^2} \left( C_1 n^{-3/2} \sqrt{\log n} \right)} \right) \]
\[ = 2 \exp \left( - \frac{C_1 \log n}{\frac{1}{8 c_0} + \frac{4 \sqrt{\log n}}{n^{3/2} c_0^2}} \right) \lesssim n^{-8 C_1 c_0^4} \]

32
Now choose $C_1$ such that $8C_1C^4_0 > 2$. We then have, by a union bound over all pairs \( \{i, i'\} \) with $i \neq i'$, that

\[
P\left( \max_{i \neq i'} \left| \sum_{i'=1}^{n} \zeta^{ii'} \right| > C_1 n^{3/2} \sqrt{\log n} \right) \lesssim (n^2 - n) \cdot n^{-8C_1C^4_0} \lesssim n^{2-8C_1C^4_0} = o(1)
\]

for sufficiently large \( n \). We thus conclude

\[
\max_{i \neq i'} \left| \sum_{i'=1}^{n} \left( \frac{p_{ii}}{p_{i}} - \frac{a_{ii}}{p_{i}} \right) \left( \frac{p_{i'i'}}{p_{i'}} - \frac{a_{i'i'}}{p_{i'}} \right) \right| \leq \max_{i \neq i'} \left| \left( \sum_{i'=1}^{n} \zeta^{ii'} \right) \right| + 2M
\]

\[
\leq O(n^{-3/2} \sqrt{\log n}) + O(n^{-2}) = O(n^{-3/2} \sqrt{\log n}).
\]

A similar argument to Eq. (A.9) shows

\[
\| \hat{W}^2 - W^2 \|_{\text{max,off}} \leq \| \Delta^{(2)}_1 \|_{\text{max}} + \| W(W - \hat{W}) \|_{\text{max}} + \delta_2^{(2,1)} + \delta_2^{(2,2)} + \max_{i \neq i'} \left| \sum_{i'=1}^{n} \left( \frac{p_{ii}}{p_{i}} - \frac{a_{ii}}{p_{i}} \right) \left( \frac{p_{i'i'}}{p_{i'}} - \frac{a_{i'i'}}{p_{i'}} \right) \right| = O(n^{-3/2} \sqrt{\log n}).
\] (A.11)

**Step 3 (Bounding \( \| \hat{W}^t - W^t \|_{\text{max}, t \geq 3} \))** We first consider \( t = 3 \). We have

\[
\| \hat{W}^3 - W^3 \|_{\text{max}} \leq \| (\hat{W}^2 - W^2) \hat{W} \|_{\text{max}} + \| W^2 (\hat{W} - W) \|_{\text{max}}
\] (A.12)

For the first term in the RHS of Eq. (A.12) we have, by Eq. (A.9), Eq. (A.11) and Lemma 15, that

\[
\| (\hat{W}^2 - W^2) \hat{W} \|_{\text{max}} \leq (n - 1) \times \| \hat{W} \|_{\text{max}} \| \hat{W}^2 - W^2 \|_{\text{max,off}} + \| \hat{W} \|_{\text{max}} \| \hat{W}^2 - W^2 \|_{\text{max,diag}}
\]

\[
\leq O_p(n^{-3/2} \sqrt{\log n}).
\]

For the second term in the RHS of Eq. (A.12) we use the same argument as that for bounding \( \| (\hat{W} - W) W \|_{\text{max}} \) in **Step 2**. In particular we have

\[
\| W^2 (\hat{W} - W) \|_{\text{max}} = O_p(n^{-3/2} \sqrt{\log n}).
\] (A.13)

Combining Eq.(A.12) through Eq. (A.13) yields

\[
\| \hat{W}^3 - W^3 \|_{\text{max}} = O_p(n^{-3/2} \sqrt{\log n}).
\] (A.14)

The case when \( t = 4 \) is analogous. More specifically,

\[
\| \hat{W}^4 - W^4 \|_{\text{max}} = \| (\hat{W}^2 - W^2) \hat{W}^2 \|_{\text{max}} + \| W^2 (\hat{W}^2 - W^2) \|_{\text{max}}
\]

\[
\leq (n - 1) \times \| \hat{W}^2 - W^2 \|_{\text{max,off}}(\| \hat{W}^2 \|_{\text{max}} + \| W^2 \|_{\text{max}})
\]

\[
+ \| \hat{W}^2 - W^2 \|_{\text{max,diag}}(\| \hat{W}^2 \|_{\text{max}} + \| W^2 \|_{\text{max}})
\]

\[
= O_p(n^{-3/2} \sqrt{\log n})
\] (A.15)
We now consider a general $t \geq 5$. We start with the decomposition
\[ \hat{W}^t - W^t = (\hat{W}^t - W^2\hat{W}^{t-2}) + (W^2\hat{W}^{t-2} - W^t) = \left( \frac{\hat{W}^2 - W^2}{\Delta_1^{(3)}} \right) \hat{W}^{t-2} + \frac{W^2(\hat{W}^{t-2} - W^{t-2})}{\Delta_2^{(3)}}. \]

We now have, by Lemma 15, Eq. (A.9), and Eq. (A.11), that
\[ \|\Delta_1^{(3)}\|_{\text{max}} \leq (n-1) \|\hat{W}^2 - W\|_{\text{max,off}} \|\hat{W}^{t-2}\|_{\text{max}} + \|\hat{W}^2 - W\|_{\text{max,diag}} \|\hat{W}^{t-2}\|_{\text{max}} = O_P(n \cdot n^{-3/2} \sqrt{\log n \cdot n^{-1}}) + O_P(n^{-1} \cdot n^{-1}) = O_P(n^{-3/2} \sqrt{\log n}). \]

(A.16)

Once again, by Lemma 15, we have
\[ ||\Delta_2^{(3)}||_{\text{max}} \leq n \cdot ||W^2||_{\text{max}} ||\hat{W}^{t-2} - W^{t-2}||_{\text{max}} = O(1) ||\hat{W}^{t-2} - W^{t-2}||_{\text{max}}. \]

Combining Eq. (A.16) and Eq. (A.17), we have
\[ ||\hat{W}^t - W^t||_{\text{max}} = O_P(n^{-3/2} \sqrt{\log n}) + O(1) ||\hat{W}^{t-2} - W^{t-2}||_{\text{max}} \]

(A.18)

As $t$ is finite, iterating the above argument yields
\[ ||\hat{W}^t - W^t||_{\text{max}} = \begin{cases} O_P(n^{-3/2} \sqrt{\log n}) + O(1) ||\hat{W}^4 - W^4||_{\text{max}} & \text{when } t \text{ is even} \\ O_P(n^{-3/2} \sqrt{\log n}) + O(1) ||\hat{W}^3 - W^3||_{\text{max}} & \text{when } t \text{ is odd} \end{cases} \]

We now recall Eq. (A.14) and Eq. (A.15) and conclude
\[ ||\hat{W}^t - W^t||_{\text{max}} = O_P(n^{-3/2} \sqrt{\log n}), \]

for $t \geq 3$, as desired.

A.2. Proof of Theorem 6

We will only present the proof for bounding $||\tilde{M}_0 - M_0||_F$ here as the bound for $||M_0||_F$ is derived in the proof of Corollary 8.

Under the dense regime of Assumption 2 and for sufficiently large $n$, the diameter of $\mathcal{G}$ is 2 with high probability, see e.g., Corollary 10.11 in Bollobás (2001). Therefore, with high probability, all entries of $\hat{W}^2$ are positive and hence
\[ \tilde{M}_0 = \log \left[ \frac{2|A|}{\gamma(L, t_L, t_U)} \sum_{t'=t_I}^{t_U} (L - t) \cdot \left(D_A^{-1}\hat{W}^t\right) \right] - \log(k)11^T \]

is well-defined with high probability.

Next recall the definition of $M_0$ given in Eq. (2.6). We have
\[ \tilde{M}_0 - M_0 = \log \left[ |A| \sum_{t'=t_I}^{t_U} (L - t) \cdot \left(D_A^{-1}\hat{W}^t\right) \right] - \log \left[ |P| \sum_{t'=t_I}^{t_U} (L - t) \cdot \left(D_P^{-1}W^t\right) \right]. \]

(A.19)
By the mean value theorem, the absolute value of $ii'$th entry in $\mathbf{M}_0 - \mathbf{M}_0$ is
\[
\frac{1}{\alpha_{ii'}} \left| \sum_{t=t_L}^{t_U} (L-t) \cdot \left[ |\mathbf{A}| \cdot \frac{w_{ii}(t)}{d_t} - |\mathbf{P}| \cdot \left( \frac{w_{ii}(t)}{p_t} \right) \right] \right|
\]
where $\alpha_{ii'} \in (I_{iA}^{ii'}, I_{iP}^{ii'})$ and $I_{iA}^{ii'}$ and $I_{iP}^{ii'}$ are the $ii'$th entry of $\mathbf{I}_A$ and $\mathbf{I}_P$, respectively. We therefore have
\[
\| \mathbf{M}_0 - \mathbf{M}_0 \|_{\text{max,off}} \leq \max_{i \neq i'} \left( \frac{1}{\alpha_{ii'}} \right) \cdot \| \mathbf{I}_A - \mathbf{I}_P \|_{\text{max,off}},
\]
\[
\| \mathbf{M}_0 - \mathbf{M}_0 \|_{\text{max,diag}} \leq \max_i \left( \frac{1}{\alpha_{ii}} \right) \cdot \| \mathbf{I}_A - \mathbf{I}_P \|_{\text{max,diag}}.
\]
We now bound the terms $\alpha_{ii'}^{-1}$, $\| \mathbf{I}_A - \mathbf{I}_P \|_{\text{max,off}}$ and $\| \mathbf{I}_A - \mathbf{I}_P \|_{\text{max,diag}}$.

**Step 1 (Bounding $\max_{ii'} \alpha_{ii'}^{-1}$)** Since $\alpha_{ii'} \in (I_{iA}^{ii'}, I_{iP}^{ii'})$, we have
\[
\max_{i,i'} \frac{1}{\alpha_{ii'}} \leq \max_{i,i'} \left( \frac{1}{I_{iA}^{ii'}}, \frac{1}{I_{iP}^{ii'}} \right).
\]
We first bound $(I_{iP}^{ii'})^{-1}$. In particular
\[
\frac{1}{I_{iP}^{ii'}} = \frac{1}{\sum_{t=t_L}^{t_U} (L-t)|\mathbf{P}| \cdot \left( \frac{w_{ii}(t)}{p_t} \right)} \leq \frac{p_i}{(L-t_L)|\mathbf{P}| \cdot w_{ii}(t_L)}.
\]
Now $c_0 \leq p_{ii'} \leq c_1$ for some constant $c_0$ and $c_1$. Then by Lemma 15, we have
\[
\max_{i,i'} \frac{1}{I_{iP}^{ii'}} \lesssim \frac{n}{n^2 \cdot \frac{1}{n}} = 1.
\]
We now consider $(I_{iA}^{ii'})^{-1}$. We have
\[
\frac{1}{I_{iA}^{ii'}} = \frac{1}{\sum_{t=t_L}^{t_U} (L-t)|\mathbf{A}| \cdot \left( \frac{w_{ii}(t)}{d_t} \right)} \leq \frac{d_i}{(L-t_L)|\mathbf{A}| \cdot w_{ii}(t_L)}.
\]
Now suppose $i \neq i'$. Then by Theorem 5, we have
\[
\max_{i \neq i'} |w_{ii}(t_L) - w_{ii'}(t_L)| = O_p(n^{-3/2} \sqrt{\log n}),
\]
which implies w.h.p.,
\[
0 \leq w_{ii}(t_L) - \max_{i \neq i'} |w_{ii}(t_L) - w_{ii'}(t_L)| \quad \text{for all } 0 \leq i, i' \leq n, i \neq i' \tag{A.20}
\]
Since $\min_{i,i'} w_{ii}(t_L), \max_{i,i'} w_{ii}(t_L) \asymp n^{-1}$. Lemma 15 then implies, for sufficiently large $n$,
\[
\max_{i \neq i'} \frac{1}{I_{iA}^{ii'}} \leq \max_{i \neq i'} \frac{d_i}{(L-t_L)|\mathbf{A}| \cdot w_{ii}(t_L) - O_p(n^{-3/2} \sqrt{\log n})} \lesssim \max_i \frac{d_i}{n d_i} \times \max_{i \neq i'} \frac{1}{w_{ii'}(t_L) - O_p(n^{-3/2} \sqrt{\log n})} = O_p(1). \tag{A.21}
\]

Now suppose that \( i = i' \). Then for \( t_L = 2 \), we have

\[
\frac{1}{\hat{w}_{ii}^{(2)}} = \frac{1}{\sum_{i'=1}^{n} \hat{w}_{ii'} \hat{w}_{i'i}} = \frac{1}{\sum_{i'=1}^{n} \frac{\alpha_{ii'} \hat{w}_{ii'}}{d_i}} = \frac{d_i}{\sum_{i'=1}^{n} \alpha_{ii'} / d_i}.\]

Once again, by Lemma 15, we have

\[
\max_i \frac{1}{\hat{w}_{ii}^{(2)}} \leq \frac{\max_i d_i}{\max_i, d_i \sum_{i'=1}^{n} \alpha_{ii'}} \leq (\max_i d_i)^2 \times \max_i \frac{1}{d_i} = \mathcal{O}(n),
\]

\[
\max_i \frac{1}{I_{ii}^{A}} \leq \max_i d_i \times \max_i \frac{1}{nd_i} \times \max_i \frac{1}{\hat{w}_{ii}^{(2)}} = \mathcal{O}(1).
\]

Now if \( t_L \geq 3 \) then Theorem 5 implies \( \max_i |\hat{w}_{ii}^{(t_L)} - \hat{w}_{ii}^{(t_L)}| = \mathcal{O}(n^{-3/2} \sqrt{\log n}) \). An identical argument to that for Eq. (A.21) also yields \( \max_i \frac{1}{I_{ii}^{A}} = \mathcal{O}(1) \). In summary, we have

\[
\max_{i,i'} \frac{1}{\alpha_{ii'}} = \mathcal{O}(1). \tag{A.22}
\]

**Step 2 (Bounding \( \|I_A - I_P\|_{\max, \text{off}} \)):** We start with the inequality

\[
\max_{i \neq i'} |I_{ii'}^{A} - I_{ii'}^{P}| = \max_{i \neq i'} \left| \sum_{t=1}^{t_L} (L - t) \cdot \left[ |A| \cdot \left( \frac{w_{ii}^{(t)}}{d_i} \right) - |P| \cdot \left( \frac{w_{ii}^{(t)}}{p_i} \right) \right] \right|
\]

\[
\leq \sum_{t=1}^{t_L} (L - t) \cdot \max_{i \neq i'} |A| \cdot \left( \frac{w_{ii}^{(t)}}{d_i} \right) - |P| \cdot \left( \frac{w_{ii}^{(t)}}{p_i} \right).
\]

We now bound each of the summand in the RHS of the above display. Consider a fixed value of \( t \geq 2 \). We have

\[
\max_{i \neq i'} |A| \cdot \frac{\hat{w}_{ii}^{(t)}}{d_i} - |P| \cdot \frac{w_{ii}^{(t)}}{p_i} \leq (|A| - |P|) \cdot \max_{i \neq i'} \frac{w_{ii}^{(t)}}{d_i} + |P| \max_{i \neq i'} \frac{(p_i - d_i) w_{ii}^{(t)}}{p_i d_i} \tag{A.23}
\]

By Lemma 15, the first term in RHS of Eq.(A.23) is bounded as

\[
(|A| - |P|) \max_{i \neq i'} \frac{w_{ii}^{(t)}}{d_i} \leq n \max_i |d_i - p_i| \times \max_{i,i'} w_{ii'}^{(t)} \times \max_i \frac{1}{d_i} = \mathcal{O}(n^{-1/2} \sqrt{\log n}) \tag{A.24}
\]

The second term in the RHS of Eq. (A.23) is also bounded by Lemma 15 as

\[
|P| \max_{i \neq i'} \left| \frac{(p_i - d_i) w_{ii}^{(t)}}{p_i d_i} \right| \leq |P| \max_i \frac{(p_i - d_i)}{p_i d_i} \times \max_{i,i'} w_{ii'}^{(t)} \tag{A.25}
\]

\[
\lesssim n^2 \times \mathcal{O}(n^{-1/2} \sqrt{\log n/n^2}) \times n^{-1} = \mathcal{O}(n^{-1/2} \sqrt{\log n}).
\]
We once again bound each summand in the above display. Similar to Eq. (A.23), we have as desired.

Combining the above terms, we have

\[
t > \]

arguments for Eq. (A.24) and Eq. (A.25). For the third term, we consider the cases \( t > 1 \) separately. For \( t = 2 \), we have

\[
\left| \mathbf{A} \cdot \max_{i \neq i'} \left| \frac{\hat{w}_{ii'}^{(t)} - w_{ii'}^{(t)}}{d_i} \right| \right| \leq |\mathbf{A}| \times \max_{i \neq i'} \left| \frac{\hat{w}_{ii'}^{(t)} - w_{ii'}^{(t)}}{d_i} \right| \times \frac{1}{d_i}
\]

\[
= \mathcal{O}_P(n^2) \times \mathcal{O}_P(n^{-3/2} \sqrt{\log n}) \times \mathcal{O}_P(n^{-1}) = \mathcal{O}_P(1).
\]

In contrast, for \( t > 2 \), we have

\[
|\mathbf{A}| \max_{i} \left| \frac{\hat{w}_{ii}^{(2)} - w_{ii}^{(2)}}{d_i} \right| = \mathcal{O}_P(n^2) \times \mathcal{O}_P(n^{-3/2} \sqrt{\log n}) \
\]

Combining the above terms, we have

\[
\| \mathbf{I}_A - \mathbf{I}_P \|_{\text{max,diag}} = \begin{cases} 
\mathcal{O}_P(1) & \text{if } t_L = 2, \\
\mathcal{O}_P(n^{-1/2} \sqrt{\log n}) & \text{if } t_L \geq 3.
\end{cases}
\]

Step 4 (Bounding \( \| \tilde{\mathbf{M}}_0 - \mathbf{M}_0 \|_F \)): Eq. (A.22), Eq. (A.26) and Eq. (A.28) imply

\[
\| \tilde{\mathbf{M}}_0 - \mathbf{M}_0 \|_{\text{max,off}} \leq \max_{i \neq i'} \left( \frac{1}{\alpha_{ii'}} \right) \cdot \| \mathbf{I}_A - \mathbf{I}_P \|_{\text{max,off}} = \mathcal{O}_P(n^{-1/2} \sqrt{\log n}),
\]

\[
\| \tilde{\mathbf{M}}_0 - \mathbf{M}_0 \|_{\text{max,diag}} \leq \max_i \left( \frac{1}{\alpha_{ii}} \right) \cdot \| \mathbf{I}_A - \mathbf{I}_P \|_{\text{max,diag}} = \mathcal{O}_P(1),
\]

\[
\| \tilde{\mathbf{M}}_0 - \mathbf{M}_0 \|_F \leq \left( n^2 \times \| \tilde{\mathbf{M}}_0 - \mathbf{M}_0 \|_{\text{max,off}}^2 + n \times \| \tilde{\mathbf{M}}_0 - \mathbf{M}_0 \|_{\text{max,diag}}^2 \right)^{1/2} = \mathcal{O}_P(n^{1/2} \sqrt{\log n}),
\]

as desired.
A.3. Proof of Corollary 8

For ease of exposition we shall assume that $n_k$, the size of community $k$, are integers with

$$n_k = nπ_k, \quad \text{for all } k \in [K].$$

The case where $k(i)$ are i.i.d. with $P(k(i) = k) = π_k$ is, except for a few slight modifications, identical. We first note that $M_0$ and $M_0$ are symmetric matrices. Indeed

$$(D_A^{-1} \hat{W}^T)^T = (D_A^{-1} A D_A^{-1} \cdots D_A^{-1} A D_A^{-1})^T = D_A^{-1} \hat{W}^T$$

and similarly for $(D_P^{-1} W^T)^T = (D_P^{-1} W^T)$. Recall from Eq. (2.3) that $P$ has a $K \times K$ blocks structure. The matrices $W^t$ therefore all have the same $K \times K$ blocks structure and hence the linear combination of these $W^t$ also has a $K \times K$ blocks structure. The elementwise logarithm used in defining $M_0$ will preserve the blocks structure. In summary we have

$$M_0 = \begin{pmatrix} ξ_{11} 1_{nπ_1} 1^T_{nπ_1} & \cdots & ξ_{1K} 1_{nπ_1} 1^T_{nπ_K} \\ \vdots & \vdots & \vdots \\ ξ_{K1} 1_{nπ_K} 1^T_{nπ_1} & \cdots & ξ_{KK} 1_{nπ_K} 1^T_{nπ_K} \end{pmatrix} = Θ \cdot (ξ_{ii'})_{K \times K} \cdot Θ^T$$

with $Θ$ as defined in Eq. 2.3 and $(ξ_{ii'})_{K \times K}$ is a symmetric matrix with rank $r_M$. Define $Ξ = (ξ_{ii'})_{K \times K}$. The non-zero eigenvalues of $M_0$ coincides with the non-zero eigenvalues of $Θ^T Θ Ξ = n \cdot \text{diag}(π_1, π_2, \ldots, π_K) \cdot Ξ$. As $(π_1, \ldots, π_K)$ and $Ξ$ are fixed, the non-zero eigenvalues of $M_0$ grows at order $n$ and hence its Frobenius norm also grows at order $n$.

By the Davis-Kahan Theorem (Davis and Kahan, 1970; Yu et al., 2015),

$$\min_{T \in O_d} \| \hat{T} - T \|_F \lesssim \frac{\| M_0 - M_0 \|_F}{\| M_0 \|_F} = O_ε(n^{1/2} \sqrt{\log n}) \times O(n^{-1}) = O_ε(n^{-1/2} \sqrt{\log n})$$

as desired.

Appendix B. Proofs under Sparse Regime

The approach used in the proofs of the theorems and corollary under the sparse regime is similar to that in the dense regime. However, if we simply use the same technique as in Appendix A then the convergence rate that we obtain for $\| \hat{W}^2 - W^2 \|_{\text{max,off}}$ and $\| \hat{W}^t - W^t \|_{\text{max}}$ for $t \geq 3$ will be $O_ε((nρ_n)^{-3/2} \sqrt{\log n})$. This convergence rate is too loose. Indeed, if we follow the same strategy as in the proof of Theorem 6 to bound $\| \hat{M}_0 - M \|_F$ under the sparse regime then Eq. (A.20) is valid only when $ρ_n = ω(n^{-1/3} \log^{1/3} n)$.

Before giving the formal proofs, we first state Lemma 16 and Lemma 18 as the main technical results for bounding $\| \hat{W}^2 - W^2 \|_{\text{max,off}}$ and $\| \hat{W}^t - W^t \|_{\text{max}}$ for $t \geq 3$ under the sparse regime. We summarize the motivation behind these lemmas below.

- Lemma 16 is the analogue of Lemma 15 in Appendix A and is used repeatedly for bounding several important terms that frequently appear in our proofs.
• In Appendix A we show that the bound for $\|\mathbf{W}^t - \hat{\mathbf{W}}^t\|_{\text{max}}$ when $t \geq 3$ is of the same magnitude, when $n$ increases, as the bound for $\|\mathbf{W}^2 - \hat{\mathbf{W}}^2\|_{\text{max},\text{off}}$. For the sparse regime these bounds are generally of different magnitude as $n$ increases and $\rho_n$ decreases. This difference is the main distinguish feature between the two regimes.

• We derive a more accurate bound for $\|\mathbf{W}^t - \hat{\mathbf{W}}^t\|_{\text{max}}$ when $t \geq 3$ in Step 4 and Step 5 of the proof of Theorem 9 presented below. The main challenge behind these steps is in controlling the term $\|\mathbf{A}^t\|_{\text{max}}$ given in Lemma 18 below. Lemma 18 is the main technical contribution of this section and might be of independent interest.

For ease of exposition we will only present the proof of Lemma 18 in this section; the proofs of the other lemmas are deferred to Appendix C.

Lemma 16 Under Assumption 3 we have, for any $t \geq 1$ and sufficiently large $n$,
\[
\|\mathbf{D}_n\| = \max_{1 \leq i \leq n} d_i = \mathcal{O}(n\rho_n), \quad \|\mathbf{D}_n^{-1}\| = \max_{1 \leq i \leq n} 1/d_i = \mathcal{O}(1/(n\rho_n)), \quad \|\mathbf{D}_n - \mathbf{D}_n^{-1}\| = \max_i |d_i - p_i| = \mathcal{O}(\sqrt{n\rho_n \log n}),
\]
\[
\|\mathbf{D}_n^{-1} - \mathbf{D}_n^{-1}\| = \max_i |d_i^{-1} - p_i^{-1}| = \mathcal{O}((n\rho_n)^{-3/2} \sqrt{\log n}),
\]
\[
\|\hat{\mathbf{W}}^t\|_{\text{max}} = \mathcal{O}(1/(n\rho_n)), \quad \max_{i,i',\eta} \omega_{ii'}^{(\eta)} \asymp 1/n, \quad \min_{i,i',\eta} \omega_{ii'}^{(\eta)} \asymp 1/n,
\] (B.1)

Remark 17 The bounds for $\|\hat{\mathbf{W}}^t\|_{\text{max}}$ given in Eq. (B.1) is generally sub-optimal for $t \geq 2$. Nevertheless we use this bound purely as a stepping stone in proving Theorem 9. Once Theorem 9 is established we can improve the bound for $\|\hat{\mathbf{W}}^t\|_{\text{max}}$ by considering $\|\mathbf{W}^t\|_{\text{max}}$ and $\|\hat{\mathbf{W}}^t - \mathbf{W}^t\|_{\text{max}}$ separately.

Lemma 18 Under Assumption 3, suppose $\rho_n$ satisfies
\[
n^{-1/2} \log \beta_2 n \preceq \rho_n \prec 1
\] (B.2)
for some $\beta_2 > 1/2$. Then for sufficiently large $n$, we have
\[
\|\mathbf{A}^2\|_{\text{max,off}} = \mathcal{O}(n\rho_n^2).
\]
Furthermore, if $t \geq 3$ and
\[
1/n^{1/2} \preceq \rho_n \prec 1,
\] (B.3)
then for sufficiently large $n$ we have
\[
\|\mathbf{A}^t\|_{\text{max}} = \mathcal{O}(n^{t-1}\rho_n^t).
\] (B.4)

Proof Suppose $\rho_n$ satisfies Eq. (B.2) for some $\beta_2 > 1/2$. Define
\[
\varsigma_{ii'}^{i,i'} = a_{ii'} + a_{ii'}^*, \quad \varsigma_{ii'} = \max_{1 \leq i,i',\leq n} \sum_{i',\neq i,i'} a_{ii'} a_{ii'}^* = \max_{1 \leq i,i',\leq n} \sum_{i',\neq i,i'} \varsigma_{ii'}^{i,i'}.
\]

Given $i \neq i'$, the $\{\varsigma_{ii'}^{i,i'}\}_{i \in [n], i',\neq i,i'}$ are a set of independent Bernoulli variables with
\[
c_2^2 \rho_n^2 \leq \mathbb{P}(\varsigma_{ii'}^{i,i'} = 1) \leq c_3^2 \rho_n^2.
\]
A similar argument to that for deriving Eq. (C.2) yields (recall that \( n^{-1/2} \log^{3/2} n \preceq \rho_n \))

\[
\log \left( P \left( \frac{\delta_{ii'}}{n - 1} \leq \frac{\varepsilon^2}{2} \rho_n^2 \right) \right) \leq \log \left( P \left( \delta_{ii'} \leq \frac{1}{2} E(\delta_{ii'}) \right) \right) \preceq -C_1 n \rho_n^2 \preceq -C_1 \log^{2/2} n \tag{B.5}
\]

for all \( i \neq i' \); here \( C_1 \geq 0 \) is a constant not depending on \( n \) or \( \rho_n \). Eq. (B.5) together with a union bound then implies

\[
P \left( \max_{i \neq i'} \delta_{ii'} \leq (n-1) \frac{\varepsilon^2}{2} \rho_n^2 \right) \leq n^2 \max_{i \neq i'} \left\{ P \left( \delta_{ii'} \leq (n-1) \frac{\varepsilon^2}{2} \rho_n^2 \right) \right\} \leq \exp \left( 2 \log n - C_1 \log^{2/2} n \right) \longrightarrow 0
\]

as \( n \rightarrow \infty \). We thus have \( \max_{i \neq i'} \sum_{i',i} a_{ii'} a_{i'i'} = O(\rho_n^2) \) and hence

\[
\|A^2\|_{\operatorname{max},\text{off}} = \max_{i \neq i'} \sum_{i',i} a_{ii'} a_{i'i'} \leq \max_{i \neq i'} \sum_{i',i} a_{ii'} a_{i'i'} + 2 = O(\rho_n^2).
\]

We next consider the case when \( t \geq 3 \) and \( \rho_n \) satisfies Eq. (B.3). We have

\[
\|A^t\|_{\max} \leq \|P^t\|_{\max} + \|A^t - P^t\|_{\max}.
\]

Under Assumption 3,

\[
\|P^t\|_{\max} \leq n\|P^{t-1}\|_{\max} \|P\|_{\max} \leq n^2\|P^{t-2}\|_{\max} \|P\|_{\operatorname{max}}^2 \leq \cdots \leq n^{t-1}\|P\|_{\max}^t = O(n^{t-1} \rho_n^t).
\]

We now focus on bounding \( \|A^t - P^t\|_{\max} \). Consider the following expansion for \( A^t - P^t \)

\[
A^t - P^t = (A^{t-1} - P^{t-1})(A - P) + P^{t-1}(A - P) + \sum_{b_0=1}^{t-1} A^{t-1-b_0}(A - P)P^{b_0}. \tag{B.6}
\]

Applying the same expansion to \( A^{t-1} - P^{t-1}, \ldots, A^2 - P^2 \), we obtain

\[
A^t - P^t = (A^{t-1} - P^{t-1})(A - P) + P^{t-1}(A - P) + \sum_{b_0=1}^{t-1} A^{t-1-b_0}(A - P)P^{b_0}
\]

\[
= \left( \sum_{b_1=0}^{t-2} A^{t-2-b_1}(A - P)P^{b_1} \right)(A - P) + P^{t-1}(A - P) + \sum_{b_0=1}^{t-1} A^{t-1-b_0}(A - P)P^{b_0}
\]

\[
= (A^{t-2} - P^{t-2})(A - P)^2 + \sum_{b_1=1}^{t-2} A^{t-2-b_1}(A - P)P^{b_1}(A - P) + \sum_{b_0=1}^{t-1} A^{t-1-b_0}(A - P)P^{b_0}
\]

\[
+ \sum_{b'=1}^{2} P^{t-b'}(A - P)^{b'}
\]

\[
= \cdots = (A - P)^t + \sum_{i=0}^{t-1} \sum_{b=1}^{t-c-1} A^{t-c-b-1}(A - P)P^{b}(A - P)^c + \sum_{b'=1}^{t-1} P^{t-b'}(A - P)^{b'}.
\tag{B.7}
\]
Now for any summand appearing in $L_1$, if both $c \neq 0$ and $t - c - b - 1 \neq 0$ then
\[
\| A^{t-c-b-1}(A - P)P^b(A - P)^c \|_{\max} \leq \| A \|_1 \times \| A^{t-c-b-2}(A - P)P^b(A - P)^c \|_{\max}
\]
\[
\leq \cdots \leq \| A \|_1^{t-c-b-1} \times \| (A - P)P^b(A - P)^c \|_{\max}
\]
\[
\leq \| A \|_1^{t-c-b-1} \times \| (A - P)P^b(A - P)^c \|_{\max} \times (\| A \|_1 + \| P \|_1)^c.
\]

Here $\| M \|_1$ denote the maximum of the absolute column sum of a matrix $M$. The bound in Eq. (B.8) also holds when $c = 0$ or $t - c - b - 1 = 0$. A similar argument yields
\[
\|(A - P)P^b\|_{\max} \leq \| P \|_{\max} \times (\| A \|_1 + \| P \|_1) \times \| P \|_{1}^{b-1}.
\]

Observe that $\| A \|_1 = \max_i d_i$ and $\| P \|_1 = \max_j \sum p_{ij}$. Combining Eq. (B.8), Eq. (B.9) and Lemma 16, we have
\[
\| L_1 \|_{\max} \leq \sum_{a=0}^{t-1} \sum_{b=1}^{t-c-1} \| A^{t-c-b-1}(A - P)P^b(A - P)^c \|_{\max} = O_{\mathbb{P}}(n^{t-1} \rho_n^t).
\]

Similarly, we also have $\| L_2 \|_{\max} = O_{\mathbb{P}}(n^{t-1} \rho_n^t)$. We therefore have
\[
\| A^t \|_{\max} \leq \| P^t \|_{\max} + \|(A - P)^t\|_{\max} + \| L_1 \|_{\max} + \| L_2 \|_{\max}
\]
\[
= \|(A - P)^t\|_{\max} + O_{\mathbb{P}}(n^{t-1} \rho_n^t).
\]

Finally we bound $\|(A - P)^t\|_{\max}$. Note that $\| M \|_{\max} \leq \| M \|_2$ for any matrix $M$. Now, under Assumption 3, the maximal expected degree of $G$ is of order $n \rho_n \geq n^{1-\beta} \gg \log^4 n$. We can thus apply the spectral norm concentration result in Lu and Peng (2013) to obtain
\[
\|(A - P)^t\|_{\max} \leq \| A - P \|_2 = O_{\mathbb{P}}((n \rho_n)^{t/2}).
\]

We thus have, after a bit of algebra, that $\|(A - P)^t\|_{\max} = O_{\mathbb{P}}(n^{t-1} \rho_n^t)$ whenever $\cdot^{\frac{2}{1-\beta}} \leq \rho_n$. Combining Eq. (B.10), we have $\| A^t \|_{\max} = O_{\mathbb{P}}(n^{t-1} \rho_n^t)$ if Eq. (B.3) holds.

**B.1. Proof of Theorem 9 and Remark 10**

The proof is organized as follows. In **Step 1** through **Step 4**, we bound $\| \hat{W}^t - W^t \|_{\max}$ under the general sparse condition as specified in Assumption 3. These arguments are generalizations of the corresponding arguments in the proof of Theorem 5 in Section A.1. In **Step 5**, we provide an improved bound for $\| \hat{W}^t - W^t \|_{\max}$ when $t \geq 3$ and $\frac{t-3}{t-1} > \beta$. For ease of exposition, we omitted some of the more mundane technical details from the current proof and refer the interested reader to Appendix C.4.

**Step 1 (Bounding $\| \hat{W} - W \|_{\max}$):** Similar to **Step 1** in the proof of Theorem 5, we have
\[
\hat{W} - W = AD_A^{-1} - PD_P^{-1} = \underbrace{AD_P^{-1}D_A^{-1}(D_P - D_A)}_{\Delta_1^{(1)}} + \underbrace{(A - P)D_P^{-1}}_{\Delta_2^{(1)}}
\]

41
and hence

$$\|\Delta_1^{(1)}\|_{\text{max}} = O_P((n\rho_n)^{-3/2}\sqrt{\log n}), \|\Delta_2^{(1)}\|_{\text{max}} = O_P((n\rho_n)^{-1}).$$

We therefore have

$$\|\hat{W} - W\|_{\text{max}} \leq \|\Delta_1^{(1)}\|_{\text{max}} + \|\Delta_2^{(1)}\|_{\text{max}} = O_P((n\rho_n)^{-3/2}\sqrt{\log n}) + O_P((n\rho_n)^{-1}) = O_P((n\rho_n)^{-1}). \quad (B.12)$$

See Appendix C.4.1 for additional details in deriving the above inequalities.

**Step 2 (Bounding \(\|\hat{W}^2 - W^2\|_{\text{max,diag}}\))**: Similar to Step 2 in the proof of Theorem 5,

$$\hat{W}^2 - W^2 = \underbrace{(\hat{W} - W)W + W(\hat{W} - W)}_{\Delta_1^{(2)}}.$$

$$\Delta_1^{(2)} = \underbrace{(A\hat{D}_F^1D_A^1(D_P - D_A) - AD_F^2(D_P - D_A))W + AD_F^2(D_P - D_A)W}_{\Delta_1^{(2,1)}} + \underbrace{(A - P)D_F^1W}_{\Delta_1^{(2,2)}}.$$

The bounds in Appendix C.4.2 then implies

$$\|\Delta_1^{(2)}\|_{\text{max}} = O_P(n^{-3/2}\rho_n^{-1/2}\sqrt{\log n}), \|W(W - \hat{W})\|_{\text{max}} = O_P(n^{-3/2}\rho_n^{-1/2}\sqrt{\log n}).$$

Furthermore, we also have

$$\|\Delta_2^{(2)}\|_{\text{max}} = \|(W - \hat{W})(\hat{W} - W) - W(W - \hat{W})\|_{\text{max}} \leq \|(W - \hat{W})^2\|_{\text{max}} + \|W(W - \hat{W})\|_{\text{max}}.$$ 

Replacing \(\|\cdot\|_{\text{max}}\) with \(\|\cdot\|_{\text{max,diag}}\) in Eq. (A.7), and following the same argument as that for Eq. (A.7) and Eq. (A.8), we have, by Lemma 16,

$$\|(W - \hat{W})^2\|_{\text{max,diag}} = O_P((n\rho_n)^{-1}), \|W^2 - \hat{W}^2\|_{\text{max,diag}} = O_P((n\rho_n)^{-1}). \quad (B.13)$$

**Step 3 (Bounding \(\|\hat{W}^2 - W^2\|_{\text{max,off}}\))**: Similar to Eq. (A.7),

$$\|(W - \hat{W})^2\|_{\text{max,off}} = \max_{i \neq i'} \sum_{i^* = 1}^{n} \left( \frac{p_{ii^*}}{p_i} - \frac{a_{ii^*}}{d_i} \right) \left( \frac{p_{i'i'}}{p_i} - \frac{a_{i'i'}}{d_i} \right) \leq \max_{i \neq i'} \sum_{i^* = 1}^{n} \left( \frac{a_{ii^*}}{p_i} - \frac{a_{ii^*}}{d_i} \right) \left( \frac{p_{i'i'}}{p_i} - \frac{a_{i'i'}}{p_i} \right) + \max_{i \neq i'} \sum_{i^* = 1}^{n} \left( \frac{p_{ii^*}}{p_i} - \frac{a_{ii^*}}{d_i} \right) \left( \frac{a_{i'i'}}{p_i} - \frac{a_{i'i'}}{d_i} \right) \delta_{2,off}^{(2,1)} \delta_{2,off}^{(2,2)} \delta_{2,off}^{(2,3)}.$$
We then have the bounds
\[ \delta^{(2,1)}_{2,\text{off}} = O_P\left((n \rho_n)^{-2} \log n \right), \quad \delta^{(2,2)}_{2,\text{off}} = O_P\left((n \rho_n)^{-2} \log n \right), \quad \delta^{(2,3)}_{2,\text{off}} = O_P\left(n^{-3/2} \rho_n^{-1} \sqrt{\log n} \right). \]

For ease of exposition we only derive the bound for \( \delta^{(2,1)}_{2,\text{off}} \) here. The bounds for \( \delta^{(2,2)}_{2,\text{off}} \) and \( \delta^{(2,3)}_{2,\text{off}} \) are derived similarly; see Appendix C.4.3 for more details.

**Claim:** \( \delta^{(2,1)}_{2,\text{off}} = O_P\left((n \rho_n)^{-2} \log n \right) \). We start by writing
\[
\sum_{i^* = 1}^{n} \left( a_{ii^*} - a_{ii'} \right) \left( \frac{p_{i^*} - p_{i'}}{p_{i^*}^2 d_{i^*}} \left( p_{i^*} - p_{i'} \right) \right) = \frac{1}{p_{i'}} \sum_{i^* = 1}^{n} a_{ii^*} \left( \frac{(d_{i^*} - p_{i^*})^2}{p_{i^*}^2 d_{i^*}} + \frac{p_{i^*} - d_{i^*}}{p_{i^*}^2} \right) \left( p_{i^*} - p_{i'} \right),
\]
and hence
\[
\delta^{(2,1)}_{2,\text{off}} \leq \max_{i \neq i'} \left[ \frac{1}{p_{i'}} \sum_{i^* = 1}^{n} a_{ii^*} \left( \frac{(d_{i^*} - p_{i^*})^2}{p_{i^*}^2 d_{i^*}} \right) \left( p_{i^*} - p_{i'} \right) \right] + \max_{i \neq i'} \left[ \frac{1}{p_{i'}} \sum_{i^* = 1}^{n} a_{ii^*} \left( \frac{d_{i^*} - p_{i^*}}{p_{i^*}^2} \right) \left( p_{i^*} - a_{ii'} \right) \right].
\]

For the term \( \xi_{ii'} \), by Lemma 16, we have
\[
\max |\xi_{ii'}| \leq \left( \max_i \frac{1}{p_i} \times \left( \max_{i'} \left| \frac{(d_{i^*} - p_{i^*})^2}{p_{i^*}^2 d_{i^*}} \right| \right) \right) \times \left( 2 + \max_{i \neq i'} \left| \left\{ i^* : a_{ii^*} = a_{ii'}, a_{ii^*} = 1 \right\} \right| + \left\| P \right\|_{\max} \times \max_{i} \left| \left\{ i^* : a_{ii^*} = 1, a_{ii'} = 0 \right\} \right| \right).
\]

For the term \( \zeta_{ii'} \), first let \( e_{ii'} = a_{ii'} - p_{ii'} \). Then expanding \( d_{i^*} - p_{i^*} \) as \( \sum_{i''} e_{i^*i''} \), we have
\[
\zeta_{ii'} = \frac{1}{p_{i'}} \sum_{i^* = 1}^{n} a_{ii^*} \left( \sum_{i'' \notin \{i', i\}} e_{i^*i''} e_{i'i'} - \sum_{i'' \in \{i', i\}} e_{i^*i''} e_{i'i'} \right).
\]
Now, for fixed \( i, i' \), since \( a_{ii} = p_{ii} = 0 \) for all \( i \), we have

\[
\frac{1}{p_{i'}} \sum_{i^* = 1}^{n} a_{ii^*} \sum_{i'' \notin \{i', i\}} e_{i^*i''} e_{i'i'} = \frac{1}{p_{i'}} \sum_{i'' \notin \{i, i'\}} \sum_{i^*} \sum_{i'' \notin \{i', i\}} \frac{1}{p_{i^*}^2} a_{ii^*} e_{i'^*} e_{i'i'} = \frac{1}{p_{i'}} \sum_{(i^*, i^*) \in T(i, i')} \left( \frac{a_{ii^*} e_{i'^*}}{p_{i^*}^2} + \frac{a_{ii^*} e_{i'^*}^*}{p_{i^*}^2} \right) e_{i'^*} \equiv \mathcal{S}(a_i, a_{ii'}). \quad (B.14)
\]

Here \( T(i, i') = \left\{ (i^*, i^*) | i^* < i^* \notin \{i, i'\}, i^* \notin \{i, i'\} \right\} \) and \( a_i = (a_{ii}) \).

Let us now, in addition to conditioning on \( P \), also condition on both \( a_i \) and \( a_{ii'} \). Then the sum for \( \mathcal{S}(a_i, a_{ii'}) \) in Eq. (B.14) is a sum of independent, mean zero random variables, i.e., once we conditioned on \( P, a_i, \) and \( a_{ii'} \), the \( e_{i^*i^*} \) for \( (i^*, i^*) \in T(i, i') \) are independent.
We then have
\[
\Var\left(\frac{1}{p_{i'}} \sum_{i \neq i'} \left( \frac{a_{ii'} e_{ii'} + a_{ii''} e_{ii''}}{p_{ii'}^2} + \frac{a_{ii'} e_{ii'} + a_{ii''} e_{ii''}}{p_{ii''}^2} \right) e_{ii''} \mid a_i, a_{i'} \right) = \frac{1}{p_{i'}} \sum_{i \neq i'} \left( \frac{a_{ii'} e_{ii'} + a_{ii''} e_{ii''}}{p_{ii'}^2} + \frac{a_{ii'} e_{ii'} + a_{ii''} e_{ii''}}{p_{ii''}^2} \right)^2 \Var[e_{ii''} | a_i, a_{i'}] \approx (n \rho_n)^{-3}. 
\]

Furthermore, we also have
\[
\left| \frac{1}{p_{i'}} \left( \frac{a_{ii'} e_{ii'} + a_{ii''} e_{ii''}}{p_{ii'}^2} + \frac{a_{ii'} e_{ii'} + a_{ii''} e_{ii''}}{p_{ii''}^2} \right) e_{ii''} \right| \approx (n \rho_n)^{-3}. 
\]

Therefore, by Bernstein inequality, for any \( c' > 0 \) there exists a constant \( C' > 0 \) such that, for all sufficiently large \( n \),
\[
\mathbb{P}\left( \left| S(a_i, a_{i'}) \right| \geq C'(n \rho_n)^{-5/2} (d_i + d_{i'})^{1/2} \sqrt{\log n} \right| a_i, a_{i'} \leq n^{-c'}
\]
We can now uncondition with respect to \( a_i \) and \( a_{i'} \). More specifically, for any \( t \), we have
\[
\mathbb{P}\left( \left| S(a_i, a_{i'}) \right| \geq t \right) \leq \mathbb{P}\left( \left| S(a_i, a_{i'}) \right| \geq t \left| \max \{d_i, d_{i'}\} < C \rho_n \} \times \mathbb{P}\left( \max \{d_i, d_{i'}\} < C \rho_n \right) 
+ \mathbb{P}\left( \max \{d_i, d_{i'}\} \geq C \rho_n \right)
\]
where \( C \) is any arbitrary positive constant. Now let \( c \) be arbitrary. Then by Lemma 16 and Eq. \( \text{(B.16)} \), together with taking \( t = C'(n \rho_n)^{-2} \sqrt{\log n} \) for some constant \( C' \), we have
\[
\mathbb{P}\left( \left| S(a_i, a_{i'}) \right| \geq C'(n \rho_n)^{-2} \sqrt{\log n} \right) \leq 2n^{-c}
\]
A union bound over the \( \binom{n}{2} \) possible choices of \( a_i \) and \( a_{i'} \) yields
\[
\max_{i \neq i'} \left| S(a_i, a_{i'}) \right| = O_P((n \rho_n)^{-2} \sqrt{\log n}) \quad (\text{B.17})
\]
Finally we also have
\[
\max_{i, i'} \left| \frac{1}{p_{i'}} \sum_{i \neq i'} \frac{a_{ii'} e_{ii'} + a_{ii''} e_{ii''}}{p_{ii'}^2} \sum_{i \neq i' \in \{i, i'\}} e_{ii''} \right| \approx \max_i \left( \frac{\max_{i'} 1/p_i}{3} \right) \approx O_P((n \rho_n)^{-2}).
\]
In summary, we have \( \max_{i' \neq i'} \zeta_{ii'} = O_P((n \rho_n)^{-2} \sqrt{\log n}) \) and hence
\[
\delta_{2, \text{off}}^{(2.1)} \leq \max_{i' \neq i} \zeta_{ii'} + \max_{i' \neq i} \zeta_{ii'} = O_P((n \rho_n)^{-2} \log n).
\]
The claim is thus verified.
Given the previous claim together with an argument similar to that for Eq. \( \text{(A.9)} \), we obtain
\[
\| W^2 - \hat{W}^2 \|_{\max, \text{off}} \leq \| \Delta_1^{(2)} \|_{\max} + \| W(W - \hat{W}) \|_{\max} + \delta_{2, \text{off}}^{(2.1)} + \delta_{2, \text{off}}^{(2.2)} + \delta_{2, \text{off}}^{(2.3)} = O_P\left( \max \left\{ (n \rho_n)^{-2} \log n, n^{-3/2} \rho_n^{-1} \sqrt{\log n} \right\} \right)
\]
\[
= \begin{cases} 
O_P(n^{-3/2} \rho_n^{-1} \sqrt{\log n}) & 0 \leq \beta < 1/2 \\
O_P((n \rho_n)^{-2} \log n) & \text{otherwise}.
\end{cases}
\]
Step 4 (Bounding $\|\hat{W}^t - W^t\|_{\max}$ for $t \geq 3$): The following argument is a generalization of the argument in Step 3 of Appendix A.1. We first consider $t = 3$. We have

$$\|\hat{W}^3 - W^3\|_{\max} \leq \|(\hat{W}^2 - W^2)\hat{W}\|_{\max} + \|W^2(\hat{W} - W)\|_{\max}$$  \hspace{1cm} (B.18)

For the first term in the RHS, by Eq. (B.13) and Lemma 16, we have

$$\|(\hat{W}^2 - W^2)\hat{W}\|_{\max} \leq \max_i d_i \times \|\hat{W}\|_{\max} \|\hat{W}^2 - W^2\|_{\max,off}$$

$$= \mathcal{O}_p(1)\|\hat{W}^2 - W^2\|_{\max,off}$$  \hspace{1cm} (B.19)

For the second term in the RHS of Eq. (B.18), we apply the similar technique for $\Delta_i^{2,1}$ in Step 2 and deduce

$$\|W^2(\hat{W} - W)\|_{\max} = \mathcal{O}_p(1)\|W^2 - \hat{W}^2\|_{\max,off}. \hspace{1cm} (B.20)$$

Combining Eq. (B.18), Eq. (B.19) and Eq. (B.20), we have

$$\|\hat{W}^3 - W^3\|_{\max} = \mathcal{O}_p\Big( \max\left\{ (n\rho_n)^{-2} \log n, n^{-3/2} \rho_n^{-1} \sqrt{\log n} \right\} \Big). \hspace{1cm} (B.21)$$

Now for $t = 4$, we have

$$\hat{W}^4 - W^4 = (\hat{W}^2 - W^2)\hat{W}^2 + W^2(\hat{W}^2 - W^2)$$

$$= (\hat{W}^2 - W^2)\hat{W}^2 + (W^2 - \hat{W}^2)(\hat{W}^2 - W^2) + \hat{W}^2 (W^2 - W^2).$$  \hspace{1cm} (B.22)

For $(\hat{W}^2 - W^2)\hat{W}^2$, by Eq. (B.19) and Lemma 16

$$\|(\hat{W}^2 - W^2)\hat{W}^2\|_{\max} \leq \max_i d_i \times \|\hat{W}\|_{\max} \times \|(\hat{W}^2 - W^2)\hat{W}\|_{\max}$$

$$= \mathcal{O}_p(n\rho_n \times (n\rho_n)^{-1}) \times \mathcal{O}_p(1)\|W^2 - \hat{W}^2\|_{\max,off}$$  \hspace{1cm} (B.23)

Similarly, we have $\|W^2(\hat{W}^2 - W^2)\|_{\max} = \mathcal{O}_p(1)\|W^2 - \hat{W}^2\|_{\max,off}$. Furthermore,

$$\|(W^2 - \hat{W}^2)^2\|_{\max} \leq (n - 1) \times \|W^2 - \hat{W}^2\|_{\max,off}^2 + \|W^2 - \hat{W}^2\|_{\max,diag}^2$$

$$= \mathcal{O}_p((n\rho_n)^{-2} \log n). \hspace{1cm} (B.24)$$

Combining Eq. (B.22), Eq. (B.23) and Eq. (B.24), we obtain

$$\|\hat{W}^4 - W^4\|_{\max} \leq \mathcal{O}_p(1)\|(W^2 - \hat{W}^2)\hat{W}^2\|_{\max} + \|(W^2 - \hat{W}^2)^2\|_{\max}$$

$$= \mathcal{O}_p\left( \max\left\{ (n\rho_n)^{-2} \log n, n^{-3/2} \rho_n^{-1} \sqrt{\log n} \right\} \right). \hspace{1cm} (B.25)$$

For any $t \geq 5$, we can write $\hat{W}^t - W^t$ as

$$\hat{W}^t - W^t = \left(\Delta_{(3)}^{(3)} \Delta_{(3)}^{(5)} + \Delta_{(3)}^{(5)} \Delta_{(3)}^{(2)}\right).$$
We first consider $\Delta_1^{(3)}$. By Lemma 16 and Eq. (B.19)
\[
\|\Delta_1^{(3)}\|_{\text{max}} \leq \max_i d_i \times \|\hat{W}\|_{\text{max}} \times \|(\hat{W}^2 - W^2)\hat{W}^{t-3}\|_{\text{max}}
\leq \cdots \leq \left( \max_i d_i \times \|\hat{W}\|_{\text{max}} \right)^{t-3} \times \|(\hat{W}^2 - W^2)\hat{W}\|_{\text{max}}
= \mathcal{O}_\mathbb{P} \left( \max \left\{ (n\rho_n)^{-2} \log n, n^{-3/2}\rho_n^{-1} \sqrt{\log n} \right\} \right).
\]
For $\Delta_2^{(3)}$, by Lemma 16
\[
\|\Delta_2^{(3)}\|_{\text{max}} \leq n \cdot \|W^2\|_{\text{max}} \|W^{t-2} - W^{t-2}\|_{\text{max}} = \mathcal{O}(1)\|\hat{W}^{t-2} - W^{t-2}\|_{\text{max}}.
\]
Similar to Eq. (A.18), we obtain
\[
\|\hat{W}^t - W^t\|_{\text{max}} = \mathcal{O}_\mathbb{P} \left( \max \left\{ (n\rho_n)^{-2} \log n, n^{-3/2}\rho_n^{-1} \sqrt{\log n} \right\} \right) + \mathcal{O}(1)\|\hat{W}^4 - W^4\|_{\text{max}},
\]
\[
\|\hat{W}^t - W^t\|_{\text{max}} = \mathcal{O}_\mathbb{P} \left( \max \left\{ (n\rho_n)^{-2} \log n, n^{-3/2}\rho_n^{-1} \sqrt{\log n} \right\} \right) + \mathcal{O}(1)\|\hat{W}^3 - W^3\|_{\text{max}},
\]
depending on whether $t$ is even or odd. Eq. (B.21) and Eq. (B.25) then implies
\[
\|\hat{W}^t - W^t\|_{\text{max}} = \mathcal{O}_\mathbb{P} \left( \max \left\{ (n\rho_n)^{-2} \log n, n^{-3/2}\rho_n^{-1} \sqrt{\log n} \right\} \right)
= \begin{cases} 
\mathcal{O}(n^{-3/2}\rho_n^{-1} \sqrt{\log n}) & \text{if } 0 \leq \beta < 1/2 \\
\mathcal{O}(\rho_n^{-2} \log n) & \text{otherwise.}
\end{cases}
\]

**Step 5 (Bounding $\|\hat{W}^t - W^t\|_{\text{max}}$ for $t \geq 4$ and $\frac{t-3}{t-1} \geq \beta$):** Similar to Eq. (B.6) and Eq. (B.7) in the proof of Lemma 18, we write
\[
\hat{W}^t - W^t = (\hat{W} - W)^t + \sum_{r=0}^{t-1} \sum_{s=1}^{r-1} \hat{W}^{t-r-s-1}(\hat{W} - W)W^s(\hat{W} - W)^r
+ \sum_{r'=1}^{t-1} W^{t-r'}(\hat{W} - W)^{r'}.
\]
In **Step 2**, we shown $\| (\hat{W} - W)W \|_{\text{max}} = \mathcal{O}_\mathbb{P} (n^{-3/2}\rho_n^{-1/2} \sqrt{\log n})$. A similar argument to that for $L_1$ in the proof of Lemma 18 yields, for $0 \leq r \leq t-1$ and $1 \leq s \leq t - r - 1$, that
\[
\|\hat{W}^{t-r-s-1}(\hat{W} - W)W^s(\hat{W} - W)^r\|
\leq \left( \max_i d_i \right)^{t-r-s-1} \|\hat{W}\|_{\text{max}}^{t-r-s-1} \|\hat{W} - W\|_{\text{max}} \cdot (n\|W\|_{\text{max}} + \max_i d_i \|\hat{W} - W\|_{\text{max}})^r
\leq \mathcal{O}_\mathbb{P}(1) \cdot \|\hat{W} - W\|_{\text{max}}^{s-1} \|W\|_{\text{max}}^{s-1}
\leq \mathcal{O}_\mathbb{P}(1) \cdot \|\hat{W} - W\|_{\text{max}} = \mathcal{O}_\mathbb{P} (n^{-3/2}\rho_n^{-1/2} \sqrt{\log n}),
\]
where the second inequality follows from Lemma 16 and Eq. (B.12) and the last inequality follows from Lemma 16. As $t$ is finite, we have
\[
\| \sum_{r=0}^{t-1} \sum_{s=1}^{r-1} \hat{W}^{t-r-s-1}(\hat{W} - W)W^s(\hat{W} - W)^r \|_{\text{max}} = \mathcal{O}_\mathbb{P} (n^{-3/2}\rho_n^{-1/2} \sqrt{\log n}).
\]
Similarly,
\[
\left\| \sum_{r'=1}^{t-1} \mathbf{W}^{t-r'} (\hat{\mathbf{W}} - \mathbf{W})^{r'} \right\|_{\max} = O_{\mathbb{P}} \left( n^{-3/2} \rho_n^{-1/2} \sqrt{\log n} \right).
\]
Recalling Eq. (B.26), we obtain
\[
\|\hat{\mathbf{W}}^t - \mathbf{W}^t\|_{\max} \leq \| (\hat{\mathbf{W}} - \mathbf{W})^t \|_{\max} + O_{\mathbb{P}} \left( n^{-3/2} \rho_n^{-1/2} \sqrt{\log n} \right).
\tag{B.27}
\]
Now we focus on \((\hat{\mathbf{W}} - \mathbf{W})^t\). We start with the polynomial expansion
\[
(\hat{\mathbf{W}} - \mathbf{W})^t = (\mathbf{A}(\mathbf{D}_A^{-1} - \mathbf{D}^{-1}_p) + (\mathbf{A} - \mathbf{P})\mathbf{D}^{-1}_p)^t = ((\mathbf{A} - \mathbf{P})\mathbf{D}^{-1}_p)^t + \sum_{c \in \{1,2\}^t \atop c \neq \{1,\cdots,1\}} \prod_{r=1}^{t} \Xi_{c_r}.
\]
Here \(c_r\) represents the \(r\)th coordinate of \(c = (c_1, \ldots, c_t) \in \{1,2\}^t\) and
\[
\Xi_{c_r} = \begin{cases} 
(\mathbf{A} - \mathbf{P})\mathbf{D}^{-1}_p, & \text{if } c_r = 1 \\
(\mathbf{A}(\mathbf{D}_A^{-1} - \mathbf{D}^{-1}_p)), & \text{if } c_r = 2.
\end{cases}
\]
We note that there are \(2^t - 1\) distinct \(c \neq \{1,1,\ldots,1\}\). Now for any given \(c\), let \(r^* = r^*(c)\) be the smallest value of \(r\) such that \(c_r = 2\). We emphasize that \(r^*\) depends on \(c\); however, for simplicity of notation we make this dependency implicit. We further denote
\[
(\Xi_{c_1}, \Xi_{c_2}) = \begin{cases} 
(\mathbf{A}\mathbf{D}_A^{-1}, -\mathbf{A}\mathbf{D}^{-1}_p), & \text{if } c_1 = 1, \\
(\mathbf{A}\mathbf{D}_A^{-1}, -\mathbf{D}^{-1}_p), & \text{if } c_1 = 2.
\end{cases}
\]
so that \(\Xi_{c_r} = \Xi_{c_1}^1 + \Xi_{c_2}^2 c_r\). Given \(c\), we could write
\[
\prod_{r=1}^{t} \Xi_{c_r} = \prod_{j<r^*} (\Xi_{c_j}^1 + \Xi_{c_j}^2) \times \prod_{k>r^*} (\Xi_{c_k}^1 + \Xi_{c_k}^2)
= \sum_{m \in \{1,2\}^{t-r^*}} \prod_{j<r^*} \Xi_{c_j}^m \times \prod_{k>r^*} \Xi_{c_k}^m.
\tag{B.28}
\]
Now for any \(c_r \in \{1,2\}^t\) and \(m \in \{1,2\}\), the \(ii'\)th entry of \(\Xi_{c_r}^m\) is \(\xi_{ii',m}^{c_r,m} \times \phi_{ii',m}^{c_r,m}\) where
\[
(\xi_{ii',m}^{c_r,m}, \phi_{ii',m}^{c_r,m}) = \begin{cases} 
(a_{ii'}, 1/d_{ii'}) & \text{if } c_r = 1 \text{ and } m = 1, \\
(a_{ii'}, -1/p_{ii'}) & \text{if } c_r = 1 \text{ and } m = 2, \\
(a_{ii'}, 1/p_{ii'}) & \text{if } c_r = 2 \text{ and } m = 1, \\
(p_{ii'}, -1/p_{ii'}) & \text{if } c_r = 2 \text{ and } m = 2.
\end{cases}
\tag{B.29}
\]
Using the above notations, we can now write the \(ii'\)th entry of \(\Xi_{c_r}^m\) as
\[
\Xi_{ii'}^{m,c} = \sum_{c \in \{1,\ldots,n\}^{t-1}} \left( \prod_{j<r^*} \xi_{ij_{j-1}i_{j-1}}^{m} \times \phi_{ij_{j-1}i_{j-1}}^{m} \right) \times a_{i_{r^*}-1i_{r^*}} \times \left( \frac{1}{d_{i_{r^*}}} - \frac{1}{p_{i_{r^*}}} \right) \times \left( \prod_{k>r^*} \xi_{ik_{k-1}i_{k-1}}^{m} \times \phi_{ik_{k-1}i_{k-1}}^{m} \right)
\]
where, with a slight abuse of notation, we denote \( i = (i_1, \ldots, i_{t-1}) \in \{1, \ldots, n\}^{t-1} \) and let \( i_0 = i \) and \( i_t = i' \). We therefore have

\[
\|\mathcal{M}_{ii'}\| \leq \sum_i a_{i_r-1,i_r} \prod_{j \neq r^*} \xi_{i_{j-1}i_j}^{c_j, m_j} \times \left| \left( \frac{1}{d_{i_r^*}} - \frac{1}{p_{i_r^*}} \right) \times \prod_{j \neq r^*} \varphi_{i_{j-1}i_j}^{c_j, m_j} \right| \\
\leq \|D_A^{-1}\|_{s(c, m)} \times \|D_P^{-1}\|_{t-1-s(c, m)} \times \|D_A^{-1} - D_P^{-1}\| \times \left( \sum_i a_{i_*-1,i_*} \prod_{j \neq r^*} \xi_{i_{j-1}i_j}^{c_j, m_j} \right).
\]

Here \( s(c, m) \) is the number of indices \( r \) with \( r \neq r^* \) and \( (c_r, m_r) = (1, 1) \). Now, by Lemma 16 and Assumption 3, we have

\[
\|\mathcal{M}_{m,c}\| \leq O_p((n \rho_n)^{-1/2-t} \sqrt{\log n}) \times \max_{ii'} \left( \sum_i a_{i_*-1,i_*} \prod_{j \neq r^*} \xi_{i_{j-1}i_j}^{c_j, m_j} \right). \tag{B.30}
\]

with the convention \( i_0 = i \) and \( i_t = i' \). Now define a matrix-valued function \( \xi_{A,P}(-) \) by

\[
\xi_{A,P}(c_r, m_r) = \begin{cases} 
A & \text{if } (c_r, m_r) \in \{(1, 1), (1, 2), (2, 1)\}, \\
P & \text{if } (c_r, m_r) = (2, 2), 
\end{cases}
\]

Also define

\[
\xi_{A,P}^{m,c} = \left( \prod_{j < r^*} \xi_{A,P}(c_j, m_j) \right) \times A \times \left( \prod_{k > r^*} \xi_{A,P}(c_k, m_k) \right).
\]

Then by the definition of the \( \xi_{si'}^{m,c} \) in Eq. (B.29), we have

\[
\|\xi_{A,P}^{m,c}\| \leq \max_{ii'} \left( \sum_i a_{i_*-1,i_*} \prod_{j \neq r^*} \xi_{i_{j-1}i_j}^{c_j, m_j} \right).
\]

We now consider two cases to bound \( \|\xi_{A,P}^{m,c}\|_{\max} \).

**Case 1:** Suppose that for the given \( m, c \), there exists at least one index \( r \neq r^* \) with \( (c_r, m_r) = 2 \), i.e., the matrix \( P \) appears at least once among the collection of \( \xi_{A,P}(c_r, m_r) \) for \( r \neq r^* \). Then \( \xi_{A,P}^{m,c} \) must have the form

\[
\begin{array}{cccc}
\cdots & \cdots & \cdots & \cdots \\
g(c,m) & PA & \cdots & t-2-g(c,m) \\
\end{array}
\text{or}
\begin{array}{cccc}
\cdots & \cdots & \cdots & \cdots \\
g(c,m) & \cdots & \cdots & t-2-g(c,m) \\
\end{array}
\tag{B.31}
\]

48
Note that it is possible that \( g(c, m) = 0 \) or \( g(c, m) = t - 2 \) as \( A \) could be either the first or last matrix in the product \( \xi_{A,P}^{m,c} \). Consider the first form in Eq. (B.31). We have

\[
\| \xi_{A,P}^{m,c} \|_{\max} = \| \begin{array}{cccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max}
\]

\[
= \begin{cases}
\max_{i,j} d_{ij} \times \| \begin{array}{ccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max} & \text{if first matrix is } A \ \\
\max_{i,j} d_{ij} \times \| \begin{array}{ccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max} & \text{if first matrix is } P
\end{cases}
\]

\[
\leq \mathcal{O}(n) \| \begin{array}{ccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max} \leq \mathcal{O}(n) \| \begin{array}{ccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max}
\]

\[
\leq \mathcal{O}(\rho_{n}) \| \begin{array}{ccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max} \leq \mathcal{O}(\rho_{n}) \| \begin{array}{ccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max}
\]

\[
\leq \mathcal{O}(\rho_{n}) \| \begin{array}{ccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max} \leq \mathcal{O}(\rho_{n}) \| \begin{array}{ccc}
\cdots & PA & \cdots \\
\end{array} \|_{\max}
\]

In the above inequality, all relationships with \( \leq \) and \( \leq \) are when we removed either \( P \) or \( A \) from before and after the term \( PA \), respectively. An identical argument also yields

\[
\| \xi_{A,P}^{m,c} \|_{\max} = \mathcal{O}(n^{t-1} \rho_{n}^t)
\]

for the second form in Eq. (B.31).

**Case 2:** Suppose now that, for all \( r \neq r^* \), we have \((c_r, m_r) \neq (2, 2)\), i.e., \( \xi_{A,P}(c_r, m_r) = A \) for all \( r \neq r^* \). Then \( \xi_{A,P}^{m,c} = A^t \) and, by the fact that \( \frac{t-3}{t-1} > \beta \) already yields Eq. (B.3) hold, we have by Lemma 18,

\[
\| \xi_{A,P}^{m,c} \|_{\max} = \| A^t \|_{\max} = \mathcal{O}(n^{t-1} \rho_{n}^t).
\]

Combining Eq. (B.28), Eq. (B.30), Eq. (B.32) and Eq. (B.33), we have

\[
\| W_n - W_n^t \|_{\max} \leq \| (W_n - W_n^t) \|_{\max} + \mathcal{O}(n^{-3/2} \rho_{n}^{-1/2} \sqrt{\log n})
\]

\[
\leq \| (A - P) D_P^{-1} \|_{\max} + \mathcal{O}(n^{-3/2} \rho_{n}^{-1/2} \sqrt{\log n})
\]

\[
\leq \| (A - P) D_P^{-1} \|_{\max} + \mathcal{O}(n^{-3/2} \rho_{n}^{-1/2} \sqrt{\log n}).
\]
The last inequality in the above display follows from the fact that the number of distinct $c$ is $2^t - 1$ which is finite and does not depend on $n$.

Finally we focus on $\|(A - P)^t P^{-1}\|_{\text{max}}$. An argument similar to that for $\|(A - P)^t\|_{\text{max}}$ in the proof of Lemma 18 yields

$$\|(A - P)^t P^{-1}\|_{\text{max}} \leq \|(A - P)^t\|_2 \times \|P^{-1}\|_2 \lesssim (n \rho_n)^{-t/2} \|A - P\|_2$$

(By Lemma 16) (By Eq. (B.11)).

It could be directly checked that the rate $(n \rho_n)^{-t/2} \lesssim n^{-3/2} \rho_n^{-1/2} \sqrt{\log n}$ when $t - 3 \rho_n > \beta$. Thus, together with Eq. (B.34), we conclude

$$\|\hat{W}^t - W^t\|_{\text{max}} = O_P((n \rho_n)^{-3/2} \rho_n^{-1/2} \sqrt{\log n})$$

for $t \geq 4$ and $t - 3 \rho_n > \beta$, as desired.

Appendix C. Additional Proofs for Completeness

C.1. Proof of Lemma 14

(1): The sum of $i$th row of $AD^{-1}$ is

$$\sum_{i=1}^n a_{ii'} d_{i'} = \sum_{i=1}^n a_{ii'} = 1.$$ 

So $1_n^T \cdot \hat{W} = 1_n^T \cdot AD^{-1} = 1_n^T$ and

$$1_n^T \cdot \hat{W}^t = 1_n^T AD^{-1} \cdots AD^{-1} = (1_n^T \cdot AD^{-1} \cdots AD^{-1})^{t-1} = 1_n^T.$$ 

With similar argument we have $1_n^T W^t = 1_n^T$.

(2) The $i$th element of $\hat{W} d$ is

$$\sum_{j=1}^n a_{ii'} d_{i'} = d_i$$ 

and hence $\hat{W} d = d$. This implies $\hat{W}^t d = d$. The same argument yields $W^t p = p$.

C.2. Proof of Lemma 15

(1) $\|D_A\|, \|D_A^{-1}\|$: With Chernoff bound, since $c_0 \leq p_{ii'} \leq c_1$ for any fixed $i$ we can get

$$\mathbb{P}\left(\frac{d_i}{n} > \frac{3c_1}{2}\right) \leq \mathbb{P}\left(d_i > \frac{3p_i}{2}\right) \lesssim \exp(-C_0 \cdot n)$$

$$\mathbb{P}\left(\frac{1}{d_i} \frac{1}{n} > \frac{2}{c_0} \frac{1}{n}\right) \leq \mathbb{P}\left(d_i \leq \frac{p_i}{2}\right) \lesssim \exp(-C_1 \cdot n)$$
for some $C_0, C_1 > 0$ by taking appropriate constant in general Chernoff bound. So we have
\[
P\left(\frac{\max_{1 \leq i \leq n} d_i}{n} \leq \frac{3c_1}{2}\right) \lesssim n \exp\left(-C_0 \cdot n\right) \to 0
\]
\[
P\left(\left(\max_{1 \leq i \leq n} \frac{1}{d_i}\right) / \frac{n}{1} \leq \frac{2}{c_0}\right) \lesssim n \exp\left(-C_1 \cdot n\right) \to 0
\]
as $n \to +\infty$ and $\|D_A\| = \max_{1 \leq i \leq n} d_i = O_P(n)$, $\|D_A^{-1}\| = \max_{1 \leq i \leq n} 1/d_i = O_P(1/n)$.

(2) $\|D_A - D_P\|$: For a given $i$, $a_{1i}, \ldots, a_{ni}$ are independent and $Ea_{ii} = p_{ii}$ for any $1 \leq i, i' \leq n$. Also $|a_{ii} - p_{ii}| \leq 1$. By Bernstein inequality, we have
\[
P\left(\left|\sum_{j=1}^n a_{ij} - \sum_{j=1}^n p_{ij}\right| > \tilde{t}\right) \leq 2 \exp\left(-\frac{\tilde{t}^2}{2\sigma_0^2 n + \frac{\tilde{t}^2}{3}}\right),
\]
where $\sigma_0^2 = \frac{1}{n} \sum_{j=1}^n \text{Var}(a_{ij} - p_{ij}) = \frac{1}{n} \sum_{j=1}^n p_{ij}(1 - p_{ij})$. So we have $\sigma_0^2 \leq \frac{1}{n} \sum_{j=1}^n (p_{ij} + 1 - p_{ii})/4 = 1/4$ and
\[
P\left(\left|\sum_{j=1}^n a_{ij} - \sum_{j=1}^n p_{ij}\right| > \tilde{t}\right) \leq 2 \exp\left(-\frac{\tilde{t}^2}{2n + \frac{2}{3}\tilde{t}}\right). \tag{C.1}
\]
Take $\tilde{t} = c\sqrt{n \log n}$ in Eq. (C.1), we have
\[
P\left(\left|\sum_{j=1}^n a_{ij} - \sum_{j=1}^n p_{ij}\right| > c\sqrt{n \log n}\right) \leq 2 \exp\left(-\frac{c^2 n \log n}{2n + \frac{2}{3}c\sqrt{n \log n}}\right)
\]
\[
\lesssim \exp\left(-2c^2 \log n\right) = \left(\frac{1}{n}\right)^{2c^2}.
\]
Combining all the events among $i \in \{1, \ldots, n\}$, we get
\[
P\left(\|D_A - D_P\| > c\sqrt{n \log n}\right) = P\left(\max_{1 \leq i \leq n} \left|\sum_{j=1}^n a_{ij} - \sum_{j=1}^n p_{ij}\right| > c\sqrt{n \log n}\right)
\]
\[
= P\left(\bigcup_{i=1}^n \left\{\left|\sum_{j=1}^n a_{ij} - \sum_{j=1}^n p_{ij}\right| > c\sqrt{n \log n}\right\}\right) \lesssim n \cdot \left(\frac{1}{n}\right)^{2c^2} = \left(\frac{1}{n}\right)^{2c^2-1},
\]
which implies $\|D_A - D_P\| = O_P(\sqrt{n \log n})$.

(3) $\|D_A^{-1} - D_P^{-1}\|$: With the results in (1) and (2), we have
\[
\|D_A^{-1} - D_P^{-1}\| = \max_i \left|\frac{1}{d_i} - \frac{1}{p_i}\right| = \max_i \left|\frac{p_i - d_i}{d_ip_i}\right|
\]
\[
\leq \|D_A - D_P\| \cdot \|D_A^{-1}\| \cdot \|D_P^{-1}\| = O_P(n^{-3/2} \sqrt{\log n}).
\]

(4) $\|\hat{W}^i\|_{\text{max}}, \max_{i,i'} w_{ii'}^{(t)} \asymp 1/n$ and $\min_{i,i'} w_{ii'}^{(t)} \asymp 1/n$: By result in (1) and $A, P$ are bounded by 1 elementwisely, it is easy to see that
\[
\|\hat{W}\|_{\text{max}} = \|AD_A^{-1}\|_{\text{max}} \leq \|A\|_{\text{max}} \cdot \|D_A^{-1}\| \leq \max_{1 \leq i \leq n} 1/d_i = O_P(n^{-1}).
\]
For any $t \geq 1$,

$$
\|\hat{W}^t\|_{\text{max}} \leq n\|\hat{W}^{t-1}\|_{\text{max}} \leq n^2\|\hat{W}^{t-2}\|_{\text{max}} \leq \cdots \leq n^{-t-1}\|\hat{W}\|_{\text{max}} = O_P(n^{-1}).
$$

Since $c_0 \leq p_{ii'} \leq c_1$, we could also see

$$
\frac{c_0 t}{c_1 n} \leq w_{ii'}^{(t)} \leq \frac{c_1 t}{c_0 n}
$$

for any given $t \geq 1$ and $1 \leq i, i' \leq n$, which implies $\min_{i,i'} w_{ii'}^{(t)}, \max_{i,i'} w_{ii'}^{(t)} \propto n^{-1}$.

### C.3. Proof of Lemma 16

**Proof (1) $\|D_\mathbf{A}\|, \|D_{\mathbf{A}}^{-1}\|$:** Since $c_2 \rho_n \leq p_{ii'} \leq c_3 \rho_n$ and $E d_i = p_i$, we have

$$
\log \left( \mathbb{P} \left( \frac{d_i}{n} > \frac{3c_3 \rho_n}{2} \right) \right) \leq \log \left( \mathbb{P} \left( d_i > \frac{3}{2} \hat{p}_i \right) \right) \leq -C_0 n \rho_n \leq -C_0 \log \beta_1 n
$$

$$
\log \left( \mathbb{P} \left( \frac{1}{d_i} \geq \frac{1}{n} \right) \right) \leq \log \left( \mathbb{P} \left( d_i \leq \frac{p_i}{2} \right) \right) \leq -C_0 n \rho_n \leq -C_0 \log \beta_1 n
$$

(\ref{eq:c3})

by Chernoff bound. Here $C_0 > 0$ is a sufficiently small and fixed constant for any $1 \leq i \leq n$. Take the exponential of both sides in Eq. (C.3) for all $i$,

$$
\mathbb{P} \left( \max_{1 \leq i \leq n} \frac{d_i}{n} > \frac{3c_3 \rho_n}{2} \right) \leq n \max_i \left\{ \mathbb{P} \left( d_i > \frac{3 \hat{p}_i}{2} \right) \right\} \leq \exp \left( \log n - C_0 \log \beta_1 n \right) \to 0
$$

as $n \to +\infty$. So $\|D_\mathbf{A}\| = \max_{1 \leq i \leq n} d_i = O_P(n \rho_n), \|D_{\mathbf{A}}^{-1}\| = \max_{1 \leq i \leq n} 1/d_i = O_P(1/(n \rho_n))$.

**Proof (2) $\|D_\mathbf{A} - D_P\|, \|D_{\mathbf{A}}^{-1} - D_{\mathbf{P}}^{-1}\|$:** For a given vertex $i$, $\{a_{i1}, \ldots, a_{in}\}$ are independent, $\mathbb{E} a_{ii'} = p_{ii'}$ for any $1 \leq i, i' \leq n$ and $|a_{ii'} - p_{ii'}| \leq 1$. By Bernstein inequality, we have

$$
\mathbb{P} \left( \left| \sum_{i'=1}^{n} a_{ii'} - \sum_{i'=1}^{n} p_{ii'} \right| > \hat{t} \right) \leq 2 \exp \left( -\frac{\hat{t}^2}{2 \sigma_0^2 + \frac{\hat{t}}{3}} \right),
$$

(\ref{eq:c4})

where $\sigma_0^2 = \frac{1}{n} \sum_{i'=1}^{n} \text{Var}(a_{ii'} - p_{ii'}) = \frac{1}{n} \sum_{i'=1}^{n} p_{ii'} (1 - p_{ii'})$. Since $\min_{i,i'} p_{ii'}, \max_{i,i'} p_{ii'} \propto \rho_n$ and $\rho_n \to 0$, we have $\sigma_0^2 \propto \rho_n$. By take $\hat{t} = C \sqrt{n \rho_n \log n}$ in Eq. (\ref{eq:c4}), we have $2\sigma_0^2(n) \propto n \rho_n, \frac{\hat{t}}{3} \propto \sqrt{n \rho_n \log n}$ and so

$$
2\sigma_0^2 / \left( \frac{\hat{t}}{3} \right) \times \sqrt{n \rho_n \log n} \to +\infty
$$

by Assumption 3, which implies

$$
\frac{\hat{t}^2}{2 \sigma_0^2 n + \frac{\hat{t}}{3}} \propto \frac{\hat{t}^2}{2 \sigma_0^2 n} \propto C^2 \log n.
$$

52
With sufficient large $C$, we could have

$$\mathbb{P}\left( \left| \sum_{j=1}^{n} a_{ii'} - \sum_{j=1}^{n} p_{ii'} \right| > C \sqrt{n \rho_n \log n} \right) \leq 2 \exp \left( -2 \log n \right)$$

and hence

$$\mathbb{P}\left( \max_{1 \leq i \leq n} \left\{ \left| \sum_{j=1}^{n} a_{ii'} - \sum_{j=1}^{n} p_{ii'} \right| \right\} > C \sqrt{n \rho_n \log n} \right) \leq n \cdot 2 \exp \left( -2 \log n \right) = \frac{2}{n} \to 0$$

as $n \to +\infty$, which implies $\|D_A - D_P\| = O_P(\sqrt{n \rho_n \log n})$. Then combine with the above results,

$$\|D_A^{-1} - D_P^{-1}\| = \max_i \left| \frac{1}{d_i} - \frac{1}{p_i} \right| = \max_i \left| \frac{p_i - d_i}{d_i p_i} \right| \leq \|D_A - D_P\| \cdot \|D_A^{-1}\| \cdot \|D_P^{-1}\| = O_P((n \rho_n)^{-3/2} \sqrt{\log n}).$$

(3) $\|\hat{W}^t\|_{\max}, \max_{i,i'} w_{ii'}^{(t)} \asymp 1/n$ and $\min_{i,i'} w_{ii'}^{(t)} \asymp 1/n$: Since $A, P$ are bounded by 1 elementwisely,

$$\|\hat{W}\|_{\max} = \|A D_A^{-1}\|_{\max} \leq \|A\|_{\max} \cdot \|D_A^{-1}\| \leq \max_{1 \leq i \leq n} 1/d_i = O_P((n \rho_n)^{-1}).$$

So for any $t \geq 1$,

$$\|\hat{W}^t\|_{\max} \leq \left( \max_i d_i \right) \cdot \|\hat{W}^{t-1}\|_{\max} \|\hat{W}\|_{\max} \leq \left( \max_i d_i \right)^2 \cdot \|\hat{W}^{t-2}\|_{\max} \|\hat{W}\|_{\max}^2 \leq \cdots \leq \left( \max_i d_i \right)^{t-1} \|\hat{W}\|_{\max}^t = O_P((n \rho_n)^{-1}).$$

Since $c_2 \rho_n \leq p_{ii'} \leq c_3 \rho_n$ for any $1 \leq i, i' \leq n$, we could see

$$\frac{c_2}{c_3 n} \leq w_{ii'}^{(t)} \leq \frac{c_3}{c_2 n}$$

for all $t \geq 1$, which implies $\min_{i,i'} w_{ii'}^{(t)}, \max_{i,i'} w_{ii'}^{(t)} \asymp n^{-1}$.  

C.4. Technical Details in Appendix B.1

C.4.1. Bounding $\|\Delta^{(1)}_1\|_{\max}$ and $\|\Delta^{(1)}_2\|_{\max}$

We could write

$$\|\Delta^{(1)}_1\|_{\max} = \max_{1 \leq i, i' \leq n} \left| \frac{a_{ii'}}{d_{i'} \cdot p_{i'}} (d_{i'} - p_{i'}) \right| \leq \max_{1 \leq i, i' \leq n} 1 \cdot \frac{d_{i'} - p_{i'}}{d_{i'}}$$

by $|a_{ii'}| \leq 1$ and Assumption 3. By Lemma 16, we have

$$\|\Delta^{(1)}_1\|_{\max} \lesssim \frac{1}{n \rho_n} \max_{1 \leq j \leq n} |d_{i'} - p_{i'}| \cdot \max_{1 \leq j \leq n} \frac{1}{d_{i'}} = O_P((n \rho_n)^{-3/2} \sqrt{\log n})$$

53
and,
\[
\|\Delta_2^{(1)}\|_{\text{max}} = \|(A - P)D_P^{-1}\|_{\text{max}} = \max_{1 \leq i, i' \leq n} \left| \frac{a_{ii'} - p_{ii'}}{p_j} \right| \leq \max_{1 \leq j \leq n} \frac{1}{p_j} = O_p\left((n\rho_n)^{-1}\right).
\]

Since \(\rho_n \gtrsim \frac{\log^3 n \log n}{n}\), we have \((n\rho_n)^{-3/2} \sqrt{\log n} \gtrsim (n\rho_n)^{-1}\).

C.4.2. Bounding \(\|\Delta_1^{(2)}\|_{\text{max}}\)

The \(ii'\)th element of \(\Delta_1^{(2,1)}\) is
\[
\sum_{i' = 1}^{n} \left( \frac{a_{ii'}}{d_{i'}} (p_{i} - d_{i'}) - \frac{a_{ii'}}{p_{ii'}} (p_{i} - d_{i'}) \right) \cdot w_{i,i'} = \sum_{i' = 1}^{n} \frac{a_{ii'} (p_{i} - d_{i'})^2}{p_{ii'}^2 d_{i'}} \cdot w_{i,i'},
\]
Take the overall maximums for each item in the sum,
\[
\|\Delta_1^{(2,1)}\|_{\text{max}} \leq \max_{1 \leq i, i' \leq n} \sum_{i' = 1}^{n} \frac{a_{ii'} (p_{i} - d_{i'})^2}{p_{ii'}^2 d_{i'}} \cdot w_{i,i'} \leq \max_i d_i \cdot \left( \max_{1 \leq i \leq n} |p_i - d_i| \right)^2 \cdot \left( \max_{1 \leq i \leq n} \frac{1}{p_i} \right)^2 \left( \max_{1 \leq i \leq n} \frac{1}{d_i} \right) \cdot \left( \max_{i,i'} w_{i,i'} \right)
\]
by Lemma 16. Similar to Eq. (A.4) and Eq. (A.5),
\[
\|\Delta_1^{(2,2)}\|_{\text{max}} = \max_{1 \leq i, i' \leq n} \left( \left| \sum_{i' = 1}^{n} \frac{a_{ii'}}{p_{ii'}} (p_{i} - d_{i'}) w_{i,i'} \right| \right)
\]
\[
\leq \max_i d_i \cdot \left( \max_{1 \leq i \leq n} |p_i - d_i| \right) \cdot \max_{i,i'} w_{i,i'} \cdot \left( \max_i \frac{1}{p_i} \right)^2 = O_p\left(n^{-3/2}(\rho_n)^{-1/2} \sqrt{\log n}\right),
\]
\[
\|\Delta_1^{(2,3)}\|_{\text{max}} = \max_{i,i'} \left| \sum_{i'} \frac{a_{ii'} - p_{ii'}}{p_{ii'}} w_{i,i'} \right| = O_p\left(n^{-3/2}(\rho_n)^{-1/2} \sqrt{\log n}\right).
\]

By Assumption 3, we have \(n^{-3/2}(\rho_n)^{-1/2} \sqrt{\log n} \gtrsim n^{-2}\rho_n^{-1} \log n\). Combine the above results,
\[
\|\Delta_1^{(2)}\|_{\text{max}} \leq \|\Delta_1^{(2,1)}\|_{\text{max}} + \|\Delta_1^{(2,2)}\|_{\text{max}} + \|\Delta_1^{(2,3)}\|_{\text{max}} = O_p\left(n^{-3/2}(\rho_n)^{-1/2} \sqrt{\log n}\right).
\]

C.4.3. Bounding \(\delta_{2,off}^{(2,2)}\) and \(\delta_{2,off}^{(2,3)}\)

(1) **Bound of \(\delta_{2,off}^{(2,2)}\):** By Lemma 16, we have
\[
\delta_{2,off}^{(2,2)} = \max_{i,i'} \left| \sum_{i' = 1}^{n} \left( \frac{p_{ii'}}{p_{i'}} - \frac{a_{ii'}}{d_{i'}} \right) \left( \frac{a_{i'i'}}{p_{i'}} - \frac{a_{i'i'}}{d_{i'}} \right) \right| \leq \max_{i,i'} \left| \sum_{i' = 1}^{n} \left( \frac{p_{ii'} - a_{ii'}}{p_{i'}} \right) a_{i'i'} \right| \times \max_{i'} \left| \frac{1}{p_{i'}} - \frac{1}{d_{i'}} \right|
\]
\[
= O_p\left(n^{-3/2} \rho_n^{-3/2} \sqrt{\log n}\right) \times \max_{i,i'} \left| \sum_{i' = 1}^{n} \left( \frac{p_{ii'} - a_{ii'}}{p_{i'}} \right) a_{i'i'} \right|.
\]
Now we focus on \( \sum_{i^* = 1}^{n} \left( \frac{p_{i^*}}{p_{i^*}} - \frac{a_{i^*}}{d_i} \right) a_{i^* i'} \). Write,

\[
\sum_{i^* = 1}^{n} \left( \frac{p_{i^*}}{p_{i^*}} - \frac{a_{i^*}}{d_i} \right) a_{i^* i'} = \sum_{i^* = 1}^{n} \left( \frac{d_i - p_{i^*}}{p_{i^*}} \right) \left( \frac{p_{i^*} d_i - a_{i^*} p_{i^*}}{p_{i^*} d_i} \right) a_{i^* i'}
\]

\[
= \sum_{i^* = 1}^{n} \frac{d_i}{p_{i^*}} \times \left[ \frac{p_{i^*} \sum_{i^*} a_{i^*} - a_{i^*} \sum_{i^*} p_{i^*}}{p_{i^*} d_i} \right] a_{i^* i'} + \sum_{i^* = 1}^{n} \left( 1 - \frac{d_i}{p_{i^*}} \right) \left( \frac{p_{i^*} d_i - a_{i^*} p_{i^*}}{p_{i^*} d_i} \right) a_{i^* i'}
\]

\[
= \sum_{i^* = 1}^{n} \sum_{i^* = 1}^{n} \left( \frac{a_{i^*} p_{i^*} - p_{i^*} a_{i^*}}{p_{i^*}^2} \right) a_{i^* i'}
\]

For the first item in RHS of Eq. (C.4), we have

\[
\sum_{i^* = 1}^{n} \sum_{i^* = 1}^{n} \left( \frac{a_{i^*} p_{i^*} - p_{i^*} a_{i^*}}{p_{i^*}^2} \right) a_{i^* i'} = \sum_{i^* = 1}^{n} \sum_{i^* = 1}^{n} \left( \frac{a_{i^*} p_{i^*} - p_{i^*} a_{i^*}}{p_{i^*}^2} \right) a_{i^* i'}
\]

\[
= \sum_{i^* = 1}^{n} \sum_{i^* = 1}^{n} \left( \frac{a_{i^*} p_{i^*} - p_{i^*} a_{i^*}}{p_{i^*}^2} \right) a_{i^* i'}
\]

Here \( T(i, i') = \{(i^*, i^*) | i^* < i^*, i^* \notin \{i, i'\}, i^* \notin \{i, i'\}\} \) and \( a_i = (a_1, \ldots, a_n) \).

Conditioning on the event that \( \|D_A\| = \max_i d_i \leq cnp_n \) with some \( c > 0 \) and \( a_i, a_{i'} \), it is easy to check the items in RHS of Eq. (C.5) are independent with 0 conditional mean, and the conditional variance of the overall summation in RHS of Eq. (C.5) has an order of \( 1/n^2 \).

Similar argument to Eq. (B.15)-Eq. (B.17) implies

\[
\max_{i, i'} \left| \sum_{i^* = 1}^{n} \left( \frac{a_{i^*} p_{i^*} - p_{i^*} a_{i^*}}{p_{i^*}^2} \right) a_{i^* i'} \right| = \mathcal{O}(n^{-1}).
\]

With Lemma 16, it is also easy to see

\[
\max_{i, i'} \left| \sum_{i^* = 1}^{n} \left( \frac{p_{i^*} d_i - a_{i^*} p_{i^*}}{p_{i^*} d_i} \right) a_{i^* i'} \right| = \mathcal{O}(nP(n)^{-1/2} \sqrt{\log n}),
\]

by noting that \( \max_{i, i'} |p_{i^*} d_i - a_{i^*} p_{i^*}| \leq \max_{i, i'} p_{i^*} d_i + p_i = \mathcal{O}(n \rho_n) \). With all the results above, we have

\[
\delta^{(2,2)}_{2, \text{off}} = \mathcal{O}((n \rho_n)^{-2} \log n).
\]

(2) **Bound of \( \delta^{(2,3)}_{2, \text{off}} \):** When the element is off-diagonal, we have

\[
\delta^{(2,3)}_{2, \text{off}} = \max_{1 \leq i \neq i' \leq n} \left| \sum_{i^* = 1}^{n} \left( \frac{p_{i^* i^*} - a_{i^* i^*}}{p_{i^*}} \right) \left( \frac{p_{i^* i^*} - a_{i^* i^*}}{p_{i^*}} \right) \right| = \max_{1 \leq i \neq i' \leq n} \left| \sum_{i^* = 1}^{n} \frac{1}{p_{i^*} p_{i^*}} \left( p_{i^* i^*} - a_{i^* i^*} \right) \left( p_{i^* i^*} - a_{i^* i^*} \right) \right|.
\]
Defining each item in Eq. (C.6) by \( \tilde{\zeta}_{ii'}^t \equiv \frac{1}{p_{ii}p_{i'i'}} (p_{ii} - a_{ii})(p_{i'i'} - a_{i'i'}) \), we could write
\[
\sum_{i^* = 1}^{n} \left( \frac{p_{ii^*}}{p_{i^*}} - a_{ii^*} \right) \left( \frac{p_{i'i'}}{p_{i'}} - a_{i'i'} \right) = \left( \sum_{i^* = 1}^{n} \frac{\tilde{\zeta}_{ii^*}^t}{p_{i}} \right) + \frac{\tilde{\zeta}_{i'i'}^t}{p_{i'}}.
\]

When \( i^* \neq i, i' \) and \( i \neq i' \), all \( \tilde{\zeta}_{ii'}^t \) are independent and mean 0, we have
\[
\frac{1}{n - 2} \sum_{i^* = 1}^{n} \text{Var}(\tilde{\zeta}_{ii'}^t) = \frac{1}{n - 2} \sum_{i^* = 1}^{n} \text{Var}\left( \left( \frac{p_{ii^*}}{p_{i^*}} - a_{ii^*} \right) \left( \frac{p_{i'i'}}{p_{i'}} - a_{i'i'} \right) \right) = \frac{1}{(n - 2)} \sum_{i^* = 1}^{n} \frac{1}{p_{i}p_{i'}^2} \text{Var}(p_{ii^*} - a_{ii}) \cdot \text{Var}(p_{i'i'} - a_{i'i'}) \asymp \frac{1}{n^2}.
\]

We also have, when \( i \neq i' \)
\[
|\tilde{\zeta}_{ii'}^t| = \left| \left( \frac{p_{ii^*}}{p_{i^*}} - a_{ii^*} \right) \left( \frac{p_{i'i'}}{p_{i'}} - a_{i'i'} \right) \right| \leq \frac{1}{n^2} \frac{1}{n^2} \approx (n^2)^{-2}
\]

An application of Bernstein inequality implies there exists some \( C' > 0 \) such that
\[
P\left( \max_{i \neq i'} \left| \sum_{i^* = 1}^{n} \tilde{\zeta}_{ii'}^t \right| > C' n^{-3/2} \rho_n^{-1} \sqrt{\log n} \right) \leq n^{-c'}
\]

as \( n \to +\infty \) with some \( c' > 0 \). So we have \( \max_{i \neq i'} \left| \sum_{i^* = 1}^{n} \tilde{\zeta}_{ii'}^t \right| = O(n^{-3/2} \rho_n^{-1} \sqrt{\log n}). \)

By Eq. (C.7), we could also see \( \max_{i,i'} |\tilde{\zeta}_{ii'}^t| = O((n^2)^{-2}) \) and thus
\[
\delta_{2,\text{off}}^{(2,3)} = \max_{i \neq i'} \left| \sum_{i^* = 1}^{n} \left( \frac{p_{ii^*}}{p_{i^*}} - a_{ii^*} \right) \left( \frac{p_{i'i'}}{p_{i'}} - a_{i'i'} \right) \right| \leq \max_{i \neq i'} \left| \left( \sum_{i^* = 1}^{n} \tilde{\zeta}_{ii'}^t \right) \right| + \max_{i,i'} |\tilde{\zeta}_{ii'}^t| + \max_{i,i'} |\tilde{\zeta}_{i'i'}^t| \leq O(n^{-3/2} \rho_n^{-1} \sqrt{\log n}) + O((n^2)^{-2}) + O((n^2)^{-2}) = O(n^{-3/2} \rho_n^{-1} \sqrt{\log n}).
\]

C.5. Proof of Theorem 3
To give a detailed analysis for all components in \( C_{ii'} \), we firstly denote \( C_{ii'}^{(t)} \) as the times that the structure,
\[
\ldots, v_i, \cdots, v_{i'}, \ldots
\]
with a fixed \( t \) appears among all random paths in \( \bigcup_{i=1}^{n} L_i \). Since \( C_{ii'} \) counts all the structures defined in Eq. (2.1), we could see
\[
C_{ii'} = \sum_{t=t_L}^{t_U} C_{ii'}^{(t)} + \sum_{t=t_L}^{t_U} C_{i'i'}^{(t)}.
\]
We then denote $C_{(t)}^{(i)}$ as the times of following structure appears among all random paths in $\bigcup_{i=1}^{n} L_i,$

$$\cdots, v_i, \cdots, v_{i'}, \cdots$$

$k$ nodes $t-1$ nodes

then it could be seen that

$$C_{i'i'}^{(t)} = \sum_{i'=0}^{L-t-1} C_{i'i'}^{(t)}.$$  

Let $\{R_i\}_{i=1}^{+\infty}$ represents a stationary simple random walk on $G$. Since all random paths are stationary and independent simple random walks over $G$, the strong law of large numbers implies

$$C_{i'i'}^{(t)} / r \xrightarrow{a.s.} \mathbb{P}(R_i^{t+1} = v_i) \cdot \mathbb{P}(R_i^{t+1} = v_i | R_i^{t+1} = v_i) = S(v_i) \cdot \mathbb{P}(R_i^{t+1} = v_i | R_i = v_i) = \frac{d_i}{2|A|} \cdot \tilde{w}_{i'i'}^{(t)}.$$ 

as $r \to \infty$. Furthermore we also have

$$C_{i'i'}^{(t)} / r = \sum_{i'=0}^{L-t-1} C_{i'i'}^{(t)} / r \xrightarrow{a.s.} (L-t) \frac{d_i}{2|A|} \tilde{w}_{i'i'}^{(t)},$$

$$C_{i'i'}^{(t)} / r = \sum_{t=t_L}^{t_U} C_{i'i'}^{(t)} / r + \sum_{t=t_L}^{t_U} C_{i'i'}^{(t)} / r \xrightarrow{a.s.} \sum_{t=t_L}^{t_U} (L-t) \cdot \left( \frac{d_i}{2|A|} \tilde{w}_{i'i'}^{(t)} + \frac{d_{i'}}{2|A|} \tilde{w}_{i'i'}^{(t)} \right).$$

Combining the above two convergences, we have

$$\sum_{i=1}^{n} C_{i'i'}^{(t)} / r \xrightarrow{a.s.} \sum_{t=t_L}^{t_U} (L-t) \cdot \left( \frac{1}{2|A|} \sum_{i=1}^{n} d_i \tilde{w}_{i'i'}^{(t)} + \frac{\sum_{i=1}^{n} \tilde{w}_{i'i'}^{(t)}}{2|A|} \right)$$

(C.8)

where we denote $\gamma(L, t_L, t_U) = \frac{(2L-t_L-t_U)(t_U-t_L+1)}{2}$). Note that the first equality in Eq. (C.8) is due to Lemma 14. Similar reasoning yields

$$\sum_{i'=1}^{n} C_{ii'}^{(t)} / r \xrightarrow{a.s.} \gamma(L, t_L, t_U) \cdot \frac{d_i}{|A|} \cdot \sum_{i=1}^{n} \sum_{i'=1}^{n} C_{ii'}^{(t)} / r \xrightarrow{a.s.} 2\gamma(L, t_L, t_U).$$

Now for $(t_L, t_U)$ satisfying Assumption 1, the $ii'$th entry in $\bar{M}(C,k)$ satisfies

$$\log \left( \frac{C_{ii'} \cdot \sum_{i',i''} C_{ii''}}{\sum_i C_{ii'} \cdot \sum_{i'} C_{ii'}} \right) - \log(k) = \log \left( \frac{(C_{ii'} / r) \cdot \left( \sum_{i,i'} C_{ii'} / r \right)}{\sum_i (C_{ii'} / r) \cdot \sum_{i'} (C_{ii'} / r)} \right) - \log(k)$$

\[ \xrightarrow{a.s.} \log \left[ \frac{2|A|}{\gamma(L, t_L, t_U)} \sum_{t=t_L}^{t_U} (L-t) \cdot \left( \frac{\tilde{w}_{ii'}^{(t)}}{d_{i'}} + \frac{\tilde{w}_{ii'}^{(t)}}{d_{i}} \right) \right] - \log(k) \]

(C.9)
where the last equality is because $\tilde{W}^t$ is a transition matrix that satisfies the detailed balance condition. Writing Eq. (C.9) in a matrix form, we obtain

$$\tilde{M}(C, k) \xrightarrow{a.s.} \log \left[ \frac{2|A|}{\gamma(L, t_L, t_U)} \sum_{t \in t_L} (L - t) \cdot (D_A^{-1} \tilde{W}^t) \right] - \log(k) \cdot 1^r.$$  

### C.6. Proof of Theorem 11

For sufficient large $t_U$ such that, w.h.p. $\tilde{M}_0$ is well defined, we use the same notations of $I_A, I_P$ as defined in Eq. (A.19). Then we also have

$$\|M_0 - M_0\|_{\text{max, off}} \leq \max_{i,i'} \left( \frac{1}{\alpha_{ii'}} \right) \cdot \|I_A - I_P\|_{\text{max, off}},$$

$$\|\tilde{M}_0 - M_0\|_{\text{max, diag}} \leq \max_{i,i'} \left( \frac{1}{\alpha_{ii'}} \right) \cdot \|I_A - I_P\|_{\text{max, diag}},$$

where $\alpha_{ii'} \in (\min\{I_A^{ii'}, I_P^{ii'}\}, \max\{I_A^{ii'}, I_P^{ii'}\})$, $I_A^{ii'}, I_P^{ii'}$ are $ii'$th entries of $I_A, I_P$. We slightly generalize the Proof of Theorem 6 to finish this proof.

#### Step 1 (Bound of $\max_{i,i'} \left( \frac{1}{\alpha_{ii'}} \right)$): We also have $\alpha_{ii'}$ is between $I_A^{ii'}$ and $I_P^{ii'}$ and $\max_{i,i'} \frac{1}{\alpha_{ii'}} \leq \max_{i,i'} \left\{ \frac{1}{I_A^{ii'}}, \frac{1}{I_P^{ii'}} \right\}$. By Lemma 16 and Assumption 3, we have

$$\max_{i,i'} \frac{1}{I_P^{ii'}} = \max_{i,i'} \frac{1}{\sum_{t \in t_L} (L - t)} \frac{1}{(\sum_{ii'} p_{ii'}) \cdot \left( \frac{w_i^{(ii')} p_i}{p_i} \right) \leq \max_{i,i'} \frac{1}{(L - t_L) (\sum_{ii'} p_{ii'}) (w_i^{(ii')} p_i)}.$$  

We also have $\frac{1}{I_A^{ii'}} \leq \frac{d_i}{(L-t_L)(\sum_{ii'} a_{ii'})(w_i^{(ii')})}$ and we consider off-diagonal and diagonal cases separately.

**1) When $i \neq i'$**: By Theorem 9 and Lemma 16, we have $\min_{i,i'} w_i^{(ii')}$, $\max_{i,i'} w_i^{(ii')} \asymp n^{-1}$ for fixed $2 \leq t_L$ and

$$\max_{i \neq i'} |w_i^{(ii')} - w_i^{(ii')}| / \min_{i,i'} w_i^{(ii')} = \begin{cases} \mathcal{O}_p \left( n^{-3/2} \rho^{-1} \sqrt{\log(n)/(1/n)} \right) = \mathcal{O}(1) & \text{ when } t_L \geq 2 \text{ and } \beta < \frac{1}{2} \\ \mathcal{O}_p \left( n^{-3/2} \rho^{-1/2} \sqrt{\log(n)/(1/n)} \right) = \mathcal{O}(1) & \text{ when } t_L \geq 4 \text{ and } \beta < \frac{t_L^{-3}}{t_L - 1} \end{cases}$$  

which implies

$$0 \leq \min_{i,i'} w_i^{(ii')} - \max_{i,i'} w_i^{(ii')} \asymp \min_{i,i'} w_i^{(ii')} = \mathcal{O}_p(1/n)$$

w.h.p. as $n \to +\infty$. And we further have

$$\max_{i \neq i'} \frac{1}{w_i^{(ii')} - \max_{i \neq i'} w_i^{(ii')} | w_i^{(ii')} - w_i^{(ii')} |} = \frac{1}{\min_{i \neq i'} (w_i^{(ii')} - \max_{i \neq i'} w_i^{(ii')} - w_i^{(ii')})} = \mathcal{O}(n)$$
and,
\[
\max_{i \neq i'} \frac{1}{I_{ii}^L} \leq \max_{i \neq i'} \frac{d_i}{(L - t_L)(\sum_{ii'} a_{ii'})(\hat{w}_{ii'}^{(t)} - w_{ii'}^{(t_L)}) + w_{ii'}^{(t_L)}} = \max_{i \neq i'} \frac{d_i}{(L - t_L)(\sum_{ii'} a_{ii'})(\hat{w}_{ii'}^{(t)} - w_{ii'}^{(t_L)}) + w_{ii'}^{(t_L)}} = O_P(1).
\]

(2) When \( i = i' \): Similarly, when \( t_L = 2 \), we have \( \max_i \frac{1}{w_{ii'}^{(2)}} = \max_i \sum_{i' = 1}^{d_i} a_{ii'}/d_i' \leq (\max_i d_i)^2 \times \max_i \frac{1}{d_i} = O_P(n \rho_n) \). Thus
\[
\max_i \frac{1}{I_{ii}^L} \leq \max_i \frac{d_i}{n d_i} \times \max_i \frac{1}{\sum_{i' = 1}^{d_i} a_{ii'}/d_i'} = \max_i \frac{1}{d_i} = O_P(1).
\]

When \( t_L \geq 3 \), since \( \|W^T - W^T\|_{\max, \text{diag}} = O_P(n^{-3/2} \rho_n^{-1/2} \sqrt{\log n}) \), a similar argument as the \( i \neq i' \) case shows
\[
\max_i \frac{1}{I_{ii}^L} \leq \max_i \frac{d_i}{(L - t_L)(\sum_{ii'} a_{ii'})(\hat{w}_{ii'}^{(t_L)} - w_{ii'}^{(t_L)}) + w_{ii'}^{(t_L)}} = O_P(1).
\]

So in summary, we have \( \max_{i, i'} \frac{1}{I_{ii'}^L} = O_P(1) \) under the conditions of Theorem 11 and deduce
\[
\max_{i, i'} \frac{1}{\alpha_{ii'}} = O_P(1). \tag{C.11}
\]

**Step 2 (Bound of \( \|I_A - I_P\|_{\max, \text{off}} \) and \( \|I_A - I_P\|_{\max, \text{diag}} \):**

**Case 1 (\( t_L \geq 2 \) and \( \beta < \frac{1}{2} \)):** When \( t \geq 2 \) we have

(a). \( \max_{i, i'} \left| \left( \sum_{ii'} a_{ii'} - \sum_{ii'} p_{ii'} \right) w_{ii'}^{(t)} / d_i \right| \leq n \max_{i, i'} \left| d_i - p_i \right| \times \max_{i, i'} w_{ii'}^{(t)} \times \max_i \frac{1}{d_i} = O_P((n \rho_n)^{-1/2} \sqrt{\log n}), \)

(b). \( \max_{i, i'} \left| \left( \sum_{ii'} p_{ii'} \left( \frac{1}{d_i} - \frac{1}{p_i} \right) w_{ii'}^{(t)} \right) / d_i \right| \leq n \max_{i, i'} \left| p_i - d_i / p_i \right| \times \max_{i, i'} w_{ii'}^{(t)} = O_P((n \rho_n)^{-1/2} \sqrt{\log n}), \)

(c). \( \max_{i, i'} \left| \left( \sum_{ii'} a_{ii'} \right) w_{ii'}^{(t)} / d_i - w_{ii'}^{(t)} \right| \leq n \max_{i, i'} \left| \hat{w}_{ii'}^{(t)} - w_{ii'}^{(t)} \right| \times \max_i \frac{1}{d_i} = O_P((n^{-1/2} \rho_n^{-1} \sqrt{\log n}), \)

(d). \( \max_{i} \left| \left( \sum_{ii'} a_{ii'} \right) w_{ii'}^{(t)} / d_i - w_{ii'}^{(t)} \right| \leq n \max_{i} \left| \hat{w}_{ii'}^{(t)} - w_{ii'}^{(t)} \right| \times \max_i \frac{1}{d_i} = O_P(\rho_n^{-1}), \)

by Lemma 16, Theorem 9 and Eq. (3.7). Thus a similar argument as the proof under dense regime gives
\[
\|I_A - I_P\|_{\max, \text{off}} = O_P(n^{-1/2} \rho_n^{-1} \sqrt{\log n}), \tag{C.12}
\]
\[
\|I_A - I_P\|_{\max, \text{diag}} = O_P(\rho_n^{-1}). \tag{C.13}
\]

59
Case 2 \((t_L \geq 4 \text{ and } \beta < \frac{t_L - 3}{t_L - 1}\) holds): We note that (a) and (b) in Eq. (C.12) do not change for \(t \geq t_L \geq 3\). By Eq. (B.4) we further have

\[
\max_{i,i'} \left| \left( \sum_{i''} a_{ii'} \right) \frac{\hat{w}_{i'i''}(t) - w_{i'i''}(t)}{d_i} \right| = \mathcal{O}_p((n \rho_n)^{-1/2} \sqrt{\log n})
\]

for all \(t \geq t_L \geq 3\), which implies

\[
\|\hat{I}_A - I_P\|_{\text{max, off}}; \|\hat{I}_A - I_P\|_{\text{max, diag}} = \mathcal{O}_p((n \rho_n)^{-1/2} \sqrt{\log n})
\]

in this case.

Step 3 (Bound of \(\|\hat{M}_0 - M_0\|_F\)): With Eq. (10), Eq. (11), Eq. (13) and Eq. (14), we finally conclude

\[
\|\hat{M}_0 - M_0\|_{\text{max, off}} = \begin{cases} \mathcal{O}_p(n^{-1/2} \rho_n^{-1} \sqrt{\log n}) & \text{when } t_L \geq 2 \text{ and } \beta < \frac{1}{7}, \\ \mathcal{O}_p((n \rho_n)^{-1/2} \sqrt{\log n}) & \text{when } t_L \geq 4 \text{ and } \beta < \frac{t_L - 3}{t_L - 1}. \end{cases}
\]

\[
\|\hat{M}_0 - M_0\|_{\text{max, diag}} = \begin{cases} \mathcal{O}_p(\rho_n^{-1}) & \text{when } t_L \geq 2 \text{ and } \beta < \frac{1}{7}, \\ \mathcal{O}_p((n \rho_n)^{-1/2} \sqrt{\log n}) & \text{when } t_L \geq 4 \text{ and } \beta < \frac{t_L - 3}{t_L - 1}. \end{cases}
\]

and so,

\[
\|\hat{M}_0 - M_0\|_F \lesssim \left( n^2 \times \|\hat{M}_0 - M_0\|_{\text{max, off}}^2 + n \times \|\hat{M}_0 - M_0\|_{\text{max, diag}}^2 \right)^{1/2}
\]

\[
= \begin{cases} \mathcal{O}_p(n^{1/2} \rho_n^{-1} \sqrt{\log n}) & \text{when } t_L \geq 2 \text{ and } \beta < \frac{1}{7}, \\ \mathcal{O}_p(n^{1/2} \rho_n^{-1/2} \sqrt{\log n}) & \text{when } t_L \geq 4 \text{ and } \beta < \frac{t_L - 3}{t_L - 1}. \end{cases}
\]

C.7. Proof of Corollary 12

Same as the Proof of Corollary 8, we assume \(n_k = n \pi_k\), for all \(k \in [K]\). From Eq. (2.6), one key observation is that when we write (1) \(B = \rho_n B_0\) and \(B_0\) is a constant matrix; (2) \(n_k = n \pi_k\), for all \(k \in [K]\); the \(M_0\) built on \(B\) or \(B_0\) are exactly same for given \(n\). Thus we still have

\[
M_0 = \begin{pmatrix} \xi_{11} 1_{n \pi_1} 1_{n \pi_1}^T & \cdots & \xi_{1K} 1_{n \pi_1} 1_{n \pi_K}^T \\ \vdots & \ddots & \vdots \\ \xi_{K1} 1_{n \pi_K} 1_{n \pi_1}^T & \cdots & \xi_{KK} 1_{n \pi_K} 1_{n \pi_K}^T \end{pmatrix} = \Theta \cdot (\xi_{ii'})_{K \times K} \cdot \Theta^T
\]

and \(\Xi \equiv (\xi_{ii'})_{K \times K}\) is fixed. Following the same argument as the Proof of Corollary 8, we still have \(\|M_0\|_F \asymp n\). Thus, combine Davis-Kahan Theorem (Davis and Kahan, 1970; Yu et al., 2015) and Theorem 11,

\[
\min_{T \in \mathcal{U}, \Theta} \|\hat{F} \cdot T - U\|_F \lesssim \frac{\|\hat{M}_0 - M_0\|_F}{\|M_0\|_F}
\]

\[
= \begin{cases} \mathcal{O}_p(n^{-1/2} \rho_n^{-1} \sqrt{\log n}) & \text{when } t_L \geq 2 \text{ and } \beta < \frac{1}{7}, \\ \mathcal{O}_p(n^{-1/2} \rho_n^{-1/2} \sqrt{\log n}) & \text{when } t_L \geq 4 \text{ and } \beta < \frac{t_L - 3}{t_L - 1}. \end{cases}
\]

as desired.
References

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

Béla Bollobás. *Random graphs*. Cambridge university press, 2001.

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

Fenxiao Chen, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo. Graph representation learning: a survey. *APSIPA Transactions on Signal and Information Processing*, 9:e15, 2020.

Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Hang Chu, Daiqing Li, David Acuna, Amlan Kar, Maria Shugrina, Xinkai Wei, Ming-Yu Liu, Antonio Torralba, and Sanja Fidler. Neural turtle graphics for modeling city road layouts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4522–4530, 2019.

Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):833–852, 2018.

C. Davis and W. Kahan. The rotation of eigenvectors by a pertubation. III. *Siam Journal on Numerical Analysis*, 7:1–46, 1970.

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

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

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017a.

William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 40:52–74, 2017b.

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

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.
Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

J. Lei. A goodness-of-fit test for stochastic block models. *Annals of Statistics*, 44:401–424, 2016.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

Linyuan Lu and Xing Peng. Spectra of edge-independent random graphs. *The Electronic Journal of Combinatorics*, 20(4):P27, 2013.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint at [http://arxiv.org/abs/1301.3781](http://arxiv.org/abs/1301.3781), 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

Enrico Palumbo, Giuseppe Rizzo, Raphaël Troncy, Elena Baralis, Michele Osella, and Enrico Ferro. Knowledge graph embeddings with node2vec for item recommendation. In *European Semantic Web Conference*, pages 117–120. Springer, 2018.

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

Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 459–467, 2018.

Jiezhong Qiu, Chi Wang, Ben Liao, Richard Peng, and Jie Tang. Concentration bounds for co-occurrence matrices of markov chains. arXiv preprint at [http://arxiv.org/abs/2008.02464](http://arxiv.org/abs/2008.02464), 2020.

Sandra L Robinson and Rebecca J Bennett. A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of management journal*, 38(2):555–572, 1995.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

P. Rubin-Delanchy, J. Cape, M. Tang, and C. E. Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph. arXiv preprint at [http://arxiv.org/abs/1709.05506](http://arxiv.org/abs/1709.05506), 2017.
P. Sarkar and P. J. Bickel. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Association, Series B*, 78:253–273, 2016.

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

Minh Tang and Carey E Priebe. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415, 2018.

Athanasios Theodoridis, Stijn Van Dongen, Anton J Enright, and Tom C Freeman. Network visualization and analysis of gene expression data using biolayout express 3d. *Nature protocols*, 4(10):1535, 2009.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *AAAI*, volume 17, pages 203–209, 2017.

Jieping Ye, Ravi Janardan, and Qi Li. Two-dimensional linear discriminant analysis. In *Advances in neural information processing systems*, pages 1569–1576, 2005.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102:315–323, 2015.

Xiaolin Zhang, Lei Chen, Zi-Han Guo, and Haiyan Liang. Identification of human membrane protein types by incorporating network embedding methods. *IEEE Access*, 7:140794–140805, 2019a.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9, 2019b.

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

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1234–1241, 2020.