Learning Representations using Spectral-Biased Random Walks on Graphs

Charu Sharma, Jatin Chauhan, Manohar Kaul
Dept. of Computer Science & Engineering
Indian Institute of Technology Hyderabad
Kandi, Sangareddy, India
Email: \{cs16resch11007, cs17btech11019, mkaul\}@iith.ac.in

Abstract—Several state-of-the-art neural graph embedding methods are based on short random walks (stochastic processes) because of their ease of computation, simplicity in capturing complex local graph properties, scalability, and interpretability. In this work, we are interested in studying how much a probabilistic bias in this stochastic process affects the quality of the nodes picked by the process. In particular, our biased walk, with a certain probability, favors movement towards nodes whose neighborhoods bear a structural resemblance to the current node’s neighborhood. We succinctly capture this neighborhood as a probability measure based on the spectrum of the node’s neighborhood subgraph represented as a normalized Laplacian matrix. We propose the use of a paragraph vector model with a novel Wasserstein regularization term. We empirically evaluate our approach against several state-of-the-art node embedding techniques on a wide variety of real-world datasets and demonstrate that our proposed method significantly improves upon existing methods on both link prediction and node classification tasks.

Index Terms—link prediction, node classification, random walks, Wasserstein regularizer

I. INTRODUCTION

Graph embedding methods have gained prominence in a wide variety of tasks including pattern recognition \[1\], low-dimensional embedding \[2, 3\], node classification \[4–6\], and link prediction \[5, 7\], to name a few. In machine learning, the task of producing graph embeddings entails capturing local and global graph statistics and encoding them as vectors that best preserve these statistics in a computationally efficient manner. Among the numerous graph embedding methods, we focus on unsupervised graph embedding models, which can be broadly classified as heuristics and random walk based models.

Heuristic based models compute node similarity scores based on vertex neighborhoods and are further categorized based on the maximum number of \(k\)-hop neighbors they consider around each vertex \[4\]. Recently, Zhang et. al. \[7\] proposed a graph neural network (GNN) based framework that required enclosing subgraphs around each edge in the graph. They showed that most higher order heuristics \((k > 2)\) are a special case of their proposed \(\gamma\)-decaying heuristic. While their method outperforms the heuristic based methods on link prediction, it nevertheless computes all walks of length at most \(k\) \((i.e., the size of the neighborhood)\) around each edge, which is quite prohibitive and results in being able to only process small and sparse graphs.

In comparison, random walk based models are scalable and have been shown to produce good quality embeddings. These methods generate several short random walks originating from each node and then embed a pair of nodes close to one another in feature space, if they co-occur more frequently in several such walks. This is achieved by treating each random walk as a sequence of words appearing in a sentence and feeding this to a word-embedding model like word2vec \[8\]. Deepwalk \[6\] first proposed this approach, after which many works \[4, 9, 10\] followed suit. Recently, WYS \[5\] presented a graph attention (GAT) model that is based on simple random walks and learning a context distribution, which is the probability of encountering a vertex in a variable sized context window, centered around a fixed anchor node. An important appeal of random walks is that they concisely capture the underlying graph structure surrounding a vertex. Yet, further important structure remains uncaptured. For example, heuristic methods rely on the intuition that vertices with similar \(k\)-hop neighborhoods should also be closer in feature space, while simple random walks cannot guarantee the preservation of any such grouping. In WYS, under certain settings of the context window size, vertices with structurally similar neighborhoods can easily be omitted and hence overlooked.

In our work, we incorporate such a grouping of structurally similar nodes directly into our random walks. Our novel methodology opens avenues to a richer class of vertex grouping schemes. To do so, we introduce biased random walks \[11, 12\] that favor, with a certain probability, moves to adjacent vertices with similar \(k\)-hop neighborhoods.

First, we capture the structural information in a vertex’s neighborhood by assigning it a probability measure. This is achieved by initially computing the spectrum of the normalized Laplacian of the \(k\)-hop subgraph surrounding a vertex, followed by assigning a Dirac measure to it. Later, we define a spectral distance between two \(k\)-hop neighborhoods as the \(p\)-th Wasserstein distance between their corresponding probability measures.

Second, we introduce a bias in the random walk, that with a certain probability, chooses the next vertex with least spectral distance to it. This allows our “neighborhood-aware” walks...
to reach nodes of interest much quicker and pack more such nodes in a walk of fixed length. We refer to our biased walks as spectral-biased random walks.

Finally, we learn embeddings for each spectral-biased walk in addition to node embeddings using a paragraph vector model \cite{le2014distributed}, such that each walk which starts at a node considers its own surrounding context within the same walk and does not share context across all the walks, in contrast to a word2vec model \cite{mikolov2013distributed}. Additionally, we also add a Wasserstein regularization term to the the objective function so that node pairs with lower spectral distance co-locate in the final embedding.

Our contributions

1) We propose a spectral-biased random walk that integrates neighborhood structure into the walks and makes each walk more aware of the quality of the nodes it visits.

2) We propose the use of paragraph vectors and a novel Wasserstein regularization term to learn embeddings for the random walks originating from a node and ensure that spectrally similar nodes are closer in the final embedding.

3) We evaluate our method on challenging real-world datasets for tasks such as link prediction and node classification. On many datasets, we significantly outperform our baseline methods. For example, our method outperforms state-of-the-art methods for two difficult datasets Power and Road by a margin of 6.23 and 6.93 in AUC, respectively.

II. RELATED WORK

Recently, several variants have been introduced to learn node embeddings for link prediction. These methods can be broadly classified as (i) heuristic, (ii) matrix factorization, (iii) Weisfeiler-Lehman based, (iv) random walks based, and (v) graph kernel based methods, which learn embeddings from a class of heuristic learning methods. They are Weisfeiler-Lehman Neural Machine (WLNM) \cite{wu2018graph} form an interesting about the graph cuts being useful for classification. However, spectral clustering (SC) \cite{shi2000normalized} also makes a strong assumption for the likelihood of links, which can be beneficial in the case methods perform well but are based on strong assumptions neighborhood of target nodes, respectively. In practice, heuristic information from the

1

first-order, second-order and higher-order methods based on using information from the 1-hop, 2-hop and k-hop (for k > 2) neighborhood of target nodes, respectively. In practice, heuristic methods perform well but are based on strong assumptions for the likelihood of links, which can be beneficial in the case of social networks, but does not generalize well to arbitrary networks.

Similarly, a matrix factorization based approach, i.e., like spectral clustering (SC) \cite{ng2001spectral} also makes a strong assumption about the graph cuts being useful for classification. However, it is unsatisfactory to generalize across diverse networks.

Weisfeiler-Lehman graph kernel (WLK) \cite{shervashidze2009weisfeiler} and Weisfeiler-Lehman Neural Machine (WLNM) \cite{wu2018graph} form an interesting class of heuristic learning methods. They are Weisfeiler-Lehman graph kernel based methods, which learn embeddings from enclosing subgraphs in which the distance between a pair of graphs is defined as a function of the number of common root subtrees between both graphs. These methods have been shown to perform much better than the aforementioned traditional heuristic methods.

Other category of random walks based methods consist of DeepWalk \cite{perozzi2014deepwalk} and Node2Vec \cite{grover2016node2vec}, which have been proven to perform well as it pushes co-occurring nodes in a walk closer to one another in the final node embeddings. Although DeepWalk is a special case of the Node2Vec model, both of these methods produce node embeddings by feeding simple random walks to a word2vec skip-gram model \cite{mikolov2013distributed}.

Finally, for both link prediction and node classification tasks, recent works are mainly graph neural networks (GNNs) based architectures. VGAE \cite{velivckovic2017graph}, WYS \cite{wu2018exploiting}, and SEAL \cite{tang2018seal} are some of the most recent and notable methods that fall under this category. VGAE \cite{velivckovic2017graph} is a variational auto-encoder with a graph convolution network \cite{defferrard2016convolutional} as an encoder. In this, the decoder is defined by a simple inner product computed at the end. It is a node-level GNN to learn node embeddings. While WYS \cite{wu2018exploiting} uses an attention model that learns context distribution on the power series of a transition matrix, SEAL \cite{tang2018seal} uses a graph-level GNN and extracts enclosing subgraphs for each edge in the graph. It learns via a decaying heuristic a mapping function for link prediction. Computing subgraphs for all edges makes it inefficient to process large and dense graphs.

III. SPECTRA OF VERTEX NEIGHBORHOODS

In this section, we describe a spectral neighborhood of an arbitrary vertex in a graph. We start by outlining some background definitions that are relevant to our study. An undirected and unweighted graph is denoted by $G = (V, E)$, where $V$ is a set of vertices and edge-set $E$ represents a set of pairs $(u, v)$, where $u, v \in V$. Additionally, $n$ and $m$ denote the number of vertices and edges in the graph, respectively. In an undirected graph $(u, v) = (v, u)$. Additionally, when edge $(u, v)$ exists, we say that vertices $u$ and $v$ are adjacent, or that $u$ and $v$ are neighbors. The degree $d_v$ of vertex $v$ is the total number of vertices adjacent to $v$. By convention, we disallow self-loops and multiple edges between the same pair of vertices. Given a vertex $v$ and a fixed integer $k > 0$, the graph neighborhood $G(v, k)$ of $v$ is the subgraph induced by the $k$ closest vertices (i.e., in terms of shortest paths on $G$) that are reachable from $v$.

Now, the graph neighborhood $G(v, k)$ of a vertex $v$ is represented in matrix form as a normalized Laplacian matrix $L^{(v)} = (l_{ij})_{i,j=1}^{k} \in \mathbb{R}^{k \times k}$. Given $L^{(v)}$, its sequence of $k$ real eigenvalues $(\lambda_1(L^{(v)}) \geq \cdots \geq \lambda_k(L^{(v)}))$ is known as the spectrum of the neighborhood $L^{(v)}$ and is denoted by $\sigma(L^{(v)})$. We also know that all the eigenvalues in $\sigma(L^{(v)})$ lie in an interval $\Omega \subset \mathbb{R}$. Let $\mu(L^{(v)})$ denote the probability measure on $\Omega$ that is associated to the spectrum $\sigma(L^{(v)})$ and is defined as the Dirac mass concentrated on each eigenvalue in the spectrum. Furthermore, let $P(\Omega)$ denote the set of probability measures on $\Omega$. We now define the $p$-th Wasserstein distance
between measures, which will be used later to define our distance between node neighborhoods.

**Definition 1.** [26] Let \( p \in [1, \infty) \) and let \( c : \Omega \times \Omega \rightarrow [0, +\infty] \) be the cost function between the probability measures \( \mu, \nu \in P(\Omega) \). Then, the \( p \)-th Wasserstein distance between measures \( \mu \) and \( \nu \) is given by the formula

\[
W_p(\mu, \nu) = \left( \inf_{\gamma} \int_{\Omega \times \Omega} c(x, y)^p \, d\gamma \right)^{\frac{1}{p}}
\]

where \( \Pi(\mu, \nu) \) is the set of transport plans, i.e., the set of all joint probabilities defined on \( \Omega \times \Omega \) with marginals \( \mu \) and \( \nu \).

We now define the spectral distance between two vertices \( u \) and \( v \), along with their respective neighborhoods \( L^{(u)} \) and \( L^{(v)} \), as

\[
W^p(u, v) := W_p(\mu_{\sigma(L^{(u)})}, \mu_{\sigma(L^{(v)})})
\]

IV. RANDOM WALKS ON VERTEX NEIGHBORHOODS

A. Simple random walk between vertices

A simple random walk on \( G \) begins with the choice of an initial vertex \( v_0 \) chosen from an initial probability distribution on \( V \) at time \( t_0 \). For each time \( t \geq 0 \), the next vertex to move to is chosen uniformly at random from the current vertex’s 1-hop neighbors. Hence, the probability of transition \( p_{ij} \) from vertex \( i \) to its 1-hop neighbor \( j \) is \( 1/d_i \) and 0 otherwise. This stochastic process is a finite Markov chain and the non-negative matrix \( P = (p_{ij})_{i,j=1}^n \in \mathbb{R}^{n \times n} \) is its corresponding transition matrix. We will focus on ergodic finite Markov chains with a stationary distribution \( \pi^T = (\pi_1, \ldots, \pi_n) \), i.e., \( \pi^T P = \pi^T \) and \( \sum_{i=1}^n \pi_i = 1 \). Let \( \{X_t\} \) denote a Markov chain (random walk) with state space \( V \). Then, the hitting time for a random walk from vertex \( i \) to \( j \) is given by \( T_{ij} = \inf\{t : X_t = j \mid X_0 = i\} \) and the expected hitting time is \( E[T_{ij}] \). In other words, hitting time \( T_{ij} \) is the first time \( j \) is reached from \( i \) in \( \{X_t\} \). By the convergence theorem [27], we know that the transition matrix \( P \) satisfies \( \lim_{n \rightarrow \infty} P^n = P_\infty \), where matrix \( P_\infty \) has all its rows equal to \( \pi \).

B. Spectral-biased random walks

We introduce a bias based on the spectral distance between vertices (as shown in Equation [28]) in our random walks. When moving from a vertex \( v \) to an adjacent vertex \( v' \) in the 1-hop neighborhood \( N(v) \) of vertex \( v \), vertices in \( N(v) \) which are most structurally similar to \( v \) are favored. The most structurally similar vertex to \( v \) is given by

\[
\min_{v' \in N(v)} W^p(v, v')
\]

Then, our spectral-biased walk is a random walk where each of its step is preceded by the following decision. Starting at vertex \( i \), the walk transitions with probability \( 1 - \epsilon \) to an adjacent vertex \( j \) in \( N(v) \) uniformly at random, and with probability \( \epsilon \), the walk transitions to the next vertex with probability \( w_{ij} \), given in the bias matrix, whose detailed construction is explained later. Informally, our walk can be likened to flipping a biased coin with probabilities \( 1 - \epsilon \) and \( \epsilon \), prior to each move, to decide whether to perform a simple random walk or choose one of \( k \) structurally similar nodes from the neighborhood. Thus, our new spectral-biased transition matrix can be written more succinctly as

\[
T = (1 - \epsilon)P + \epsilon W
\]

where \( P \) is the original transition matrix for the simple random walk and \( W \) contains the biased transition probabilities we introduce to move towards a structurally similar vertex.

C. Spectral bias matrix construction

It is well known that the spectral decomposition of a symmetric stochastic matrix produces real eigenvalues in the interval \([-1, 1]\). In order to build a biased transition matrix \( W \) which allows the spectral-biased walk to take control with probability \( \epsilon \) and choose among \( k \) nearest neighboring vertices with respect to the spectral distance between them, we must construct this bias matrix in a special manner. Namely, it should represent a reversible Markov chain, so that it can be “symmetrized”. For brevity, we omit a detailed background necessary to understand symmetric transformations, but we refer the reader to [28]. A Markov chain is said to be reversible [29], when it satisfies the detailed balance condition \( \pi_j p_{ij} = \pi_i p_{ji} \), i.e., on every time interval of the stochastic process the distribution of the process is the same when it is run forward as when it is run backward.

Recall, the 1-hop neighborhood of vertex \( i \) is denoted by \( N(i) \). Additionally, we define \( N_k(i) \subseteq N(i) \) to be the \( k \)-closest vertices in spectral distance to \( i \) among \( N(i) \).

We then define a symmetric \( k \)-closest neighbor set \( S_k(i) \) as a union of all the members of \( N_k(i) \) and those vertices \( j \in N(i) \setminus N_k(i) \), who have vertex \( i \) in \( N_k(j) \). More formally,

\[
S_k(i) := N_k(i) \cup \bigcup_{j \in N(i) \setminus N_k(i)} 1_{N_k(j)}(i)
\]

where the indicator function \( 1_A(x) := 1 \), if \( x \in A \) or 0 if \( x \notin A \).

In accordance to property 7.1.1 in [30], we construct a transition matrix as follows to form a reversible Markov chain which satisfies the detailed balance condition and hence is symmetrizable. Our bias matrix \( W = (w_{ij})_{i,j=1}^n \in \mathbb{R}^{n^2} \) is a stochastic transition probability matrix in \( \mathbb{R}^{n^2} \), whose elements are given by

\[
w_{ij} = \left\{ \begin{array}{ll}
1 - \frac{W^p(i, j)}{\sum_{m \in S_k(i)} W^p(i, m)} & \text{if } i, j \in N_k(i) \\
0 & \text{otherwise}
\end{array} \right.
\]

The rows of the spectral bias matrix \( W \) in Equation [6] are scaled appropriately to convert it into a transition matrix.

D. Time complexity of our spectral walk

Given \( n \) nodes in a graph, we first pre-compute the spectra of every vertex’s neighboring subgraph (represented as a normalized Laplacian). This spectral computation per vertex
includes spectral decomposition of the Laplacian around each vertex, which has a time complexity of $O(k^2n^{2.376})$ (where, $k$ is the size of each vertex neighborhood, typically of $O(10)$, which is very fast to compute) using the Coppersmith and Winograd algorithm for matrix multiplication, which is the most dominant cost in decomposition. This amounts to a total pre-computation time complexity of $O(nk^{2.376})$.

In the worst case, a spectral-biased walk of length $l$ will be biased at each step and hence would compute the spectral distance among its $k$ neighbors at each step (i.e., a total of $kl$ times). The Wasserstein distance between the spectra of the neighborhoods has an empirical time complexity of $O(d^2)$, where $d$ is the order of the histogram of spectra $\sigma(L^{(v)})$. Thus the time-complexity of our online spectral-biased walk is $O(kld^2)$. Although, in practice, we use the Python OT library based on entropic regularized OT, which uses the Sinkhorn algorithm on a GPU and thus computing Wasserstein distances are extremely fast and easy.

E. Empirical analysis of expected hitting time and cover time of spectrally similar vertices

In this section, we empirically study the quality of the random walks produced by our spectral-bias random walk method. In order to accomplish this, we start with a given vertex $v$ and measure the walk quality under two popular quality metrics associated with random walks, namely their expected hitting time and cover time of nodes with structurally similar neighborhoods to that of node $v$. It is important to note here that the consequence of packing more nodes of interest in each random walk, boosts the quality of training samples (i.e., walks setup as sentences) in our neural language model that is described later in Section [V].

**Expected hitting time:** To study the expected hitting times of our spectral-biased and simple random walks, we first randomly sampled 1000 ordered vertex pairs $(s,t)$ with structurally similar neighborhoods, where $s$ and $t$, denoted the start and target vertices, respectively. Next, we considered all the random walks (both spectral-biased and simple) initiated from the start vertex $s$ and ranked the appearance of the target vertex $t$ in a fixed length walk, for both the types of walks. Our ranking results where averaged over all the walks and $(s,t)$ pairs considered. In our experiments on real-world datasets (shown in Figures [1a] and [1c]), we found the target vertex $t$ to appear earlier in our spectral-biased walks, i.e., we had a lower expected hitting time from $s$ to $t$.

Furthermore, we also studied the packing density of spectrally similar nodes in fixed-length walks generated by both the spectral-bias and simple random walk methods. Figures [1b] and [1f] clearly show that our spectral-biased walk packs a higher number of spectrally similar nodes.

**Cover time:** After having empirically studied the spectral-
biased walk’s expected hitting time, it naturally leads to study the \textit{cover time} of our walk, which is the first time when all vertices that are spectrally similar to a start vertex have been visited.

We begin by defining a \textit{Wasserstein ball} around an arbitrary vertex \( v \) that encompasses the set of vertices whose \textit{spectral distance} from \( v \) is less than a constant \( c \).

\textbf{Definition 2.} A Wasserstein ball of radius \( c \) centered at vertex \( v \), denoted by \( B_w(v; c) \), is defined as

\[
B_w(v; c) = \{ u \in V \mid W^p(u, v) \leq c \} \quad (7)
\]

Given a start vertex \( s \), a user-defined fixed constant \( c \), and its surrounding Wasserstein ball \( B_w(s; c) \), we found that our spectral-bias walk covers all spectrally similar vertices in the ball with much \textit{shorter} walks than simple random walks, as is shown in Figures 1c and 1f.

\textbf{V. OUR NEURAL LANGUAGE MODEL WITH WASSERSTEIN REGULARIZATION}

Our approach of learning node embeddings is to use a shallow neural network. This network takes spectral-biased walks as input and predicts either the node labels for node classification or the likelihood of an edge / link between a pair of nodes for the link prediction task.

We leverage the similarity of learning paragraph vectors in a document from NLP to learn our spectral-biased walk embeddings. In order to draw analogies to NLP, we consider a vertex as a word, a walk as a paragraph / sentence, and the entire graph as a document. Two walks are said to \textit{co-occur} when they originate from the same node. Originating from each node \( v \in V \), we generate \( K \) co-occurring spectral-biased random walks \( W^{(v)} = (W^{(v)}_1, \ldots, W^{(v)}_K) \), each of fixed length \( T \). A family of all \( W^{(v)} \) for all \( v \in V \) is analogous to a collection of paragraphs in a document.

In our framework, each vertex \( v \) is mapped to a unique word vector \( w \), represented as a column in a matrix \( W \). Similarly, each biased walk \( w \) is mapped to a unique paragraph vector \( p \) stored as a column in a matrix \( P \). Given a spectral-biased walk as a sequence of words \( w_1, w_2, \ldots, w_T \), our objective is to minimize the following cross-entropy loss

\[
L_{par} = -\frac{1}{T} \sum_{t=c}^{T-c} \log p(w_t | w_{t-c}, \ldots, w_{t+c}) \quad (8)
\]

As shown in [13], the probability is typically given by the softmax function

\[
p(w_t | w_{t-c}, \ldots, w_{t+c}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_{w_i}}} \quad (9)
\]

Each \( y_{w_t} \) is the unnormalized log probability for \( w_t \), given as \( y_{w_t} = b + U h(w_{t-c}, \ldots, w_{t+c}, P, W) \), where \( U, b \) are softmax parameters, and \( h \) is constructed from \( W \) and \( P \). A paragraph vector can be imagined as a word vector that is cognizant of the context information encoded in its surrounding paragraph, while a normal word vector \textit{averages} this information across all paragraphs in the document. For each node \( v \in V \), we apply 1d-convolution to all the paragraphs / walks in \( W^{(v)} \), to get a unique vector \( x_v \).

Our goal is to learn node embeddings which best preserve the underlying graph structure along with clustering structurally similar nodes in feature space. With this goal in mind and inspired by the work of Mu et al. [31] on negative skip-gram sampling with quadratic regularization, we construct the following loss function with a Wasserstein regularization term

\[
L_{ov} = L_{par} + L_{class} + \gamma W^2(\sigma(x_v), \sigma(y_v)) \quad (10)
\]

\textit{2-Wasserstein regularizer}

Here, \( x_v \) is the node embedding learned from the paragraph vector model and \( y_v \) is the 1d-convolution of node \( v \)’s 1-hop neighbor embeddings. \( L_{class} \) represents a task-dependent \textit{classifier loss} which is set to \textit{mean-square error (MSE)} for \textit{link prediction} and \textit{cross-entropy loss} for \textit{node classification}.

We convert the node embedding \( x_v \) and its combined 1-hop neighborhood embedding \( y_v \) into probability distributions via the softmax function, denoted by \( \sigma(s) \) in Equation 10.

Our regularization term is the 2-Wasserstein distance between the two probability distributions, where \( \gamma \) is the regularization parameter. This regularizer penalizes neighboring nodes whose neighborhoods do not bear structural similarity with the neighborhood of the node in question. Finally, the overall loss \( L_{ov} \) is minimized across all nodes in \( G \) to arrive at final node embeddings.

\textbf{VI. EXPERIMENTAL RESULTS}

We conduct exhaustive experiments to evaluate our spectral-biased walk method\textsuperscript{2}. Network datasets were sourced from SNAP and Network Repository. We picked ten datasets for \textit{link prediction} experiments, as can be seen in Table I and three datasets (i.e., Cora, Citeseer, and Pubmed) for \textit{node classification} evaluation. The dataset statistics are outlined in more detail in Section VI-A. We performed experiments by making 90\% - 10\% train-test splits on both positive (existing edges) and negative (non-existent edges) samples from the graphs, following the split ratio outlined in SEAL [7]. We borrow notation from WYS [5] and similarly denote our set of edges for training and testing as \( E_{train} \) and \( E_{test} \), respectively.

\textbf{A. Datasets}

We used ten datasets for link prediction experiments and three datasets for node classification experiments. Datasets for both the experiments are described with their statistics in Table II. Power [32] is the electrical power network of US grid, C.elegans [32] is the neural network of the nematode worm \textit{C.elegans}, USAir [33] is an infrastructure network of US Airlines, Road-Euro and Road-Minnesota [33] are road networks (sparse), Bio-SC-GT [33] is a biological network of WormNet, Infect-hyper [33] is a proximity network, PPI [34] is a network of protein-protein interactions, HepTh is a citation network and Facebook is a social network. Cora, Citeseer and Pubmed datasets for node classification are citation networks of publications [32].

\textsuperscript{2} Our Method
We also analyze sensitivity of hyper-parameters in Figure 2 to walk in Figure 3.

Table II: Datasets for link prediction and node classification tasks.

| Datasets    | Nodes | Edges | Mean Degree | Median Degree |
|-------------|-------|-------|-------------|---------------|
| Power       | 3041  | 6594  | 2.66        | 4             |
| C. elegans  | 297   | 2148  | 14.46       | 24            |
| USAir       | 332   | 2126  | 12.8        | 10            |
| Road-Euro   | 1174  | 1417  | 2.41        | 4             |
| Road-Minnesota | 2642 | 3303  | 2.5         | 4             |
| Bio-SC-GT   | 1716  | 33987 | 39.61       | 41            |
| Infect-hyper| 113   | 2196  | 38.86       | 74            |
| PPI         | 3852  | 37841 | 19.64       | 18            |
| Facebook    | 4039  | 88234 | 43.69       | 50            |
| HepTh       | 8637  | 24805 | 5.74        | 6             |
| Cora        | 2708  | 5278  | 3.89        | 6             |
| Pubmed      | 19717 | 44324 | 4.49        | 4             |
| Citeseer    | 3327  | 4732  | 2.77        | 4             |

B. Training

We now turn our attention to a two-step procedure for training. First, we construct a 2-hop neighborhood around each node for spectra computation. Probability \( p \) is set to 0.6, walk length \( W = 100 \) with 50 walks per node in step one of the spectral-biased walk generation. Second, the context window size \( C = 10 \) and regularization term \( \gamma \) ranges from \( 10^{-6} \) to \( 10^{-8} \) for all the results provided in Table IV. The model for link prediction task to compute final AUC is trained for 100 to 200 epochs depending on the dataset. The dimension of node embeddings is set to 128 for all the cases and a model is learned with a single-layer neural network as a classifier. We also analyze sensitivity of hyper-parameters in Figure 2 to show the robustness of our algorithm. Along with sensitivity, we also discuss how probability \( p \) affects the quality of our walk in Figure 3.

C. Baselines

Our baselines are based on graph kernels (WLK) [22], GNNs (WYS) [5], SEAL [7], VGAE [24], and WLNM [23]) and random walks (Node2Vec) [4]. We use available codes for all the methods and evaluate them by computing the area under curve (AUC). WYS [5] learns context distribution by using an attention model on the power series of a transition matrix. On the other hand, SEAL [7] extracts a local subgraph around each link and learns via a decaying heuristic a mapping function to predict link. VGAE [24] is a graph based variational auto-encoder (VAE) with a graph convolutional network (GCN) [25] as an encoder and simple inner product computed at the decoder side. A graph kernel based approach is the Weisfeiler-Lehman graph kernel (WLK) [22], where the distance between a pair of graphs is defined as a function of the number of common rooted subtrees between both graphs. Weisfeiler-Lehman Neural Machine (WLNM) [23] is neural network training model based on the WLK algorithm. Node2Vec [4] produces node embeddings based on generated simple random walks that are fed to a word2vec skip-gram model for training.

D. Link prediction

This task entails removing links/edges from the graph and then measuring the ability of an embedding algorithm to infer such missing links. We pick an equal number of existing edges (“positive” samples) \( E^{+}_{\text{train}} \) and non-existent edges (“negative” samples) \( E^{-}_{\text{train}} \) from the training split \( E_{\text{train}} \) and similarly pick positive \( E^{+}_{\text{test}} \) and negative \( E^{-}_{\text{test}} \) test samples from the test split \( E_{\text{test}} \). Consequently, we use \( E^{+}_{\text{train}} \cup E^{-}_{\text{train}} \) for training our model selection and use \( E^{+}_{\text{test}} \cup E^{-}_{\text{test}} \) to compute the AUC evaluation metric. We report results averaged over 10 runs along with their standard deviations in Table IV. Our node embeddings based on spectral-biased walks outperform the state of the art methods with significant margins on most of the datasets. Our method better captures not only the adjacent nodes with structural similarity, but also the ones that are farther out, due to our walk’s tendency to bias such nodes, and hence pack more such nodes in the context window.

Among the baselines, we find that SEAL [7] and WYS [5] have comparable results for few datasets such as SEAL performs better on dense than sparse datasets and an argument can’t be generalized for WYS since its performance is better only for few datasets and not to any specific kind of datasets.

E. Sensitivity Analysis

We test sensitivity towards the following three hyper-parameters. Namely, the spectral-biased walk length \( W \), the context window size \( C \), and the regularization parameter \( \gamma \) in our Wasserstein regularizer. We measure the AUC (y-axis) by varying \( W \) and \( C \) over two values each, namely \( \{50, 100\} \) and \( \{5, 10\} \), respectively, spanning across four different values of \( \gamma \) (in x-axis), as shown in Figure 2. We conducted the sensitivity

\[ a^2 + b^2 = c^2 \]

3 WYS 4 SEAL 5 VGAE 6 WLNM 7 Node2Vec
Table III: Node classification results in accuracy (%). Bold indicate best and underline indicate second best results.

| Algorithms   | DeepWalk | Node2Vec | Our Method |
|--------------|----------|----------|------------|
| Citeeseer    | 41.56 ± 0.01 | 42.60 ± 0.01 | **51.8 ± 0.25** |
| Cora         | 66.54 ± 0.01 | 67.90 ± 0.52  | **70.4 ± 0.30** |
| Pubmed       | 69.98 ± 0.12 | 70.30 ± 0.15  | **71.4 ± 0.80** |

analysis on two dense datasets (i.e., Celegans and USAir) and on two sparse datasets (i.e., Power and Road-minnesota).

Our accuracy metrics lie within a range of 2% and are always better than baselines (WYS and Node2vec), i.e., are robust to various settings of hyper-parameters. Furthermore, even with shorter walks (W = 50), our method boasts a stable AUC, indicating that our expected hitting times to structurally similar nodes is quite low in practice.

F. Node Classification

In addition to link prediction, we also demonstrate the efficacy of our node embeddings, via node classification experiments on three citation networks, namely Pubmed, Citeeseer, and Cora. We produce node embeddings from our algorithm and perform classification of nodes without taking node attributes into consideration. We ran experiments on the train-test data splits already provided by [35]. Results are compared against Node2vec and Deepwalk, as other state-of-the-art methods for node classification assumed auxiliary node features during training. Results in Table III show that our method beats the baselines.

G. Effect of Probability, p

In earlier sections of the paper, we showed that our algorithm picks the next node in the walk from nodes with similar neighborhoods, with probability p. Thus, we conducted an experiment to show the effect of p on the final result (AUC) of link prediction. Here, p ranges from 0 to 1, where p = 0 implies that the next node is picked completely at random from the 1-hop neighborhood (as in simple random walk) and p = 1 indicates that every node is picked from the top-k structurally similar nodes in the neighborhood.

As we move towards greater values of p, we tend to select more spectrally similar nodes in the walk. Results are shown in Figure 3 for four datasets Celegans, USAir, Power and Infect-hyper. Figure 3 shows that there is an improvement in AUC for 5% and 2% on an average for dense and sparse datasets respectively. Increase in AUC is recorded when p increases up to a certain value of p ranges between 0.4 to 0.8.

VII. Conclusions

We introduced node embeddings based on spectral-biased random walks, rooted in an awareness of the neighborhood structures surrounding the visited nodes. We further empirically studied the quality of the spectral-biased random walks by comparing their expected hitting time between pairs of spectrally similar nodes, packing density of fixed-sized walks, and the cover time to hit all the spectrally similar nodes within a fixed Wasserstein ball defined by us. We found our spectral-biased walks outperformed simple random walks in all the aforementioned quality parameters.

Motivated by our findings and in an attempt to break away from word vector models, we proposed a paragraph vector model along with a novel Wasserstein regularizer. Experimentally, we showed that our method significantly outperformed existing state-of-the-art node embedding methods on a large and diverse set of graphs, for both link prediction and node classification.

We believe that there does not exist a “one-size-fits-all” graph embedding for all applications and domains. Therefore, our future work will primarily focus on generalizing our biased walks to a broader class of functions that could possibly capture graph properties of interest to the applications at hand.
A. Statistics of spectral-biased walk

In order to understand the importance of our spectral-biased random walk, we perform few more experiments similar to those mentioned in our main paper. We conducted mainly two experiments based on walk length $W$ and probability $p$ for spectral-biased walk (SW) and random walk (RW).

a) Spectrally similar nodes: We compare our walk, SW with RW in terms of covering spectrally similar nodes in walks of varying lengths from 40 to 200. We randomly sample 100 nodes in each of the six datasets, generate both the walks with walk length, $W = \{40, 80, 120, 160, 200\}$ and average it over 100 runs. Both the walks are compared with the percentage of packing spectrally similar nodes in their walks. We computed results of SW for two values of probability $p$. Figure 5 shows the plots for six datasets considering both sparse and dense datasets. Our plots show that our walk, SW packs more number of spectrally similar nodes in its walk even for different values of $p$ than random walk, RW. In addition, our walk with shorter length is still able to outperform RW. We also find that some datasets have large margins between both the walks in covering spectrally similar nodes like Cellegans, Power, etc. However, SW always covers more spectrally similar nodes in the walk than RW for all the datasets.

Algorithm 1 decides which vertex to transition to next

Require: vertex $v$, $A \in \mathbb{R}^{n \times n}$

1: function MOVE_TO($v$, $A$)
2: - Compute transition probability vector $[p(v_1), \ldots, p(v_n)] \leftarrow A[v]$
3: - Divide $[0, 1]$ into intervals
4: where $I_j := F(v_{j-1}), F(v_j)]$ for $j = (1, \ldots n)$
5: - Find $j$ s.t. random number $x \sim [0, 1] \in I_j$
6: return $v_j$

B. Algorithms

In this section, we describe in detail the algorithms used. Algorithm 1 serves as a helper sub-routine to Algorithm 2 to decide the next vertex to move to depending on the transition matrix passed to it as input.

Our algorithm is initiated with the given transition matrices, walk length, parameter $\epsilon$ and an initial vertex $v_0$. In each iteration, the current vertex $v_{curr}$ take a move to the next adjacent vertex $v_{next}$ using one of the transition matrices $W$ and $P$. This decision is based on the value of $x$ taken uniformly at random and compared with the probability parameter $\epsilon$. With probability $\epsilon$, the next adjacent vertex would be picked up using biased transition matrix $W$, and with probability $1 - \epsilon$, next adjacent vertex is taken uniformly at random using transition matrix $P$.

Algorithm 2 Spectral-Biased Random Walk

Require: initial vertex $v_0$, transition matrices $P, W \in \mathbb{R}^{n \times n}$, walk length $k$, parameter $\epsilon$

1: function SPECTRAL_WALK($v_0$, $P, W, k, \epsilon$)
2: $v_{curr} \leftarrow v_0$
3: for $i \leftarrow 1$ to $k - 1$ do
4: $x \sim U[0, 1]$
5: if $x \leq \epsilon$ then
6: $v_{next} \leftarrow \text{move_to}(v_{curr}, W)$
7: else
8: $v_{next} \leftarrow \text{move_to}(v_{curr}, P)$
9: $v_{curr} \leftarrow v_{next}$
10: Walk $w \leftarrow$ append $v_{curr}$ to $w$
11: return $w$

C. Training

The training setup of our method is explained in the main paper where parameters are as follows: probability $p = 0.6$, walk length $W = 100$ with 50 walks per node, context window size $C = 10$, regularization term $\gamma$ ranges from $1e - 6$ to $1e - 8$, node embedding dimension is 128 with 100 to 200 epochs to train the model. We observe that our model is quite fast, including the time for pre-processing step of spectral-biased walk generation. We also analyze sensitivity of hyper-parameters in Figure 2 in main paper and Figure 3 to show the robustness of our algorithm. Along with sensitivity, we also discuss how probability $p$ affects the quality of our walk and finally show the results of our method in Figure 3 in main paper. In order to understand the importance of our spectral-biased walk, we perform few experiments for which results are shown in Figures 5 and 6.

D. Ablative Study

In this section, we illustrate the results of link prediction task for 50% - 50% train-test split in Table IV. Results for 90% - 10% split is provided in Table I of main paper. The observation we can draw here is that with partial data of randomly picking 50% existing (positive) links and 50% non-existent (negative) links for training, we outperform existing methods for majority of the datasets. We can see that dense datasets are not affected by large margins whereas sparse datasets have an effect of partial data in the results. From the baseline methods, SEAL is comparable to our results for almost 6/10 datasets.

While WYS shows more stable AUC results (with no drops) for sparse datasets, it still suffers from a huge standard deviation in its reported AUC values and is hence not as stable. We also observe that WLNM and WLK perform comparably to SEAL and our method, for sparse datasets. But, the drawback of kernel based methods is that they are not efficient in terms of memory and computation time, as they compute pairwise kernel matrix between the subgraphs of all the nodes in the graph.
TABLE IV: Link prediction results with 50% training links. "-" represents incomplete execution due to either out of memory errors or computation time exceeding 20 hours. Bold indicate best and underline indicate second best results.

| Algorithms       | Node2Vec | VGAE     | WLK     | WLNM    | SEAL    | WYS      | Our Method |
|------------------|----------|----------|---------|---------|---------|----------|------------|
| Power            | 52.24 ± 0.31 | 51.26 ± 0.87 | -       | 39.91 ± 0.12 | 59.91 ± 0.12 | 88.14 ± 10.43 | 53.22 ± 0.05 |
| Celegans         | 58.98 ± 0.43 | 76.50 ± 0.79 | 66.58 ± 0.29 | 65.99 ± 0.34 | 82.35 ± 0.41 | 74.67 ± 8.01 | 83.22 ± 0.11 |
| USAir            | 75.21 ± 0.59 | 92.00 ± 0.10 | 83.49 ± 0.39 | 84.39 ± 0.48 | 95.14 ± 0.10 | 93.81 ± 3.65 | 95.31 ± 0.35 |
| Road-Euro        | 51.27 ± 0.98 | 48.53 ± 0.40 | 60.37 ± 0.41 | 63.61 ± 0.52 | 49.29 ± 0.92 | 77.56 ± 20.33 | 52.09 ± 0.15 |
| Road-Minnesota   | 50.94 ± 0.67 | 50.72 ± 1.11 | 68.15 ± 0.44 | 67.78 ± 0.41 | 57.43 ± 0.13 | 76.07 ± 20.12 | 53.65 ± 0.21 |
| Bio-SC-GT        | 82.21 ± 0.54 | 85.64 ± 0.21 | -       | 85.64 ± 0.21 | -       | 90.07 ± 0.03 | 87.84 ± 7.01 | 96.54 ± 0.12 |
| Infect-hyper     | 61.38 ± 0.28 | 76.29 ± 0.20 | 60.99 ± 0.59 | 63.60 ± 0.61 | 75.11 ± 0.05 | 78.32 ± 0.31 | 81.34 ± 0.24 |
| PPI              | 60.14 ± 0.18 | 88.60 ± 0.11 | -       | -       | 91.47 ± 0.04 | 85.01 ± 0.32 | 90.38 ± 0.18 |
| Facebook         | 95.93 ± 0.11 | -       | -       | -       | 80.11 ± 0.23 | 98.78 ± 0.10 | 99.13 ± 0.01 |
| HepTh            | 75.21 ± 0.21 | 82.60 ± 0.25 | -       | -       | 90.47 ± 0.07 | 93.45 ± 3.67 | 90.13 ± 0.03 |

E. Sensitivity Analysis

For sensitivity, we run our algorithm for three different hyper-parameters in order to check robustness of our method. These hyper-parameters are spectral-biased walk length $W$, the context window size $C$ and the regularization term $\gamma$ for our Wasserstein regularizer. We perform experiments on eight datasets for link prediction, out of which results for four datasets are mentioned in our main paper. Figure 4 shows the results for four datasets with $C = 5, 10, W = 50, 100$ and regularization term on $x$-axis. Here, Road is a sparse dataset and other three datasets (Infect-Hyper, Bio-SC-GT and PPI) are dense datasets. We see that our results lie within the range of 2% and are stable for different hyper-parameters. We observe that our results are always above the baseline methods (WYS and Node2Vec).

a) Ranking of target nodes: We also studied the hitting time of target node $j$ to appear in the walk when a walk is initiated from a node $i$ for both the walks SW and RW. For this experiment, we pick 100 unseen pair of nodes $(i, j)$. Both the walks are generated starting from node $i$ for varying length from 40 to 200, considering two values of probability $p$ for our walk. We compute the number of times $j$ appears early in $i$'s walk for SW and RW and we average the results over 100 runs. We conducted this experiment for six datasets covering both sparse and dense datasets. The results are shown in Figure 6. The analysis from our results represents the lower hitting time for our walk (i.e., SW) in ranking of a target node appearance in a walk starting from a node $i$ on all the datasets. We found that the percentage of target nodes appears early in SW is more than in RW.
Fig. 4: Sensitivity of window size, $C$ and walk length, $W$ with respect to regularization term, $\gamma$ is measured in AUC for four datasets for link prediction task.
Fig. 5: Percentage of spectrally-similar nodes packed in walks of varying length for random walk (RW) and spectral-biased walk (SW) for six datasets.
Fig. 6: Average ranking of target nodes encountered by simple random walk (RW) and our spectral-biased walk (SW) of varying length for six datasets.