Structure of Nonlinear Node Embeddings in Stochastic Block Models

Christopher Harker
University of Utah

Aditya Bhaskara
University of Utah

Abstract

Nonlinear node embedding techniques such as DeepWalk and Node2Vec are used extensively in practice to uncover structure in graphs. Despite theoretical guarantees in special regimes (such as the case of high embedding dimension), the structure of the optimal low dimensional embeddings has not been formally understood even for graphs obtained from simple generative models. We consider the stochastic block model and show that under appropriate separation conditions, the optimal embeddings can be analytically characterized. Akin to known results on eigenvector based (spectral) embeddings, we prove theoretically that solution vectors are well-clustered, up to a sublinear error.

1 INTRODUCTION

Graphs are a classic abstraction for studying relationships between objects. Graph data sets consisting of billions of nodes have become commonplace, requiring efficient and scalable algorithms to analyze them. An increasingly popular class of algorithms is node embedding algorithms, which aim to obtain vector representations of nodes that capture node similarities and help understand the structural properties of a graph. These vector representations can be used as inputs to various machine learning tasks, such as community detection (Wang et al., 2017; Ng et al., 2001; Von Luxburg, 2007; Rohe et al., 2011; Belkin & Niyogi, 2001), node classification (Hamilton et al., 2017; Perozzi et al., 2014; Grover & Leskovec, 2016; Tang et al., 2015) and link prediction (Grover & Leskovec, 2016; Tang et al., 2015; Backstrom & Leskovec, 2011). Embeddings of nodes into geometric spaces is also a classic topic in theoretical computer science, which has led to the development of algorithms for clustering and partitioning of graphs (Linial et al., 1994; Spielman & Teng, 2007; Arora et al., 2009).

A classic technique for node embeddings is the so-called spectral embedding, which uses the top eigenvectors of the Laplacian matrix to find node embeddings. Starting with the work of Hall (1970); Alon (1986); Sinclair & Jerrum (1989); Spielman & Teng (2007); McSherry (2001) on the algorithmic side and works such as Belkin & Niyogi (2001); Ng et al. (2001); Rohe et al. (2011); Von Luxburg (2007), properties of spectral embeddings have been studied extensively. More recently, non-linear embedding methods (Grover & Leskovec, 2016; Perozzi et al., 2014; Tang et al., 2015) have been shown to improve upon spectral embeddings along multiple axes: they provide better performance on partitioning tasks, they can better encode structural properties of the graphs, and they can be computed/updated *locally* when the graph is modified. These methods rely on random walks on the graph. Roughly speaking, they are based on the skip-gram model (Mikolov et al., 2013a,b) developed originally in natural language processing. Random walks are viewed as the analog of n-grams in sentences, enabling the application of word embedding methods to networks.

Given their experimental advantages, an important theoretical direction has been to understand the power of these non-linear embedding methods. For the embeddings produced by DeepWalk (Perozzi et al., 2014) and node2vec (Grover & Leskovec, 2016), existing theoretical results establish a connection between the algorithms and matrix factorization: Levy and Goldberg (Levy & Goldberg, 2014) show that when the embedding dimension is at least as large as the number of nodes in the graph, then skip-gram with negative sampling (SGNS) is implicitly performing matrix factorization of a shifted point-wise mutual information matrix. In a similar vein, Qiu et al. (2018) show that LINE, DeepWalk, and node2vec can also be viewed as matrix factorization problems when the embedding dimension is large. Chanpuriya & Musco (2020) later simplified the expression derived by Qiu et al. (2018). However, the characterization obtained holds only when the embedding dimension is large, which is not usually the case in practice (Perozzi et al., 2014).

However, even for very simple graph classes, it is not

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

known if low dimensional embeddings obtained by algorithms like DeepWalk and node2vec *provably* exhibit good structural properties. In this work, we consider graphs generated by stochastic block models (SBMs). SBMs (see Section 2 for a more detailed introduction) are a classic way to model well clustered graphs arising in different applications. They have also acted as a testbed for reasoning about embedding algorithms. For example, many theoretical works on the efficacy of spectral embeddings focused on graphs produced by block models (see McSherry (2001); Abbe (2017) and references therein). In this paper, we ask:

> Can we prove theoretical guarantees on the embeddings obtained by DeepWalk and node2vec on graphs generated by an SBM?

One key challenge turns out to be the non-linearity of the objective used by the embedding algorithms. Similar difficulties are known to arise in proving theoretical properties of methods like t-SNE, which was only done recently in the works of Linderman & Steinerberger (2019); Arora et al. (2018). One tractable case, however, is the one where the embedding dimension is large ($> n$, the number of nodes in the graph). In this case, as shown by Levy & Goldberg (2014), the optimal solution has a closed form. The recent works of Zhang & Tang (2021); Barot et al. (2021) analyzed low rank properties of this closed form solution on graphs generated by SBM. But note that this is not the same as analyzing the embedding algorithms directly (because the closed form only holds for large dimension).

Our goal in this paper is to establish structural properties of the optimal solutions of the objectives used in algorithms like DeepWalk. We study both the classic formulation, as well as the formulation with negative sampling (SGNS; see Section 2 for details). This is also in contrast with prior theoretical works that focus on negative sampling.

Results. Our main contributions are as follows:

1. We first consider the case where the co-occurrence matrix used to obtain the embeddings (see Section 2) has a block structure. Here we show that, as expected, optimal solutions to the DeepWalk and SGNS objectives also have a block structure. [Theorems 3 and 4.]

2. Next we show (Theorem 5) that when the co-occurrence matrix is obtained using random walks on a graph drawn from an SBM, the optimum values of the DeepWalk and SGNS objectives are approximately equal to the objectives computed on an appropriate block matrix.

3. As our main result, we show (Theorem 8) that for a symmetric SBM with two components, the optimal solution vectors (i.e., embeddings) are well-clustered, and further have the structure that vertices in the two components approximately map to antipodal vectors.

4. We finally consider the case of block matrices with more than two blocks (where a simple antipodal structure is not optimal), and experimentally characterize the structure of the optimal solution as the sizes of the components vary. We also observe that for SGNS, the "number" of negative samples significantly affects the embeddings obtained, which suggests that a careful tuning is important in practice.

We will introduce the necessary background on SBMs and the DeepWalk algorithm (along with SGNS) in Section 2. We then consider the simplest case of block structured co-occurrence matrices in Section 3. Then in Section 4 to connect the optimal solution value in an SBM to that of a block matrix using appropriate concentration inequalities. This is then used in Section 5 to show our main result, which argues that for a two-block SBM, the optimal embedding vectors are necessarily well-clustered.

2 BACKGROUND AND PROBLEM SETUP

Common Notation. For a matrix M we use the notation $|M_{\cdot j}| = \sum_{i\in[n]} M_{ij}$, $|M_{i\cdot}| = \sum_{j\in[n]} M_{ji}$, and $|M| = \sum_{i,j\in[n]} M_{ij}$. We denote by $\|M\|_1$ its L_1 norm after flattening it along the columns (and similarly $\|M\|_\infty$). For two matrices M and X, we define $M \cdot X = \sum_{i,j} M_{ij}X_{ij}$.

For an graph $G(V, E)$ with n nodes and $|E|$ edges, the matrix A denotes the adjacency matrix of the graph G. The degree of node i is $d_i = \sum_{j=1}^{n} A_{ij}$ and the matrix D is a diagonal matrix where $D_{ii} = d_i$. We let $W = D^{-1}A$ denote the transition matrix of the graph.

2.1 Stochastic Block Model

The stochastic block model (SBM) (Holland et al., 1983) is a popular generative framework used in the theoretical analysis of community recovery algorithms. SBMs have been widely studied (see, e.g., Abbe (2017) and references therein), and may be viewed as a natural generalization of Erdős-Renyi random graphs.

A K block stochastic block model (SBM) generates a random graph $G = (V, E)$ by first assigning each node in the graph to one of K blocks. These blocks are also referred to as categories or communities. We refer to V_k as the set of vertices that belong to community k and define a community membership matrix $Z \in \{0, 1\}^{n\times K}$ as $Z_{ik} = \mathbb{1}\{i \in V_k\}$, whose entries are 1 if node $i \in V_k$ and 0 otherwise. Let $B \in [0, 1]^{K\times K}$ be a symmetric matrix of probabilities of full rank whose entries B_{ij} denote the probability of a node in cluster i and a node in cluster j being connected. We define the matrix P as $P := ZBZ^\top$, which is simply a block matrix of probability values defined by the partitioning Z.

The edges of the graph $G(V, E)$ are then generated as $A_{ij} \sim \text{Bern}(P_{ij})$, $A_{ij} = A_{ji}$ for all $i < j$. We also assume that graphs do not have self loops, i.e., $A_{ii} = 0$ for all i. In a *symmetric* SBM, we further assume that each cluster is of equal size, i.e., $|V_k| = \frac{n}{K}$. We also assume that the entries of B are p on the diagonal and q off-diagonal. Thus P is a block matrix with K equal-sized blocks. Finally, we assume that the number of communities K is fixed and that the number of nodes n tends to infinity.

2.2 DeepWalk

In this section we describe the Perozzi's DeepWalk graph embedding algorithm (Perozzi et al., 2014). Let $G = (V, E)$ be an undirected, connected graph with n nodes. The DeepWalk algorithm consists of two key steps. The first step generates r random walks of length L on the graph and uses these walks to create an $n \times n$ co-occurrence matrix C whose entries count the number of times that two nodes appear a certain distance to each other in these random walks. The second step uses the co-occurrence matrix to compute the node representations by solving an optimization problem.

Generating Random Walks. The algorithm first generates r random walks of length L and slides a window of size T over each random walk. Let $w^{(m)}$ denote the m^{th} random walk represented as a sequence of L nodes. Let $w_k^{(m)}$ denote the k^{th} step of the m^{th} random walk. To be consistent with prior works, we assume that the starting node of each walk is sampled from a stationary distribution on G

$$\Pr\left[w_1^{(m)} = i\right] = \frac{d_i}{2|E|}$$

Using the sample random walks, the algorithm creates the co-occurrence matrix C whose entries C_{ij} contains the number of times that a node j appears in a size T window around a node i. Formally,

$$C_{ij} = \sum_{t=1}^{T}\sum_{m=1}^{r}\sum_{k=1}^{L-t}\mathbb{1}\{w_k^m = i, w_{k+t}^m = j\} + \sum_{t=1}^{T}\sum_{m=1}^{r}\sum_{k=1}^{L-t}\mathbb{1}\{w_k^m = j, w_{k+t}^m = i\}$$

Many of the prior theoretical works involve studying limiting cases of this co-occurrence matrix as the length of the walk L or the number of walks r goes to ∞ (Zhang & Tang, 2021; Barot et al., 2021; Qiu et al., 2018), or as the window size T goes to ∞ (Chanpuriya & Musco, 2020). A few works focus solely on obtaining concentration bounds on this matrix (Qiu et al., 2020; Kloepfer et al., 2021).

Borrowing from prior works, we will make use of the following lemma in our analysis (Zhang & Tang, 2021; Barot et al., 2021). The proof can be found in Appendix A.1

Lemma 1. *Let A be an adjacency matrix of a fixed graph G and let $w_k^{(m)}$ denote the k^{th} step of the m^{th} random walk generated by the DeepWalk algorithm. Let $\pi_i = \frac{d_i}{2|E|}$ and let $(W^t)_{ij} = \Pr[w_{t+1} = j | w_1 = i]$. Then as $r \to \infty$,*

$$\frac{C_{ij}}{r} \xrightarrow{a.s} 2\sum_{t=1}^{T}(L-t)\cdot\pi_i(W^t)_{ij} \tag{1}$$

Note that this limiting matrix can be constructed explicitly from the adjacency matrix A without needing to sample with random walks.

Computing Node Representations. Given a co-occurrence matrix C, DeepWalk uses the skip-gram model to learn a matrix of node embeddings $X \in \mathbb{R}^{d\times n}$ and a matrix of context embeddings $Y \in \mathbb{R}^{d\times n}$. The d-dimensional node embedding $\boldsymbol{x}_i \in \mathbb{R}^d$ for node i is the i^{th} column of X, while the d-dimensional context embedding $\boldsymbol{y}_i \in \mathbb{R}^d$ for node i is the i^{th} column of Y. For a co-occurrence matrix C, the node and context embeddings (X, Y) are computed by optimizing the following objective function:

$$\mathcal{L}_{\text{dw}}(C; X, Y) = \sum_{i,j=1}^{n} C_{ij}\log\left(\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{y}_j}}{\sum_{k=1}^{n} e^{\boldsymbol{x}_i^\top \boldsymbol{y}_k}}\right) \tag{2}$$

Skip-gram with Negative Sampling Most works analyze the more computationally efficient skip-gram with negative sampling objective function with some simplifications (Zhang & Tang, 2021; Barot et al., 2021; Levy & Goldberg, 2014; Qiu et al., 2018). In this case, the objective function is

$$\mathcal{L}_{\text{ns}}(C; X, Y) = \sum_{i,j=1}^{n} C_{ij}\log\sigma\left(\boldsymbol{x}_i^T\boldsymbol{y}_j\right) + \sum_{i,j=1}^{n} C_{ij}s_n\mathbb{E}_{l\sim P_C}[\log\sigma\left(-\boldsymbol{x}_i^T\boldsymbol{y}_l\right)]$$

where s_n is the "number" of negative samples (a coefficient that will affect the embeddings, as we will see), $\sigma(\cdot)$ is the sigmoid function, and P_C is the probability distribution on $[n]$ that samples j with probability $|C_{\cdot j}|/|C|$.

Note that the loss can be simplified as

$$\mathcal{L}_{\text{ns}}(C; X, Y) = \sum_{i,j=1}^{n} C_{ij}\log\sigma(\boldsymbol{x}_i^T\boldsymbol{y}) + \sum_{ij=1}^{n} s_n\frac{|C_{i\cdot}||C_{\cdot j}|}{|C|}\log\sigma(-\boldsymbol{x}_i^T\boldsymbol{y}_j) \tag{3}$$

3 BLOCK STRUCTURED CO-OCCURRENCES

To motivate community recovery using DeepWalk embeddings, we first explore a special case where the co-occurrence matrix is a block-diagonal matrix, whose on- and off-diagonal matrices are multiples of the all-ones matrix. This is reminiscent of work on spectral clustering and non-negative matrix factorization (Ng et al., 2001; Paul & Chen, 2016; Von Luxburg, 2007). Indeed, the classical analysis of spectral clustering on SBMs starts by showing that the adjacency matrix is spectrally similar to the block-diagonal matrix of edge probabilities. Similarly in our setting, the block structure allows us to understand the geometric properties of optimal DeepWalk and SGNS embeddings.

When the co-occurrence matrix C has a block structure, we show that there exists an optimum solution to (2) in which the context vectors have a block structure defined as follows:

Definition 2 (Block Structure)**.** *Let C be an $n \times n$ matrix. We say that C has a block structure with blocks $V_1, V_2, \ldots, V_K$ if $\{V_i\}$ forms a partition of $[n]$, and further, for all $i, j \in [K]$, the submatrix of C induced by the rows V_i and columns V_j has constant entries. (The constant can depend on i and j).*

We say that a solution X, Y *(each in $\mathbb{R}^{d\times n}$) has a block structure with blocks $V_1, \ldots, V_K$ if $\{V_i\}$ are disjoint, and moreover, for every k, we have $\boldsymbol{x}_i = \boldsymbol{x}_j$ for all $i, j \in V_k$, and the same holds for the columns of Y. (Note that the $\boldsymbol{x}_k$ need not be related to $\boldsymbol{y}_k$ in any way.)*

In the following theorems, we use the term *symmetric* block structure for the matrix C. This simply means that the entries in the off-diagonal blocks are all equal. In other words, if $i \sim j$ denotes whether i and j are in the same community or not, then we say that C has symmetric block structure if there exist scalars a, b such that for all $i, j \in [n]$, we have $C_{ij} = a$ if $i \sim j$ and b otherwise.

We note that our proofs can directly be adapted to deal with general block structure, but we prove them for the symmetric case due to simplicity, and because it is the setting that we will use later.

Theorem 3. *Let C be a matrix with a symmetric block structure with blocks $V_1, V_2, \ldots, V_K$. Then there exists an optimal solution (X, Y) maximizing (2) whose node and context embeddings X, Y also have a block structure defined by $\{V_i\}$. Moreover, this also holds for constrained maximization over any domain $\mathcal{D}$, where the columns of X, Y are required to belong to $\mathcal{D}$.*

The proof uses a natural convexity argument is deferred to Appendix A.2. A similar result can be shown for the negative sampling objective.

Theorem 4. *Let C be a matrix with a symmetric block structure defined by $V_1, V_2, \ldots, V_K$. Then there exists an optimal solution (X, Y) maximizing (3) whose node and context embeddings X, Y have block structures defined by the same $\{V_i\}$. This also holds for constrained optimization of equation 3 over any domain $\mathcal{D}$.*

The proof for Theorem 4 is in Appendix A.3.

4 FROM BLOCK MATRICES TO SYMMETRIC SBM

We now consider graphs sampled from the *symmetric* stochastic block model that is defined by three parameters (K, p, q). The vertex set V is divided into K clusters $V_1, V_2, \ldots, V_K$, each of size n/K, and the edge probabilities are p and q, between vertices within and across clusters, respectively.

4.1 DeepWalk Solution Value for Symmetric SBM

Let G be a graph drawn from the symmetric SBM with parameters (K, p, q), and let M be the co-occurrence matrix obtained from G by performing random walks with parameters $L \geq 3$, window length $T = 2$ and $r \to \infty$, as described in Section 2. Using Lemma 1 the co-occurrence matrix M (in the limit) satisfies

$$M = \frac{(L-1)}{|E|}A + \frac{(L-2)}{|E|}AD^{-1}A, \tag{4}$$

where D is the diagonal degree matrix and the length of the random walk L is treated as a constant. We will argue that the solutions obtained by solving the DeepWalk and SGNS optimization problems using M are close to the solutions obtained using the "expected" matrix $\overline{M}$. The matrix $\overline{M}$ is defined as:

$$\overline{M} = \frac{2(L-1)}{n\Delta}\mathbb{E}[A] + \frac{2(L-2)}{n\Delta^2}\mathbb{E}[A^2] \tag{5}$$

where $\Delta = \frac{np}{K} + (K-1)\frac{nq}{K}$ and $\mathbb{E}[A]$, $\mathbb{E}[A^2]$ are the expected values of the adjacency matrix and its square (respectively) when G is drawn from the symmetric SBM. Note that $\overline{M}$ is not, strictly speaking, the expected value of M because the denominator terms of M are also dependent on the graph.

Theorem 5. *Suppose the matrix M is obtained by performing random walks on a graph drawn from the symmetric SBM with parameters K, p, q as described above. Let $\mathcal{L}_{dw}(M; X, Y)$ and $\mathcal{L}_{ns}(M; X, Y)$ be the DeepWalk and SGNS objectives respectively. Suppose also that p, q are such that $\Delta > (20 \log n)\sqrt{\frac{nd}{\epsilon}}$, for some desired error parameter $\epsilon > 0$. Then with probability at least $1 - n^{-4}$, we*

have

$$|\mathcal{L}_{dw}(M;X,Y) - \mathcal{L}_{dw}(\overline{M};X,Y)| \leq O\left(L\epsilon + L\sqrt{\frac{\log^3 n}{\Delta}}\right), \quad (6)$$

$$|\mathcal{L}_{ns}(M;X,Y) - \mathcal{L}_{ns}(\overline{M};X,Y)| \leq O\left(L\epsilon + L\sqrt{\frac{\log n}{\Delta}}\right).$$

for all solutions X, Y whose columns have length ≤ 1.

Remark. While the theorem treats ϵ as a free parameter, it is natural to set it so as to make the two terms on the RHS equal. (In a sense, this is the "least value" of ϵ.) For DeepWalk, this corresponds to $\epsilon = \Delta^{-1/2}\log^{3/2} n$, in which case the lower bound on Δ becomes $c(nd)^{2/3}\log^{1/3} n$, for some (large enough) constant c. Likewise, in the SGNS case, the condition becomes $\Delta > c(nd)^{2/3}\log n$. Requiring such a degree lower bound is a limitation of our analysis and is discussed further in the following section.

4.2 Proof of Theorem 5

The proof proceeds along similar lines for both the loss functions. We define an intermediate matrix M' as follows

$$M' = \frac{2(L-1)}{n\Delta}A + \frac{2(L-2)}{n\Delta^2}A^2,$$

where $\Delta = \frac{np}{K} + (K-1)\frac{nq}{K}$ as defined earlier. M' turns out to be more amenable than M, since it does not have denominator terms that vary with G (i.e., terms such as $|E|$ and D^{-1}). The following lemma relates M with M'.

Lemma 6. *For M, M' defined as above, we have that with probability $\geq 1 - n^{-5}$,*

$$\|M - M'\|_1 \leq O\left(L\sqrt{\frac{\log n}{\Delta}}\right).$$

The proof is deferred to Appendix B.2. The main technical lemma is then the following:

Lemma 7. *Let $\epsilon, \beta > 0$ be parameters, and let Λ be any fixed $n \times n$ matrix satisfying $\|\Lambda\|_\infty \leq \beta$. Then for $n > 8\beta/\epsilon$, we have*

$$\Pr[|(M' - \overline{M})\cdot\Lambda| \geq 2\epsilon L] \leq 4\exp\left(-\min\left\{\frac{\epsilon^2\Delta^4}{8n\beta^2}, \frac{\epsilon\Delta^2}{4\beta}\right\}\right).$$

The lemma is proved using concentration inequalities for quadratic forms of random variables, and is deferred to Appendix B.3. We now sketch the analysis for the case of the DeepWalk loss function. The details, as well as the SGNS case, are deferred to Appendix B.

We start by observing that for any candidate solution X, Y,

$$\mathcal{L}_{\text{dw}}(M;X,Y) = \sum_{i,j} M_{ij}\alpha_{ij}$$

where

$$\alpha_{ij} = \log\frac{e^{\boldsymbol{x}_i^T\boldsymbol{y}_j}}{\sum_k e^{\boldsymbol{x}_i^T\boldsymbol{y}_k}}.$$

Define Λ (which is dependent on X, Y) to be the matrix whose (i,j)th entry is α_{ij}, so we have $\mathcal{L}_{\text{dw}}(M;X,Y) = M\cdot\Lambda$. Then the goal of the theorem is to prove that w.h.p., $|(M - \overline{M})\cdot\Lambda|$ is small for all X, Y.

Proof outline. The proof starts by using Lemma 6 to show that $(M - M')\cdot\Lambda$ is small for all X, Y. Next, we argue using Lemma 7 that for any *fixed* solution X, Y, equation 6 holds with an exponentially small failure probability. Finally, we use an epsilon-net argument (a standard technique in random matrix theory, e.g., Vershynin (2018) to take a union bound over a fine enough "net" over the possible X, Y. This lets us obtain the conclusion of the theorem for all X, Y.

4.3 Discussion

Spectral analyses of the SBM usually require the expected degree to be poly-logarithmic (McSherry, 2001; Rohe et al., 2011; Abbe, 2017; Vu, 2014). Theorem 5, however, assumes that the graph be much denser, requiring that the degree be greater than $n^{\frac{1}{2}}\log n$. This limitation arises in the proof of Lemma 7, which relies on the use of the Hanson-Wright inequality where the degree must be large enough in order to achieve such a strong bound. We conjecture, however, that the overall results hold even for sparse graphs and an interesting open problem is to see whether this large degree requirement can be avoided, potentially by arguing that a union bound over a smaller net would be sufficient.

The ϵ-net argument also requires the degree to increase depending on d. While this term is insignificant for us (as we focus on the low dimensional case where $d = O(1)$), an interesting open question is to obtain an analysis that does not have this requirement.

We also note that M can be constructed explicitly from the adjacency matrix A without needing to sample with random walks. In practice, however, the co-occurrence matrix is constructed by conducting walks. Assuming the random walks are independent, the convergence rate can be obtained as in Kloepfer et al. (2021). The error due to constructing this "empirical" co-occurrence matrix can be incorporated into the results just as we handle the terms $(M - M')$ in the proof of Theorem 5.

Finally, we believe that the arguments in this section can be generalized to other generative graph models, such as the degree-corrected SMB or the random dot product graph,

as long as the degree remains large enough. However, the analysis will need to be more involved as Δ has to be replaced by an appropriate diagonal matrix. This is an interesting avenue for future work.

5 NEAR OPTIMAL SOLUTIONS TO TWO BLOCK SBM

We now show that for the case $K = 2$, it is possible to analytically characterize an (approximate) maximizer of the DeepWalk or SGNS objectives. Before defining our main result, we introduce some parameters. As before, the walk parameters will be $L \geq 3, T = 2, r \to \infty$, and the graph is drawn from a symmetric SBM with $K = 2$ blocks, and edge probabilities p, q.

$$a = \frac{2(L-1)}{n\Delta}p + \frac{(L-2)}{n\Delta^2}(p^2+q^2)n$$
$$b = \frac{2(L-1)}{n\Delta}q + \frac{(L-2)}{n\Delta^2}(2pq)n. \tag{7}$$

As a rough intuition, we can think of a and b as being proportional to p and q respectively. This is because we can think of $\Delta \approx np/2$ (when $q < p/4$, say), and thus the second terms above can be simplified to look similar to the corresponding first terms. Our main result is then the following.

Theorem 8. *Let M be a matrix obtained from random walks on a graph drawn from the symmetric SBM as above, and suppose $\Delta > (nd)^{2/3}\log n$ is chosen so that Theorem 5 applies.*[1] *Let $\textsc{Opt}_{dw}(M)$ and $\textsc{Opt}_{ns}(M)$ denote the optimum objective value of $\mathcal{L}_{dw}(M; X, Y)$ and $\mathcal{L}_{ns}(M; X, Y)$, and let $\delta > 0$ be a parameter. Then with probability at least $1 - n^{-4}$, we have*

1. *Suppose $a/b > e^2$, and let X, Y be a feasible solution (vectors of length ≤ 1) such that $\mathcal{L}_{dw}(M; X, Y) \geq \textsc{Opt}_{dw}(M) - \delta$. Then there exist unit vectors $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ and a constant $\gamma > 0$ such that $\boldsymbol{x}_1^* = -\boldsymbol{x}_2^*$, and for $k \in \{1, 2\}$,*

$$\frac{1}{|V_k|}\sum_{i\in V_k}\|\boldsymbol{x}_i - \boldsymbol{x}_k^*\|^2 \leq \frac{1}{an^2\gamma}\cdot O\left(\delta + \sqrt{\frac{\log^3 n}{\Delta}}\right).$$

2. *Suppose the negative sampling parameter s_n satisfies $\frac{eb}{a+b} < s_n < \frac{a}{(a+b)e}$ and let X, Y be a feasible solution such that $\mathcal{L}_{ns}(M; X, Y) \geq \textsc{Opt}_{ns}(M) - \delta$. Then there exist antipodal unit vectors $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*$ as before, and a parameter $\gamma > 0$ such that for $k \in \{1, 2\}$,*

$$\frac{1}{|V_k|}\sum_{i\in V_k}\|\boldsymbol{x}_i - \boldsymbol{x}_k^*\|^2 \leq \frac{1}{n\gamma}\cdot O\left(\delta + \sqrt{\frac{\log n}{\Delta}}\right).$$

[1] We can also introduce an additional parameter ϵ' and use Theorem 5 to obtain weaker results as long as $\Delta > \sqrt{nd}$, but we skip this for the sake of simplicity of the presentation.

We note that the parameters γ, a, b will ensure that the terms $1/(an^2\gamma)$ and $1/(n\gamma)$ are constants, so that the average *clustering error* is only $O(\delta)$ plus an additional sublinear term. Also, there are two ways of interpreting the condition $\frac{eb}{a+b} < s_n < \frac{a}{(a+b)e}$. The first is that as long as $e^2 b < a$, there exists a choice of s_n that yields a solution with a good cluster structure. Alternatively, for a fixed s_n small enough (say $1/2e$), the solution satisfies a structured property as long as b/a is small enough (around s_n/e). We empirically demonstrate the sensitivity of the solution to s_n in our experiments.

We make two further remarks before moving to the proof. The first is on the robustness. Theorem 8 says that even if an algorithm finds some "near optimal" solution (within δ in objective value), we will have an approximate cluster structure if δ is small enough. This robustness is important because in continuous, non-convex optimization, we usually never find an exact optimum.

Secondly, the Theorem can be used to bound the number of misclassified vertices. Assuming that *k*-means that is solved (nearly) exactly on the resulting embeddings, we can first argue that the *k*-means clusters are very close to each x_i^*. Since the x_i^* are separated, Theorem 8 then implies that the number of misclassifications is small.

We now sketch the proof of the theorem, assuming lemmas that will be shown later. Let $\overline{M}$ be a matrix whose rows and columns are divided into two blocks V_1, V_2. Suppose that a, b denote the entries in the diagonal and off-diagonal blocks respectively.

Proof sketch. Let us focus on the DeepWalk loss function (SGNS will be similar). We first apply Theorem 5 to conclude that a solution (X, Y) that approximately optimizes $\mathcal{L}_{\text{dw}}(M; X, Y)$ also approximately optimizes $\mathcal{L}_{\text{dw}}(\overline{M}; X, Y)$. Then, we observe that $\overline{M}$ (up to an $O(1/n)$ error resulting from the diagonal terms A_{ii} being zero) is a block matrix with entries a, b as defined in equation 7. Then we apply Lemma 9 (the key lemma of this section) with ϵ in the lemma statement set to $O(\delta + \sqrt{\frac{\log^3 n}{\Delta}})$ to show that X is well clustered, and this completes the proof of the theorem. □

We now outline the main technical lemmas described above. For DeepWalk, we have:

Lemma 9. *Suppose the parameters of the matrix $\overline{M}$ (defined using V_1, V_2, a, b, as above) satisfy $\frac{a}{b} > e^2$. Let (X, Y) be any feasible solution that satisfies*

$$\mathcal{L}_{dw}(\overline{M}; X, Y) \geq \textsc{Opt} - \epsilon,$$

where Opt *is the optimum value of the objective, and $\epsilon \geq 0$ is some parameter. Define $\gamma = \sigma(2)\left(e^{-2} - b/a\right)$. Then there exist unit vectors $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ such that $\boldsymbol{x}_2^* = -\boldsymbol{x}_1^*$ and*

for $k \in \{1, 2\}$,

$$\frac{1}{|V_k|} \sum_{i \in V_k} \|\boldsymbol{x}_i - \boldsymbol{x}_k^*\|^2 \leq \frac{8\epsilon}{an^2\gamma}. \tag{8}$$

Note that we showed in Theorem 3 that the optimum solution has a block structure, but the lemma above is stronger in two ways: (a) it argues that the vectors corresponding to the blocks are actually antipodal, and (b) it shows that an approximate block structure holds when we consider an approximately optimal solution.

Proof. Consider any near optimal solution (X, Y). We begin by appealing to the proof of Theorem 3. There, we showed that given any X, there exists a block structured Y^* such that

$$\mathcal{L}_{\text{dw}}(\overline{M}; X, Y^*) \geq \mathcal{L}_{\text{dw}}(\overline{M}; X, Y).$$

Note that Y^* may not be unique, but we can consider any Y^* for our argument. Suppose $\boldsymbol{y}_1^*$ and $\boldsymbol{y}_2^*$ are the vectors corresponding to the blocks V_1 and V_2 in the solution. We can then write the objective $\mathcal{L}_{\text{dw}}(\overline{M}; X, Y^*)$ as $\sum_i g(\boldsymbol{x}_i)$, where for any $i \in V_1$, $g(\boldsymbol{x}_i)$ is given by:

$$g(\boldsymbol{x}_i) = \frac{n}{2}\left(a \cdot \log \frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_1^*}}{\frac{n}{2}\left(e^{\boldsymbol{x}_i^T \boldsymbol{y}_1^*} + e^{\boldsymbol{x}_i^T \boldsymbol{y}_2^*}\right)}\right) + \frac{n}{2}\left(b \cdot \log \frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_2^*}}{\frac{n}{2}\left(e^{\boldsymbol{x}_i^T \boldsymbol{y}_1^*} + e^{\boldsymbol{x}_i^T \boldsymbol{y}_2^*}\right)}\right).$$

Writing $t_i = \boldsymbol{x}_i^T(\boldsymbol{y}_1^* - \boldsymbol{y}_2^*)$, we can write $g(\boldsymbol{x}_i)$ as

$$g(\boldsymbol{x}_i) = -\frac{n}{2}a \log\left(1 + e^{-t_i}\right) - \frac{n}{2}b \log\left(1 + e^{t_i}\right) - \frac{n}{2}(a+b)\log\left(\frac{n}{2}\right).$$

Let us fix an $i \in V_1$ and drop the subscript in t_i for a moment. Then $g(\boldsymbol{x}_i)$ is only a function of the scalar variable t. Taking the derivative with respect to t,

$$\frac{dg(\boldsymbol{x}_i)}{dt} = \frac{n}{2}a\left(\frac{e^{-t}}{1+e^{-t}}\right) - \frac{n}{2}b\left(\frac{1}{1+e^{-t}}\right) = \frac{na}{2}\left(\frac{1}{1+e^{-t}}\right)\left(e^{-t} - \frac{b}{a}\right).$$

It is easy to see that this is a decreasing function of t. Hence for any $t \in [-2, 2]$, we have

$$\frac{dg(\boldsymbol{x}_i)}{dt} \geq \frac{na}{2}\left(\frac{1}{1+e^{-2}}\right)\left(e^{-2} - \frac{b}{a}\right) = \frac{na}{2} \cdot \gamma.$$

In the last step, we used the definition of γ from the statement of the Lemma. By our assumption that $a/b > e^2$, γ is positive, and consequently, g is maximized when t is as large as possible, i.e. when $t = 2$ (because $t_i \leq \|\boldsymbol{x}_i\|\|\boldsymbol{y}_1^* - \boldsymbol{y}_2^*\| \leq 2$). Furthermore, if we denote by g^* the value at $t = 2$, we have by the intermediate value theorem, $g^* - g(\boldsymbol{x}_i) \geq g'(\eta)(2 - t_i)$, where η is some real number between t_i and 2. Since the derivative is a decreasing function, we have $g'(\eta) \geq g'(2)$. Next, let $\boldsymbol{z}$ denote the vector of length 2 parallel to $\boldsymbol{y}_1^* - \boldsymbol{y}_2^*$.

Case 1. $t_i \geq 0$. In this case, we have

$$\begin{aligned} g^* - g(\boldsymbol{x}_i) &\geq g'(2)(2 - t_i) \\ &\geq g'(2)(2 - \langle \boldsymbol{x}_i, \boldsymbol{z}\rangle) \\ &\geq g'(2)\|\boldsymbol{x}_i - \frac{\boldsymbol{z}}{2}\|^2 \end{aligned}$$

In the last step, we used also the property that $\|\boldsymbol{x}_i\| \leq 1$.

Case 2. $t_i < 0$. In this case,

$$g^* - g(\boldsymbol{x}_i) \geq 2g'(2) \geq \frac{g'(2)}{2}\|\boldsymbol{x}_i - \frac{\boldsymbol{z}}{2}\|^2,$$

simply because $\|\boldsymbol{x}_i - \frac{\boldsymbol{z}}{2}\| \leq 2$.

Now if we perform the same argument with any $i \in V_2$, we obtain $g^* - g(\boldsymbol{x}_i) \geq \frac{g'(2)}{2}\|\boldsymbol{x}_i + \frac{\boldsymbol{z}}{2}\|^2$.

Thus, setting $\boldsymbol{x}_1^* = \frac{\boldsymbol{z}}{2}$ and $\boldsymbol{x}_2^* = -\boldsymbol{x}_1^*$, we have (plugging in $g'(2) = \frac{na}{2}\gamma$)

$$\text{OPT} - \mathcal{L}_{\text{dw}}(\overline{M}; X, Y^*) \geq \frac{na\gamma}{4}\left[\sum_{i \in V_1}\|\boldsymbol{x}_i - \boldsymbol{x}_1^*\|^2 + \sum_{i \in V_2}\|\boldsymbol{x}_i - \boldsymbol{x}_2^*\|^2\right].$$

(In obtaining the above, we are also implicitly using the fact that it is possible to achieve a value of g^* for every term, by setting $\boldsymbol{x}_i^*$ appropriately.) Thus, if we started with the assumption that the LHS is at most ϵ (which holds because $\mathcal{L}_{\text{dw}}(\overline{M}; X, Y^*) \geq \mathcal{L}_{\text{dw}}(\overline{M}; X, Y)$), we have that each of the terms on the RHS is bounded, as desired. □

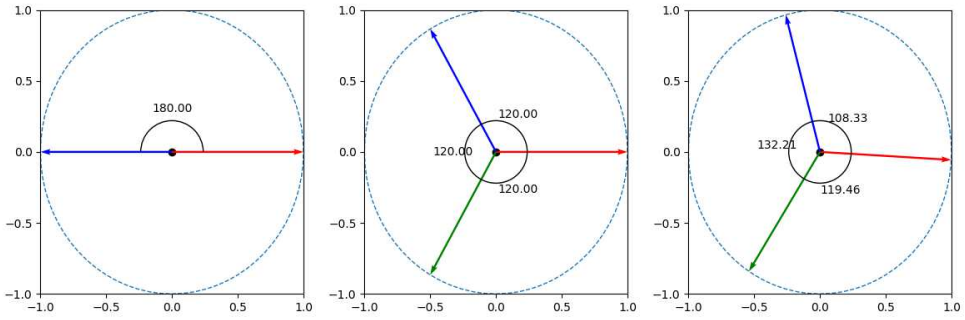
We have an analogous lemma for the SGNS objective. The proof can be found in Appendix C

Lemma 10. *Suppose the parameters of the matrix $\overline{M}$ (defined using V_1, V_2, a, b, as above) satisfy $\frac{eb}{a+b} < s_n < \frac{a}{(a+b)e}$. Let (X, Y) be any feasible solution that satisfies*

$$\mathcal{L}_{dw}(\overline{M}; X, Y) \geq \text{OPT} - \epsilon,$$

where OPT *is the optimum value of the objective, and $\epsilon \geq 0$ is some parameter. Define*

$$\gamma = \min\left\{\frac{na}{2(1+e)}\left(1 - \frac{(a+b)s_n}{a}e\right), \frac{ns_n(a+b)}{2(1+e)}\left(1 - \frac{b}{(a+b)s_n}e\right)\right\}.$$



(a) (b) (c)

Figure 1: Embeddings produced by optimizing over all possible orientations of (a) $K = 2$ block solutions and (b) $K = 3$ block solutions. Plot (c) shows $K = 3$ block solutions with clusters of varying sizes, namely $|V_1| = 25, |V_2| = 50$ and $|V_3| = 100$. Only the node embedding vectors are shown because the context embedding vectors are equivalent.

Then there exist unit vectors $\boldsymbol{x}_1^*$ *and* $\boldsymbol{x}_2^*$ *such that* $\boldsymbol{x}_2^* = -\boldsymbol{x}_1^*$ *and for* $k \in \{1, 2\}$,

$$\frac{1}{|V_k|} \sum_{i \in V_k} \|\boldsymbol{x}_i - \boldsymbol{x}_k^*\|^2 \leq \frac{4\epsilon}{n\gamma}. \tag{9}$$

Remark. Note that our proof of Theorem 8 is tailored to the case $K = 2$. For $K = 3$, we can no longer reduce to a single variable optimization problem, which makes the analysis intricate. Further, for $K = 2$, the ambient dimension of the embeddings does not matter in the analysis, while this is not the case for $K \geq 3$ (i.e., $d = 2$ and $d = 3$ makes a difference). Analysis of the structure of optimal solutions for higher K remains an open problem.

6 EXPERIMENTS

This section presents experimental results that support our theoretical analysis. The first section presents the results of a brute force search of the optimal orientation of node and context embeddings with block structure on symmetric block co-occurrence matrices with two and three blocks. The second experiment shows how the orientation of embedding vectors is strongly influenced by the number of negative samples.

6.1 Optimal Orientations

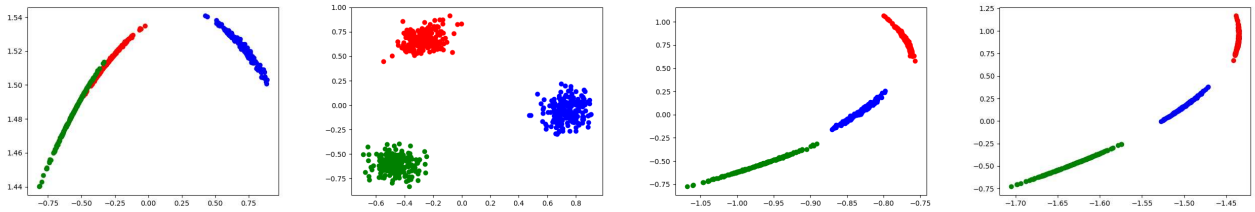
This section provides experimental results supporting the analytical results presented in 5. Assuming a co-occurrence matrix with a symmetric block structure and block-structured solutions as defined in Definition 2, the DeepWalk objective 2 was maximized by conducting a brute force search over possible unit vectors for $K = 2$ blocks. The off-diagonal block entries of $b = 0.1$ and on block-diagonal entries of $a = 8b$ were used. We fix $y_1 = (1, 0)$ and for every value of y_2 on the unit circle, in increments of one degree, we find the pair (x_1, x_2) that maximizes the objective. Finally, we take the pairs (x_1, x_2) and (y_1, y_2) that produce the largest objective value.

The results, shown in Figure 1, show that for large enough spread between a and b, the optimal node embedding vectors x_1^* and x_2^* are anti-polar to each other. They are also equivalent to their respective context embeddings y_1^*, y_2^*.

A similar experiment was conducted for $K = 3$ blocks using on and off-block values of $b = 0.1$ and $a = 8b$. The brute-force search was conducted over a grid of angles in increments of five degrees. The optimal orientation is one in which the vectors x_1^*, x_2^* and x_3^* are as spread out as possible, i.e. 120 degrees between each vector. Similar to the $K = 2$ case, they are also equivalent to their respective context embedding vectors y_1^*, y_2^* and y_3^*. This experiment suggests a similar theoretical treatment for the $K = 3$ case may be possible. Interestingly, if the community sizes are not equal, however, the optimal orientation of the vectors deviate from the 120 degrees that was optimal in the equal community size case. This suggests that more general results might depend on the community sizes.

6.2 Sensitivity to Negative Samples

The SGNS result in Theorem 8 for the $K = 2$ case relies on the sufficient but not necessary condition $\frac{eb}{a+b} < s_n < \frac{a}{(a+b)e}$ where s_n is the number of negative samples. Indeed, Mimno & Thompson (2017) experimentally show that angles between word embedding vectors produced with negative sampling largely depend on the number of negative samples used. Similarly, Arora et al. (2019) suggests that the performance of SGNS embeddings on downstream classification tasks can be negatively impacted if the number of negative samples is too large. We conduct a simple experiment in the $K = 3$ case to show that the number of negative samples impacts the geometry of SGNS embeddings of an SBM with $K = 3$ blocks, suggesting that conditions on the number of negative samples



(a) (b) (c) (d)

Figure 2: Embeddings produced by optimizing the skip-gram with negative sampling objective using (a) $s_n = 1/10$, (b) $s_n = 1$, (c) $s_n = 2$ and (d) $s_n = 10$ negative samples.

similar to the $K = 2$ case may be necessary in general in order to achieve satisfactory performance on downstream tasks.

We sampled a graph from a stochastic block model with $K = 3$ equally sized communities each with $|V_k| = 200$ nodes with parameters $q = 0.1, p = 4q$. The co-occurrence matrix was constructed by performing $r = 100$ random walks of length $L = 100$ for each node as a starting node. The window length was $T = 3$. SGNS embeddings were produced by optimizing (3) using gradient descent for $s_n = 1/10, 1, 2, 10$ negative samples.

The results can be seen if Figure 2. The embeddings produced by optimizing the SGNS objective differ depending on the number of negative samples used, especially among the smaller values. A fractional number of negative samples $s_n = 1/10$, which can be interpreted as taking one negative sample for every 10 positive samples, produces embeddings for two of the three communities that partly overlap each other. Using $s_n = 1$ negative sample produces embeddings that separate those belonging to each community well. In addition, the embeddings for each community appear to be evenly spaced from those of the other communities. The $s_n = 2$ and $s_n = 10$ cases produce very similar embeddings. However, they are quite different than those produced by $s_n = 1$ despite being able to recover each community. These results seem to justify the need for some conditions on s_n in order to achieve a satisfactory clustering performance. The similarity between the embeddings for $s_n = 2, 10$, however, suggest that there may exist conditions on the number of negative samples in which the embeddings might begin to stabilize.

7 CONCLUSION

We prove guarantees on the embeddings obtained by DeepWalk and SGNS on graphs generated by an SBM, establishing structural properties of their optimal solutions. We show that the optimal values of the DeepWalk and SGNS objectives are close to the optimal values computed on an appropriate block co-occurrence matrix. As a result, for a symmetric SBM with two communities, we are able to provide conditions in which the optimal solution vectors are well-clustered. Unlike previous works, we analyze the low dimensional embeddings directly. The generalization of our results to SBMs with more than two components and to components of unequal size, is left for future work.

Acknowledgements

Aditya Bhaskara acknowledges support from NSF grants CCF-2008688 and CCF-2047288.

References

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

Noga Alon. Eigenvalues and expanders. *Combinatorica*, 6 (2):83–96, June 1986. ISSN 0209-9683. doi: 10.1007/BF02579166.

Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56(2), apr 2009. ISSN 0004-5411. doi: 10.1145/1502793.1502794. URL https://doi.org/10.1145/1502793.1502794.

Sanjeev Arora, Wei Hu, and Pravesh K Kothari. An analysis of the t-sne algorithm for data visualization. In *Conference On Learning Theory*, pp. 1455–1462. PMLR, 2018.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *CoRR*, abs/1902.09229, 2019. URL http://arxiv.org/abs/1902.09229.

Lars Backstrom and Jure Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pp. 635–644, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450304931.

doi: 10.1145/1935826.1935914. URL https://doi.org/10.1145/1935826.1935914.

Aman Barot, Shankar Bhamidi, and Souvik Dhara. Community detection using low-dimensional network embedding algorithms. *arXiv preprint arXiv:2111.05267*, 2021.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pp. 585–591, Cambridge, MA, USA, 2001. MIT Press.

Sudhanshu Chanpuriya and Cameron Musco. Infinitewalk: Deep network embeddings as laplacian embeddings with a nonlinearity. *CoRR*, abs/2006.00094, 2020. URL https://arxiv.org/abs/2006.00094.

Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1):79–127, 2006.

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*, pp. 855–864, 2016.

Kenneth M. Hall. An r-dimensional quadratic placement algorithm. *Manage. Sci.*, 17(3):219–229, nov 1970. ISSN 0025-1909. doi: 10.1287/mnsc.17.3.219. URL https://doi.org/10.1287/mnsc.17.3.219.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Paul Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5: 109–137, 1983.

Dominik Kloepfer, Angelica I Aviles-Rivero, and Daniel Heydecker. Delving into deep walkers: A convergence analysis of random-walk-based vertex embeddings. *arXiv preprint arXiv:2107.10014*, 2021.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/feab05aa91085b7a8012516bc3533958-Paper.pdf.

George C Linderman and Stefan Steinerberger. Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, 2019.

N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pp. 577–591, 1994. doi: 10.1109/SFCS.1994.365733.

F. McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pp. 529–537, 2001. doi: 10.1109/SFCS.2001.959929.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Empirical Methods in Natural Language Processing*, 2017.

A. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

Subhadeep Paul and Yuguo Chen. Orthogonal symmetric non-negative matrix factorization under the stochastic block model. *arXiv preprint arXiv:1605.05349*, 2016.

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*, pp. 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*, pp. 459–467, 2018.

Jiezhong Qiu, Chi Wang, Ben Liao, Richard Peng, and Jie Tang. A matrix chernoff bound for markov chains and its application to co-occurrence matrices. *Advances in Neural Information Processing Systems*, 33:18421–18432, 2020.

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

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013. doi: 10.1214/ECP.v18-2865. URL https://doi.org/10.1214/ECP.v18-2865.

Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing markov chains. *Information and Computation*, 82(1):93–133, 1989. ISSN 0890-5401. doi: https://doi.org/10.1016/0890-5401(89)90067-9. URL https://www.sciencedirect.com/science/article/pii/0890540189900679.

Daniel A. Spielman and Shang-Hua Teng. Spectral partitioning works: Planar graphs and finite element meshes. *Linear Algebra and its Applications*, 421(2):284–305, 2007. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2006.07.020. URL https://www.sciencedirect.com/science/article/pii/S0024379506003454. Special Issue in honor of Miroslav Fiedler.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pp. 1067–1077, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741093. URL https://doi.org/10.1145/2736277.2741093.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Van Vu. A simple svd algorithm for finding hidden partitions, 2014. URL https://arxiv.org/abs/1404.3918.

Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 203–209. AAAI Press, 2017.

Yichi Zhang and Minh Tang. Consistency of random-walk based network embedding algorithms. *arXiv preprint arXiv:2101.07354*, 2021.

A DETAILED PROOFS FROM SECTIONS 2 AND 3

A.1 Proof of Lemma 1

The proof proceeds by a direct computation of the terms C_{ij}. This is similar to some of the analysis in prior works (Zhang & Tang, 2021; Barot et al., 2021). We denote the degree of vertex i by d_i, and write D to denote the diagonal matrix of degrees.

Proof. Recall that the entries of the co-occurrence matrix, can be written as

$$\begin{aligned}\frac{C_{ij}}{r} &= \sum_{t=1}^{T}\sum_{k=1}^{L-t}\sum_{m=1}^{r}\left[\frac{\mathbb{1}\{w_k^{(m)}=i, w_{k+t}^{(m)}=j\}}{r} + \frac{\mathbb{1}\{w_k^{(m)}=j, w_{k+t}^{(m)}=i\}}{r}\right] \\ &= \sum_{t=1}^{T}\sum_{k=1}^{L-t}\left[\sum_{m=1}^{r}\left(\frac{\mathbb{1}\{w_k^{(m)}=i, w_{k+t}^{(m)}=j\}}{r}\right) + \sum_{m=1}^{r}\left(\frac{\mathbb{1}\{w_k^{(m)}=j, w_{k+t}^{(m)}=i\}}{r}\right)\right]\end{aligned}$$

Taking the limit as the number of random walks $r \to \infty$ and by the large of law numbers we have

$$\begin{aligned}\frac{C_{ij}}{r} &\to \sum_{t=1}^{T}\sum_{k=1}^{L-t}\left(\Pr\left[w_k^{(1)}=i, w_{k+t}^{(1)}=j\right] + \Pr\left[w_k^{(1)}=j, w_{k+t}^{(1)}=i\right]\right) \\ &= \sum_{t=1}^{T}(L-t)\left(\Pr\left[w_{t+1}=j|w_1=i\right]\Pr\left[w_1=i\right] + \Pr\left[w_{t+1}=i|w_1=j\right]\Pr\left[w_1=j\right]\right) \\ &= \sum_{t=1}^{T}(L-t)\left(\Pr\left[w_{t+1}=j|w_1=i\right]\frac{d_i}{2|E|} + \Pr\left[w_{t+1}=i|w_1=j\right]\frac{d_j}{2|E|}\right) \\ &= \sum_{t=1}^{T}(L-t)\left(2\Pr[w_{t+1}=j|w_1=i]\frac{d_i}{2|E|}\right) \\ &= 2\sum_{t=1}^{T}(L-t)\cdot\pi_i(W^t)_{ij}.\end{aligned}$$

where the second to last step is because the powers of the transition matrix $(D^{-1}A)$ are symmetric after left multiplication by D. □

A.2 Proof of Theorem 3

Proof. As in the definition of symmetric block structure, let a, b be parameters for which the matrix C can be defined as having $C_{ij} = a$ if $i \sim j$ and $C_{ij} = b$ otherwise.

We first show that given any fixed X, there exists an optimal solution in which Y has a block structure. Suppose nodes 1 and node 2 are in the same cluster, and suppose we start with an optimal solution with $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$. We consider two candidate solutions: one in which $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ is replaced with $(\boldsymbol{y}_1, \boldsymbol{y}_1)$ and another where it is replaced with $(\boldsymbol{y}_2, \boldsymbol{y}_2)$, keeping the rest of the solution the same. We argue that at least one of these solutions has an objective value at least as high as the original (X, Y). Note that this process maintains feasibility (belonging to some domain $\mathcal{D}$), which is why our argument also applies to constrained optimization.

Let us denote the log-sum-exp function, parameterized by $\boldsymbol{x}_i$ and $\{\boldsymbol{y}_j\}_{k>2}$, as

$$\mathrm{lse}\left(\boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{x}_i, \{\boldsymbol{y}_j\}_{j>2}\right) = \log\left[e^{\boldsymbol{x}_i^T\boldsymbol{y}_1} + e^{\boldsymbol{x}_i^T\boldsymbol{y}_2} + \sum_{k>2} e^{\boldsymbol{x}_i^T\boldsymbol{y}_k}\right]$$

Note that lse corresponds to the denominator of our loss function. We will usually ignore the parameters $\boldsymbol{x}_i$ and $\{\boldsymbol{y}_j\}_{k>2}$ and just write $\mathrm{lse}(\boldsymbol{y}_1, \boldsymbol{y}_2)$. For simplicity and to emphasize $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, we write the objective as $\mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_2)$. Then

$$\mathcal{L}\left(\boldsymbol{y}_1, \boldsymbol{y}_2\right) = \sum_{i=1}^{n} \left[a\boldsymbol{x}_i^T \left(\boldsymbol{y}_1 + \boldsymbol{y}_2\right) + \sum_{j \in V_1 \setminus \{1,2\}} a\boldsymbol{x}_i^T \boldsymbol{y}_j + \sum_{j \notin V_1} b\boldsymbol{x}_i^T \boldsymbol{y}_j - \operatorname{lse}(\boldsymbol{y}_1, \boldsymbol{y}_2) \left(\sum_{j \in V_1} a + \sum_{j \notin V_1} b \right) \right]$$

Taking the average of $\mathcal{L}\left(\boldsymbol{y}_1, \boldsymbol{y}_1\right)$ and $\mathcal{L}\left(\boldsymbol{y}_2, \boldsymbol{y}_2\right)$ we see that

$$\begin{aligned} \frac{\mathcal{L}\left(\boldsymbol{y}_1, \boldsymbol{y}_1\right) + \mathcal{L}\left(\boldsymbol{y}_2, \boldsymbol{y}_2\right)}{2} = \sum_{i=1}^{n} \Bigg[& a\boldsymbol{x}_i^T \left(\boldsymbol{y}_1 + \boldsymbol{y}_2\right) + \sum_{j \in V_1 \setminus \{1,2\}} a\boldsymbol{x}_i^T \boldsymbol{y}_j + \sum_{j \notin V_1} b\boldsymbol{x}_i^T \boldsymbol{y}_j \\ & - \frac{1}{2} \left(\operatorname{lse}(\boldsymbol{y}_1, \boldsymbol{y}_1) + \operatorname{lse}(\boldsymbol{y}_2, \boldsymbol{y}_2) \right) \left(\sum_{j \in V_1} a + \sum_{j \notin V_1} b \right) \Bigg] \\ \geq \mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_2) & \end{aligned}$$

where the last line uses an inequality of the form $\log(2A + Z) + \log(2B + Z) \leq 2\log(A + B + Z)$, for appropriate choices of A, B, Z. This is a consequence of the AM-GM inequality (or equivalently, concavity of the logarithm).

This implies that either $\mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_1) \geq \mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_2)$ or that $\mathcal{L}(\boldsymbol{y}_2, \boldsymbol{y}_2) \geq \mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_2)$. Regardless, when $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$ the objective function is no larger than when $\boldsymbol{y}_1 = \boldsymbol{y}_2$. It follows that there exists a solution that maximizes the objective whose context embeddings Y have a block structure.

The proof for X is actually easier. Once we have a block structured solution Y, since the objective splits based on the $\boldsymbol{x}_i$, we will be solving exactly the same optimization problem for every i in a block, and thus there exists a block structured solution X. This concludes the proof. □

A.3 Proof of Theorem 4

Proof. As in the proof in Section A.2, we first show that there exists an optimal solution whose context embeddings Y have a block structure. Start with some optimal solution, and assume that node 1 and node 2 are vertices in the same cluster with $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$. As before, we write the objective as $\mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_2)$. Let us define

$$h(\boldsymbol{y}_1, \boldsymbol{y}_2) = \sum_{i=1}^{n} \left[C_{i1} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_1) + C_{i2} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_2) + \sum_{j>2} C_{ij} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_j) \right]$$

$$g(\boldsymbol{y}_1, \boldsymbol{y}_2) = \sum_{i=1}^{n} \left[s_n \frac{|C_{i\cdot}|}{|C|} \left(|C_{\cdot 1}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_1) + |C_{\cdot 2}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_2) + \sum_{j>2} |C_{\cdot j}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j) \right) \right]$$

We can write $\mathcal{L}$ as

$$\mathcal{L}\left(\boldsymbol{y}_1, \boldsymbol{y}_2\right) = h(\boldsymbol{y}_1, \boldsymbol{y}_2) + g(\boldsymbol{y}_1, \boldsymbol{y}_2)$$

Since node 1 and node 2 are in the same cluster, we have $C_{i1} = C_{i2}$. Notice that

$$h(\boldsymbol{y}_1, \boldsymbol{y}_1) = \sum_{i=1}^{n} \left[2C_{i1} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_1) + \sum_{j>2} C_{ij} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_j) \right]$$
$$h(\boldsymbol{y}_2, \boldsymbol{y}_2) = \sum_{i=1}^{n} \left[2C_{i2} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_2) + \sum_{j>2} C_{ij} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_j) \right]$$

$$\begin{aligned} h(\boldsymbol{y}_1, \boldsymbol{y}_1) + h(\boldsymbol{y}_2, \boldsymbol{y}_2) &= \sum_{i=1}^{n} \left[2C_{i1} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_1) + 2C_{i2} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_2) + 2 \sum_{j>2} C_{ij} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_j) \right] \\ &= 2h(\boldsymbol{y}_1, \boldsymbol{y}_2) \end{aligned}$$

Similarly, since node 1 and node 2 are in the same cluster, $|C_{\cdot 1}| = |C_{\cdot 2}|$. So we have

$$g(\boldsymbol{y}_1, \boldsymbol{y}_1) = \sum_{i=1}^{n} \left[s_n \frac{|C_{i\cdot}|}{|C|} \left(2|C_{\cdot 1}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_1) + \sum_{j>2} |C_{\cdot j}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j) \right) \right]$$
$$g(\boldsymbol{y}_2, \boldsymbol{y}_2) = \sum_{i=1}^{n} \left[s_n \frac{|C_{i\cdot}|}{|C|} \left(2|C_{\cdot 2}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_2) + \sum_{j>2} |C_{\cdot j}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j) \right) \right]$$

$$\begin{aligned} g(\boldsymbol{y}_1, \boldsymbol{y}_1) + g(\boldsymbol{y}_2, \boldsymbol{y}_2) &= \sum_{i=1}^{n} \left[s_n \frac{|C_{i\cdot}|}{|C|} \left(2|C_{\cdot 1}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_1) + 2|C_{\cdot 2}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_2) + 2 \sum_{j>2} |C_{\cdot j}| \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j) \right) \right] \\ &= 2g(\boldsymbol{y}_1, \boldsymbol{y}_2) \end{aligned}$$

It follows that

$$\frac{\mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_1) + \mathcal{L}(\boldsymbol{y}_2, \boldsymbol{y}_2)}{2} = \mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_2)$$

This implies that either $\mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_1) \geq \mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_2)$ or that $\mathcal{L}(\boldsymbol{y}_2, \boldsymbol{y}_2) \geq \mathcal{L}(\boldsymbol{y}_1, \boldsymbol{y}_2)$. Regardless, when $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$ the objective function is no larger than when $\boldsymbol{y}_1 = \boldsymbol{y}_2$. It follows that there exists a solution that maximizes the objective whose context embeddings Y has a block structure.

Once we have a block structured solution Y, since the objective splits based on the $\boldsymbol{x}_i$, we will be solving exactly the same optimization problem for every i in a block, and thus there exists a block structured solution X. This concludes the proof.

□

B DETAILS OF SECTION 4

B.1 Proof of Theorem 5

B.1.1 DeepWalk Objective Function

We start by observing that for any candidate solution X, Y,

$$\mathcal{L}_{\text{dw}}(M; X, Y) = \sum_{i,j} M_{ij} \alpha_{ij}, \quad \text{where } \alpha_{ij} := \log \frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_j}}{\sum_k e^{\boldsymbol{x}_i^T \boldsymbol{y}_k}}.$$

Define Λ (which is dependent on X, Y) to be the matrix whose (i, j)th entry is α_{ij}, so we have $\mathcal{L}_{\text{dw}}(M; X, Y) = M \cdot \Lambda$. Then the goal of the theorem is to prove that w.h.p., $|(M - \overline{M}) \cdot \Lambda|$ is small for all X, Y.

Outline. The proof starts by using Lemma 6 to show that $(M - M') \cdot \Lambda$ is small for all X, Y. Next, we argue using Lemma 7 that for any *fixed* solution X, Y, equation 6 holds with an exponentially small failure probability. Finally, we use an epsilon-net argument (a standard technique in random matrix theory, e.g., Vershynin (2018)) to take a union bound over a fine enough "net" over the possible X, Y. This lets us obtain the conclusion of the theorem for all X, Y.

Proof. For bounding $(M - M') \cdot \Lambda$, we observe that for any X, Y with $\|\boldsymbol{x}_i\|, \|\boldsymbol{y}_j\| \leq 1$, we have $e^{\boldsymbol{x}_i^T \boldsymbol{y}_j} \in [\frac{1}{e}, e]$ for all i, j, and thus

$$\alpha_{ij} \in \left[\log\left(\frac{1}{ne^2}\right), \log\left(\frac{e^2}{n}\right)\right] = [-\log n - 2, -\log n + 2].$$

This implies that $\|\Lambda\|_\infty \leq 2\log n$, assuming $n > 10$. Thus, we can use Lemma 6 to conclude that with probability $\geq 1 - n^{-5}$, for all candidate solutions X, Y,

$$(M - M') \cdot \Lambda \leq \|M - M'\|_1 \|\Lambda\|_\infty \leq O\left(L\sqrt{\frac{\log^3 n}{\Delta}}\right). \tag{10}$$

Next, we bound $(M' - \overline{M}) \cdot \Lambda$ for all X, Y. We do this via an epsilon-net argument. Let δ be a parameter we will choose later, and let S_δ be a set of matrix pairs (X', Y') such that for every (X, Y) with columns of length at most 1, there exists $(X', Y') \in S_\delta$ such that $\|X - X'\|_F + \|Y - Y'\|_F \leq \delta$. Using standard bounds on the sizes of epsilon nets (e.g., Corollary 4.2.13 of Vershynin (2018)), we have that there exists a net with

$$|S_\delta| \leq \left(\frac{6n}{\delta}\right)^{2nd}.$$

(Indeed, a net of this size exists for the radius-$2n$ ball in a $2nd$ dimensional space, which is a superset of the matrix pairs (X, Y) that we are interested in.)

Next, we apply Lemma 7 with parameters $\beta = 2\log n$ and an ϵ that satisfies the condition:

$$\min\left\{\frac{\epsilon^2\Delta^4}{32n\log^2 n}, \frac{\epsilon\Delta^2}{8\log n}\right\} \geq 5\log n + 2nd\log\left(\frac{6n}{\delta}\right)$$

Then, we can use the union bound to conclude that with prob. $\geq 1 - 4n^{-5}$, for *all* $(X', Y') \in S_\delta$,

$$|\mathcal{L}_{\text{dw}}(M'; X', Y') - \mathcal{L}_{\text{dw}}(\overline{M}; X', Y')| \leq 2\epsilon L.$$

Now, if we set $\delta = \frac{\epsilon}{n^4}$, we have that if (X, Y) is any pair of matrices with columns of norm ≤ 1 and (X', Y') is the pair closest to it in S_δ,

$$|\mathcal{L}_{\text{dw}}(M'; X, Y) - \mathcal{L}_{\text{dw}}(M'; X', Y')| \leq \epsilon L/2,$$

and the same inequality holds with $\overline{M}$. Thus, using triangle inequality, we have that with probability at least $1 - 4n^{-5}$, for all pairs (X, Y),

$$|\mathcal{L}_{\text{dw}}(M'; X, Y) - \mathcal{L}_{\text{dw}}(\overline{M}; X, Y)| \leq 3\epsilon L.$$

Finally, combining this step with equation 10, we have that with probability at least $1 - n^{-4}$, for all (X, Y),

$$|\mathcal{L}_{\text{dw}}(M; X, Y) - \mathcal{L}_{\text{dw}}(\overline{M}; X, Y)| \leq O\left(L\epsilon + L\sqrt{\frac{\log^3 n}{\Delta}}\right),$$

as long as we have the inequality

$$\min\left\{\frac{\epsilon^2\Delta^4}{32n\log^2 n}, \frac{\epsilon\Delta^2}{8\log n}\right\} \geq 5\log n + 2nd\log\left(6n^4\right).$$

By a straightforward calculation, we see that this condition holds as long as

$$\Delta^2 \geq \frac{400nd\log^2 n}{\epsilon}.$$

This completes the proof of the Theorem, for the DeepWalk objective.

□

B.1.2 Negative Sampling Objective Function

Recall that the SGNS loss function corresponding to matrix M has two components:

$$\mathcal{L}_{\text{ns}}(M; X, Y) = \sum_{i,j} M_{ij} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_j) + s_n \sum_{i,j} \frac{|M_{i\cdot}||M_{\cdot j}|}{|M|} \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j),$$

where M is defined in equation 4.

Proof outline. We can split $\mathcal{L}_{\text{ns}}(M; X, Y) - \mathcal{L}_{\text{ns}}(\overline{M}; X, Y)$ into two terms as above. The first one, $\sum_{i,j}(M_{ij} - \overline{M}_{ij}) \log \sigma(\boldsymbol{x}_i^T \boldsymbol{y}_j)$, is handled exactly as we did for the DeepWalk objective function. The second term can be handled more directly and shown to be small.

Proof. Let us begin by looking at the second term, as it is unique to the SGNS loss. Note that by definition,

$$|E| \cdot M_{ij} = (L-1)A_{ij} + (L-2) \sum_r \frac{A_{ir}A_{rj}}{d_r},$$

where d_r is the degree of the rth vertex. Thus we have

$$\begin{aligned} |E| \cdot |M_{i\cdot}| = |E| \cdot \sum_j M_{ij} &= (L-1)d_i + (L-2) \sum_r \frac{A_{ir}}{d_r} \sum_j A_{rj} \\ &= (L-1)d_i + \sum_r \frac{A_{ir}}{d_r} \cdot d_r = (2L-3)d_i. \end{aligned}$$

Similarly, we can see that $|E| \cdot |M_{\cdot j}| = (2L-3)d_j$. This then implies that $|M| = \sum_j |M_{\cdot j}| = 2(2L-3)$. So we have

$$\sum_{i,j} \frac{|M_{i\cdot}||M_{\cdot j}|}{|M|} \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j) = \sum_{i,j} \frac{(2L-3)d_i d_j}{2|E|^2} \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j).$$

For any vectors $\boldsymbol{x}_i, \boldsymbol{y}_j$ of norm ≤ 1, the $\sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j)$ term is $\Theta(1)$ (and in fact, it is at most 1 in magnitude). Thus, we have

$$\left| \sum_{i,j} \frac{(2L-3)d_i d_j}{2|E|^2} \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j) - \sum_{i,j} \frac{2(2L-3)\Delta^2}{n^2\Delta^2} \log \sigma(-\boldsymbol{x}_i^T \boldsymbol{y}_j) \right| \leq 2(2L-3) \sum_{i,j} \left| \frac{d_i d_j}{2|E|^2} - \frac{2\Delta^2}{n^2\Delta^2} \right|.$$

As in the proof of Lemma 6, we have that for the range of Δ of interest, for every i, j, the magnitude of the difference is $\leq O\left(\frac{\sqrt{\log n}}{n^2\sqrt{\Delta}}\right)$ with probability $\geq 1 - n^{-5}$. This implies that the error in total is $O\left(L\sqrt{\frac{\log n}{\Delta}}\right)$, and this shows that with high probability, the difference between the second term in the SGNS loss for M and for $\overline{M}$ differ by at most $O\left(L\sqrt{\frac{\log n}{\Delta}}\right)$ for all $\boldsymbol{x}_i, \boldsymbol{y}_j$.

Finally, we note that the first term in the SGNS loss can be bounded exactly as before (with the sigmoid terms playing the role of the softmax). Indeed, we can use the slightly improved value of $\beta = 2$, because the sigmoid is bounded by a constant. This completes the proof of the theorem.

□

B.2 Proof of Lemma 6

Proof. First, we argue that the following bounds hold with probability $\geq 1 - n^{-5}$ when a graph G is drawn from a symmetric SBM with parameters (K, p, q) and $p > n^{-1/2}$.

$$\left| |E| - \frac{n\Delta}{2} \right| \leq O\left((n\Delta)^{1/2}(\log n)^{1/2}\right), \tag{11}$$

$$\forall i, \;\; |d_i - \Delta| \leq O(\sqrt{\Delta \log n}). \tag{12}$$

These inequalities follow from the standard Chernoff bound (see, e.g., Theorem 2.3.1 of Vershynin (2018)) on the sum of independent Bernoulli random variables.

Next, whenever G satisfies the above conditions, we show that for all X, Y with columns of length at most 1, $\mathcal{L}_{\text{dw}}(M; X, Y) - \mathcal{L}_{\text{dw}}(M'; X, Y)$ is small. To this end, note that it is sufficient to bound $\sum_{ij} |M_{ij} - M'_{ij}|$. Writing it out for a fixed i, j,

$$M_{ij} - M'_{i,j} = (L-1)A_{ij}\left(\frac{1}{|E|} - \frac{2}{n\Delta}\right) + (L-2)\left(\frac{\sum_k A_{ik}A_{kj}/d_k}{|E|} - \frac{2\sum_k A_{ik}A_{kj}}{n\Delta^2}\right). \tag{13}$$

Let us bound the two terms in the parentheses separately. By using equation 11, the first term can be bounded by $O\left((n\Delta)^{-3/2}(\log n)^{1/2}\right)$. Let us consider the second term. We split it into two terms as follows:

$$\frac{\sum_k A_{ik}A_{kj}/d_k}{|E|} - \frac{2\sum_k A_{ik}A_{kj}}{n\Delta^2} = \left(\frac{\sum_k A_{ik}A_{kj}/d_k}{|E|} - \frac{2\sum_k A_{ik}A_{kj}/d_k}{n\Delta}\right) + \left(\frac{2\sum_k A_{ik}A_{kj}/d_k}{n\Delta} - \frac{2\sum_k A_{ik}A_{kj}}{n\Delta^2}\right).$$

For the first term, we again use equation 11 along with triangle inequality and equation 12 to bound its magnitude by

$$\frac{O((\log n)^{1/2})}{(n\Delta)^{3/2}} \sum_k \frac{A_{ik}A_{kj}}{d_k} \leq \frac{O((\log n)^{1/2})}{\Delta(n\Delta)^{3/2}} \sum_k A_{ik}A_{kj}.$$

The second term can be bounded similarly using equation 12 by

$$\frac{2}{n\Delta}\sum_k A_{ik}A_{kj}\left|\frac{1}{d_k} - \frac{1}{\Delta}\right| \leq \frac{O(\sqrt{\Delta \log n})}{n\Delta^3}\sum_k A_{ik}A_{kj}.$$

Note that this bound is strictly worse (by a factor $n^{1/2}$) than our bound for the first term, so we only keep this one. Plugging all of these into equation 13 and summing over i, j, we obtain

$$\sum_{ij} |M_{ij} - M'_{ij}| \leq O(L\sqrt{\log n})\left(\frac{\sum_{ij} A_{ij}}{(n\Delta)^{3/2}} + \sum_{ij}\frac{1}{n\Delta^{5/2}}\sum_k A_{ik}A_{kj}\right).$$

Now, $\sum_{ij} A_{ij} \leq 2n\Delta$, as it is the number of edges. Likewise, $\sum_{ij}\sum_k A_{ik}A_{kj}$ counts the total number of length-2 paths in the graph, which is also equal to sum of the squares of the degrees. This is thus bounded by $2n\Delta^2$. Putting everything together, the above bound simplifies as

$$\sum_{ij} |M_{ij} - M'_{ij}| \leq O(L\sqrt{\log n})\left((n\Delta)^{-1/2} + \Delta^{-1/2}\right).$$

Since the last term dominates, this completes the proof of the lemma. □

B.3 Proof of Lemma 7

Proof of Lemma 7. Recalling the definitions of matrices M' and $\overline{M}$, we have

$$|(M' - \overline{M}) \cdot \Lambda| \leq \left|\frac{2(L-1)}{n\Delta}(A - \mathbb{E}[A]) \cdot \Lambda\right| + \left|\frac{2(L-2)}{n\Delta^2}(A^2 - \mathbb{E}[A^2]) \cdot \Lambda.\right| \tag{14}$$

If the LHS is $\geq t$, at least one of the terms on the RHS must be $\geq t/2$. We will thus bound the probabilities of these two events separately. For convenience, we will call them the linear and quadratic terms respectively. We also denote by λ_{ij} the (i,j)th entry of Λ. The technical issue in bounding the probabilities of the events comes from (a) the symmetry conditions $A_{ij} = A_{ji}$, and (b) the quadratic term A^2.

The linear term on the RHS of equation 14 turns out to be straightforward. Due to the conditions $A_{ij} = A_{ji}$ and $A_{ii} = 0$, we can write $(A - \mathbb{E}[A]) \cdot \Lambda = \sum_{i<j} A_{ij}(\lambda_{ij} + \lambda_{ji})$. Using the Bernstein inequality for the sum of independent but non-identical Bernoulli variables (e.g., Theorem 3.3 of Chung & Lu (2006)) and using the fact that each of the coefficients is at most 2β, we have that for all $s > 0$

$$\Pr\left[|(A - \mathbb{E}[A]) \cdot \Lambda| \geq s\right] \leq 2e^{-\frac{s^2}{4\beta^2 n\Delta + 2s\beta}}. \tag{15}$$

Setting $s = \frac{\epsilon n \Delta}{2}$, the RHS simplifies to $2\exp\left(-\frac{\epsilon^2 n\Delta}{4(4\beta^2+\epsilon\beta)}\right)$.

Next, let us focus on obtaining a large deviation bound for the quadratic term, i.e., analyzing $\Pr[|(A^2 - \mathbb{E}[A^2]) \cdot \Lambda| > s]$. Since λ_{ij} are fixed, by the linearity of expectation, we are effectively looking to bound $\Pr[|Y - \mathbb{E}[Y]| \geq s]$, where

$$Y = \sum_{i,j} \sum_{k} A_{ik} A_{kj} \lambda_{ij}.$$

Y is a quadratic polynomial in the random variables $\{A_{ij}\}_{i<j}$ (again, because of symmetry, and because $A_{ii} = 0$). Thus we will apply the Hanson and Wright inequality (e.g. Theorem 6.2.1 of Vershynin (2018)) in order to obtain a tail bound. This requires understanding the matrix B that defines the quadratic form Y, and more specifically, bounding its Frobenius and spectral norms, $\|B\|_F$ and $\|B\|_2$. Writing $Y = \sum_k \sum_{i,j\neq k} A_{ki} A_{kj} \lambda_{ij}$, we see that B is an $\binom{n}{2} \times \binom{n}{2}$ matrix, whose entries are defined by pairs of indices $\{i,j\}$ and $\{k,l\}$. We see that if $\{i,j\} \cap \{k,l\} = \emptyset$, the corresponding entry of B is zero. If the intersection is one, i.e., the indices are $\{i,j\}$ and $\{i,k\}$, the corresponding entry of B is $\lambda_{jk} + \lambda_{kj}$. Finally, for the diagonal terms (corresponding to say $\{i,j\}, \{i,j\}$), the entry is $\lambda_{ii} + \lambda_{jj}$.

First, we can bound the Frobenius norm of B directly: using the basic fact that $(a+b)^2 \leq 2(a^2+b^2)$ for all a, b, we have

$$\|B\|_F^2 \leq \sum_i \sum_{j,k} 2(\lambda_{jk}^2 + \lambda_{kj}^2) + \sum_{i<j} 2(\lambda_{ii}^2 + \lambda_{jj}^2).$$

This is clearly upper bounded by $2n\|\Lambda\|_F^2$.

Bounding the spectral norm turns out to be a bit more tricky. Let us define B' to be the "asymmetric" version of B, i.e., an $n(n-1) \times n(n-1)$ matrix whose rows and columns are indexed by ordered pairs (i,j), $i \neq j$. The diagonal entry in the (i,j)th row is set to be $2\lambda_{jj}$. Further, the entry corresponding to the (i,j)th row and (i,k)th column is $(\lambda_{jk} + \lambda_{kj})$. The rest of the entries in the matrix are 0. In other words, we can can think of B' as a matrix with diagonal blocks, where for every i, we have an $(n-1) \times (n-1)$ block of $\Lambda + \Lambda^T$, excluding the ith row and column.

The key observation is that we can relate the spectral norms $\|B\|_2$ and $\|B'\|_2$ in a simple manner. First, for any vector $x \in \mathbb{R}^{\binom{n}{2}}$, consider the vector $x' \in \mathbb{R}^{n(n-1)}$ where $x'_{(i,j)} = x'_{(j,i)} = x_{\{i,j\}}$. By construction, we have $\|x'\|^2 = 2\|x\|^2$.

Claim. We have $\|Bx\|^2 \leq 2\|B'x'\|^2$. This is because for any $x \in \mathbb{R}^{\binom{n}{2}}$, by definition, the $\{i,j\}$th entry of Bx is precisely

$$\sum_{k\neq i} (\lambda_{jk} + \lambda_{kj}) x_{\{i,k\}} + \sum_{k \neq j} (\lambda_{ik} + \lambda_{ki}) x_{\{j,k\}}. \tag{16}$$

On the other hand, the (i,j)th entry of $B'x'$ is $\sum_{k\neq i} (\lambda_{jk} + \lambda_{kj}) x_{(i,k)}$. Thus the expression in equation 16 is exactly the sum of the (i,j)th and (j,i)th entries of $B'x'$. This implies that $\|Bx\|^2 \leq 2\|B'x'\|^2$. Moreover, since $\|x'\|^2 = 2\|x\|^2$, we have

$$\frac{\|Bx\|^2}{\|x\|^2} \leq \frac{\|B'x'\|^2}{\|x'\|^2}.$$

Thus the spectral norm of B is at most the spectral norm of B'. The latter is easy to bound, as it is a block diagonal matrix. We have

$$\|B'\|_2^2 \leq \|\Lambda + \Lambda^T\|_F^2 \leq 4\|\Lambda\|_F^2.$$

Since $\|\Lambda\|_F^2 \leq n^2\beta^2$, we have that

$$\|B\|_F^2 \leq 2n^3\beta^2 \qquad \text{and} \qquad \|B\|_2^2 \leq 4n^2\beta^2.$$

Equipped with these bounds, we can use the Hanson-Wright inequality (indeed, the stronger form by Rudelson & Vershynin (2013) which also applies to non-identical random variables). Noting that the sub-Gaussianity constant of a Bernoulli random variable is ≤ 1, this implies that $\forall\ s > 0$,

$$\Pr[|Y - \mathbb{E}[Y]| > s] \leq 2\exp\left(-\min\left\{\frac{s^2}{2n^3\beta^2}, \frac{s}{2n\beta}\right\}\right).$$

Setting $s = \frac{\epsilon n \Delta^2}{2}$, this simplifies to $2\exp(-\min\{\frac{\epsilon^2\Delta^4}{8n\beta^2}, \frac{\epsilon\Delta^2}{4\beta}\})$.

We have thus obtained bounds on the probability of the linear and quadratic terms in equation 14 each exceeding ϵL. Observe that as long as $n > 8\beta/\epsilon$, the bound from the quadratic term is strictly larger than the bound from the linear term. Thus, by using the union bound, we obtain the conclusion of the lemma. □

C DETAILS OF SECTION 5

C.1 Proof of Lemma 10

Proof. The proof is similar to that of Lemma 9. We appeal to Theorem 4 to conclude that there exists a block structured Y^* (with components $\boldsymbol{y}_1^*$ and $\boldsymbol{y}_2^*$, say), such that

$$\mathcal{L}_{\text{ns}}(\overline{M}; X, Y^*) \geq \mathcal{L}_{\text{ns}}(\overline{M}; X, Y).$$

Once again, we decompose the loss as a sum $g(\boldsymbol{x}_i)$, where for some $i \in V_1$, we have

$$g(\boldsymbol{x}_i) = \frac{na}{2}\log\sigma(t_i) + \frac{s_n(a+b)n}{2}\log\sigma(-t_i) + \frac{nb}{2}\log\sigma(t_i') + \frac{s_n(a+b)n}{2}\log\sigma(-t_i'),$$

where we have defined $t_i = \boldsymbol{x}_i^T\boldsymbol{y}_1^*$ and $t_i' = \boldsymbol{x}_i^T\boldsymbol{y}_2^*$. We next observe that under our assumptions, this quantity is maximized when $t_i = 1$ and $t_i' = -1$.

To do this, suppose we define $h_{A,B}(t) = A\log\sigma(t) + B\log\sigma(-t)$, where A, B are parameters that satisfy $A/B \geq 2$. Then,

$$\begin{aligned}\frac{dh_{A,B}}{dt} &= A\left(\frac{e^{-t}}{1+e^{-t}}\right) - B\left(\frac{1}{1+e^{-t}}\right)\\ &= \left(\frac{1}{1+e^t}\right)\left(A - Be^t\right)\\ &= \frac{A}{1+e^t}\left(1 - \frac{B}{A}e^t\right).\end{aligned}$$

Once again, this is a decreasing function of t, so over the interval $[-1, 1]$, the minimum value is attained at $t = 1$. Further, if $A/B > e$, the derivative is positive.

As before, denote by g^* the value of $g(\boldsymbol{x}_i)$ obtained by setting $t_i = 1$ and $t_i' = -1$ (note that this may not actually be attainable by some $\boldsymbol{x}_i$, for instance, if y_1^* and y_2^* are not antipodal unit vectors). By the observation about the derivatives of $h_{A,B}$, the mean value theorem, and the choice of γ, we have that for every i,

$$g^* - g(\boldsymbol{x}_i) \geq \gamma(1 - t_i) + \gamma(t_i' + 1) = \gamma(2 - \boldsymbol{x}_i^T(\boldsymbol{y}_1^* - \boldsymbol{y}_2^*)).$$

Now we are in exactly the same situation as in Lemma 9 (with γ playing the role of $g'(2)$). Using the same argument (and once again, noticing that OPT achieves the value g^* for every vertex), we have that

$$\textsc{Opt} - \mathcal{L}_{\text{ns}}(\overline{M}; X, Y^*) \geq \frac{\gamma}{2}\left[\sum_{i\in V_1}\|\boldsymbol{x}_i - \boldsymbol{x}_1^*\|^2 + \sum_{i\in V_2}\|\boldsymbol{x}_i - \boldsymbol{x}_2^*\|^2\right].$$

As before, this completes the proof of the Lemma. □