

A Simple Spectral Failure Mode for Graph Convolutional Networks

Carey E. Priebe, Cencheng Shen, Ningyuan (Teresa) Huang, Tianyi Chen

Abstract—Neural networks have achieved remarkable successes in machine learning tasks. This has recently been extended to graph learning using neural networks. However, there is limited theoretical work in understanding how and when they perform well, especially relative to established statistical learning techniques such as spectral embedding. In this short paper, we present a simple generative model where unsupervised graph convolutional network fails, while the adjacency spectral embedding succeeds. Specifically, unsupervised graph convolutional network is unable to look beyond the first eigenvector in certain approximately regular graphs, thus missing inference signals in non-leading eigenvectors. The phenomenon is demonstrated by visual illustrations and comprehensive simulations.

Index Terms—Graph Embedding, Spectral Embedding, Node Classification, Convolutional Neural Network

1 INTRODUCTION

GRAPH embedding aims to learn a low-dimensional representation for each node in a given graph, on which subsequent inferences can be directly performed. Recently, graph neural networks have emerged an attractive graph embedding solution due to the representation power: they can learn nonlinear mappings from the graph space to the embedding space, while traditional methods such as matrix factorization are restricted to linear mappings [1]. On the other hand, powerful graph neural networks can fail if their architectures do not align with the downstream tasks [2].

The natural question is how well graph neural networks perform: Are they always the preferred graph embedding solution? And if not, can we characterize when and how they are sub-optimal? To answer the question, we present a simple and intuitive random graph generative model, where unsupervised graph convolutional network (GCN) fails while adjacency spectral embedding (ASE) – a matrix factorization method – succeeds.

Specifically, we propose a latent position graph model where the classification information lies on the second eigenvector and perpendicular to the first eigenvector. The graph is approximately regular, and unsupervised GCN embedding mainly recovers the leading eigenvector of the graph adjacency matrix. Namely, unsupervised GCN can be

sub-optimal for an approximately regular graph where the signal of the downstream task lies on non-leading eigenvectors, while ASE can succeed under proper dimension choice.

To our best knowledge, our results provide the first geometry characterization of GCN limitation under random graph models, which is an important aspect of unsupervised graph convolutional network that requires further investigation and improvement. In the following, we first introduce the necessary background, followed by the failure geometry, and extensive simulations to demonstrate the failure mode.

2 BACKGROUND

2.1 Classification with Empirical Risk Minimization

The classical statistical formulation of the classification problem consists of

$$(X, Y), (X_1, Y_1), \dots, (X_m, Y_m) \stackrel{iid}{\sim} F_{XY}$$

where $\{(X_i, Y_i)\}_{i \in \{1, \dots, m\}}$ is the training data, and (X, Y) represents the to-be-classified test observation X with true-but-unobserved class label Y . We consider the simple setting in which X is a feature vector in finite-dimensional Euclidean space and Y is a binary class label in $\{0, 1\}$. We assume that X has a density, and denote its marginal and class-conditional densities by f_X and $f_{X|Y}$. We denote the Bayes optimal classifier $g^*(\cdot)$ and its associated Bayes optimal probability of misclassification by $L^* = P[g^*(X) \neq Y]$.

The goal is to learn a classification rule $g_m(\cdot)$ from the training data, which maps feature vectors to class labels such that the probability of misclassification $L(g_m) = P(g_m(X) \neq Y)$ is small (see [3], page 2). We will evaluate performance via the expectation (taken over the training data) of probability of misclassification, $\mathbb{E}[L(g_m)]$. In this setting, choosing $g_m \in \mathcal{C}$ via empirical risk minimization (ERM) yields $\mathbb{E}[L(g_m)] \rightarrow L^*$ as $m \rightarrow \infty$ (see [3], Theorem 4.5).

- Carey E. Priebe is with the Department of Applied Mathematics and Statistics (AMS), the Center for Imaging Science (CIS), and the Mathematical Institute for Data Science (MINDS), Johns Hopkins University. E-mail: cep@jhu.edu
- Cencheng Shen is with the Department of Applied Economics and Statistics, University of Delaware. E-mail: shenc@udel.edu
- Ningyuan (Teresa) Huang and Tianyi Chen are with the Department of Applied Mathematics and Statistics, Johns Hopkins University. E-mail: nhuang19@jhu.edu, tchen94@jhu.edu

This work was supported in part by the Defense Advanced Research Projects Agency under the D3M program administered through contract FA8750-17-2-0112, the National Science Foundation HDR TRIPODS 1934979, the National Science Foundation DMS-2113099, the University of Delaware Data Science Institute Seed Funding Grant, and by funding from Microsoft Research. The authors thank Wade Shen for providing the motivation for this investigation, and also thank the editor and reviewers for valuable comments and suggestions that significantly improved the exposition of the paper.

2.2 A Latent Position Graph

The generative model of interest in this paper involves

$$(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{iid}{\sim} F_{XY}$$

with $(X_i, Y_i) \in \mathbb{R}^d \times \{0, 1\}$. It is required that $X_i^\top X_j \in [0, 1]$, i.e., the inner product is a valid probability. Rather than observing the X_i 's, we observe a latent position graph $G = (V, E)$ on n nodes with (binary, symmetric, hollow) adjacency matrix A , that is generated by

$$A_{ij} \stackrel{ind}{\sim} \text{Bernoulli}(X_i^\top X_j)$$

for $i < j$. This is the so-called random dot product graph (RDPG); see [4] for a recent survey.

For node classification, we observe the $n \times n$ matrix A , and the class labels Y_i are observed for only m nodes with $m < n$. The task is to classify the remaining $n - m$ nodes. This is a simple case of the problem studied in [5].

2.3 Graph Embedding

Given the $n \times n$ matrix A and $\{Y_1, \dots, Y_m\}$, we approach the node classification task via unsupervised graph embedding followed by subsequent linear classification. Specifically, we consider embedding, denoted $h : A_{n \times n} \rightarrow (\mathbb{R}^{d'})^n$, via two approaches: the adjacency spectral embedding (the scaled leading eigenvectors $U_{d'} | \Sigma_{d'} |^{1/2}$ of A) [6], and the graph convolutional network (a stack of layers, each of which consists of a graph convolution followed by a pointwise nonlinearity; see Section 4.1 for details) [7], [8]. Other graph embedding methods, not germane to our investigation but popular nonetheless, include Laplacian spectral embedding [9], node2vec [10], non-negative matrix factorization [11], [12], etc. See [13] for an overview.

For either ASE or unsupervised GCN, we denote $\hat{X} = \{\hat{X}_1, \dots, \hat{X}_n\}$ as the embedding. A classifier $\hat{g}(\cdot)$ is trained on $\{(\hat{X}_i, Y_i)\}_{i \in \{1, \dots, m\}}$. For $i > m$ the classified label is $\hat{g}(\hat{X}_i)$, and performance is measured via $\mathbb{E}[L(\hat{g})]$. Note that d' is the embedding dimension we choose (e.g., the embedding dimension in ASE, or the neuron size of the last layer in GCN), while d as in Section 2.2 is the true latent position dimension.

Here $\mathbb{E}[L(g_m)]$ is the target performance – the performance we would achieve if we observed the latent positions themselves. We cannot hope to do better, via an argument reminiscent of the data processing lemma, as the inner products $X_i^\top X_j$ are corrupted through the Bernoulli noise channel into A_{ij} . RDPG theory for ASE into dimension $d' = \text{rank}(\mathbb{E}[A])$ (specifically the $2 \rightarrow \infty$ norm convergence result presented in [14]) implies that $\mathbb{E}[L(\hat{g})] \rightarrow \mathbb{E}[L(g_m)]$ as $n \rightarrow \infty$. In other words, ASE is consistent for the classification task.

However, to the best of our knowledge, there is no such guarantee of consistency for GCN despite its empirical success in applications. Theoretical work for GCN has been focused on its expressive power [2], [15], [16], [17], [18], [19], [20], [21], [22], stability [23] and transferability [24], [25]. As pointed out in [26], most impossibility results (such as [16], [17], [19]) are based on *anonymous* GCN where nodes have the same attributes. In contrast, *non-anonymous* GCN with discriminative node features are provably more powerful,

which could be Turing universal given sufficient capacity and the right learning procedure [26]. Indeed, the learning procedure is crucial, as some graph neural networks generalize better than others [2]. Recent studies also suggest that graph neural networks can be simplified as linear models without negatively impacting their performance [27]. The GCN algorithm studied in this paper can be viewed as *non-anonymous* GCN embeddings trained from two different learning procedures: unsupervised versus semisupervised.

3 FAILURE GEOMETRY

In this section we present a latent position model such that unsupervised GCN fails. We shall start with an auxiliary variable Z , then construct the latent position variable X using *srt* (scale-rotate-translate) transformation. The *srt* transformation is the key to the desired failure geometry, where the first eigenvector of A is orthogonal to the support of X (i.e., the classification signal). It changes the graph topology but preserves the latent distribution.

3.1 The *srt* Transformation

Given a univariate $Z \in [0, 1]$, we define the *srt* transformation $\text{srt}(\cdot) : [0, 1] \rightarrow \mathbb{R}^2$:

$$X = \text{srt}(Z) = \begin{bmatrix} \cos(r) & -\sin(r) \\ \sin(r) & \cos(r) \end{bmatrix} \begin{bmatrix} sZ \\ 0 \end{bmatrix} + t,$$

where $s \in \mathbb{R}_+$ is the scalar term, $r \in [0, 2\pi)$ is the rotation angle, and $t \in \mathbb{R}^2$ is the translation term. For X to be a valid latent position for RDPG (namely $X_i^\top X_j \in [0, 1]$), the range of (s, r, t) needs to be a proper subset of $\mathbb{R}_+ \times [0, 2\pi) \times \mathbb{R}^2$.

As such, we may assume without loss of generality that the marginal density f_X is supported on $S \subset \mathbb{R}_+^2 \cap B(0, 1)$ (unit ball of dimension 2). Note that $\mathbb{E}[A_{\text{srt}}]$ is positive semidefinite and, except for very special values for (s, r, t) for which S is on a ray from the origin, we have $\text{rank}(\mathbb{E}[A_{\text{srt}}]) = 2$. Both L^* and $\mathbb{E}[L(g_m)]$ remain unchanged under the *srt* transformation.

3.2 Visualize the Latent X

Figure 1 illustrates the failure mode for GCN. The density f_X is supported on the line segment $S \subset \mathbb{R}_+^2 \cap B(0, 1)$. There are two angles of interest: $\theta_\perp$ is the angle between $\mathbb{R}_+$ and the ray from the origin perpendicular to S ; θ_E is the angle between $\mathbb{R}_+$ and the ray from the origin to $\mathbb{E}[X]$. When $\theta_E = \theta_\perp$, the first eigenvector of $\mathbb{E}[A_{\text{srt}}]$ is orthogonal to S and has no classification signal.

In Figure 1, we let $f_{X|0} = \text{srt}(\text{Beta}(10, 3))$, $f_{X|1} = \text{srt}(\text{Beta}(4, 3))$, $(s, r, t) = (1.1, 53\pi/32, [0.13, 0.97]^\top)$, and class prior probability being 0.5 (so each class is equally likely). In this case, $\theta_\perp = 5\pi/32$ and $\theta_E = 3\pi/16$ and hence $|\theta_E - \theta_\perp| = \pi/32$, which is designed to be almost orthogonal but not exactly.

The *srt*-transformed graph is approximately regular with almost constant node degree. As shown in the simulation section, unsupervised GCN (and also ASE into dimension $d' = 1$) is only able to recover the first eigenvector, thus performs poorly for A_{srt} . This phenomenon may be generalized to other approximately regular graphs whose inference signal lies on non-leading eigenvectors.

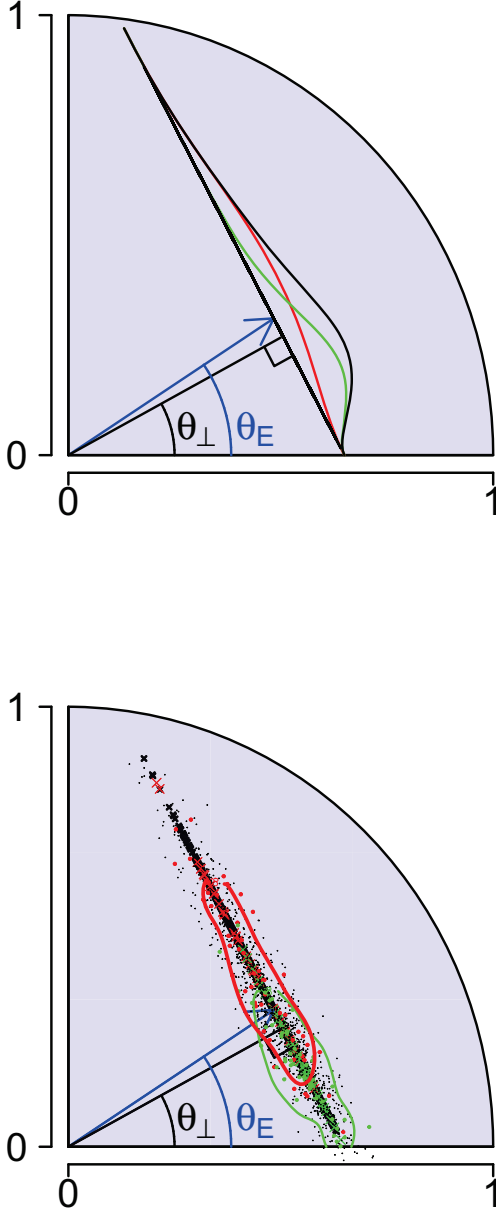


Fig. 1. Geometry for the canonical case where ASE succeeds but GCN fails. The top panel shows the density: the black line represents density of f_X , while the green line and red line represent the class-conditional density of $f_{X|0}$ and $f_{X|1}$. The bottom panel shows one sample data generation with $n = 1000$ and $m = 100$: the green crosses and red crosses are training points from $f_{X|0}$ and $f_{X|1}$ respectively with known label, while the black crosses are the remaining test points. Round dots are ASE after Procrustes with same color setting. Contours are from kernel density estimation of red dots and green dots. For this case, indicated by the vertical line at $|\theta_E - \theta_\perp| = \pi/32$ in Figure 2, both ASE into dimension $d' = 1$ and unsupervised two-layer GCN perform poorly while ASE into two dimensions has nearly optimal performance.

4 FAILURE SIMULATION

In this section, we first introduce GCN in both unsupervised and semi-supervised formulations, followed by the node

classification results under A_{srt} with varying parameter settings, then provide simulations to demonstrate eigenvector recovery in approximately regular graphs.

4.1 Graph Convolutional Networks

In this section we describe the architecture of unsupervised GCN and semisupervised GCN. The former is directly comparable with ASE as they produce graph embedding without utilizing any additional node information, thus the focus of comparison here.

We consider the two-layer GCN model [7] designed as

$$GCN(\mathbf{X}, A) = \tilde{A} \text{ReLU}(\tilde{A} \mathbf{X} W_0) W_1 \quad (1)$$

where $\mathbf{X}$ is the node feature matrix, $\tilde{A} = D^{-1/2}(A + I)D^{-1/2}$ is the symmetrically normalized adjacency matrix, and D is the node degree matrix. Since our latent position graph does not have node features, the input $\mathbf{X}$ is set to the $n \times n$ identity matrix by default, which is equivalent to a *non-anonymous* GCN. Alternatively, by replacing ReLU activation by the identity activation [27], it becomes a linear GCN.

Unsupervised GCN

Unsupervised GCN (Variational Graph Auto-Encoder [8]) estimates the latent graph embedding matrix $\hat{\mathbf{X}}$ using a variational auto-encoder architecture without seeing any labels. The variational inference model of $\hat{\mathbf{X}}$ is given by

$$q(\hat{\mathbf{X}}|\mathbf{X}, A) = \prod_{i=1}^n q(\hat{X}_i|\mathbf{X}, A)$$

with $q(\hat{X}_i|\mathbf{X}, A) = \mathcal{N}(\hat{X}_i|\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2))$. In our featureless setting where $\mathbf{X} = I$, the parameters of the inference model $\boldsymbol{\mu}, \boldsymbol{\sigma}$ (the matrix of mean vectors $\boldsymbol{\mu}_i$ and variance vectors $\boldsymbol{\sigma}_i$ respectively) are estimated by

$$\boldsymbol{\mu} = GCN_{\boldsymbol{\mu}}(\mathbf{X}, A) = \tilde{A} \text{ReLU}(\tilde{A} W_0) W_1^{\boldsymbol{\mu}} \quad (2)$$

$$\log \boldsymbol{\sigma} = GCN_{\boldsymbol{\sigma}}(\mathbf{X}, A) = \tilde{A} \text{ReLU}(\tilde{A} W_0) W_1^{\boldsymbol{\sigma}} \quad (3)$$

where $GCN_{\boldsymbol{\mu}}$ and $GCN_{\boldsymbol{\sigma}}$ share the first layer weights W_0 and learn the second layer weights $W_1^{\boldsymbol{\mu}}, W_1^{\boldsymbol{\sigma}}$ independently.

Unsupervised GCN is trained by reconstructing the input graph. It optimizes the variational lower bound $\mathcal{L}$ with respect to $W_0, W_1^{\boldsymbol{\mu}}, W_1^{\boldsymbol{\sigma}}$:

$$\mathcal{L}_{unsup} = \mathbb{E}_{q(\hat{\mathbf{X}}|\mathbf{X}, A)} [\log p(A|\hat{\mathbf{X}})] - KL[q(\hat{\mathbf{X}}|\mathbf{X}, A)||p(\hat{\mathbf{X}})] \quad (4)$$

where the $KL[q(\cdot)||p(\cdot)]$ term represents the Kullback-Leibler divergence between $q(\cdot)$ and $p(\cdot)$. We assume a Gaussian prior $p(\hat{\mathbf{X}}) = \prod_i \mathcal{N}(\hat{X}_i|0, 1)$. In the training stage, the embedding is estimated using the reparameterization trick [28] as $\hat{\mathbf{X}} = \boldsymbol{\mu} + \epsilon \boldsymbol{\sigma}$. After training, the graph embedding $\hat{\mathbf{X}} = \boldsymbol{\mu}$ is used for node classification, as described in Section 2.3.

Semisupervised GCN

The semisupervised GCN produces the class probabilities as

$$P = \text{softmax}(GCN(\mathbf{X}, A)), \quad (5)$$

trained with m labels via cross-entropy loss

$$\mathcal{L}_{semisup} = - \sum_{i=1}^m \sum_{k=1}^K Y_{ij} \log P_{ij} \quad (6)$$

and evaluated on the remaining $n - m$ labels. Besides deep learning, other types of semisupervised learning methods exist, typically involved constructing the graph from the input data [29]; and see [30] for a survey.

4.2 Experiment 1

Here we consider the latent graph A_{srt} as in Section 3 and Figure 1. Everything else being the same, we vary the srt transformation as follows: Let $(s, r, t) = (1.1, 53\pi/32, t)$, where t ranges linearly from $[0.25, 0.97]^\top$ to $[0.02, 0.97]^\top$ and preserves the inner product constraint for RDPG. As a result, $|\theta_E - \theta_\perp|$ ranges from 0 to $\pi/14$, i.e., from exact orthogonality to less orthogonal between the first eigenvector of $E[A_{srt}]$ and the support of X . Figure 2 shows the classification performance of each method at $m = 100$ and $n = 1000$ and increasing angle difference.

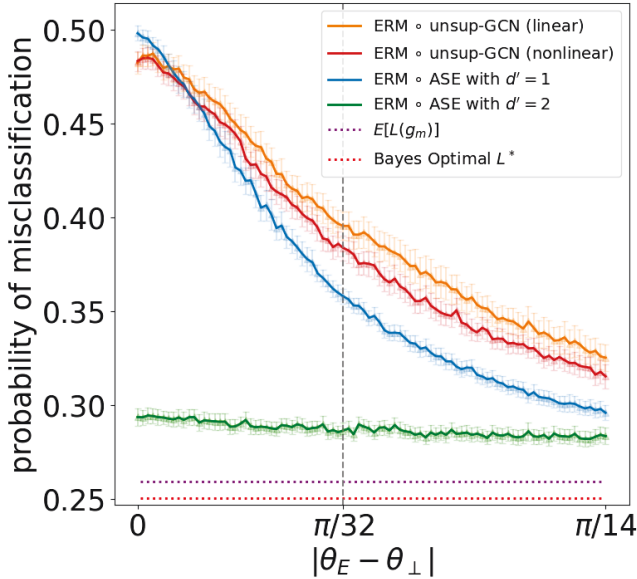


Fig. 2. Performance for the canonical case where ASE succeeds but two-layer unsupervised GCN fails. ERM in the legend stands for empirical risk minimization classifier. We run 100 Monte Carlo replicates and report the average classification error.

For the given A_{srt} , the optimal Bayes error is $L^* \approx 0.25$ and $\mathbb{E}[L(g_{100})] \approx 0.26$ regardless of t . ASE into two dimensions always works well, while both unsupervised GCN and ASE into one dimension yield much worse performance when $|\theta_E - \theta_\perp| \approx 0$. As the angle becomes less orthogonal, the first eigenvector of A_{srt} captures more classification signal, and all methods perform better.

Note that nonlinear GCN (Relu activation) performs similarly as linear GCN (linear activation). In this experiment, unsupervised GCN (either linear or nonlinear) uses the two-layer GCN model with hidden size 2 (because the underlying latent variable lies in $\mathbb{R}^2$), learning rate 0.01, no weight decay, and a maximum of 500 epochs with early stopping at a patience of 100 epochs.

4.3 Experiment 2

The same phenomenon always holds regardless of the training data ratio m/n . Figure 3 summarizes the classification error at fixed $n = 500$, increasing m/n in $\{0.02, 0.05, 0.1, 0.3, 0.5, 0.7\}$, using three different model geometries with 100 Monte Carlo replicates. As m increases, every method has slightly better classification error, yet ASE at $d = 2$ exhibits consistent advantages against unsupervised GCN.

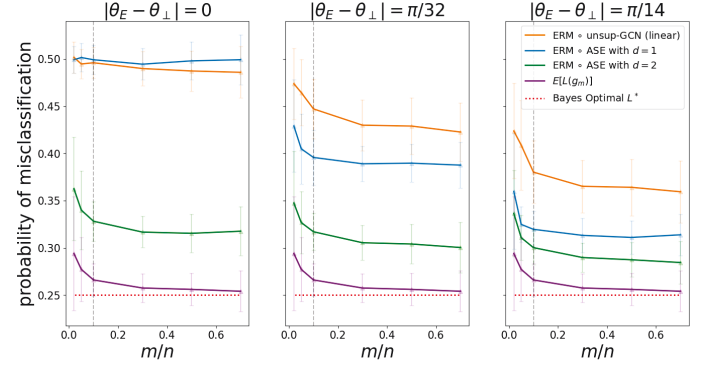


Fig. 3. Effects of changing m/n on the classification performance for different methods. Vertical dash line indicates $m/n = 0.1$ used in Figure 2.

Note that ASE into $d' = 2$ always performs well because it selects the correct embedding representation under the latent position model. ASE into $d' = 1$ is sub-optimal because it only selects the first eigenvector — the same reason why unsupervised GCN fails (see Section 4.5). As angle increases, ASE with $d' = 1$ slightly outperforms unsupervised GCN, because neural networks typically has larger estimation variance. Moreover, Relu activation performs slightly better than linear activation for the same reason: as the graph itself and the first eigenvector are all non-negative in each entry, Relu shall have less estimation variance than linear.

4.4 Experiment 3

The unsupervised GCN results presented in Section 4.2 are obtained without hyper-parameter search. To show the phenomenon is invariant to parameters, we further present the results using cross validation on the hyperparameters. We also include semisupervised GCN in this experiment, which improves on the unsupervised version but does not outperform ASE with $d' = 2$ (again due to the larger estimation variance of neural networks).

We consider the following parameter in GCN:

- Hidden size: $[2, 4, 8, 16, 32]$
- Learning rate: $[0.01, 0.001]$.

The hyper-parameter search is repeated on 3 different weight initializations, and 20% validation set from the training data. We pick the model with the best average training accuracy, and report the testing error.

The results are presented in Figure 4: there is little change to the performance of unsupervised GCN after optimizing for hyper-parameters (orange stars). The semisupervised GCN performs slightly worse than ASE into two dimensions

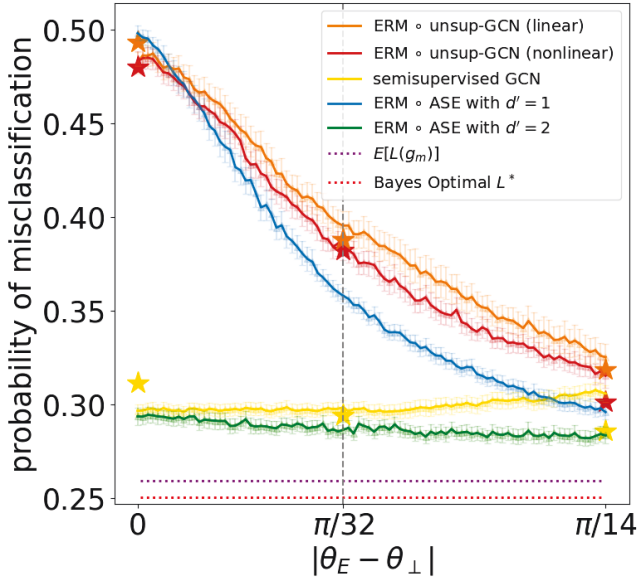


Fig. 4. Adding parameter cross validation results for unsupervised GCN, and include semisupervised GCN. The curves for GCN models are obtained without hyper-parameter search, while the stars represent the best results after hyper-parameter search.

(yellow line for the no cross validation case, and yellow star for cross validated case).

4.5 Experiment 4

In this experiment, we numerically demonstrate that unsupervised GCN only recovers the first eigenvector in certain approximately regular graphs. We consider two settings: the first setting is the same latent position graph A_{srt} as in Figure 1; then the second setting uses a simple Erdos-Renyi graph where $A_{ij} \stackrel{i.i.d.}{\sim} \text{Bernoulli}(0.1)$, for which the degree of each node is about $\frac{n}{10}$.

In each setting, we compute the unsupervised GCN embedding at output and hidden dimension 4, then calculate the angle difference between each embedding dimension and each eigenvector of the graph adjacency. The results are summarized in Table 1 for the top three eigenvectors: all GCN dimensions are very close to the leading eigenvector (which is a scalar multiple of ones), and almost exactly orthogonal to the remaining eigenvectors. As ASE at $d' = 1$ is the first eigenvector (up-to a scalar), and ASE at $d' = 2$ includes the first two eigenvectors, Table 1 can also be viewed as the angle difference between unsupervised GCN and ASE, i.e., GCN embedding is always very similar to ASE at $d' = 1$.

The phenomenon is the same throughout small to large d , and the embedding is always orthogonal to all remaining eigenvectors of A . This illustrates why unsupervised GCN fails in the proposed latent position model A_{srt} , and potentially any approximately regular graph whose inference signal lies beyond the first eigenvector.

5 DISCUSSION

The purpose of this short paper is to characterize a spectral failure mode for graph convolutional networks in the

TABLE 1
Computing the angle difference (rounded to two digits) between unsupervised GCN embedding and the eigenvectors of approximately regular graphs at $n = 1000$. GCN embedding dimensions are always very close to the first eigenvector, and almost orthogonal to later eigenvectors.

The Latent Position Graph				
GCN	Dim 1	Dim 2	Dim 3	Dim 4
Eig 1	0.98	0.98	0.99	0.98
Eig 2	90	90	90	90
Eig 3	90	90	90	90
An Erdos-Renyi Graph				
GCN	Dim 1	Dim 2	Dim 3	Dim 4
Eig 1	2.8	2.8	2.9	2.8
Eig 2	90	90	90	90
Eig 3	90	90	90	90

context of the classical statistical learning framework. The failure mode in a latent position graph comes down to: 1) the classification signal lies mostly beyond the leading eigenvector of the graph adjacency matrix; 2) the graph embedding mainly recovers the leading eigenvector. The fundamental idea and geometry insights are presented in its fullest simplicity.

The failure mode can be generalized to a wide variety of settings, such as multivariate X , more complicated generative model and discriminant boundary, etc. Take multivariate X as an example: one may consider the random variable $X \in \mathbb{R}^d$ (after srt transformation) supported on $S \subset \mathbb{R}_+^d \cap B(0, 1)$ (unit ball of dimension d). V_E is a ray from the origin to $E(Z)$. When $V_E \perp S$, the srt -transformed random dot product graph is still an approximately regular graph with classification signal on the non-leading eigenvectors.

The stochastic block model [31] offers another special case of our RDPG followed by srt model, and from [32] Lemma 3 we see that our model encompasses the mixed membership stochastic block model, as well as the degree corrected stochastic block model [33]. Furthermore, the failure mode can easily be extended to the case of $\text{rank}(E[A]) > 2$ with multiple non-leading eigenvectors, and to weighted and/or directed graphs.

Moreover, the use case considered herein — a large number of nodes n but a small number of labeled nodes m — is relevant in many applications, such as text classification [7], traffic prediction [34], molecular property prediction [35]. The failure mode has significant implications in molecular applications, as many molecules are regular graphs, e.g., decaprismane $C_{20}H_{20}$ and dodecahedrane $C_{20}H_{20}$ are examples of 3-regular graphs, see Figure 3 in [36].

Finally, it is not surprising that there exist specific tasks for which GCN perform worse than other methods — every embedding method has its own bias, and fails when the bias of the task at hand does not coincide with the methods bias. For example, the Laplacian embedding and the adjacency embedding are both consistent, but one may work better than another depending on the graph structure [37]. Moreover, when spectral embedding provides a good representation and the training sample size is small, then

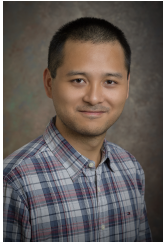
it is likely that the added GCN embedding complexity will prove to be a hindrance in terms of the bias-variance trade-off. As noted in [3]: “Simple rules survive.” This short paper provides the foundation for further study: 1) design GCN architectures and learning procedure that are provably better; 2) investigate more general settings to compare GCN and spectral embedding approaches.

REFERENCES

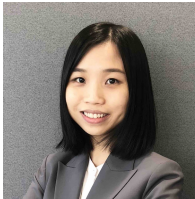
- [1] P. Goyal and E. Ferrara, “Graph embedding techniques, applications, and performance: A survey,” *Knowledge-Based Systems*, vol. 151, pp. 78–94, 2018.
- [2] K. Xu, J. Li, M. Zhang, S. S. Du, K. ichi Kawarabayashi, and S. Jegelka, “What can neural networks reason about?” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=rjxbjeHFPS>
- [3] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, ser. Stochastic Modelling and Applied Probability. Springer New York, 1997.
- [4] A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, Y. Qin, and D. L. Sussman, “Statistical inference on random dot product graphs: a survey,” *Journal of Machine Learning Research*, vol. 18, no. 226, pp. 1–92, 2018.
- [5] M. Tang, D. L. Sussman, and C. E. Priebe, “Universally consistent vertex classification for latent positions graphs,” *Annals of Statistics*, vol. 41, no. 3, pp. 1406–1430, 2013.
- [6] D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe, “A consistent adjacency spectral embedding for stochastic blockmodel graphs,” *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1119–1128, 2012.
- [7] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
- [8] —, “Variational graph auto-encoders,” *arXiv preprint arXiv:1611.07308*, 2016.
- [9] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
- [10] 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*, 2016, pp. 855–864.
- [11] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [12] Q. Wang, X. He, X. Jiang, and X. Li, “Robust bi-stochastic graph regularized matrix factorization for data clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [13] W. L. Hamilton, “Graph representation learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 14, no. 3, pp. 1–159.
- [14] J. Cape, M. Tang, and C. E. Priebe, “The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics,” *Annals of Statistics*, vol. 47, no. 5, pp. 2405–2439, 10 2019.
- [15] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “Computational capabilities of graph neural networks,” *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 81–102, 2009.
- [16] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?” in *Proc. ICLR*, 2019, pp. 1–17.
- [17] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, “Weisfeiler and leman go neural: Higher-order graph neural networks,” *Association for the Advancement of Artificial Intelligence*, 2019.
- [18] Z. Chen, S. Villar, L. Chen, and J. Bruna, “On the equivalence between graph isomorphism testing and function approximation with gnns,” in *Advances in Neural Information Processing Systems*, 2019, pp. 15868–15876.
- [19] Z. Chen, L. Chen, S. Villar, and J. Bruna, “Can graph neural networks count substructures?” *arXiv preprint arXiv:2002.04025*, 2020.
- [20] S. S. Du, K. Hou, R. R. Salakhutdinov, B. Póczos, R. Wang, and K. Xu, “Graph neural tangent kernel: Fusing graph neural networks with graph kernels,” in *Advances in Neural Information Processing Systems*, 2019, pp. 5723–5733.
- [21] Q. Li, Z. Han, and X.-M. Wu, “Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning,” in *The Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI, 2018.
- [22] K. Oono and T. Suzuki, “Graph neural networks exponentially lose expressive power for node classification,” in *International Conference on Learning Representations*, 2019.
- [23] F. Gama, J. Bruna, and A. Ribeiro, “Stability properties of graph neural networks,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 5680–5695, 2020.
- [24] R. Levie, W. Huang, L. Bucci, M. M. Bronstein, and G. Kutyniok, “Transferability of spectral graph convolutional neural networks,” 2020.
- [25] L. Ruiz, L. F. O. Chamon, and A. Ribeiro, “Graphon neural networks and the transferability of graph neural networks,” 2020.
- [26] A. Loukas, “What graph neural networks cannot learn: depth vs width,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=B112bp4YwS>
- [27] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, “Simplifying graph convolutional networks,” in *International Conference on Machine Learning*, 2019, pp. 6861–6871.
- [28] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [29] Y. Yuan, X. Li, Q. Wang, and F. Nie, “A semi-supervised learning algorithm via adaptive laplacian graph,” *Neurocomputing*, vol. 426, pp. 162–173, 2021.
- [30] X. J. Zhu, “Semi-supervised learning literature survey,” 2005.
- [31] P. Holland, K. Laskey, and S. Leinhardt, “Stochastic blockmodels: First steps,” *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.
- [32] P. Rubin-Delanchy, C. E. Priebe, and M. Tang, “Consistency of adjacency spectral embedding for the mixed membership stochastic blockmodel,” *arXiv preprint arXiv:1705.04518*, 2017.
- [33] Y. Zhao, E. Levina, and J. Zhu, “Consistency of community detection in networks under degree-corrected stochastic block models,” *Annals of Statistics*, vol. 40, no. 4, pp. 2266–2292, 2012.
- [34] A. Rahimi, T. Cohn, and T. Baldwin, “Semi-supervised user geolocation via graph convolutional networks,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2009–2019. [Online]. Available: <https://www.aclweb.org/anthology/P18-1187>
- [35] Z. Hao, C. Lu, Z. Huang, H. Wang, Z. Hu, Q. Liu, E. Chen, and C. Lee, “Asgn: An active semi-supervised graph neural network for molecular property prediction,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 731–752.
- [36] R. Sato, “A survey on the expressive power of graph neural networks,” *arXiv preprint arXiv:2003.04078*, 2020.
- [37] C. Priebe, Y. Parker, J. Vogelstein, J. Conroy, V. Lyzinski, M. Tang, A. Athreya, J. Cape, and E. Bridgeford, “On a ‘two truths’ phenomenon in spectral graph clustering,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 13, pp. 5995–5600, 2019.



Carey E. Priebe received the BS degree in mathematics from Purdue University in 1984, the MS degree in computer science from San Diego State University in 1988, and the PhD degree in information technology (computational statistics) from George Mason University in 1993. From 1985 to 1994 he worked as a mathematician and scientist in the US Navy research and development laboratory system. Since 1994 he has been a professor in the Department of Applied Mathematics and Statistics at Johns Hopkins University. His research interests include computational statistics, kernel and mixture estimates, statistical pattern recognition, model selection, and statistical inference for high-dimensional and graph data. He is a Senior Member of the IEEE, an Elected Member of the International Statistical Institute, a Fellow of the Institute of Mathematical Statistics, and a Fellow of the American Statistical Association.



Cencheng Shen received the BS degree in Quantitative Finance from National University of Singapore in 2010, and the PhD degree in Applied Mathematics and Statistics from Johns Hopkins University in 2015. He is assistant professor in the Department of Applied Economics and Statistics at University of Delaware. His research interests include testing independence, correlation measures, dimension reduction, and statistical inference for high-dimensional and graph data.



Ningyuan (Teresa) Huang received the BS degree in Statistics from the University of Hong Kong, the MS degree in Data Science from New York University, and is currently pursuing her PhD in the Department of Applied Mathematics and Statistics at Johns Hopkins University. She is interested in representation learning, the theory of deep learning, and inter-disciplinary research in data science.



Tianyi Chen received the BS degree in Statistics from Renmin University of China, and is currently pursuing his PhD in the Department of Applied Mathematics and Statistics at Johns Hopkins University.