

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 vertex¹. 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 γ -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

¹ “vertex” and “node” will be used interchangeably.

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 [13], 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 wordvec model [8]. 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 neural network (GNN) methods.

Common neighbors (CN), Adamic-adar (AA) [14], PageRank [15], SimRank [16], resource allocation (RA) [17], preferential attachment (PA) [18], Katz and resistance distance are some popular examples of heuristic methods. These methods compute a heuristic similarity measure between nodes to predict if they are likely to have a link [19] [20] between them or not. Heuristic methods can be further categorized into *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) [21] 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) [22] and Weisfeiler-Lehman Neural Machine (WLNLM) [23] 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 rooted 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 [6] and Node2Vec [4], 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 [8].

Finally, for both link prediction and node classification tasks, recent works are mainly graph neural networks (GNNs) based architectures. VGAE [24], WYS [5], and SEAL [7] are some of the most recent and notable methods that fall under this category. VGAE [24] is a variational auto-encoder with a graph convolution network [25] 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 [5] uses an attention model that learns context distribution on the power series of a transition matrix, SEAL [7] 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* connecting the same pair of vertices. Given a vertex v and a fixed integer $k > 0$, the *graph neighborhood* $\mathcal{G}(v, k)$ 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 $\mathcal{G}(v, k)$ of a vertex v is represented in matrix form as a *normalized Laplacian matrix* $\mathbf{L}^{(v)} = (l_{ij})_{i,j=1}^k \in \mathbb{R}^{k^2}$. Given $\mathbf{L}^{(v)}$, its sequence of k real eigenvalues $(\lambda_1(\mathbf{L}^{(v)}) \geq \dots \geq \lambda_k(\mathbf{L}^{(v)}))$ is known as the *spectrum* of the neighborhood $\mathbf{L}^{(v)}$ and is denoted by $\sigma(\mathbf{L}^{(v)})$. We also know that all the eigenvalues in $\sigma(\mathbf{L}^{(v)})$ lie in an interval $\Omega \subset \mathbb{R}$. Let $\mu_{\sigma(\mathbf{L}^{(v)})}$ denote the probability measure on Ω that is associated to the spectrum $\sigma(\mathbf{L}^{(v)})$ and is defined as the Dirac mass concentrated on each eigenvalue in the spectrum. Furthermore, let $\mathcal{P}(\Omega)$ denote the set of probability measures on Ω . 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 $\mu, \nu \in \mathcal{P}(\Omega)$. Then, the p -th Wasserstein distance between measures μ and ν is given by the formula

$$W_p(\mu, \nu) = \left(\inf_{\gamma} \int_{\Omega \times \Omega} c(x, y)^p d\gamma \mid \gamma \in \Pi(\mu, \nu) \right)^{\frac{1}{p}} \quad (1)$$

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 μ and ν .

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

$$W^P(u, v) := W_p(\mu_{\sigma(\mathbf{L}^{(u)})}, \mu_{\sigma(\mathbf{L}^{(v)})}) \quad (2)$$

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* $\pi^T = (\pi_1, \dots, \pi_n)$, 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 $\mathbb{E}[T_{ij}]$. 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_∞ has all its rows equal to π .

B. Spectral-biased random walks

We introduce a *bias* based on the *spectral distance* between vertices (as shown in Equation 2) in our random walks. When moving from a vertex v to an adjacent vertex v' in the 1-hop neighborhood $\mathbf{N}(v)$ of vertex v , vertices in $\mathbf{N}(v)$ which are most *structurally similar* to v are *avored*. The most structurally similar vertex to v is given by

$$\min_{v' \in \mathbf{N}(v)} W^P(v, v') \quad (3)$$

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 $\mathbf{N}(v)$ uniformly at random, and with probability ϵ , 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 ϵ , 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

$$T = (1 - \epsilon)P + \epsilon W \quad (4)$$

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 ϵ 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* $\pi_i p_{ij} = \pi_j p_{ji}$, 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 $\mathbf{N}(i)$. Additionally, we define $\mathbf{N}_k(i) \subseteq \mathbf{N}(i)$ to be the k -*closest* vertices in spectral distance to i among $\mathbf{N}(i)$.

We then define a *symmetric k -closest neighbor set* $\mathbf{S}_k(i)$ as a union of all the members of $\mathbf{N}_k(i)$ and those vertices $j \in \mathbf{N}(i) \setminus \mathbf{N}_k(i)$, who have vertex i in $\mathbf{N}_k(j)$. More formally,

$$\mathbf{S}_k(i) := \mathbf{N}_k(i) \cup \left\{ \bigcup_{j \in \mathbf{N}(i) \setminus \mathbf{N}_k(i)} \mathbb{1}_{\mathbf{N}_k(j)}(i) \right\} \quad (5)$$

where the indicator function $\mathbb{1}_A(x) := 1$, 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 $W = (w_{ij})_{i,j=1}^n$ is a stochastic transition probability matrix in \mathbb{R}^{n^2} , whose elements are given by

$$w_{ij} = \left\{ 1 - \frac{W^P(i, j)}{\sum_{m \in \mathbf{S}_k(i)} W^P(i, m)} \right\} \quad (6)$$

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

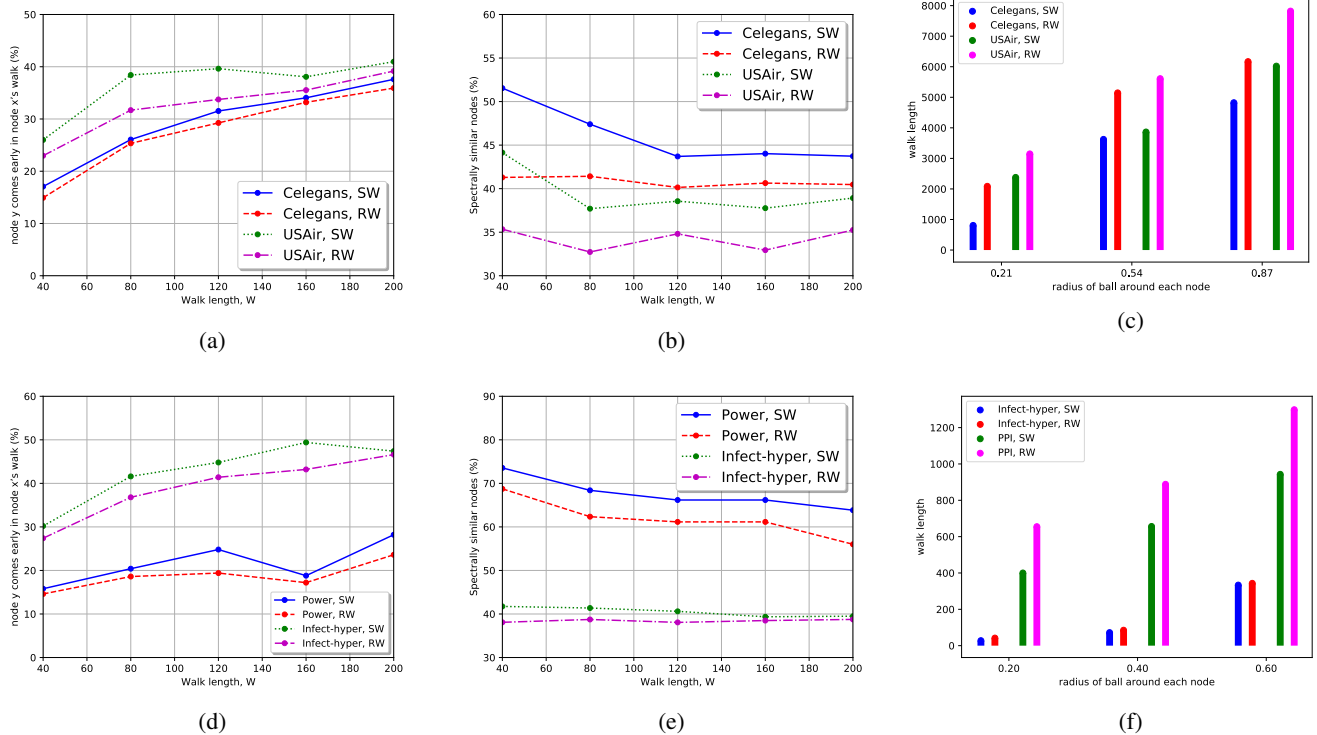


Fig. 1: (a) and (d) Average ranking of target nodes encountered by simple random walk (RW) and our spectral-biased walk (SW). (b) and (e) Percentage of spectrally-similar nodes packed in walks of varying length for RW versus. SW. (c) and (f) Walk lengths to cover entire ball of vertices on Celegans, USAir, Infect-hyper and PPI for our spectral biased walk (SW) and simple random walk (RW).

includes spectral decomposition of the Laplacian around each vertex, which has a time complexity of $O(k^2)$ (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 $O(nk^2)$.

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 $\sigma(\mathbf{L}^{(v)})$. 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 were averaged over all the walks and (s, t) pairs considered. In our experiments on real-world datasets (shown in Figures 1a and 1d), 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 1e, 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 *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 *Wassertein ball* around an arbitrary vertex v that encompasses the set of vertices whose *spectral distance* from v is less than a constant c .

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 *shorter* walks than simple random walks, as is shown in Figures 1c and 1f.

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 *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 $\mathbf{W}^{(v)} = (\mathbf{W}_1^{(v)}, \dots, \mathbf{W}_K^{(v)})$, each of fixed length T . A family of all $\mathbf{W}^{(v)}$ 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, \dots, w_T , our objective is to minimize the following cross-entropy loss

$$L_{par} = -\frac{1}{T} \sum_{t=c}^{T-c} \log p(w_t \mid w_{t-c}, \dots, w_{t+c}) \quad (8)$$

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

$$p(w_t \mid w_{t-c}, \dots, w_{t+c}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_{w_i}}} \quad (9)$$

Each y_{w_i} is the unnormalized log probability for w_i , given as $y_{w_t} = b + Uh(w_{t-c}, \dots, w_{t+c}; P, W)$, 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 *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 $\mathbf{W}^{(v)}$, 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

$$L_{ov} = L_{par} + L_{class} + \underbrace{\gamma W^2(\sigma^{(s)}(x_v), \sigma^{(s)}(y_v))}_{2\text{-Wasserstein regularizer}} \quad (10)$$

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 *classifier loss* which is set to *mean-square error* (MSE) for *link prediction* and *cross-entropy loss* for *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 γ 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.

VI. EXPERIMENTAL RESULTS

We conduct exhaustive experiments to evaluate our spectral-biased walk method². Network datasets were sourced from **SNAP** and **Network Repository**. We picked ten datasets for *link prediction* experiments, as can be seen in Table I, and three datasets (i.e., Cora, Citeseer, and Pubmed) for *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.

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, Celegans [32] is the neural network of the nematode worm *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].

² Our Method

TABLE I: Link prediction results (AUC). ”-” for incomplete execution due to either *out of memory errors* or runtime exceeding 20 hours. Bold indicate best and underline indicate second best results.

Algorithms	Node2Vec	VGAE	WLK	WLNLM	SEAL	WYS	Our Method
Power	78.37 ± 0.23	77.77 ± 0.95	-	-	74.69 ± 0.21	89.37 ± 0.21	95.60 ± 0.25
Celegans	69.85 ± 0.89	74.16 ± 0.78	73.27 ± 0.41	70.64 ± 0.57	85.53 ± 0.15	74.97 ± 0.19	87.36 ± 0.10
USAir	84.90 ± 0.41	93.18 ± 1.46	87.98 ± 0.71	87.01 ± 0.42	<u>96.9 ± 0.37</u>	94.01 ± 0.23	97.40 ± 0.21
Road-Euro	50.35 ± 1.05	68.94 ± 5.23	61.17 ± 0.28	65.95 ± 0.33	60.89 ± 0.22	80.42 ± 0.11	87.35 ± 0.33
Road-Minnesota	67.12 ± 0.63	67.36 ± 2.33	75.15 ± 0.16	74.91 ± 0.19	86.92 ± 0.52	75.33 ± 2.77	91.16 ± 0.15
Bio-SC-GT	88.39 ± 0.79	86.76 ± 1.41	-	-	97.26 ± 0.13	87.72 ± 0.47	<u>97.16 ± 0.32</u>
Infect-hyper	66.66 ± 0.51	80.89 ± 0.21	65.39 ± 0.39	67.68 ± 0.41	<u>81.94 ± 0.11</u>	78.42 ± 0.15	85.25 ± 0.24
PPI	71.51 ± 0.09	<u>88.19 ± 0.11</u>	-	-	-	84.12 ± 1.27	91.16 ± 0.30
Facebook	96.33 ± 0.11	-	-	-	-	98.71 ± 0.14	99.14 ± 0.05
HepTh	88.18 ± 0.21	90.78 ± 1.15	-	-	97.85 ± 0.39	93.63 ± 2.36	<u>97.40 ± 0.25</u>

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

Datasets	Nodes	Edges	Mean Degree	Median Degree
Power	4941	6594	2.66	4
Celegans	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 γ ranges from $1e - 6$ to $1e - 8$ for all the results provided in Table I. 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 WLNLM [23]) and random walks (Node2Vec [4]). We use available codes for all the methods and evaluate the methods by computing the *area under curve* (AUC). WYS [5] learns context distribution by using an attention model on the power series of a transition matrix³. On the other hand, SEAL [7] extracts a local subgraph around each link and learns via a *decaying heuristic* a mapping variation to predict links⁴. 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 side⁵. 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* (WLNLM) [23] is neural network training model based on the WLK algorithm⁶. Node2Vec [4] produces node embeddings based on generated simple random walks that are fed to a word2vec skip-gram model for training⁷.

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_{train}^+ and *non-existent edges* (“negative” samples) E_{train}^- from the training split E_{train} and similarly pick positive E_{test}^+ and negative E_{test}^- test samples from the test split E_{test} . Consequently, we use $E_{train}^+ \cup E_{train}^-$ for training our model selection and use $E_{test}^+ \cup E_{test}^-$ to compute the AUC evaluation metric. We report results averaged over 10 runs along with their standard deviations in Table I. 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 γ 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 γ (in x-axis), as shown in Figure 2. We conducted the sensitivity

³ WYS ⁴ SEAL

⁵ VGAE ⁶ WLNLM ⁷ Node2Vec

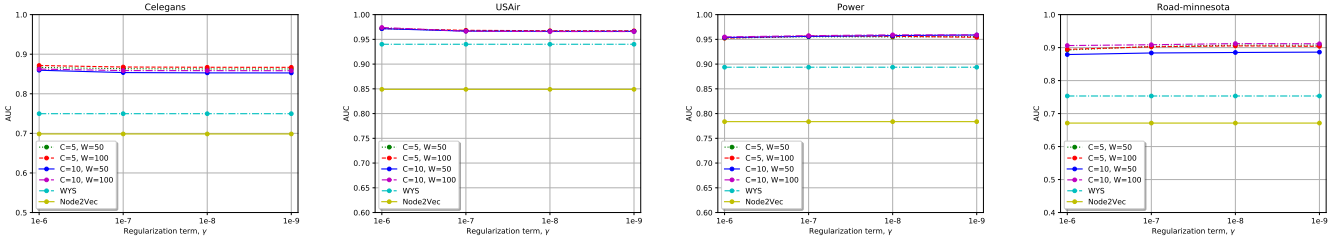


Fig. 2: Sensitivity of window size, C and walk length, W with respect to regularization term, γ is measured in AUC for four datasets of link prediction.

TABLE III: Node classification results in accuracy (%). Bold indicate best and underline indicate second best results.

Algorithms	DeepWalk	Node2Vec	Our Method
Citeseer	41.56 \pm 0.01	42.60 \pm 0.01	51.8 \pm 0.25
Cora	66.54 \pm 0.01	67.90 \pm 0.52	70.4 \pm 0.30
Pubmed	69.98 \pm 0.12	70.30 \pm 0.15	71.4 \pm 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, Citeseer, 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

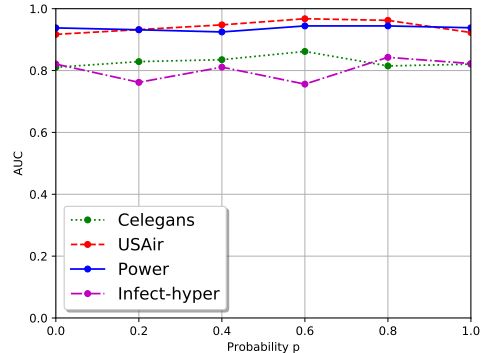


Fig. 3: Effect of p on AUC for four datasets for link prediction task.

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.

REFERENCES

- [1] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model cnns," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 5425–5434.
- [2] M. Belkin and P. 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*, ser. NIPS'01, 2001, pp. 585–591.
- [3] M. Brand and K. Huang, "A unifying theorem for spectral embedding and clustering," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, AISTATS 2003, Key West, Florida, USA, January 3-6, 2003*, 2003.
- [4] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.
- [5] S. Abu-El-Hajja, B. Perozzi, R. Al-Rfou, and A. A. Alemi, "Watch your step: Learning node embeddings via graph attention," in *Advances in Neural Information Processing Systems*, 2018, pp. 9180–9190.
- [6] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.
- [7] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 5165–5175.
- [8] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: <http://arxiv.org/abs/1301.3781>
- [9] L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, "Struc2vec: Learning node representations from structural identity," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, 2017, pp. 385–394.
- [10] H. Chen, B. Perozzi, Y. Hu, and S. Skiena, "HARP: hierarchical representation learning for networks," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 2127–2134.
- [11] M. Bonaventura, V. Nicosia, and V. Latora, "Characteristic times of biased random walks on complex networks," *Phys. Rev. E*, vol. 89, 2014.
- [12] Y. Azar, A. Z. Broder, A. R. Karlin, N. Linial, and S. J. Phillips, "Biased random walks," *Combinatorica*, vol. 16, no. 1, pp. 1–18, 1996.
- [13] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14, 2014, pp. II–1188–II–1196.
- [14] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.
- [15] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer networks*, vol. 56, no. 18, pp. 3825–3833, 2012.
- [16] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 538–543.
- [17] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.
- [18] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [19] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.
- [20] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: statistical mechanics and its applications*, vol. 390, no. 6, pp. 1150–1170, 2011.
- [21] L. Tang and H. Liu, "Leveraging social media networks for classification," *Data Mining and Knowledge Discovery*, vol. 23, no. 3, pp. 447–478, 2011.
- [22] N. Shervashidze, P. Schweitzer, E. J. v. Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels," *Journal of Machine Learning Research*, vol. 12, no. Sep, pp. 2539–2561, 2011.
- [23] M. Zhang and Y. Chen, "Weisfeiler-lehman neural machine for link prediction," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 575–583.
- [24] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *NIPS Workshop on Bayesian Deep Learning*, 2016.
- [25] —, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.
- [26] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.
- [27] D. Aldous and J. A. Fill, *Reversible Markov Chains and Random Walks on Graphs*, 2002. [Online]. Available: <http://www.stat.berkeley.edu/~aldous/RWG/book.html>
- [28] J. R. Norris, *Markov chains.*, 1998.
- [29] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Mathematical Society, 2006.
- [30] Q. Jiang, "Construction of transition matrices of reversible markov chains," Ph.D. dissertation, Windsor, Ontario, Canada, 2009.
- [31] C. Mu, G. Yang, and Z. Yan, "Revisiting skip-gram negative sampling model with rectification," 2018.
- [32] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.
- [33] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. [Online]. Available: <http://networkrepository.com>
- [34] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D535–D539, 2006.
- [35] Z. Yang, W. Cohen, and R. Salakhudinov, "Revisiting semi-supervised learning with graph embeddings," in *International Conference on Machine Learning*, 2016, pp. 40–48.