A FLEXIBLE PAGERANK-BASED GRAPH EMBEDDING FRAMEWORK CLOSELY RELATED TO SPECTRAL EIGENVECTOR EMBEDDINGS

Disha Shur · Yufan Huang · David F. Gleich

July 21, 2022

We study a simple embedding technique based on a matrix of personalized PageRank vectors seeded on a random set of nodes. We show that the embedding produced by the element-wise logarithm of this matrix (1) are related to the spectral embedding for a class of graphs where spectral embeddings are significant, and hence useful representation of the data, (2) can be done for the entire network or a smaller part of it, which enables precise local representation, and (3) uses a relatively small number of PageRank vectors compared to the size of the networks. Most importantly, the general nature of this embedding strategy opens up many emerging applications, where eigenvector and spectral techniques may not be well established, to the PageRank-based relatives. For instance, similar techniques can be used on PageRank vectors from hypergraphs to get "spectral-like" embeddings.

1 INTRODUCTION

The eigenvectors of the graph Laplacian are among the most widely used algorithmic measures of a graph. They are used to find cuts and clusters in a variety of settings [Shi and Malik, 2000; Chung, 1992; Pothen et al., 1990]. They give a signal basis for a graph [Hammond et al., 2011; Donnat et al., 2018]. And one of their original uses was to draw informative pictures of graphs in a low dimensional space [Hall, 1970; Koren, 2003]. These are all related to the idea of embedding the graph into a low dimensional space and recent uses have closely studied this embedding framework.

Likewise, PageRank is itself a widely used algorithmic measure on a graph [Brin and Page, 1998]. The uses are extremely diverse [Gleich, 2015]. Relationships between PageRank and spectral clustering are also known [Andersen et al., 2006; Mahoney et al., 2012; Kloster and Gleich, 2014]. These exist because both techniques can be related to random walks, and seeded PageRank is a localized type of random walk, or random walk with restart [Tong et al., 2006].

In this manuscript, we study a particular type of relationship between a matrix of seeded PageRank vectors and the eigenvectors of the Laplacian matrix. Our log-PageRank embedding uses the singular vectors of the elementwise log of a random collection of seeded PageRank vectors. An example is in Figure 1. This shows that log-PageRank embeddings resemble spectral clustering.

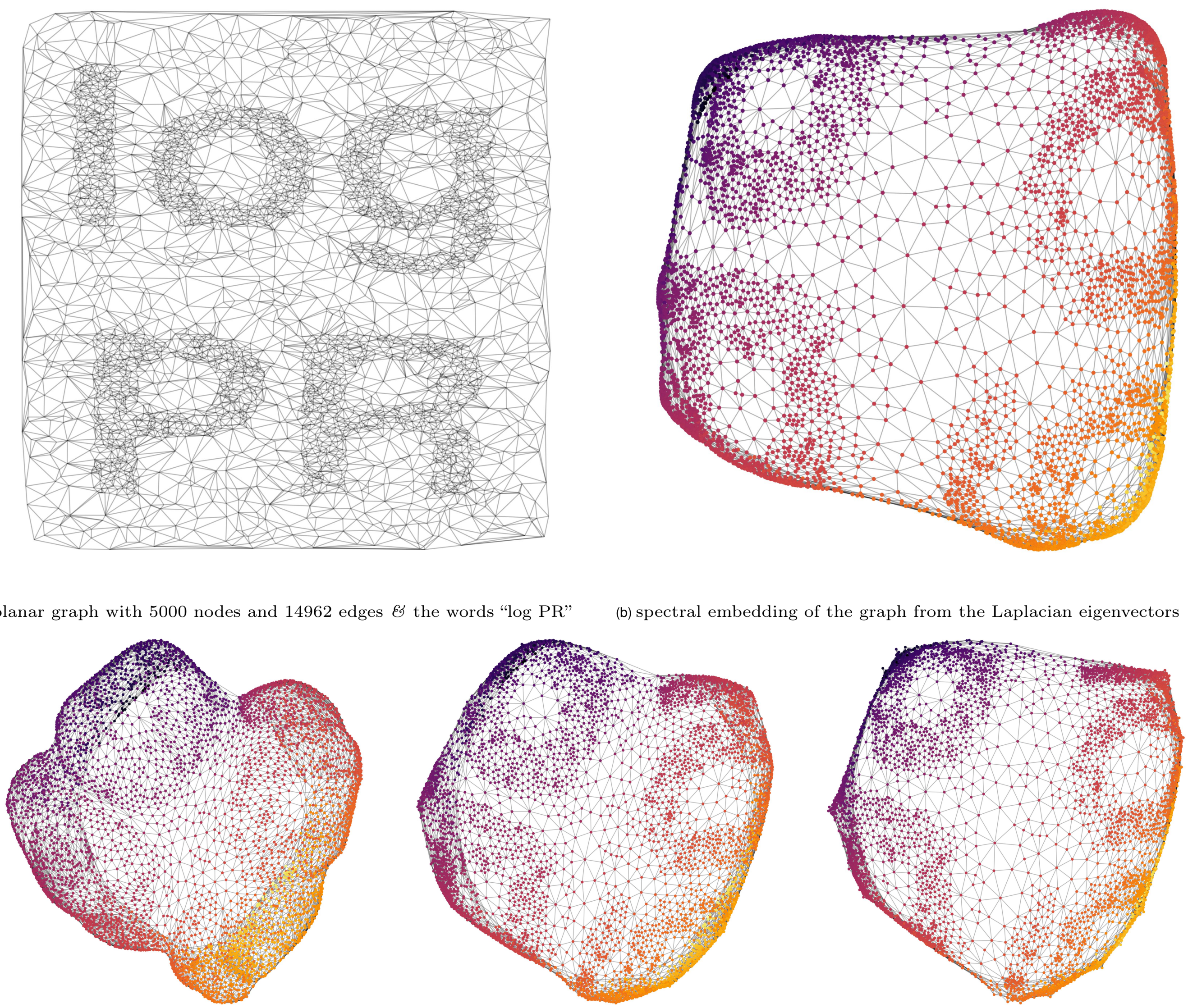
Our manuscript shows that this relationship is expected for degree-regular graphs through an approximation argument (Section 5). This builds from a study on chain graphs that closely characterizes the log-PageRank values (Section 4).

PageRank vectors have long been viewed in terms of log-scaling. When Google published PageRank scores for websites, they were understood to represent an approximation of the log of Google's internal metrics [Bar-Yossef and Mashiach, 2008]. When PageRank was used in spam analysis, log scaling was used by Becchetti et al. [2008, Section 6.4]. So a log-scaled analysis is not surprising.

That said, analyzing the singular vectors of personalized PageRank vectors under log-scaling presents interesting challenges from a technical perspective. As

All authors
Purdue University Computer Science
{dshur,huan1754,dgleich}@purdue.edu

Both Shur and Huang contributed equally. Research supported in part by NSF CCF-1909528, IIS-2007481, NSF Center for Science of Information STC, CCF-0939370.



(a) planar graph with 5000 nodes and 14962 edges & the words "log PR" (b) spectral embedding of the graph from the Laplacian eigenvectors

(c) log-PageRank embedding for $\alpha = 0.85$ (d) log-PageRank embedding for $\alpha = 0.99$ (e) log-PageRank embedding for $\alpha = 0.999$.

FIGURE 1 – The embedding pictures have all the nodes colored with the same values to show relative position. The log-PageRank embedding uses singular vectors of the element-wise logarithm of seeded PageRank vectors. Our paper argues that the similarity between (b), (d) and (e) is expected through an approximation analysis. The advantage the log-PageRank embeddings is that they can be deployed in many emerging data scenarios where spectral embeddings and eigenvectors are not as well established or may be computationally expensive but where analogues of random walks or PageRank may be possible, as in hypergraphs (see Figure 9b).

such, we are only able to get an approximate, although compelling, understanding of the relationship illustrated in the figure.

These log-PageRank embeddings offer a different set of computational tradeoffs compared with eigenvectors. First, they only require random samples of a diffusion process on the graph. Indeed, a closely related methodology to these log-PageRank embeddings was previously used in Fountoulakis et al. [2020] to compare spectral clustering with alternatives. This shows how the ideas behind the log-PageRank embeddings give more flexible structures to help users study their datasets. For instance, it is easy to study a variety of localized log-PageRank embeddings that only work in a subset of a graph. Yet, these can also be designed to pull in other nearby regions as suggested by the PageRank vectors instead of more brittle Dirichlet eigenvector approximations [Chung et al., 2011]. In this paper, we briefly explore using log-PageRank embeddings on hypergraphs to visualize their structure as well.

Moreover, the idea of customizing embeddings is highly relevant to the ongoing use of graph embeddings for ML algorithms. Embedding development is the primary task in problems pertaining to network analysis, language processing, image processing or any problem that seeks to understand and use data. Literature on embedding is filled with learning models based on functions of spectral entities [Yadati et al., 2019; Perozzi et al., 2014; Grover and Leskovec, 2016; Tudisco et al., 2021b; Wang and Leskovec, 2020; Huang et al., 2021]. Like our log-scaling, these embeddings often involve nonlinearities such as sigmoid.

PageRank or diffusion based techniques have previously been used for learning graph embedding (or clustering) [Donnat et al., 2018; Klicpera et al., 2019; Yang et al., 2020; Takai et al., 2020; Liu et al., 2021; Carletti et al., 2020] where the personalized PageRank vector based on a set of nodes, called the **seed set** is used for further computations. Interestingly, although each methodology comes with its own set of merits, all of these methods boil down to a function of random walk on the graph developed from the available data. For example, Chung [2007] and Sahai et al. [2011] and Donnat et al. [2018] and Tsitsulin et al. [2018] are based on different functions of the random walk matrix. PageRank too can be expressed as an infinite geometric sum of the random walk matrix [Chung, 2007].

In summary, the contributions and remainder of this paper discuss:

- the log-PageRank embedding framework (Section 3)
- a study of log-PageRank values on a chain graph that shows how log-PageRank values are related to graph distance (Section 4)
- an approximation analysis between log-PageRank embeddings and spectral clustering on d-regular graphs (Section 5)
- a computational study of similarities and differences between spectral and log-PageRank embeddings (Section 6)
- examples of log-PageRank embeddings in hypergraphs using hypergraph PageRank [Liu et al., 2021] (Section 7).

2 PRELIMINARIES

In this manuscript we consider a connected weighted or unweighted undirected graph $G = (V, E)$ where V is the vertex set with n vertices and E is the edge set with m edges. Let $\boldsymbol{A}$ and $\boldsymbol{D}$ denote the adjacency matrix and degree matrix of a graph G correspondingly. The Laplacian matrix $\boldsymbol{L}$ of a graph G is $\boldsymbol{D} - \boldsymbol{A}$ and the normalized Laplacian matrix $\boldsymbol{\mathcal{L}}$ is $\boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$. Let $\boldsymbol{W}$ denote the lazy random walk $\frac{\boldsymbol{I}+\boldsymbol{A}\boldsymbol{D}^{-1}}{2}$. For a column-stochastic matrix $\boldsymbol{P}$, a stationary distribution $\boldsymbol{\pi}$ is any solution to the eigensystem $\boldsymbol{P}\boldsymbol{\pi} = \boldsymbol{\pi}$ where $\boldsymbol{\pi}$ is non-negative and sums to 1. This is an eigenvector of $\boldsymbol{P}$ corresponding to the eigenvalue 1. The stationary distribution of $\boldsymbol{P}$ is unique if the underlying graph is connected.

We use subscript to index entries of a matrix or a vector: let $\boldsymbol{A}_i$ denote the ith column of matrix $\boldsymbol{A}$, $\boldsymbol{A}_{ij}$ denote the (i, j)th entry of matrix $\boldsymbol{A}$ and $\boldsymbol{A}_{i:j}$ denote the matrix of columns $\boldsymbol{A}_i, \ldots, \boldsymbol{A}_j$; let $\mathbf{x}_i$ denote the ith entry of vector $\mathbf{x}$ and $\mathbf{x}_{i:j}$ denote the vector of entries $\mathbf{x}_i, \ldots, \mathbf{x}_j$. Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ denote the columns of the identity matrix and the n standard basis vectors of $\mathbb{R}^n$ and $\mathbf{e}$ be the all-ones vector. We use $\log.$ to denote the element-wise log operator applied to a vector.

PageRank The classical PageRank problem is defined as follows

DEFINITION 1 (see for example Gleich [2015]) *Let $\boldsymbol{P}$ be a column-stochastic matrix and $\mathbf{v}$ be a column-stochastic vector, then PageRank problem is to find the solution $\mathbf{x}$ to the linear system*

$$(\boldsymbol{I} - \alpha\boldsymbol{P})\mathbf{x} = (1 - \alpha)\mathbf{v} \tag{1}$$

where the solution $\mathbf{x}$ is called the PageRank vector, $\alpha \in (0, 1)$ is the teleportation parameter and $\mathbf{v}$ is the teleportation distribution over all vertices.

By the definition above and the fact that all eigenvalues of a column-stochastic matrix have magnitude at most 1, we have $\boldsymbol{I} - \alpha \boldsymbol{P}$ is non-singular and the PageRank vector can be written as $\mathbf{x} = (1-\alpha)(\boldsymbol{I} - \alpha\boldsymbol{P})^{-1}\mathbf{v}$. When the teleportation distribution $\mathbf{v}$ has support size 1, the PageRank problem is also called personalized PageRank problem and the corresponding solution $\mathbf{x}$ is personalized PageRank vector or a seeded PageRank vector. For convenience, let $\boldsymbol{X}(\alpha)$ denote $(1-\alpha)(\boldsymbol{I} - \alpha\boldsymbol{P})^{-1}$, $\mathbf{x}(u,\alpha)$ denote the personalized PageRank vector seeded on vertex u, i.e. $\mathbf{x}(u,\alpha) = \boldsymbol{X}(\alpha)\mathbf{e}_u = (1-\alpha)(\boldsymbol{I} - \alpha\boldsymbol{P})^{-1}\mathbf{e}_u$.

3 LOG-PAGERANK EMBEDDING

The authors of Fountoulakis et al. [2020] used the linearity of PageRank and the relationship with an expectation to study spectral-like embeddings of nonlinear operators.

This inspiration led to our study of the log-PageRank embedding detailed in Algorithm 1. It takes as input the graph $G = (V, E)$ and outputs the k-dimensional node embeddings. Our technique offers freedom in the algorithm being used for calculation of PageRank vector.

We randomly sample nodes of the graph, compute personalized PageRank vectors, and then compute an elementwise log of the resulting vectors. Then we compute an SVD of the overall set of vectors. The non-dominant vectors give us our log-PageRank embedding. Note that a personalized PageRank vector has mathematically non-negative entries for a connected graph, so computing the log is always mathematically well defined. However, numerically, some of the elements may be sufficiently close to zero to cause an algorithm to return a floating point zero. For this reason, we often replace any zero entries with a value smaller than the smallest non-zero element returned before taking the log. This only occurs for small values of α and tends not to happen once α is close enough to one.

Parameters The user chosen parameters in this technique are the dimension of embedding, k, the teleportation parameter, α, and the number of samples s. The dimension is entirely at a user's discretion. For the number of samples, we suggest a result using a simple coupon collector-like bound that would be common in randomized matrix computations. For the teleportation parameter, we suggest use $\alpha > 0.9$, such as $\alpha = 0.99$ or $\alpha = 0.999$. Because we use many PageRank computations with large values of α, we find it pragmatic to compute a single sparse LU decomposition of the matrix $\boldsymbol{I} - \alpha\boldsymbol{P}$ to repeatedly solve systems. Apart from the PageRank computation, the runtime depends on the SVD of the PageRank matrix, for which any type of randomized SVD computation could be used to make it more efficient.

Intuition and Analysis The idea behind the algorithm is that the matrix of samples should have substantial information from other eigenspaces beyond the dominant one and the SVD will return this information. Our study of this algorithm revealed that the log is essential to getting qualitatively *similar* pictures such as those in Figure 1. We show in Section 5 that as α approaches 1, the log-PageRank embedding approximates the eigenvectors of the lazy random walk matrix $\boldsymbol{W}$. We illustrate a simple example that motivates a relationship between log-PageRank values and a notion of distance.

4 LOG-PAGERANK ON THE CHAIN GRAPH

We developed a closed form expression for the personalized PageRank on chain graph and observed a linear dependence between the element-wise log of PageRank and the graph distance. For a chain graph of size $n > 2$, we solve the linear

Algorithm 1 Log-PageRank Embedding

Input: Graph adjacency matrix $\boldsymbol{A}$, Dimension of embedding k, Number of samples $s \geq k+1$ (we suggest $s = (10+k)\log n$), Teleportation parameter α

Output: Graph embedding $\boldsymbol{Z} \in \mathbb{R}^{n \times k}$

for $i = 1 \to s$ **do**
  $u \leftarrow$ `random sample of 1 to` n
  $\boldsymbol{X}_i \leftarrow$ `pagerank on` $\boldsymbol{A}$ `with seed` u`, teleportation param` α
    ▷ We use a single sparse LU on $\boldsymbol{I} - \alpha \boldsymbol{P}$ to compute PageRank
end for
$\boldsymbol{Y} \leftarrow \log .(\mathbf{X})$ ▷ Apply element-wise log on $\boldsymbol{X}$
$\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V} \leftarrow$ SVD of $\boldsymbol{Y}$
$\boldsymbol{Z} \leftarrow \boldsymbol{U}_{2:k+1}$
return $\boldsymbol{Z}$ ▷ Return left singular vectors of $\boldsymbol{Y}$

system defined in equation (1) with k as the seed node to obtain the following closed form expression in terms of k, n, α.

$$\mathbf{x}_i = \begin{cases} cf(i), i \in \{2, \dots, k-1\} \\ \frac{c\alpha}{2}(f(k-1) + g(k+1)), i = k \\ cg(i), i \in \{k+1, \dots, n-1\} \end{cases} \tag{2}$$

where

$$f(i) = \frac{(+)^{i-1} + (-)^{i-1}}{(+)^{k-1} + (-)^{k-1}}, g(i) = \frac{(+)^{n-i} + (-)^{n-i}}{(+)^{n-k} + (-)^{n-k}},$$
$$c = \sqrt{\frac{1-\alpha}{1+\alpha}}, (+) = \frac{1+\sqrt{1-\alpha^2}}{\alpha}, (-) = \frac{1-\sqrt{1-\alpha^2}}{\alpha}.$$

Notice that when α is far from 1, $(-) \approx 0$ for high powers, and when α is close to 1, $(+) \approx (-)$. In both ways we get the same approximation that

$$\mathbf{x}_i \approx \begin{cases} c(+)^{-|i-k|}, i \in [n] \setminus \{k\} \\ c\frac{\alpha}{(+)}, i = k \end{cases} .$$

Then the logarithm of PageRank expression for $\mathbf{x}_i$ can be written as,

$$\log \mathbf{x}_i \approx -|k-i| \log((+)) + \log(\sqrt{\frac{1-\alpha}{1+\alpha}})$$

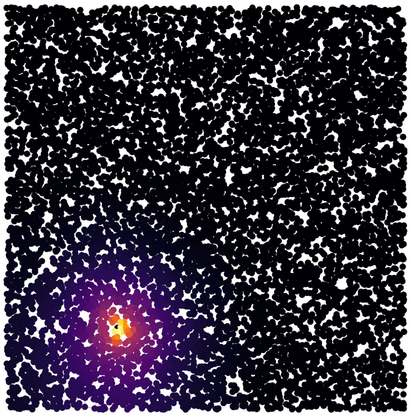
The above formulation indicates a linear relation between the log-PageRank and the distance from the seed node. This hints at log-PageRank being a good measure of the structure of the network around the seed node.

We quickly verify that log-PageRank resembles the notion of “distance” in a geometric graph. The graph is created by randomly sampling points and connecting every point to its 6 nearest neighbors. The difference between PageRank and log-PageRank in this context is illustrated in Figure 2.

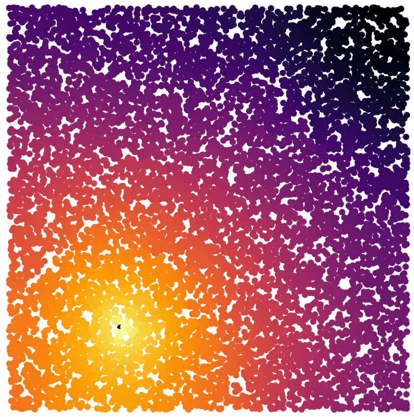
5 RELATION BETWEEN LOG-PAGERANK EMBEDDING AND SPECTRAL EMBEDDING

In this section, we theoretically illustrate the relation between log-PageRank Embedding and Spectral Embedding on a special class of graphs, d-regular graphs.

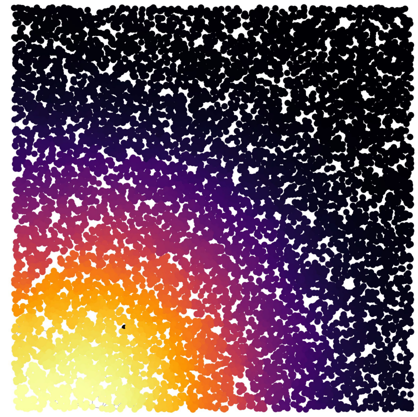
Recall that the lazy random walk matrix is $\boldsymbol{W} = \frac{\boldsymbol{I} + \boldsymbol{A}\boldsymbol{D}^{-1}}{2}$. Our use of the lazy walk matrix is due to the simplicity in analyzing powers of the matrix because



(a) PageRank values for a 10000 node graph with 6 nearest neighbour with $\alpha = 0.999$



(b) Log of PageRank values for a 10000 node graph with 6 nearest neighbour with $\alpha = 0.999$



(c) Normalized Adjacency powers for the seed used in PageRank above and $p = 2000$

FIGURE 2 – Distance effect created by log of PageRank in a geometric graph. The normalized adjacency matrix power is $(\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2})^p\mathbf{v}$ where $\mathbf{v}$ is the same as the PageRank seed. Note the stronger similarity of (b) and (c). Also note the difference is at the boundary. The boundary is where we tend to see the biggest differences between log-PageRank and spectral embeddings.

it is fundamentally aperiodic. A more intricate analysis would likely be able to remove the aperiodicity.

Let the transition probability matrix $\boldsymbol{P}$ of PageRank be the lazy random walk matrix $\boldsymbol{W}$. By a variety of existing analyses [Serra-Capizzano, 2005; Gleich, 2009], we know that $\lim_{\alpha\to 1^-} \mathbf{x}(u,\alpha) = \lim_{\alpha\to 1^-}(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{W})\mathbf{e}_u = \boldsymbol{\pi}$. This extends to log by continuity. Thus, $\lim_{\alpha\to 1^-} \log.(\mathbf{x}(u,\alpha)) = \log.(\boldsymbol{\pi})$.

We continue our study on d-regular graphs. Because for d-regular graphs, $\boldsymbol{A}$ shares the same eigenvectors with $\boldsymbol{W}$, instead of analyzing the eigenvectors of $\boldsymbol{A}$ as Spectral Embedding does, we analyze the eigenvectors of $\boldsymbol{W}$ and connect them with log-PageRank Embedding.

We first prove a result which characterizes the relation between PageRank vectors after element-wise log and columns of high power of $\boldsymbol{W}$.

LEMMA 2 *For a connected d-regular graph G, let $\mathbf{x}(u,\alpha)$ be the personalized PageRank vector seeded on vertex u with parameter α, we have*

$$\lim_{\alpha\to 1^-} \frac{\log.(\mathbf{x}(u,\alpha))}{\|\log.(\mathbf{x}(u,\alpha))\|} = \lim_{k\to\infty} \frac{\boldsymbol{W}^k\mathbf{e}_u}{\|\boldsymbol{W}^k\mathbf{e}_u\|}$$

PROOF For d-regular graphs, the stationary distribution $\boldsymbol{\pi}$ of $\boldsymbol{W}$ is $\frac{\mathbf{e}}{n}$. Thus

$$\lim_{\alpha\to 1^-} \frac{\log.(\mathbf{x}(u,\alpha))}{\|\log.(\mathbf{x}(u,\alpha))\|} = \frac{\log.(\boldsymbol{\pi})}{\|\log.(\boldsymbol{\pi})\|} = \frac{\mathbf{e}}{\sqrt{n}}.$$

Let $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ be the eigenvalue decomposition of $\boldsymbol{W}$. From Perron–Frobenius theory [Perron, 1907; Frobenius, 1912], since G is connected and $\boldsymbol{W}$ models a walk with self-loop, we have that $\lambda_1 > |\lambda_i|$ for all $i \neq 1$, and

$$\lim_{k\to\infty} \frac{\boldsymbol{W}^k\mathbf{e}_u}{\|\boldsymbol{W}^k\mathbf{e}_u\|} = \boldsymbol{Q}_1 = \frac{\mathbf{e}}{\sqrt{n}} \qquad \blacksquare$$

Further, the approximation result below connects the eigenvectors of $\boldsymbol{W}$ with the left singular vectors of the matrix composed of randomly sampled columns of $\boldsymbol{W}^k$.

Approximation Result 1 *For a graph G and $m \leq n$, let $k_1,\dots,k_m$ be m large integers and $i_1,\dots,i_m$ be m indices uniformly sampled from $[n]$, in expectation left singular vectors of $\boldsymbol{B}$ equal eigenvectors of $\boldsymbol{W}$ where $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{W}^{k_1}\mathbf{e}_{i_1} & \boldsymbol{W}^{k_2}\mathbf{e}_{i_2} & \dots & \boldsymbol{W}^{k_m}\mathbf{e}_{i_m}\end{bmatrix}$.*

Justification As we know, the left singular vectors of $\boldsymbol{B}$ are the eigenvectors of $\boldsymbol{B}\boldsymbol{B}^T$. Let $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ be the eigenvalue decomposition of $\boldsymbol{W}$, then we have $\boldsymbol{W}^{k_i} = \boldsymbol{Q}\boldsymbol{\Lambda}^{k_i}\boldsymbol{Q}^T$, and we have

$$\begin{aligned}\mathbb{E}_{i_1,i_2,\ldots,i_m}\boldsymbol{B}\boldsymbol{B}^T &= \mathbb{E}_{i_1,i_2,\ldots,i_m}\left(\sum_{j=1}^{m}\boldsymbol{Q}\boldsymbol{\Lambda}^{k_j}\boldsymbol{Q}^T\mathbf{e}_{i_j}\mathbf{e}_{i_j}^T\boldsymbol{Q}\boldsymbol{\Lambda}^{k_j}\boldsymbol{Q}^T\right)\\ &= \left(\sum_{j=1}^{m}\boldsymbol{Q}\boldsymbol{\Lambda}^{k_j}\boldsymbol{Q}^T\frac{\boldsymbol{I}}{n}\frac{\boldsymbol{I}^T}{n}\boldsymbol{Q}\boldsymbol{\Lambda}^{k_j}\boldsymbol{Q}^T\right)\\ &= \frac{1}{n^2}\boldsymbol{Q}\left(\sum_{j=1}^{m}\boldsymbol{\Lambda}^{k_j}\right)\boldsymbol{Q}^T.\end{aligned}$$

Therefore in expectation the first m eigenvectors of $\boldsymbol{B}\boldsymbol{B}^T$ are $\boldsymbol{Q}_1,\ldots,\boldsymbol{Q}_m$. Notice that when one eigenvalue of $\boldsymbol{W}$ has multiplicity larger than 1, the corresponding eigenvectors of $\boldsymbol{B}\boldsymbol{B}^T$ may undergo one orthogonal transformation but the spaces they span are invariant. ■

The approximation result below states that the low-rank log-PageRank Embedding is expected to approximate the Spectral Embedding for degree-regular graphs.

Approximation Result 2 *For a d-regular graph G and $m \leq n$, let $i_1,\ldots,i_m$ be m indices randomly sampled from $[n]$, for α close to 1, left singular vectors of $\boldsymbol{C} = \frac{1}{\sqrt{n}}\begin{bmatrix}\frac{\log.(\mathbf{x}(i_1,\alpha))}{\|\log.(\mathbf{x}(i_1,\alpha))\|} & \frac{\log.(\mathbf{x}(i_2,\alpha))}{\|\log.(\mathbf{x}(i_2,\alpha))\|} & \cdots & \frac{\log.(\mathbf{x}(i_n,\alpha))}{\|\log.(\mathbf{x}(i_n,\alpha))\|}\end{bmatrix}$ approximates the eigenvectors of $\boldsymbol{W}$.*

Justification By Lemma 2, we know that $\lim_{\alpha\to 1^-}\frac{\log.(\mathbf{x}(i_j,\alpha))}{\|\log.(\mathbf{x}(i_j,\alpha))\|} = \lim_{k_j\to\infty}\frac{\boldsymbol{W}^{k_j}\mathbf{e}_{i_j}}{\|\boldsymbol{W}^{k_j}\mathbf{e}_{i_j}\|}$. Let $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ be the eigenvalue decomposition for $\boldsymbol{W}$, we have for d-regular graphs, $\boldsymbol{Q}_1 = \frac{\mathbf{e}}{\sqrt{n}}$ and $\lambda_1 = 1$. Thus, for large enough k_j, $\boldsymbol{W}^{k_j}\mathbf{e}_{i_j} \approx \lambda_1^{k_j}\boldsymbol{Q}_1\boldsymbol{Q}_1^T\mathbf{e}_{i_j} = \boldsymbol{Q}_1\boldsymbol{Q}_1^T\mathbf{e}_{i_j}$ and $\|\boldsymbol{W}^{k_j}\mathbf{e}_{i_j}\| \approx \frac{1}{\sqrt{n}}$. Therefore let $k_1,\ldots,k_m$ be m randomly sampled large integers, we have

$$\boldsymbol{C} = \frac{1}{\sqrt{n}}\begin{bmatrix}\frac{\log.(\mathbf{x}(i_1,\alpha))}{\|\log.(\mathbf{x}(i_1,\alpha))\|} & \frac{\log.(\mathbf{x}(i_2,\alpha))}{\|\log.(\mathbf{x}(i_2,\alpha))\|} & \cdots & \frac{\log.(\mathbf{x}(i_n,\alpha))}{\|\log.(\mathbf{x}(i_n,\alpha))\|}\end{bmatrix}$$

approximates

$$\boldsymbol{B} = \begin{bmatrix}\boldsymbol{W}^{k_1}\mathbf{e}_{i_1} & \boldsymbol{W}^{k_2}\mathbf{e}_{i_2} & \ldots & \boldsymbol{W}^{k_m}\mathbf{e}_{i_m}\end{bmatrix}$$

Further, by Approximation Result 1, we have in expectation left singular vectors of $\boldsymbol{B}$ approximates the eigenvectors of $\boldsymbol{W}$, thus left singular vectors of $\boldsymbol{C}$ approximates the eigenvectors of $\boldsymbol{W}$ as well. ■

6 EMPIRICAL COMPARISON RESULTS

We study the log-PageRank embedding on synthetic and real world graphs. We focus on those where the spectral embedding gives a good picture of the graph, as spectral embeddings may fail to give useful pictures for many real-world networks [Lang, 2005]. We will analyse its performance on nearest neighbour graphs and the following graphs with a strong geometry.

6.1 IMPLEMENTATION

We implemented the log-PageRank embedding in Julia. We use the built in sparse LU solver to factorize the PageRank matrix $\boldsymbol{I} - \alpha\boldsymbol{P}$ to solve linear systems for large values of α. And we use the built in dense SVD solver. There are many alternatives we could use here, but our focus was on understanding the log-PageRank embeddings rather than optimizing the speed at which we can compute them.

(a) Graph for the word "log PR" with 5000 nodes, 14962 edges

(b) The Minnesota road network with 2640 nodes, 3302 edges

(c) The Tapir [Bern et al., 1994] graph with 1024 nodes, 2846 edges

FIGURE 3 – Our geometric graphs

6.2 QUALITATIVE AND QUANTITATIVE ERROR AND APPROXIMATION

Given a spectral embedding and a log-PageRank embedding, we first simply look at the differences in the pictures. Our results say these should look similar, not that they should be exactly the same as the graphs we study are not in the class where we expect sharp approximations. Let the second singular vectors of the log-PageRank embedding be $\mathbf{u}_2$, and the second eigenvector of the Laplacian be $\mathbf{z}_2$. Likewise for the 3rd vectors. So the spectral embedding is $\mathbf{z}_2, \mathbf{z}_3$ and the log-PageRank embedding is $\mathbf{u}_2, \mathbf{u}_3$.

We quantitatively measure the error by evaluating the relative difference between the Rayleigh quotient with respect to the vectors used for embedding. This gives us the following measure:

$$\text{approximation error} = \frac{s - p}{s} \tag{3}$$

where

$$s = \frac{\mathbf{z}_2^T \boldsymbol{\mathcal{L}} \mathbf{z}_2}{\mathbf{z}_2^T \mathbf{z}_2}, \qquad p = \frac{\mathbf{u}_2^T \boldsymbol{\mathcal{L}} \mathbf{u}_2}{\mathbf{u}_2^T \mathbf{u}_2}.$$

The second way we evaluate the embeddings is by looking at the joint plot of $\mathbf{u}_2$ vs. $\mathbf{z}_2$ and $\mathbf{z}_3$ vs. $\mathbf{u}_3$. If the embeddings are close, these should look like a straight line, or at least a very highly correlated relationships.

6.3 EVALUATION ACROSS GRAPHS

We record the error according to equation (3) in Table 1 for the following types of graphs.

Nearest Neighbor Graphs For a graph named $n - k$ nearest neighbor, there are n points randomly distributed in the unit square and connected to k nearest neighbors.

Chain Graphs These are simply the chain graphs we had from the analysis in Section 4.

Graphs with Strong Geometry These are the graphs from Figure 3.

Stochastic Block Models A graph named $\text{sbm}(n, k, p, q)$ has k groups of n vertices with inter-group probability p and between group probability q. These show the worst approximation results and largest differences.

Graph	$\alpha = 0.99$		$\alpha = 0.9999$	
	raw	log	raw	log
30-6 nearest neighbour	3.27%	0.06%	2.89%	0.05%
3000-6 nearest neighbour	47.6%	0.37%	5.06%	2.88%
10000-6 nearest neighbour	170.75%	2.13%	13.5%	1.76%
30 chain	26.88%	0.47%	28.42%	6.02%
3000 chain	2858.82%	1.06%	30.38%	0.75%
Minnesota $n = 2640$	16.07%	1.97%	11.15%	0.44%
Tapir $n = 1024$	10.17%	1.13%	15.41%	0.66%
LogPR $n = 5000$	19.95%	0.15%	4.76%	0.34%
sbm(50,60,0.001,0.005)	51.77%	15.22%	51.32%	67.25%
sbm(1000,3,0.001,0.005)	47.35%	16.93%	45.78%	89.39%
sbm(50,60,0.25,0.005)	17.88%	15.22%	90.13%	402.27%
sbm(1000,3,0.25,0.001)	53.7%	1.04%	16.21%	15.73%

TABLE 1 – Error between PageRank embedding and spectral embedding for different graphs at a low teleportation probability, $\alpha = 0.99$ and at a higher one $\alpha = 0.99999$ both without log (raw) and with log.

6.4 A FEW EXAMPLES

Figure 5 and Figure 6 show the embeddings as α varies both with and without the nonlinear *log* operation for the graphs in figures 3b and 3c. The result on the "log PR" graph from Figure 3a was in the introduction.

On nearest neighbor graphs, such as Figure 7, these embeddings show a clear rotational ambiguity that might arise with other evaluations of this strategy. (This occured with the other graphs too.) Put plainly, the eigenvectors are almost in 2d invariant subspace. Consequently, when we randomize the method, we can only capture this near 2d subspace up to rotation. However, this will not show high error with respect to the approximation error measure as the results are all near eigenvectors.

We show one of the examples of the stochastic block model in Figure 8. Although this has bad approximation with respect to the spectral embedding, the result for the log-PageRank embedding for $\alpha = 0.99$ is arguably better than the spectral embedding.

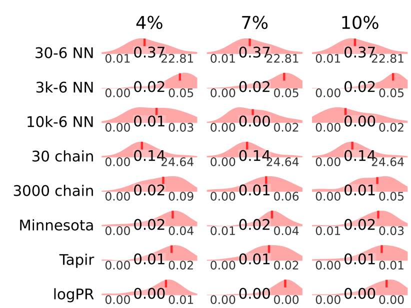


FIGURE 4 – Error variation with column for log of PageRank with $\alpha = 0.99$. The percentage indicated in the column headings are the fraction of the nodes as seeds. Each entry is the variance, the maximum and the minimum for 50 trials.

6.5 EMBEDDING ERROR VARIANCE

We studied the dependence of this error on the number of randomly sampled nodes and which sampled nodes. We record this for log-PageRank at $\alpha = 0.99$ in Figure 4. This shows the distribution of errors as a density estimate, along with the max/min values (small) and the median value (big).

As expected, there are largely minimal effects. This occurs because the sensitivity of PageRank to the seed vector, $\mathbf{v}$, is a function of α [Langville and Meyer, 2006].

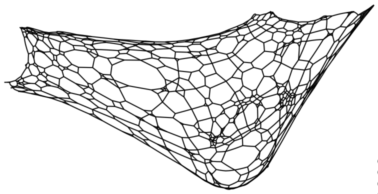
$$\frac{d\mathbf{x}}{d\mathbf{v}} = (1-\alpha)(\boldsymbol{I} - \alpha\boldsymbol{P})^{-1}$$

which satisfies $\|\frac{d\mathbf{x}}{d\mathbf{v}}\|_1 = 1$. Further, for $\alpha \to 1$, dependence of the PageRank values on the $\mathbf{v}$ reduces. Our experiments confirm the same as the minimum, maximum and variance of error over 50 trials show negligible change.

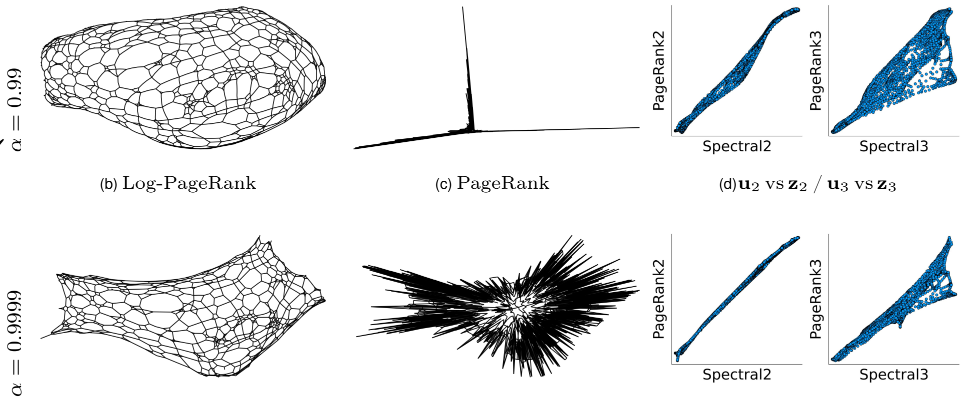
7 HYPERGRAPH EMBEDDINGS

One driving reason for our study of the log-PageRank embedding is to support similar embedding strategies for different types of data, such as those studied in [Fountoulakis et al., 2020]. In this section, we use the log-PageRank embedding technique on five hypergraphs: Yelp (`https://www.yelp.com/dataset`), Walmart

FIGURE 5 – Comparison of embeddings for the Minnesota network.



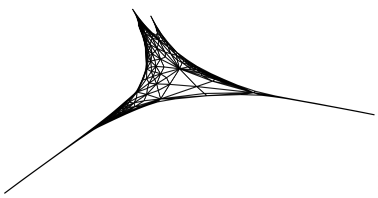
(a) Spectral Embedding



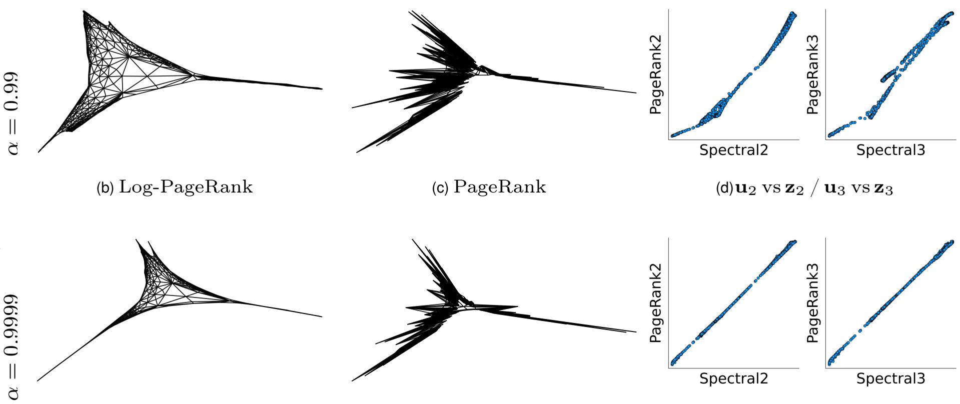
(b) Log-PageRank (c) PageRank (d) $\mathbf{u}_2$ vs $\mathbf{z}_2$ / $\mathbf{u}_3$ vs $\mathbf{z}_3$

(e) Log-PageRank (f) PageRank (g) $\mathbf{u}_2$ vs $\mathbf{z}_2$ / $\mathbf{u}_3$ vs $\mathbf{z}_3$

FIGURE 6 – Comparison of embeddings for the Tapir graph.



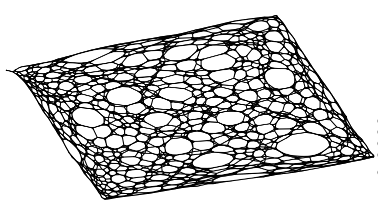
(a) Spectral Embedding



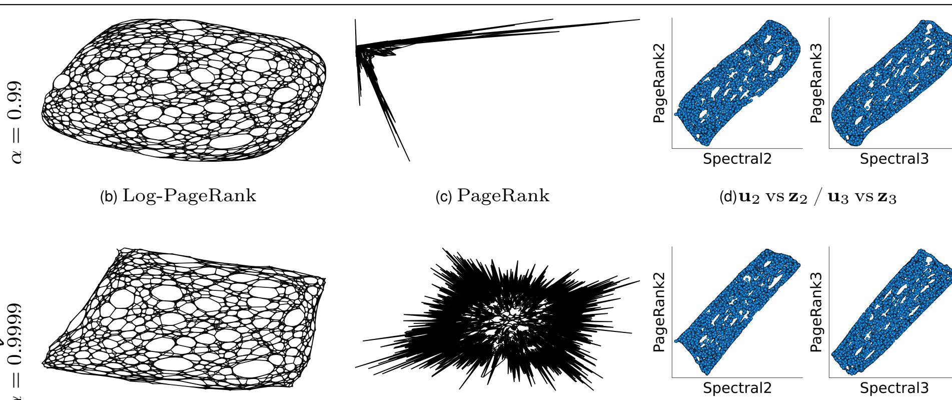
(b) Log-PageRank (c) PageRank (d) $\mathbf{u}_2$ vs $\mathbf{z}_2$ / $\mathbf{u}_3$ vs $\mathbf{z}_3$

(e) Log-PageRank (f) PageRank (g) $\mathbf{u}_2$ vs $\mathbf{z}_2$ / $\mathbf{u}_3$ vs $\mathbf{z}_3$

FIGURE 7 – Embedding for 10000 node graph with 6 nearest neighbours. Note that log-PageRank and spectral embeddings are similar after a rotation. This is expected because the eigenvalues are close in magnitude.



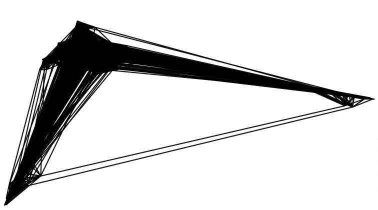
(a) Spectral Embedding



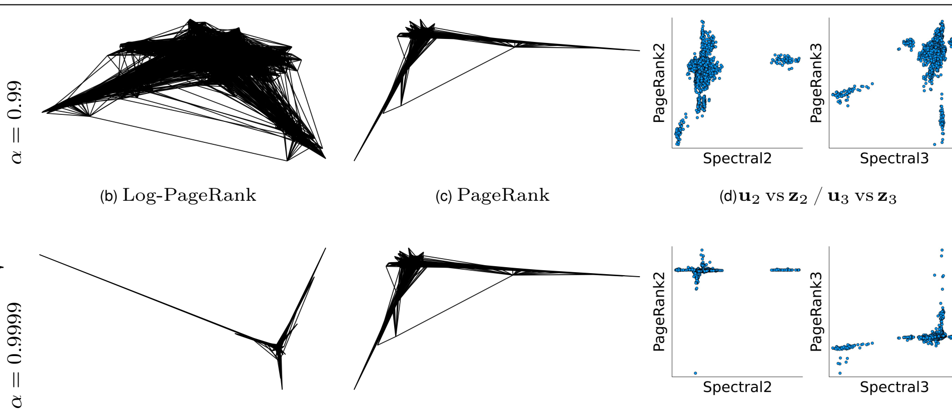
(b) Log-PageRank (c) PageRank (d) $\mathbf{u}_2$ vs $\mathbf{z}_2$ / $\mathbf{u}_3$ vs $\mathbf{z}_3$

(e) Log-PageRank (f) PageRank (g) $\mathbf{u}_2$ vs $\mathbf{z}_2$ / $\mathbf{u}_3$ vs $\mathbf{z}_3$

FIGURE 8 – Comparison of the embedding techniques on the planted partition model with good conductance cuts. This is one case where the technique does not seem to work.

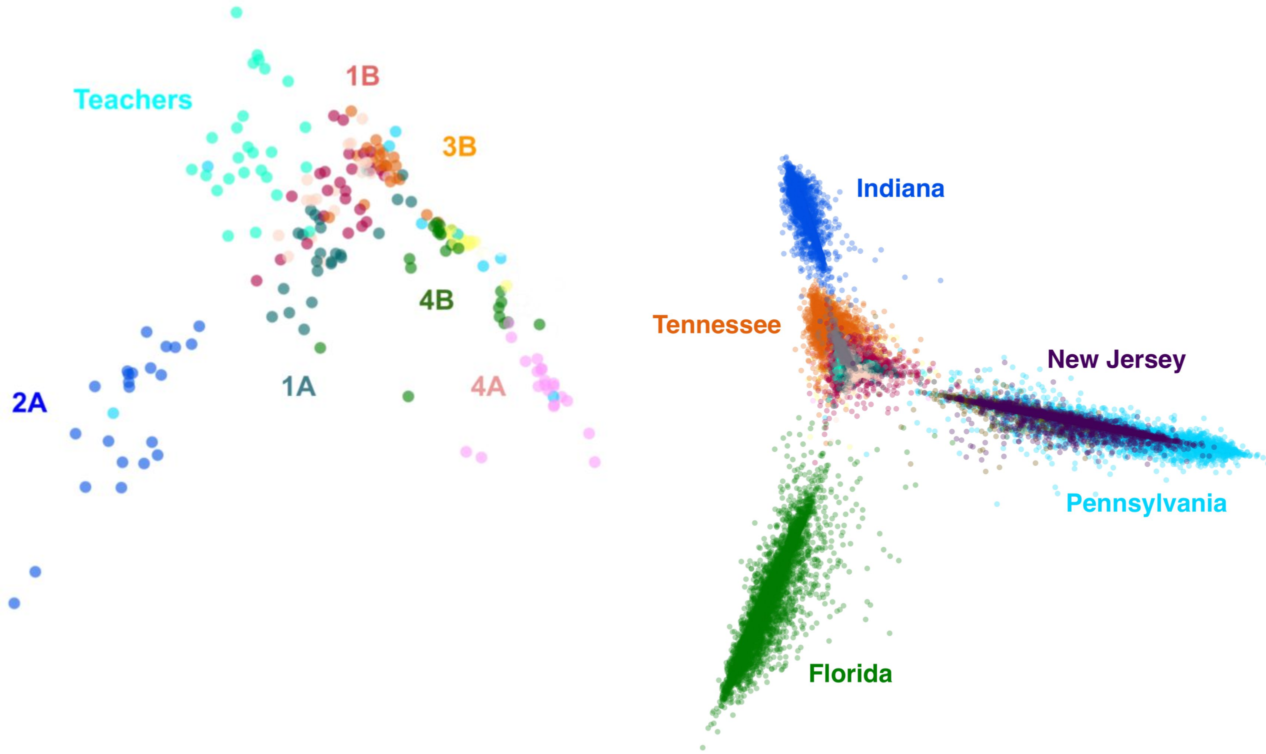


(a) Spectral Embedding



(b) Log-PageRank (c) PageRank (d) $\mathbf{u}_2$ vs $\mathbf{z}_2$ / $\mathbf{u}_3$ vs $\mathbf{z}_3$

(e) Log-PageRank (f) PageRank (g) $\mathbf{u}_2$ vs $\mathbf{z}_2$ / $\mathbf{u}_3$ vs $\mathbf{z}_3$



(a) Contact Primary School

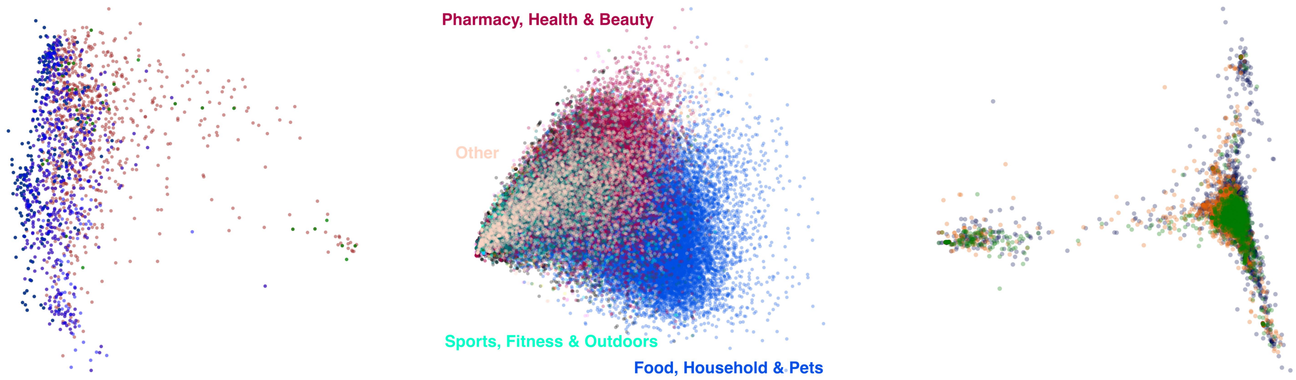
(b) Yelp Restaurants

FIGURE 9 – Log-PageRank embedding of hypergraphs. The Contact Primary School dataset has 242 nodes and 12704 hyperdges. Nodes are colored by classroom and teachers, which form cohesive groups due to the contact structure. The Yelp Restaurant dataset has 52260 nodes and 597261 hyperedges. Nodes are colored by one of 14 states used for analysis, which show clear geographic relationships.

Trips [Amburg et al., 2020] , a contact tracing network [Benson et al., 2018; Stehlé et al., 2011], posts on Math Overflow [Veldt et al., 2020a], and a Drug Abuse network (DAWN) [Amburg et al., 2020]. The only modification to our algorithm is that we replace seeded PageRank in Algorithm 1 with the Local Quadratic PageRank (LQPR), a method proposed in [Liu et al., 2021]. Specifically we use LQHD with a 2-norm penalty with $\rho = 0.5$ for all experiments For the Yelp and Walmart trips network, we set $\kappa = 0.000025$ and $\gamma = 1.0$ while for the Math Overflow network, with the same sparsity factor $\kappa = 0.000025$ and $\gamma = 0.001$. For Contact Primary School and DAWN, we set $\kappa = 0.0025$ and $\gamma = 0.001$. These choices were made arbitrarily, there are small differences that result when changing them.

Figure 9a shows our 2d embedding on Contact Primary School dataset where each node represents a student or a teacher, each hyperedge represents a group of people who are spatially close at a given time. Each node of the graph is colored as a teacher or as classroom for the student. Note that each classroom forms a cohesive group in the plot. The nodes for the teachers are not a good cluster because different teachers go to different classrooms. Moreover, we observe that the students from the same grade, e.g. students colored red (1B) and dark green (1A), share some spatial proximity, which is due to the fact that their classrooms are close.

Figure 9b shows our embedding on Yelp Review data. Following Veldt et al. [2020b], we build one hypergraph with each restaurant being a node and each user being a hyperedge. We show the state associated with each location as the color. We can clearly see that our embedding captures the geographic information of the underlying hypergraph. For example, the nodes labeled dark blue are those restaurants from state Indiana, which are close to the orange nodes from state Tennessee. Also the green nodes from state Florida are quite well-separated from nodes with other colors, which is due to the fact that none of other 13 states we plot is close to Florida.



(a) DAWN

(b) Walmart Trips

(c) Math Overflow

FIGURE 10 – Log-PageRank embeddings of the (a) DAWN dataset [Amburg et al., 2020] with 2109 nodes (drugs) and 87104 hyperedges where each hyperedge is an individual and the consisting nodes are the drugs consumed by them; the plot shows 3 out 10 labels (b) the Walmart dataset [Amburg et al., 2020] with 88860 nodes and 69906 hyperedges where each node is a product and each hyperedge consists of products purchased in one trip to Walmart and (c) Math Overflow [Veldt et al., 2020a] dataset with 73851 nodes, 5446 hyperedges, where each hyperedge multiple labels associated with a question; Although there is structure evident in the plots, it does not strongly correlate with the known labels on the data (some of the plotting makes the structure look more present than it is).

In addition, we show log-PageRank embeddings of three other hypergraphs in Figure 10a, 10b and 10c. We were unable to identify obvious relationships between these embeddings and the existing groups, which means the embeddings likely show a different type of structure. The promising results on all the datasets above show that our simple algorithm is capable of generating good embeddings even on higher order graphs.

8 RELATED WORK, CONCLUSION, AND FUTURE DIRECTIONS

The key finding of this paper is the elementwise log of a matrix of seeded PageRank vector approximates the spectral embedding of the Laplacian in some scenarios. This analysis can be transparently mapped to new scenarios such as hypergraphs given a PageRank-like primitive. This greatly simplifies the scenario compared with nonlinear spectral methods on hypergraphs [Tudisco et al., 2021c,a; Tudisco and Higham, 2021; Nguyen et al., 2017].

The idea of using the log of a PageRank vector originated in Google's initial use of these for their PageRank scores. That said, the elementwise log emerged in other scenarios as well. For example, Levy and Goldberg [2014] detail a similar analysis between SkipGram [Mikolov et al., 2013], a popular representation learning framework, and the SVD of the element wise log of a probability transition matrix developed from the data. Following that, multiple papers [Chanpuriya and Musco, 2020; Qiu et al., 2018] showed relationships between embedding techniques [Grover and Leskovec, 2016; Tang et al., 2015a,b; Perozzi et al., 2014] and asymptotic matrix expressions.

We believe our framework offers a successful technique for structural embedding and opens up some nontrivial research problems. Our code to compute these embeddings for these examples is available: `https://github.com/dishashur/log-pagerank`. For example, although we provide an approximate analysis as to why the log function and PageRank-like matrices converge to the spectral embedding and hence generated meaningful representation, there does not yet exist a quantifiable expression between the strength of this relation and the elements of the structure, such as its conductance, sparsity, degree distribution. We believe this work lays the foundation for a reliable structural representation and the generalizability of this technique offers ample ground for new results.

REFERENCES

[Amburg et al., 2020] I. Amburg, N. Veldt, and A. R. Benson. *Clustering in graphs and hypergraphs with categorical edge labels*. In *Proceedings of the Web Conference*. 2020. Cited on pages 11 and 12.

[Andersen et al., 2006] R. Andersen, F. Chung, and K. Lang. *Local graph partitioning using pagerank vectors*. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 475–486. 2006. Cited on page 1.

[Bar-Yossef and Mashiach, 2008] Z. Bar-Yossef and L.-T. Mashiach. *Local approximation of PageRank and reverse PageRank*. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pp. 279–288. 2008. doi:10.1145/1458082.1458122. Cited on page 1.

[Becchetti et al., 2008] L. Becchetti, C. Castillo, D. Donato, R. Baeza-Yates, and S. Leonardi. *Link analysis for web spam detection*. ACM Trans. Web, 2 (1), pp. 1–42, 2008. doi:10.1145/1326561.1326563. Cited on page 1.

[Benson et al., 2018] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg. *Simplicial closure and higher-order link prediction*. Proceedings of the National Academy of Sciences, 2018. doi:10.1073/pnas.1800683115. Cited on page 11.

[Bern et al., 1994] M. Bern, S. Mitchell, and J. Ruppert. *Linear-size nonobtuse triangulation of polygons*. In *Proceedings of the Tenth Annual Symposium on Computational Geometry*, p. 221–230. 1994. doi:10.1145/177424.177974. Cited on page 8.

[Brin and Page, 1998] S. Brin and L. Page. *The anatomy of a large-scale hypertextual web search engine*. Computer Networks and ISDN Systems, 30 (1), pp. 107–117, 1998. Proceedings of the Seventh International World Wide Web Conference. doi:https://doi.org/10.1016/S0169-7552(98)00110-X. Cited on page 1.

[Carletti et al., 2020] T. Carletti, F. Battiston, G. Cencetti, and D. Fanelli. *Random walks on hypergraphs*. Phys. Rev. E, 101, p. 022308, 2020. doi:10.1103/PhysRevE.101.022308. Cited on page 3.

[Chanpuriya and Musco, 2020] S. Chanpuriya and C. Musco. *InfiniteWalk: Deep Network Embeddings as Laplacian Embeddings with a Nonlinearity*, p. 1325–1333. Association for Computing Machinery, New York, NY, USA, 2020. Cited on page 12.

[Chung, 2007] F. Chung. *The heat kernel as the pagerank of a graph*. Proceedings of the National Academy of Sciences, 104 (50), pp. 19735–19740, 2007. arXiv:https://www.pnas.org/content/104/50/19735.full.pdf, doi:10.1073/pnas.0708838104. Cited on page 3.

[Chung et al., 2011] F. Chung, A. Tsiatas, and W. Xu. *Dirichlet pagerank and trust-based ranking algorithms*. In *Algorithms and Models for the Web Graph*, pp. 103–114. 2011. Cited on page 2.

[Chung, 1992] F. R. L. Chung. *Spectral Graph Theory*, American Mathematical Society, 1992. Cited on page 1.

[Donnat et al., 2018] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec. *Learning structural node embeddings via diffusion wavelets*. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1320–1329. 2018. doi:10.1145/3219819.3220025. Cited on pages 1 and 3.

[Fountoulakis et al., 2020] K. Fountoulakis, M. Liu, D. F. Gleich, and M. W. Mahoney. *Flow-based algorithms for improving clusters: A unifying framework, software, and performance*. 2020. arXiv:2004.09608. Cited on pages 2, 4, and 9.

[Frobenius, 1912] G. Frobenius. *Über matrizen aus nicht negativen elementen*. Königliche Akademie der Wissenschaften Sitzungsber, Kön, pp. 456–477, 1912. Cited on page 6.

[Gleich, 2009] D. F. Gleich. *Models and Algorithms for PageRank Sensitivity*. Ph.D. thesis, Stanford University, 2009. Cited on page 6.

[Gleich, 2015] ———. *Pagerank beyond the web*. SIAM Rev., 57 (3), p. 321–363, 2015. doi:10.1137/140976649. Cited on pages 1 and 3.

[Grover and Leskovec, 2016] 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*, p. 855–864. 2016. doi:10.1145/2939672.2939754. Cited on pages 3 and 12.

[Hall, 1970] K. M. Hall. *An r-dimensional quadratic placement algorithm*. Management science, 17 (3), pp. 219–229, 1970. Cited on page 1.

[Hammond et al., 2011] D. K. Hammond, P. Vandergheynst, and R. Gribonval. *Wavelets on graphs via spectral graph theory*. Applied and Computational Harmonic Analysis, 30 (2), pp. 129–150, 2011. doi:https://doi.org/10.1016/j.acha.2010.04.005. Cited on page 1.

[Huang et al., 2021] Q. Huang, H. He, A. Singh, S.-N. Lim, and A. Benson. *Combining label propagation and simple models out-performs graph neural networks*. In *International Conference on Learning Representations*. 2021. Cited on page 3.

[Klicpera et al., 2019] J. Klicpera, A. Bojchevski, and S. Günnemann. *Combining neural networks with personalized pagerank for classification on graphs*. In *International Conference on Learning Representations*. 2019. Cited on page 3.

[Kloster and Gleich, 2014] K. Kloster and D. F. Gleich. *Heat kernel based community detection*. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1386–1395. 2014. Cited on page 1.

[Koren, 2003] Y. Koren. *On spectral graph drawing*. In *Computing and Combinatorics*, pp. 496–508. 2003. Cited on page 1.

[Lang, 2005] K. Lang. *Fixing two weaknesses of the spectral method*. Advances in Neural Information Processing Systems, 18, 2005. Cited on page 7.

[Langville and Meyer, 2006] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, USA, 2006. Cited on page 9.

[Levy and Goldberg, 2014] O. Levy and Y. Goldberg. *Neural word embedding as implicit matrix factorization*. Advances in neural information processing systems, 27, 2014. Cited on page 12.

[Liu et al., 2021] M. Liu, N. Veldt, H. Song, P. Li, and D. F. Gleich. *Strongly local hypergraph diffusions for clustering and semi-supervised learning*. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pp. 2092–2103. 2021. doi:10.1145/3442381.3449887. Cited on pages 3 and 11.

[Mahoney et al., 2012] M. W. Mahoney, L. Orecchia, and N. K. Vishnoi. *A local spectral method for graphs: With applications to improving graph partitions and exploring data graphs locally*. The Journal of Machine Learning Research, 13 (1), pp. 2339–2365, 2012. Cited on page 1.

[Mikolov et al., 2013] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient estimation of word representations in vector space*. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. 2013. Cited on page 12.

[Nguyen et al., 2017] Q. Nguyen, F. Tudisco, A. Gautier, and M. Hein. *An efficient multilinear optimization framework for hypergraph matching.* IEEE transactions on pattern analysis and machine intelligence, 39 (6), pp. 1054–1075, 2017. Cited on page 12.

[Perozzi et al., 2014] 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*, p. 701–710. 2014. doi:10.1145/2623330.2623732. Cited on pages 3 and 12.

[Perron, 1907] O. Perron. *Zur theorie der matrices.* Mathematische Annalen, 64 (2), pp. 248–263, 1907. Cited on page 6.

[Pothen et al., 1990] A. Pothen, H. D. Simon, and K.-P. Liou. *Partitioning sparse matrices with eigenvectors of graphs*. SIAM J. Matrix Anal. Appl., 11, pp. 430–452, 1990. doi:10.1137/0611030. Cited on page 1.

[Qiu et al., 2018] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. 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*, p. 459–467. 2018. doi:10.1145/3159652.3159706. Cited on page 12.

[Sahai et al., 2011] T. Sahai, A. Speranzon, and A. Banaszuk. *Hearing the clusters in a graph: A distributed algorithm.* 2011. arXiv:0911.4729. Cited on page 3.

[Serra-Capizzano, 2005] S. Serra-Capizzano. *Jordan canonical form of the google matrix: A potential contribution to the pagerank computation*. SIAM Journal on Matrix Analysis and Applications, 27 (2), pp. 305–312, 2005. arXiv:https://doi.org/10.1137/S0895479804441407, doi:10.1137/S0895479804441407. Cited on page 6.

[Shi and Malik, 2000] J. Shi and J. Malik. *Normalized cuts and image segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (8), pp. 888–905, 2000. doi:10.1109/34.868688. Cited on page 1.

[Stehlé et al., 2011] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. V. den Broeck, C. Régis, B. Lina, and P. Vanhems. *High-resolution measurements of face-to-face contact patterns in a primary school*. PLoS ONE, 6 (8), p. e23176, 2011. doi:10.1371/journal.pone.0023176. Cited on page 11.

[Takai et al., 2020] Y. Takai, A. Miyauchi, M. Ikeda, and Y. Yoshida. *Hypergraph Clustering Based on PageRank*, p. 1970–1978. Association for Computing Machinery, New York, NY, USA, 2020. Cited on page 3.

[Tang et al., 2015a] J. Tang, M. Qu, and Q. Mei. *Pte: Predictive text embedding through large-scale heterogeneous text networks*. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1165–1174. 2015a. doi:10.1145/2783258.2783307. Cited on page 12.

[Tang et al., 2015b] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. *Line: Large-scale information network embedding*. In *Proceedings of the 24th International Conference on World Wide Web*, p. 1067–1077. 2015b. doi:10.1145/2736277.2741093. Cited on page 12.

[Tong et al., 2006] H. Tong, C. Faloutsos, and J.-Y. Pan. *Fast random walk with restart and its applications.* In *Sixth international conference on data mining (ICDM'06)*, pp. 613–622. 2006. Cited on page 1.

[Tsitsulin et al., 2018] A. Tsitsulin, D. Mottin, P. Karras, A. Bronstein, and E. Müller. *Netlsd: Hearing the shape of a graph*. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018. doi:10.1145/3219819.3219991. Cited on page 3.

[Tudisco et al., 2021a] F. Tudisco, A. R. Benson, and K. Prokopchik. *Nonlinear higher-order label spreading*. In *Proceedings of the Web Conference 2021*, pp. 2402–2413. 2021a. doi:10.1145/3442381.3450035. Cited on page 12.

[Tudisco and Higham, 2021] F. Tudisco and D. J. Higham. *Node and edge eigenvector centrality for hypergraphs.* arXiv:2101.06215, 2021. Cited on page 12.

[Tudisco et al., 2021b] F. Tudisco, K. Prokopchik, and A. R. Benson. *A nonlinear diffusion method for semi-supervised learning on hypergraphs.* 2021b. arXiv:2103.14867. Cited on page 3.

[Tudisco et al., 2021c] ———. *A nonlinear diffusion method for semi-supervised learning on hypergraphs.* arXiv:2103.14867, 2021c. Cited on page 12.

[Veldt et al., 2020a] N. Veldt, A. R. Benson, and J. Kleinberg. *Minimizing localized ratio cut objectives in hypergraphs.* In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2020a. Cited on pages 11 and 12.

[Veldt et al., 2020b] ———. *Minimizing localized ratio cut objectives in hypergraphs.* In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1708–1718. 2020b. Cited on page 11.

[Wang and Leskovec, 2020] H. Wang and J. Leskovec. *Unifying graph convolutional neural networks and label propagation*. 2020. Cited on page 3.

[Yadati et al., 2019] N. Yadati, M. Nimishakavi, P. Yadav, V. Nitin, A. Louis, and P. Talukdar. *Hypergcn: A new method of training graph convolutional networks on hypergraphs.* 2019. arXiv:1809.02589. Cited on page 3.

[Yang et al., 2020] R. Yang, J. Shi, X. Xiao, Y. Yang, and S. S. Bhowmick. *Homogeneous network embedding for massive graphs via reweighted personalized pagerank*. Proc. VLDB Endow., 13 (5), p. 670–683, 2020. doi:10.14778/3377369.3377376. Cited on page 3.