
SpeAr: A Spectral Approach for Zero-Shot Node Classification

Ting Guo¹, Da Wang², Jiye Liang^{2*}, Kaihan Zhang¹, Jianchao Zeng¹

1. Data Science and Technology, North University of China, Taiyuan, China.
2. School of Computer and Information Technology, Shanxi University, Taiyuan, China.

Abstract

Zero-shot node classification is a vital task in the field of graph data processing, aiming to identify nodes of classes unseen during the training process. Prediction bias is one of the primary challenges in zero-shot node classification, referring to the model's propensity to misclassify nodes of unseen classes as seen classes. However, most methods introduce external knowledge to mitigate the bias, inadequately leveraging the inherent cluster information within the unlabeled nodes. To address this issue, we employ spectral analysis coupled with learnable class prototypes to discover the implicit cluster structures within the graph, providing a more comprehensive understanding of classes. In this paper, we propose a **Spectral Approach** for zero-shot node classification (SpeAr). Specifically, we establish an approximate relationship between minimizing the spectral contrastive loss and performing spectral decomposition on the graph, thereby enabling effective node characterization through loss minimization. Subsequently, the class prototypes are iteratively refined based on the learned node representations, initialized with the semantic vectors. Finally, extensive experiments verify the effectiveness of the SpeAr, which can further alleviate the bias problem.

1 Introduction

Graph data is widely used to reveal interactions between various entities, such as citation networks [1], social networks [2], recommendation systems [3], etc. In graph data structures, various entities are abstractly represented through the form of nodes, while their complex relationships are precisely depicted through the connections of edges. Node classification [4, 5], an essential task in graph data analysis, focuses on predicting labels for unlabeled nodes by harnessing the relational structure of the graph and a select subset of labeled nodes. The advent of graph neural network (GNN) technology, as delineated in [6, 7, 8], represents a monumental stride in the domain of node classification. This development has accelerated advancements in the analysis of graph-structured data.

Nevertheless, the ongoing dynamics of real-world networks, characterized by the continual emergence of new nodes and edges, facilitate the discovery of new classes, thereby introducing a series of innovative challenges for node classification tasks [9]. Traditional GNN-based node classification models struggle to handle the identification of new classes in dynamic and open environments. For instance, as citation networks continue to grow, the emergence of new research papers promotes the rise of new academic fields. A key challenge is the effective integration of newly emerging nodes into the pre-existing foundational classes or nascent ones currently in development. Therefore,

*Corresponding author. Email: ljy@sxu.edu.cn. This work is supported by the National Science and Technology Major Project (2020AAA0106102), the National Natural Science Foundation of China (62376141), and the Natural Science Foundation of Shanxi Province, China (202203021222075).

constructing a model that can recognize nodes of unseen classes offers significant practical utility for the analysis of graph data.

Zero-shot node classification (ZNC) [10] presents a potent strategy for tackling the recognition of unseen classes that are absent from the labeled nodes. It endeavors to harness external categorical knowledge (e.g. semantic vectors) and comprehensive graph data (e.g. node attributes and structural information), to classify nodes belonging to unseen classes. In ZNC, prediction bias is a significant issue, where models incorrectly predict unseen class nodes as seen classes. Several studies concentrate on constructing optimized node representations and ensuring their alignment with semantic vectors in the latent space. This alignment is achieved by minimizing the distance between nodes and their corresponding class semantic vectors. DGN [10] captures the local and global embedding of graph nodes, devising a distance loss function that bridges the gap between graph representations at varying scales and semantic vectors. MVE [11] underscores the insufficiency of modeling node features from a single view in describing the comprehensive essence of a node, thus advocating for a multi-view approach to augment node features. In addition, the utilization of inter-class relationships is also crucial in ZNC. DBiGCN [12] incorporates class relationships into the node embedding process to facilitate the model's transfer to unseen classes. GraphCEN [13] enhances model adaptation to unseen classes through a dual contrastive loss mechanism that captures node-class dependencies.

Existing research predominantly leverages external knowledge, such as semantic vectors and the category relationships stemming from them, to mitigate prediction bias. However, most methods insufficiently exploit the cluster information inherent within unlabeled nodes. Strategically leveraging this latent information can reveal hidden class cluster structures, thereby facilitating a comprehensive understanding of the categories present within the graph. In this paper, we present a novel method, termed **Spectral Approach for ZNC (SpeAr)**, that incorporates spectral analysis and learnable class prototypes to discover the intrinsic cluster structure embedded within the graph. Specifically, we introduce a spectral contrastive loss for optimizing node embedding and posit that minimizing this loss is approximate to conducting a spectral decomposition of the graph. The spectral contrastive loss can elucidate the intrinsic structure of graph data, ensuring a high degree of discriminability among classes in the latent space. In addition, we initiate the class prototypes by class semantic vectors and iteratively refine these prototypes based on the node embeddings.

In summary, our contribution is three-fold: (1) We propose a spectral contrastive loss to update nodes and establish an approximate relationship between the node embeddings obtained from this loss and the feature vectors derived from spectral decomposition. (2) Taking class semantic vectors as the initial class prototypes, we iteratively refine these prototypes building upon the node embeddings. (3) Experiments demonstrate that our proposed method, SpeAr, can alleviate the bias issue that occurs in existing ZNC tasks, thereby enhancing the model's ability to recognize unseen classes.

2 Related Work

Zero-shot learning. Zero-shot learning [14] has gained attention for its potential for a wide range of applications in fields such as image processing [15], speech recognition [16], and natural language processing [17]. In zero-shot learning models, commonly used external semantic knowledge includes attribute vectors and semantic vectors, among others. In particular, text vectors play an important role in representing textual content, and they translate words, sentences, and even whole paragraphs into vector representations through techniques such as Word2Vec [18], GloVe [19], and BERT [20]. Existing methods focus on establishing mapping relationships between samples and external knowledge for effective knowledge transfer to unseen classes. These include direct mapping methods [21, 22], which work by constructing a mapping function between samples and external knowledge and optimizing it; generation-based methods [15, 23], which train semantic-visual generators to generate samples for the unseen classes; and local embedding-based methods [24, 25], which use the attributes to guide the learning and migration of discriminative local embeddings. While the aforementioned methods have exhibited noteworthy progress in managing regular data types like images, they encounter obstacles in handling ZNC. The distinct structural attributes of graph data often hinder traditional zero-shot learning techniques from capturing and addressing these crucial properties, thereby rendering them ineffective in resolving the ZNC issue.

Node classification. Node classification represents a pivotal task within the domain of graph data analysis. The introduction of GNNs has catalyzed a transformative evolution in the advancement of

node classification techniques [7]. GNNs possess the capability to concurrently process the adjacency structures and feature information of nodes. A multitude of methods predicated on GNNs [26, 27] have been introduced, progressively emerging as the predominant technologies for addressing node classification challenges. These conventional approaches are largely predicated on the assumption that the class labels present in the training data encompass all possible classes. However, the reality is that new categories continually emerge. Regrettably, existing GNN frameworks exhibit notable deficiencies when it comes to effectively handling unlabeled nodes from unseen classes.

3 A Spectral Approach for Zero-shot Node Classification

3.1 Problem Formalization

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ denote a **graph**. \mathcal{V} is the set of nodes with $|\mathcal{V}| = N$, \mathcal{E} is the set of edges. In the graph, the feature matrix is $X \in \mathbb{R}^{N \times d}$, d is the dimension of features. $A \in \{0, 1\}^{N \times N}$ denotes the adjacency matrix, where $a_{ij} = 1$ if $(v_i, v_j) \in \mathcal{E}$, otherwise $a_{ij} = 0$.

For **zero-shot node classification**, $\mathcal{C} = \{c_1, \dots, c_{|\mathcal{C}|}\}$ is the class set, $\mathcal{C}_s = \{c_1, \dots, c_{|\mathcal{C}_s|}\}$ is the seen class set, $\mathcal{C}_u = \{c_1, \dots, c_{|\mathcal{C}_u|}\}$ is the unseen class set, $\mathcal{C}_s \cup \mathcal{C}_u = \mathcal{C}$ and $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$. Given an $x \in X$, we use \mathcal{P} to denote the marginal distributions of all data that $x \sim \mathcal{P}$, and let \mathcal{P}_{c_i} denote the distribution of labeled samples with class label $c_i \in \mathcal{C}$ and \mathcal{P}_u denote those of unlabeled data. In addition, every class has a distinct class semantic vector $s_i \in \mathbb{R}^{d_1}$, $|\mathcal{C}|$ class semantic vectors (CSVs) (Further details are provided in Section 4.1) can be formed into a matrix $S \in \mathbb{R}^{|\mathcal{C}| \times d_1}$. In graph \mathcal{G} , there are N_s labeled nodes from seen classes, forming the label matrix $Y_s \in \mathbb{R}^{N_s \times |\mathcal{C}_s|}$. For ZNC, we predict labels for N_u unlabeled nodes from unseen classes, with predictions in $\bar{Y}_u \in \mathbb{R}^{N_u \times |\mathcal{C}_u|}$. Generalized zero-shot node classification (GZNC) predicts labels for N_{su} unlabeled nodes, including both seen and unseen classes, with predictions in $\bar{Y}_{su} \in \mathbb{R}^{N_{su} \times |\mathcal{C}|}$.

For **node embedding**, we employ the GNN [7] model for node representations. Specifically, given the adjacency matrix A and the feature matrix X , aggregating information from neighboring nodes, the node embedding is:

$$g(X) = \sigma(D^{-1/2}AD^{-1/2}XW_1), \quad (1)$$

where $D_{ii} = \sum A_i$ is the diagonal matrix of node degrees. W_1 denotes the parameters of a single-layer neural network, and $\sigma(\cdot)$ is activation function. In addition to aggregating information about neighboring nodes, this study also emphasizes the importance of information about the nodes themselves in category mining. Therefore, we input each node into the neural network to obtain its representation, the embedding of nodes in the latent space is formulated as follows:

$$f(X) = \sigma(D^{-1/2}AD^{-1/2}XW_1) + \sigma(XW_2), \quad (2)$$

where W_2 denotes the parameters of a single-layer neural network.

3.2 Preliminaries of Spectral Decomposition

Given a graph structure, we primarily employ spectral decomposition as a key technique to derive principled embeddings. Spectral decomposition, a robust mathematical instrument, has long exhibited its unparalleled value in clustering algorithms. Specifically, the spectral clustering methodology [28, 29, 30] utilizes spectral decomposition [31, 32] to learn the embeddings of samples in a designated space, subsequently enabling efficient cluster analysis through algorithms like K -means [28].

The spectral decomposition can be succinctly expressed as $\tilde{A} = Q\Lambda Q^T$, wherein $\tilde{A} = D^{-1/2}AD^{-1/2}$, Q denotes an orthogonal matrix encapsulating the complete set of eigenvectors of \tilde{A} . Λ signifies a diagonal matrix whose diagonal entries are constituted by the eigenvalues of \tilde{A} , corresponding to the eigenvectors in Q . Let $\lambda_1, \lambda_2, \dots, \lambda_k$ denote the top- k eigenvalues, and q_1, q_2, \dots, q_k represent the corresponding top- k eigenvectors. We define $F^* = [q_1, q_2, \dots, q_k]^T \in \mathbb{R}^{N \times k}$ as the matrix of eigenvector moments, which serves as a novel, condensed representation of the sample. These eigenvectors embody the most significant directions within the data. Let z_i be the i^{th} row of the matrix F^* . It turns out that z_i can serve as desirable node embeddings of x_i .

3.3 Spectral Zero-shot Node Representation Learning

Spectral decomposition, by leveraging the relationships between data points, effectively unearths the implicit class cluster structures within a graph, thereby obtaining node representations with class discriminability. One of the fundamental challenges of ZNC is identifying unlabeled nodes of unseen categories. These unlabeled nodes contain cluster information, and utilizing the information is important for enhancing the model's recognition and understanding of unseen categories. By revealing the latent class cluster structures in the graph through spectral decomposition, we offer a novel perspective for the ZNC.

It is imperative to consider not only the relationships between nodes within the unlabeled set but also the labeled information in the seen classes. On this basis, we reconfigure the adjacency matrix A by integrating the labeled information, so that $A = \alpha A^l + \beta A$, where A^l represents the label relation matrix. We use $A_{xx'}$ to denote the entries of the reshaped A , where $A_{xx'} = 0$ represents that x and x' neither belong to the same category nor have an adjacency relationship. Otherwise, $A_{xx'} \neq 0$ indicates that at least one of the two conditions is met.

$$A_{xx'} = \alpha \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x' \sim \mathcal{P}_{c_i}} A_{xx'}^l + \beta \mathbb{E}_{x \sim \mathcal{P}_u, x' \sim \mathcal{P}_u} A_{xx'}, \quad (3)$$

and thus, we have $A_x = \sum_{x' \in X} A_{xx'}$. In this section, based on the reshaped adjacency matrix A and its normalization \tilde{A} , we refer to the Eckart-Young-Mirsky theorem [33], where the loss of solving F^* can be formalized as:

$$\min_{F \in \mathbb{R}^{N \times k}} \mathcal{L}_{sd}(F, \tilde{A}) = \|\tilde{A} - FF^\top\|_F. \quad (4)$$

Now, we view \mathbf{f}_x^\top of F as a scaled version of learned feature embedding $f : X \mapsto R^k$. \mathcal{L}_{sd} can be formulated as a variant of the contrastive learning objective, enabling us to theoretically establish the approximation between the learned node representations and the top- k singular vectors of \tilde{A} . We formalize the approximation in Theorem 3.1.

Theorem 3.1. *We define $\mathbf{f}_x = \sqrt{A_x} f(x)$ for some function f , α, β are hyper-parameters. Then minimizing the loss function $\mathcal{L}_{sd}(F, \tilde{A})$ is equivalent to minimizing the following loss function for f , which we term spectral contrastive loss,*

$$\mathcal{L}_{scl}(f) \triangleq -2\alpha \mathcal{L}_1(f) - 2\beta \mathcal{L}_2(f) + \alpha^2 \mathcal{L}_3(f) + 2\alpha\beta \mathcal{L}_4(f) + \beta^2 \mathcal{L}_5(f), \quad (5)$$

where

$$\begin{aligned} \mathcal{L}_1(f) &= \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x^+ \in \{x' | A_{xx'} \neq 0, x' \sim \mathcal{P}_{c_i}\}} f(x)^\top f(x^+), \\ \mathcal{L}_2(f) &= \mathbb{E}_{x \sim \mathcal{P}_u, x^+ \in \{x' | A_{xx'} \neq 0, x' \sim \mathcal{P}_u\}} f(x)^\top f(x^+), \\ \mathcal{L}_3(f) &= \sum_{c_i \in \mathcal{C}} \sum_{c_j \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x^- \in \{x' | A_{xx'} = 0, x' \sim \mathcal{P}_{c_j}\}} [(f(x)^\top f(x^-))]^2, \\ \mathcal{L}_4(f) &= \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x^- \in \{x' | A_{xx'} = 0, x' \sim \mathcal{P}_u\}} [(f(x)^\top f(x^-))]^2, \\ \mathcal{L}_5(f) &= \mathbb{E}_{x \sim \mathcal{P}_u, x^- \in \{x' | A_{xx'} = 0, x' \sim \mathcal{P}_u\}} [(f(x)^\top f(x^-))]^2. \end{aligned}$$

Proof. (sketch) We can expand $\mathcal{L}_{sd}(F, A)$ and obtain

$$\begin{aligned} \mathcal{L}_{sd}(F, A) &= \sum_{x, x' \in X} \left(\frac{A_{xx'}}{\sqrt{A_x A_{x'}}} - \mathbf{f}_x^\top \mathbf{f}_{x'} \right)^2 \\ &= \text{const} + \sum_{x, x' \in X} \left(-2A_{xx'} f(x)^\top f(x') + A_x A_{x'} (f(x)^\top f(x'))^2 \right). \quad (6) \end{aligned}$$

The first term is a constant. The form of $\mathcal{L}_{scl}(f)$ is derived from plugging $A_{xx'}$ and A_x . \square

Analysis of Theorem 3.1. The \mathcal{L}_1 loss is tailored to reduce the distances between embeddings of label nodes sharing identical class labels, thereby enhancing intra-class compactness. The \mathcal{L}_2 loss targets the unlabeled node pair with the highest adjacency probability as a positive pair, amplifying the discriminative capacity of the embedding space. Conversely, the \mathcal{L}_3 , \mathcal{L}_4 , and \mathcal{L}_5 losses are dedicated to the strategic dispersion of embeddings associated with negative pairs. The \mathcal{L}_3 loss function is calibrated to induce a distinct separation between embeddings of labeled nodes with different class labels. The \mathcal{L}_4 loss treats labeled nodes and unlabeled nodes counterparts as negative pairs. Finally, the \mathcal{L}_5 loss further refines the embedding space by considering all remaining unlabeled node pairs, apart from those identified as positive in the \mathcal{L}_2 loss, as negative pairs. Above all, these intricately designed loss $\mathcal{L}_{scl}(f)$ facilitates the aggregation of similar nodes and the segregation of dissimilar ones within the embedding space significantly enhancing the representational efficacy of the embeddings.

The Theorem 3.1 of this section is primarily adapted from Theorem 4.1 in [29] and Theorem 3.1 in [34]. The main differences from the original theorems are reflected in the following two aspects: First, the types of data objects processed are different; the proposed method focuses on graph data objects that inherently possess structural characteristics. Second, during the construction of the loss function, we define an unlabeled node pair exhibiting the highest adjacency probability as a positive pair. These adjustments make the Theorem 3.1 more aligned with the specific context and requirements of the research presented in this paper.

3.4 Class Prototype Learning

The proposed method employs a meticulously crafted spectral contrastive loss function to ensure the inter-class separability of the embedding vectors obtained in the latent space. To achieve classification of unseen class nodes, we utilize valuable pseudo-label information to perform iterative updates on class prototypes, which serve as the centers for each class. These prototypes are initially instantiated with semantic vectors S that are rich in categorical information. Utilizing these vectors, we perform preliminary classification predictions for the nodes, thereby creating pseudo-labels. Based on these pseudo-labels, the unlabeled nodes with pseudo-label predictions exceeding a preset threshold q are selected to update unseen class prototypes. For labeled nodes with pseudo-label predictions matching the ground-truth label, we use them to refine prototypes for seen classes. Our update rule is defined as:

$$p_c = \begin{cases} (1 - \mu) \cdot p_c + \mu \cdot z & \text{if } \bar{y} = c \text{ and } \Pr(\bar{y}) > q, \bar{y} \text{ is the pseudo-label of } z \\ p_c & \text{otherwise.} \end{cases} \quad (7)$$

Here, we use $\Pr(\bar{y}) > q$ to denote the selected nodes whose pseudo-label predicted probability exceeds threshold q . μ is the updated parameter. This strategy not only significantly enhances the model's generalization capability for unseen classes but also ensures that the classification accuracy for seen classes is maintained and enhanced.

3.5 Training and Testing

Considering the significant modal differences between the initial class prototypes (i.e., semantic vectors) and node embedding, we design a carefully planned two-stage model training approach. This strategy aims to gradually narrow the inter-modal gap through a careful optimization process, leading to improved and refined class prototypes. In the first phase, we focus on roughly tuning the prototype to establish a solid starting point. Subsequently, in the second phase, labeling and adjacency-guided refinement of the prototypes ensures the quality of prototype learning. Through this staged training, our model can achieve a more accurate representation of the class prototypes.

In the first phase, the backbone is pre-trained using unsupervised spectral contrastive loss \mathcal{L}_{uscl} . \mathcal{L}_{uscl} means that the positive samples are the nodes themselves, and the negative samples are selected from other nodes in the graph. **In the second phase**, the model is trained using \mathcal{L}_{scl} .

During the test phase, the embedding z_i of x_i in the latent space is obtained by the network. Based on the learned class prototypes, we predict its label by :

$$c^* = \arg \max_{c \in \mathcal{C}_u / \mathcal{C}} (z \times p_c), \quad (8)$$

$c \in \mathcal{C}_u / \mathcal{C}$ corresponds to ZNC/GZNC tasks.

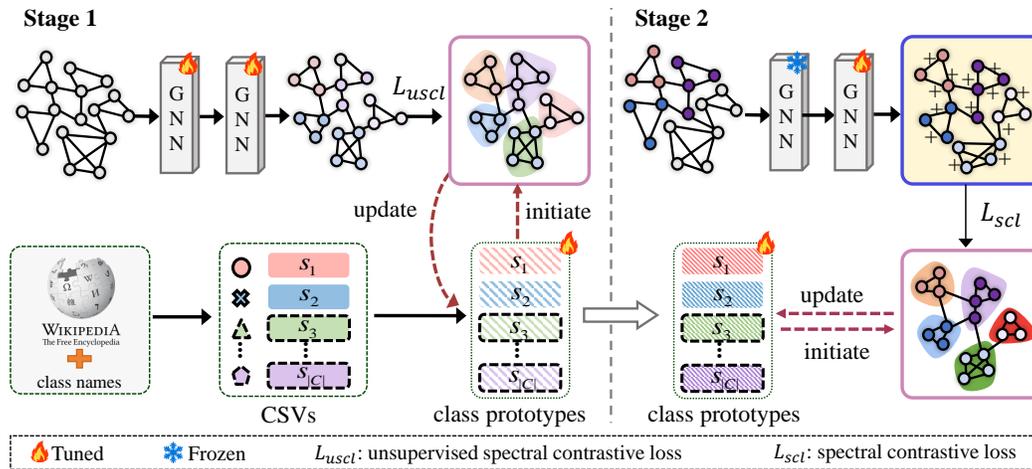


Figure 1: The overall framework for SpeAr. The whole training process consists of two stages.

4 Experiment

4.1 Experimental Setup

Datasets. Following the methods DGN [10], DBiGCN [12], and GraphCEN [13], we seek to substantiate the validity of our proposed SpeAr through its application to three public datasets: Cora [1], Citeseer [1], C-M10M [35]. To ensure equitable comparison, the data partitioning strategy mirrors that of the aforementioned methods. The dataset details are shown in Table 1. For Class Split I and II, the evaluation criterion is classification accuracy. In addition to that, we also give the Class Split III to validate the effectiveness of SpeAr in handling GZNC. The evaluation criterion is H , defined as $H = 2 \times (\text{seen} \times \text{unseen}) / (\text{seen} + \text{unseen})$. The *seen* and *unseen* are the classification accuracies of seen and unseen classes, respectively.

For class semantic vectors (CSVs), DGN has delineated two primary categories: label-based CSVs, which are word embeddings derived from class names, and text-based CSVs, which are document embeddings textual descriptions related to the class. These vectors are extracted utilizing the esteemed natural language processing model, Bert-Tiny [36]. In our experiments, we mainly use text-based CSVs because text contains richer information. Furthermore, we examine the different impacts of employing distinct CSVs on the experimental outcomes.

Baselines Methods. DGN draws upon a suite of comparative methods rooted in traditional zero-shot learning. We follow DGN and list algorithms such as DAP [37], ESZSL [21], ZS-GCN [38], WDVSc [39], Hyperbolic-ZSL [40], DAP and ZS-GCN, as well as their CNN-enhanced counterparts, DAP(CNN) and ZS-GCN(CNN). These conventional methods, predominantly tailored for visual domain data (e.g., image data) prediction are often ill-equipped to handle the graph data, thereby exhibiting limitations in effectively addressing the ZNC problem. Our study places a particular emphasis on evaluating the performance of our proposed SpeAr model against these established methods such as DGN [10], DBiGCN [12], GraphCEN [13], underscoring the superior effectiveness of SpeAr in the context of ZNC.

Parameter Settings. In SpeAr, we employ a two-stage training strategy. During the first phase, the model is trained by the loss \mathcal{L}_{uscl} , which is governed by parameters α and β . In this stage, samples are utilized for updating category prototypes only if the predicted probability of pseudo-labels exceeds a predefined threshold q , with q restricted to the range $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. To ensure the stability of the update process, the prototype update parameter μ is cautiously set to 0.1 to prevent the introduction of erroneous information that might arise from higher values. In the second phase, the loss function \mathcal{L}_{scl} continues with the parameters α and β established in the first phase. Here, we refine the prototype update process by selecting the top- s nodes with the highest pseudo-label probabilities, with s uniformly set to 1000 across all datasets. Furthermore, to achieve fine-tuning of the prototypes, the update parameter μ is further reduced to 0.01 in this phase.

Table 1: Summary of the datasets.

Dataset	Nodes	Edges	Features	Classes	ZNC		GZNC
					Class Split I [Train/Val/Test]	Class Split II [Train/Val/Test]	Class Split III [Train/Val/Test]
Cora	2708	5429	1433	7	[3/0/4]	[2/2/3]	[3/0/7]
Citeseer	3327	4732	3703	6	[2/0/4]	[2/2/3]	[2/0/6]
C-M10M	4464	5804	128	6	[3/0/3]	[2/2/3]	[3/0/6]

4.2 Experimental Results

Table 2 records the ZNC accuracies corresponding to Class Split I and II. A thorough analysis of Table 2 yields the following insights:

a) For Class Split I, SpeAr achieves a notable accuracy enhancement on the Cora, Citeseer, and C-M10M datasets when compared to existing methods. This enhancement underscores the potency of SpeAr’s spectral decomposition and prototype updating tactics. By delving deeper into the intrinsic class cluster structures embedded within unlabeled data, SpeAr significantly amplifies its comprehension and identification of unseen classes, while ensuring good separability between different classes.

(b) For Class Split II, SpeAr shows a significant advantage over the state-of-the-art methods. In this scenario, model parameters are determined using the validation set. This approach is justified within the methodology of our study, as our objective is to thoroughly explore the information across all classes. We believe that a model optimized using validation sets can achieve a comprehensive understanding of the classes and is applicable to some extent to the recognition of unseen classes.

Table 2: Zero-shot node classification accuracy (%).

	Cora	Citeseer	C-M10M	
Class Split I	RandomGuess	25.35	24.86	33.21
	DAP	26.56	34.01	38.71
	DAP (CNN)	27.80	30.45	32.97
	ESZSL	27.35	30.32	37.00
	ZS-GCN	25.73	28.62	37.89
	ZS-GCN (CNN)	16.01	21.18	36.44
	WDVSc	30.62	23.46	38.12
	Hyperbolic-ZSL	26.36	34.18	35.80
	DGPN	33.78	38.02	41.98
	DBiGCN	45.14	40.97	45.45
	GraphCEN	48.43	40.77	44.17
SpeAr (Ours)	60.48	59.72	54.22	
Improve	24.88%	45.77%	19.30%	
Class Split II	RandomGuess	32.69	50.48	49.73
	DAP	30.22	53.30	46.79
	DAP (CNN)	29.83	50.07	46.29
	ESZSL	38.82	55.32	56.07
	ZS-GCN	29.53	52.22	55.28
	ZS-GCN (CNN)	33.20	49.27	51.37
	WDVSc	34.13	52.70	46.26
	Hyperbolic-ZSL	37.02	46.27	55.07
	DGPN	46.40	61.90	62.46
	DBiGCN	49.20	60.11	71.86
	GraphCEN	50.61	60.47	70.83
SpeAr (Ours)	58.20	75.13	79.00	
Improve	15.00%	21.37%	9.94%	

Table 3 records the GZNC accuracies corresponding to Class Splits III. A thorough analysis of Table 3 yields the following insights: We record the performance of the DGPN and DBiGCN in addressing the GZNC problem. Table 3 shows that both DGPN and DBiGCN exhibit suboptimal performance when tackling GZNC. The classification accuracies for unseen classes are less than satisfactory and there is a significant prediction bias. In contrast, SpeAr concurrently learns the prototypes for both seen and unseen classes during its training process. Experimental results indicate that SpeAr achieves a significant enhancement in classification accuracy of unseen classes compared to DGPN and DBiGCN, and gained a better H value. This enhancement indicates that SpeAr bolsters the model’s comprehension of unseen classes through more intensive mining of the clustering information embedded within unlabeled nodes, thereby enhancing inter-class separability and partially mitigating prediction bias. However, we also note that despite the progress made by SpeAr, there is still room for improving its performance.

Table 3: Generalized zero-shot node classification accuracy (%).

	Cora			Citeseer			C-M10M		
	<i>seen</i>	<i>unseen</i>	<i>H</i>	<i>seen</i>	<i>unseen</i>	<i>H</i>	<i>seen</i>	<i>unseen</i>	<i>H</i>
DGPN	94.26	0	0	89.76	0	0	96.71	0	0
DBiGCN	66.39	1.27	2.50	35.70	4.83	8.50	30.04	4.30	7.52
SpeAr(Ours)	34.84	26.32	29.98	27.17	25.78	26.45	26.75	18.95	22.19

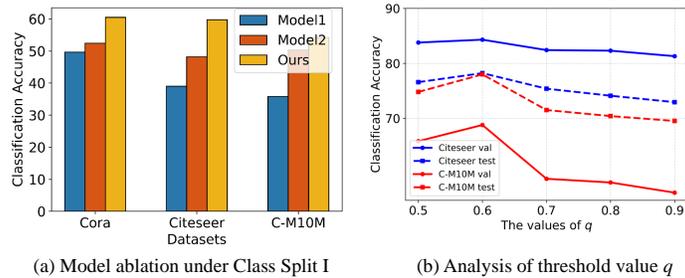


Figure 2: Ablation study and parametric analysis of threshold value q .

Future work will further explore how to optimize the model to achieve superior GZNC classification results.

4.3 Ablation Study

In SpeAr, the loss L_{scl} plays an essential role, as the node embeddings optimized by it approximate the eigenvectors obtained through spectral decomposition. By optimizing the L_{scl} , we not only capture the local connectivity information of the graph but also uncover the implicit class cluster structures within the unlabeled nodes, thereby achieving a comprehensive of the class information depicted in the graph. In Figure 2 (a), the Model1 indicates that upon the removal of the L_{scl} loss from SpeAr, its performance experiences a significant degradation, especially on the Citeseer dataset. This result strongly demonstrates that the L_{scl} , by effectively leveraging the relationships between nodes in the graph, deeply mines the categorical information on the graph and enhances the discriminability between classes.

Furthermore, the mechanism of prototype updating is equally crucial. Random initialization of prototypes can lead to instability in the results. Therefore, this study employs semantic vectors as the initial values for prototypes and devises a two-stage training strategy. In the first phase, we utilize an unsupervised spectral contrastive loss for node embedding. In the second phase, we take the updated prototypes from the first phase as the initial values and further optimize the representations of nodes and prototypes using a supervised spectral contrastive loss. In Figure 2 (a), Model2 is the model that omits the first phase and directly uses semantic vectors as the initial prototypes for the second phase. The results decline across three datasets. This phenomenon further confirms the effectiveness of our designed two-stage training strategy.

4.4 Parametric Analysis

The threshold value q is instrumental in determining the subset of nodes that contribute to the prototype update process. When a sample's pseudo-label probability exceeds the threshold q , it updates the class prototypes. The choice of q depends on model and dataset: low q risks noisy pseudo-labels, degrading performance, high q ensures accuracy but limits unlabeled data utilization. In SpeAr, select q using the validation set. As shown in Figure 2 (b), across three datasets, as q varies, the accuracy changes in both the test and validation sets are roughly consistent, confirming the strategy's effectiveness and reliability.

In SpeAr, L_{scl} includes two pivotal parameters α and β , deciding the relative significance of labeled and unlabeled samples, respectively. As illustrated in the Figure 3, the selection range for α and β is constrained to the set $\{0.25, 0.5, 0.75, 1\}$. We record the effects of α and β on Cora, Citeseer, and C-M10M. The α and β regulate the contribution of labeled and unlabeled samples in overall loss. The

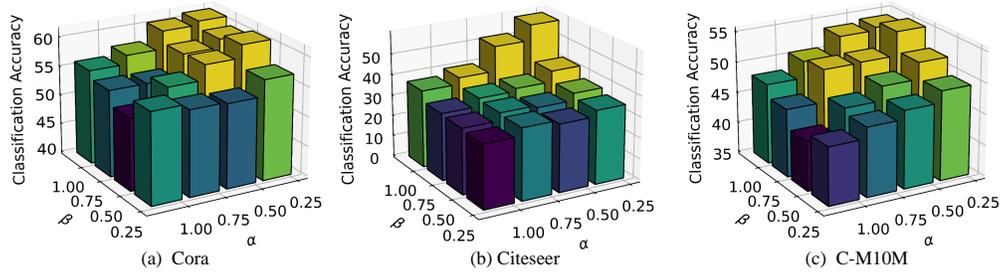


Figure 3: The effects of α and β on Cora, Citeseer, and C-M10M.

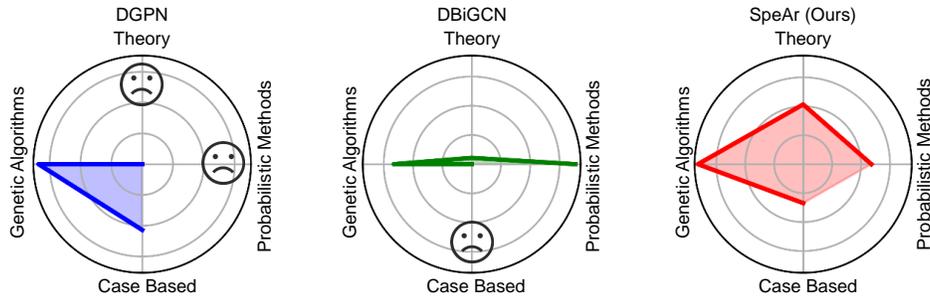


Figure 4: An example of the SpeAr’s effectiveness in mitigating prediction bias on Cora dataset.

results reveal a notable trend: when the value of β is larger than α , the model generally exhibits better performance, with this advantage being particularly prominent when $\beta = 1$. The conclusions meet our expectations. A larger β can reduce overfitting to seen classes and better exploit the clustering information embedded in unlabeled nodes. Thus, the parameters need to satisfy $\beta > \alpha$. This suggests that the α and β are not sensitive to some extent.

4.5 Further Analysis.

Table 3 demonstrates the significant effectiveness of the SpeAr model in mitigating predictive bias, with marked improvements in classification accuracy for both seen and unseen classes. In addition, for Cora, Citeseer, and C-M10M, we find that existing models fail to recognize some of the categories. Initially, we define recall as the proportion of samples correctly classified into a class relative to the total number of samples in that class. As shown in Figure 4, the recall of certain classes is extremely low or even zero for DGNP and DBiGCN on Cora dataset. For instance, the unseen classes “Probabilistic Methods” and “Theory” of Cora exhibit a zero recall rate when predicted by the DGNP model. This reveals the prediction bias issue inherent in current methods when addressing the ZNC problem, where some classes are not correctly identified, and unseen class nodes are erroneously concentrated in a few incorrect classes. In contrast, the SpeAr model provides more accurate classification outcomes for all unseen classes. These results further confirm that the SpeAr model can improve the recognition of unseen classes and enhance class separability by utilizing and mining the class cluster information in unlabeled nodes.

SpeAr has demonstrated exceptional performance in addressing the ZNC problem. We argue that this methodology is equally efficacious when applied to conventional zero-shot learning endeavors, which are predominantly concerned with the identification of unseen classes within the Euclidean space. When applying SpeAr to such tasks, the central challenge lies in effectively constructing the adjacency relationships between samples and integrating them with the loss function proposed in this paper.

5 Conclusion

This paper introduces a spectral method designed to address the challenge of ZSNC, ensuring the effective unearthing of class cluster structures within graphs and enhancing the separability

between classes. Unlike existing approaches that focus on leveraging external knowledge to mitigate predictive bias, our novel approach SpeAr accentuates the exploration of the inherent, implicit cluster information encapsulated within the data of unlabeled nodes. SpeAr optimizes node embeddings by minimizing spectral contrast loss and iteratively updates the class prototypes with semantic vectors as initialization. Empirical results demonstrate that SpeAr achieves significant accuracy improvements in tackling ZNC and GZNC problems, effectively alleviating predictive bias. In future research, we plan to extend the concepts of this study to other tasks with similar training data, thereby further validating its universality and efficacy.

Acknowledgement

Research is supported by the National Science and Technology Major Project (2020AAA0106102), the National Natural Science Foundation of China (62376141), and the Natural Science Foundation of Shanxi Province, China (202203021222075). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements either expressed or implied, of the sponsors.

References

- [1] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.
- [2] Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):226–238, 2023.
- [3] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568, 2020.
- [4] Smriti Bhagat, Graham Cormode, and S Muthukrishnan. Node classification in social networks. *Social network data analytics*, pages 115–148, 2011.
- [5] Qitian Wu, Wentao Zhao, Zenan Li, David P Wipf, and Junchi Yan. Nodeformer: A scalable graph structure learning transformer for node classification. *Advances in Neural Information Processing Systems*, 35:27387–27401, 2022.
- [6] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and deep locally connected networks on graphs. In *International Conference on Learning Representations*, 2014.
- [7] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.
- [8] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2024.
- [9] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.
- [10] Zheng Wang, Jialong Wang, Yuchen Guo, and Zhiguo Gong. Zero-shot node classification with decomposed graph prototype network. In *Proceedings of the ACM SIGKDD conference on knowledge discovery & data mining*, pages 1769–1779. ACM, 2021.
- [11] Jiahui Wang, Likang Wu, Hongke Zhao, and Ning Jia. Multi-view enhanced zero-shot node classification. *Information Processing & Management*, 60(6):103479, 2023.

- [12] Qin Yue, Jiye Liang, Junbiao Cui, and Liang Bai. Dual bidirectional graph convolutional networks for zero-shot node classification. In *Proceedings of the ACM SIGKDD conference on knowledge discovery & data mining*, pages 2408–2417. ACM, 2022.
- [13] Wei Ju, Yifang Qin, Siyu Yi, Zhengyang Mao, Kangjie Zheng, Luchen Liu, Xiao Luo, and Ming Zhang. Zero-shot node classification with graph contrastive embedding network. *Transactions on Machine Learning Research*, 2023.
- [14] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.
- [15] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551. IEEE, 2018.
- [16] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the international conference on machine learning*, volume 162, pages 2709–2720. PMLR, 2022.
- [17] Pratyay Banerjee and Chitta Baral. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–162, 2020.
- [18] Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.
- [19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1532–1543, 2014.
- [20] Vincent Micheli, Martin Martin d’Hoffschmidt, and François Fleuret. On the importance of pre-training data volume for compact language models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 7853–7858, 2020.
- [21] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR, 2015.
- [22] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174. IEEE, 2015.
- [23] Hyeonwoo Yu and Beomhee Lee. Zero-shot learning via simultaneous generating and learning. In *Advances in Neural Information Processing Systems*, pages 46–56. MIT Press, 2019.
- [24] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *Proceedings of the european conference on computer vision*, pages 562–580. Springer, 2020.
- [25] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE international conference on computer vision*, pages 4482–4492. IEEE.
- [26] Lecheng Kong, Jiarui Feng, Hao Liu, Dacheng Tao, Yixin Chen, and Muhan Zhang. Mag-gnn: Reinforcement learning boosted graph neural network. In *Advances in Neural Information Processing Systems*, volume 36, pages 12000–12021. MIT Press, 2023.
- [27] Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, and Doina Precup. In *Advances in Neural Information Processing Systems*, volume 36, pages 28748–28760. MIT Press, 2023.

- [28] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001.
- [29] Yiyun Sun, Zhenmei Shi, and Yixuan Li. A graph-theoretic framework for understanding open-world semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 23934–23967. MIT Press, 2023.
- [30] Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011, 2021.
- [31] Canyi Lu, Jiashi Feng, Zhouchen Lin, Tao Mei, and Shuicheng Yan. Subspace clustering by block diagonal representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):487–501, 2018.
- [32] Ery Arias-Castro, Gilad Lerman, and Teng Zhang. Spectral clustering based on local pca. *The Journal of Machine Learning Research*, 18(1):253–309, 2017.
- [33] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
- [34] Yiyun Sun, Zhenmei Shi, Yingyu Liang, and Yixuan Li. When and how does known class help discover unknown ones? provable understanding through spectral analysis. In *Proceedings of the international conference on machine learning*, volume 202, pages 33014–33043. PMLR, 2023.
- [35] Shirui Pan, Jia Wu, Xingquan Zhu, Chengqi Zhang, and Yang Wang. Tri-party deep network representation. In *International Joint Conference on Artificial Intelligence*, pages 1895–1901, 2016.
- [36] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.
- [37] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013.
- [38] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018.
- [39] Ziyu Wan, Dongdong Chen, Yan Li, Xingguang Yan, Junge Zhang, Yizhou Yu, and Jing Liao. Transductive zero-shot learning with visual structure constraint. In *Advances in Neural Information Processing Systems*, pages 9972–9982. MIT Press, 2019.
- [40] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9270–9278. IEEE, 2020.
- [41] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, 2020.

6 Appendix material

6.1 Proof of Theorem 3.1

Proof. We can expand $\mathcal{L}_{sd}(F, A)$ and obtain

$$\begin{aligned} \mathcal{L}_{sd}(F, A) &= \sum_{x, x' \in X} \left(\frac{A_{xx'}}{\sqrt{A_x A_{x'}}} - \mathbf{f}_x^\top \mathbf{f}_{x'} \right)^2 \\ &= \text{const} + \sum_{x, x' \in X} \left(-2A_{xx'} f(x)^\top f(x') + A_x A_{x'} (f(x)^\top f(x'))^2 \right), \end{aligned} \quad (9)$$

where $\mathbf{f}_x = \sqrt{A_x} f(x)$ is a re-scaled version of $f(x)$. At a high level, we follow the proof in [30], while the specific form of loss varies with the different definitions of positive/negative pairs. The form of $\mathcal{L}_{scl}(f)$ is derived from plugging $A_{xx'}$ and A_x .

Recall that $A_{xx'}$ is defined by

$$A_{xx'} = \alpha \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x' \sim \mathcal{P}_{c_i}} A_{xx'}^l + \beta \mathbb{E}_{x \sim \mathcal{P}_u, x' \sim \mathcal{P}_u} A_{xx'}, \quad (10)$$

thus we have,

$$\begin{aligned} & -2 \sum_{x, x' \in X} A_{xx'} f(x)^\top f(x') \\ &= -2\alpha \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x' \sim \mathcal{P}_{c_i}} A_{xx'} f(x)^\top f(x') - 2\beta \mathbb{E}_{x \sim \mathcal{P}_u, x' \sim \mathcal{P}_u} A_{xx'} f(x)^\top f(x') \\ &= -2\alpha \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x' \sim \mathcal{P}_{c_i}, A_{xx'} \neq 0} f(x)^\top f(x') - 2\beta \mathbb{E}_{x \sim \mathcal{P}_u, x' \sim \mathcal{P}_u, A_{xx'} \neq 0} f(x)^\top f(x') \\ &= -2\alpha \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x^+ \in \{x' | A_{xx'} \neq 0, x' \sim \mathcal{P}_{c_i}\}} f(x)^\top f(x^+) - 2\beta \mathbb{E}_{x \sim \mathcal{P}_u, x^+ \in \{x' | A_{xx'} \neq 0, x' \sim \mathcal{P}_u\}} f(x)^\top f(x^+) \\ &= -2\alpha \mathcal{L}_1(f) - 2\beta \mathcal{L}_2(f). \end{aligned}$$

The penultimate equation is derived from the following lemma:

$$\begin{aligned} \mathcal{L}_{sd}(F) &= \mathcal{L}(f) + \text{const} \\ \text{where } \mathcal{L}(f) &\triangleq -2 \cdot \mathbb{E}_{x, x^+} [f(x)^\top f(x^+)] + \mathbb{E}_{x, x^-} [(f(x)^\top f(x^-))^2]. \end{aligned} \quad (11)$$

Recall that A_x is given by

$$A_x = \sum_{x' \in X} A_{xx'} \quad (12)$$

$$= \alpha \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}} A_x + \beta \mathbb{E}_{x \sim \mathcal{P}_u} A_x. \quad (13)$$

thus plugging A_x and $A_{x'}$ we have,

$$\begin{aligned}
& \sum_{x, x' \in X} A_x A_{x'} (f(x)^\top f(x'))^2 \\
&= \sum_{x, x' \in X} \left(\alpha \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}} A_x + \beta \mathbb{E}_{x \sim \mathcal{P}_u} A_x \right) \cdot \left(\alpha \sum_{c_j \in \mathcal{C}} \mathbb{E}_{x' \sim \mathcal{P}_{c_j}} A_{x'} + \beta \mathbb{E}_{x' \sim \mathcal{P}_u} A_{x'} \right) (f(x)^\top f(x'))^2 \\
&= \alpha^2 \sum_{x, x' \in X} \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}} A_x \sum_{c_j \in \mathcal{C}} \mathbb{E}_{x' \sim \mathcal{P}_{c_j}} A_{x'} (f(x)^\top f(x'))^2 \\
&\quad + 2\alpha\beta \sum_{x, x' \in X} \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}} A_x \mathbb{E}_{x' \sim \mathcal{P}_u} A_{x'} (f(x)^\top f(x'))^2 \\
&\quad + \beta^2 \sum_{x, x' \in X} \mathbb{E}_{x \sim \mathcal{P}_u} A_x \mathbb{E}_{x' \sim \mathcal{P}_u} A_{x'} (f(x)^\top f(x'))^2 \\
&= \alpha^2 \sum_{c_i \in \mathcal{C}} \sum_{c_j \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x^- \in \{x' | A_{xx'} = 0, x' \sim \mathcal{P}_{c_j}\}} [(f(x)^\top f(x'))^2] \\
&\quad + 2\alpha\beta \sum_{c_i \in \mathcal{C}} \mathbb{E}_{x \sim \mathcal{P}_{c_i}, x^- \in \{x' | A_{xx'} = 0, x' \sim \mathcal{P}_u\}} [(f(x)^\top f(x'))^2] \\
&\quad + \beta^2 \mathbb{E}_{x \sim \mathcal{P}_u, x^- \in \{x' | A_{xx'} = 0, x' \sim \mathcal{P}_u\}} [(f(x)^\top f(x'))^2] \\
&= \alpha^2 \mathcal{L}_3(f) + 2\alpha\beta \mathcal{L}_4(f) + \beta^2 \mathcal{L}_5(f).
\end{aligned}$$

□

The proof of Theorem 3.1 is finished.

6.2 Additional Experiments

6.2.1 Zero-shot node classification for large-scale data

For SpeAr, the spectral contrastive loss computes the similarities between samples, with a time complexity of $O(N_s N_s^+ + N_u N_u^+ + N_s N_u + N_s N_s^- + N_u N_u^-)$. Let N_s be the count of labeled nodes, N_s^+ be the count of positive nodes of labeled nodes, and N_s^- the negative nodes. N_u is the count of unlabeled nodes, with N_u^+ and N_u^- representing the count of positive and negative nodes, respectively. This complexity reveals a substantial demand for computational resources, presenting a notable challenge for processing large-scale graph data.

Following GraphCEN [13], we validate the efficacy of the SpeAr on large-scale dataset, such as ogbn-arxiv [41]. Ogbn-arxiv has 169343 nodes, and 2484941 edges. The feature dimension is 128 and the total class number is 40. Class split I is [20/0/20], 20 seen classes as training set, 20 unseen classes as testing set. Class split II is [13/13/14], 13 seen classes as training set, 13 unseen classes as validation set, and 14 unseen classes as testing set. Confronted with memory limitations, we adopt a multi-round subgraph extraction strategy. Specifically, in each round, we extract subgraphs that encompass both seen and unseen class nodes and execute the SpeAr algorithm on these subgraphs. Through this iterative process of extraction, we aim to progressively accumulate performance gains that mirror the execution of SpeAr on the entire graph, all while maintaining computational efficiency. As shown in Table 4, our proposed method SpeAr shows significant improvement in performance metrics compared to existing methods. The comparative analysis in the table highlights the superiority of our method in capturing class-discriminative information in graph structures.

Table 4: A comparative performance analysis of DGPN, DBiGCN, and ours SpeAr for zero-shot node classification on ogbn-arxiv. (%)

	DGPN	DBiGCN	GraphCEN	SpeAr(Ours)
Class Split I	22.37	21.40	23.96	30.45
Class Split II	21.95	25.92	28.36	32.20

Table 5: The Comparison of zero-shot node classification accuracy (%) using the different CSDs.

	Cora			Citeseer			C-M10M		
	TEXT	LABEL	Decline rate	TEXT	LABEL	Decline rate	TEXT	LABEL	Decline rate
DAP	26.56	25.34	-4.59 %	34.01	30.01	-11.76%	38.71	32.67	-15.60%
ESZSL	27.35	25.79	-5.70%	30.32	28.52	-5.94%	37.00	35.02	-5.35%
ZS-GCN	25.73	23.73	-7.77%	28.62	26.11	-8.77%	37.89	33.32	-12.06%
WDVSc	30.62	18.73	-38.83%	23.46	19.70	-16.02%	38.12	30.82	-19.15%
Hyperbolic-ZSL	26.36	25.47	-3.38%	34.18	21.04	-38.44%	35.80	34.49	-3.66%
DGPN	33.78	32.55	-3.64%	38.02	31.83	-16.28%	41.98	35.05	-16.51%
DBiGCN	45.14	39.05	-13.49%	40.97	39.10	-3.10%	45.45	43.71	-3.83%
SpeAr(Ours)	60.48	49.52	-18.12%	59.72	48.88	-18.15%	54.22	47.05	-13.22%

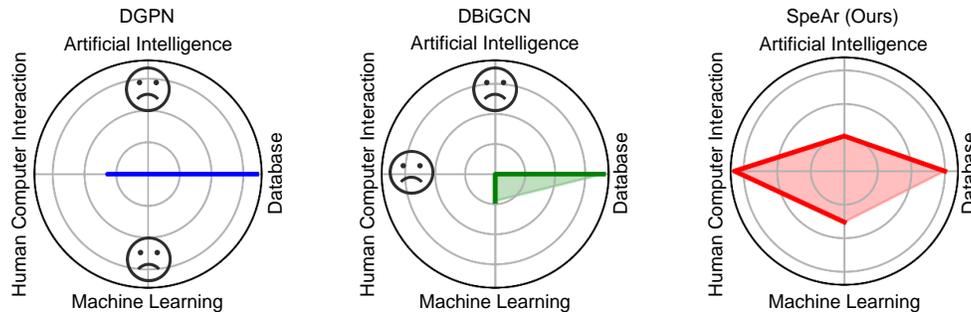


Figure 5: An example of the SpeAr model’s effectiveness in mitigating prediction bias on Citeseer.

6.2.2 Discussion on Different CSVs

The impact of external knowledge from different sources on model outcomes is significantly varied. In Table 5, we individually examined the effects of LABEL-based CSVs and TEXT-based CSVs as external knowledge. Given that TEXT data encapsulates a richer set of categorical information, the SpeAr model utilizing text-based CSVs demonstrates superior performance. Indeed, when employing LABEL-based CSVs as the input external knowledge, SpeAr also outperforms existing methods, further corroborating the efficacy of the spectral contrastive loss and prototype updating mechanisms proposed in this paper for excavating and identifying categories on graphs. This series of results underscore that our approach significantly enhances the discriminability between different classes, thereby elevating the model’s overall recognition capability.

6.2.3 Discussion on SpeAr model’s effectiveness in mitigating prediction bias

We verify the benefits of SpeAr in mitigating prediction bias on the dataset Citeseer. As shown in Figure 5, the recall for certain classes is extremely low or even zero. For instance, the unseen classes “Human Computer Interaction” and “Artificial Intelligence” exhibit a zero recall rate when predicted by the DBiGCN. In contrast, the SpeAr model provides more accurate classification outcomes for all unseen classes.

7 Limitation

Although the SpeAr model shows excellent performance on the ZNC task, its relatively high computational complexity may become a challenge when dealing with large-scale graph data. Especially in application scenarios with limited resources or high real-time requirements, the high computational cost may limit the usefulness of the model. Therefore, we effectively alleviate this problem by adopting the strategy of multi-round subgraph training. The model can gradually learn and integrate information from different subgraphs, thus realizing effective processing of large graph data while maintaining computational efficiency.

8 Experiments Compute Resources

Computation resources: We execute our code on a computer with NVIDIA GeForce RTX 3090 (GPU) and Intel Xeon Gold 6254 (CPU).

9 Societal Impacts

The introduction of SpeAr has made a significant contribution to the advancement of zero-shot node classification tasks. It demonstrates tremendous potential in the field of data analysis, aiding researchers in uncovering new insights and knowledge. There are no negative societal impacts on our work.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Both the abstract and introduction include the claims made in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations of the proposed method in terms of time complexity, see Appendix 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: This paper provide the full set of assumptions and a complete (and correct) proof. This paper proposes Theorem 3.1 in Subsection 3.3 and a complete proof in Appendix 6.1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide details of the algorithm and its implementation in Section 3 and 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the relevant code at github.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: This paper gives all the training and test details in Subsection 4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: This paper reports the zero-shot node classification precision, generalized zero-shot classification results H in Subsection 4.2, and class recall for the Cora dataset in Subsection 4.5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the experiments compute resources of our work in Appendix 8.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed in Appendix 9.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original paper that produced the code package.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide details of new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.