Supplemental Material of Paper "SPINE: Structural Identity Preserved Inductive Network Embedding"

Anonymous Author(s)

Algorithm 1 RootedRandomWalk

Input: the graph G, the present node v_i, the continuation probability $\beta_{\text{RPR}} \in (0, 1)$, the walk length l
Output: a rooted random walk sequence P_s

Initialize a list $P_s = [v_i]$
for $j = 1$ to $l - 1$ **do**
  $v_{cur} \leftarrow$ the last element of P_s
  if random$(0, 1) < \beta_{\text{RPR}}$ **then**
    Randomly select a node v_j from the neighbors of v_{cur}
    Append v_j to the end of P_s
  else
    Append v_i to the end of P_s
  end if
end for
return P_s

Dataset	# Classes	# Nodes	# Edges	# Features
Citeseer	6	3327	4732	3703
Cora	7	2708	5429	1433
Pubmed	3	19717	44338	500
PPI	121	56944	818716	50
FB-686	-	168	3312	63

Table 1: Dataset statistics

1 Rooted Random Walk

Please refer to Algorithm 1 for details of the Monte Carlo approximation of rooted random walk.

2 Dataset Details

We test the proposed model on four benchmark datasets. As most existing methods are transductive, we adopt three static datasets and one across network dataset to measure the transductive and inductive performance of SPINE respectively. The statistics of datasets are summarized in Table 1. Among them, the static datasets are Citation Networks:

Citeseer contains 3312 publications of 6 different classes and 4732 edges between them. Nodes represent papers and edges represent citations. Each paper is described by a one-hot vector of 3703 unique words.

Cora includes 2708 publications of machine learning with 7 different classes. Similar to Citeseer, Cora contains 5429 citation links between them. And each paper is described by a one-hot vector of 1433 unique words.

Pubmed consists of 19717 scientific publications pertaining to diabetes classified into one of three classes. It contains 44338 citation links and each document is described by a TFIDF weighted word vector from a dictionary of 500 unique words.

To test the performance of SPINE while generalizing across networks, we further introduce the PPI dataset:

PPI contains various protein-protein interactions, where each graph corresponds to a different human tissue. Nodes represent proteins with 121 different cellular functions from gene ontology as their labels.

FB-686 is a subset of Facebook dataset, which consists of 168 users with 3312 links between them. Each user is described by a 0-1 vector of 63 features.

3 Hyperparameter Settings

We keep the following settings for all tasks and datasets: ratios λ_1 and λ_2 are set to 0.4 and 0.2 respectively, while the structural rate α and the restart rate β_{RPR} of rooted random walk are both set to 0.5. The learning rate of the Adam optimizer is set to 0.001. For methods that leverage random walks, we set the number of repeats for each node to 10, the length of random walks to 40 and the window size to 5 for a fair comparison.

4 Parameter Study

In order to evaluate the influence of parameters on the performance of SPINE, we conduct experiments on a subset of PPI datasets, denoted as subPPI, which contains 3 training networks and 1 test network. We first investigate the effect of the structural rate α, then test the robustness of SPINE with respect to varying λ_1 and λ_2 values, both on node classification.

4.1 Effect of α

We vary the value of the structural rate α from 0 to 1 with an interval of 0.2, and report the results in Figure 1. As expected, with increasing α values, the classification performance increases first and then decreases when α gets too large. The results show that only considering the structural identity or

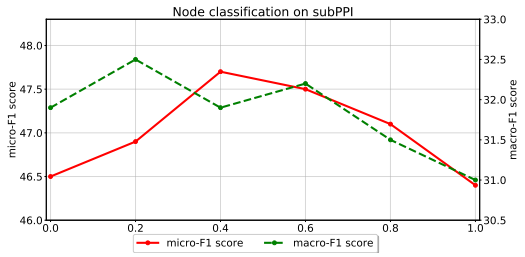

Figure 1: Classification results with different α on subPPI. The left and right vertical axes indicates micro-F1 and macro-F1 score respectively. Both are in percentage.

Table 2: Results of node classification on subPPI with varying λ_1 and λ_2 values (in percentage)

λ_1	λ_2	micro-F1	macro-F1
0.2	0.2	47.1	32.0
0.2	0.4	47.0	31.4
0.2	0.6	46.6	30.5
0.4	0.2	47.2	**32.1**
0.4	0.4	**47.8**	30.5
0.6	0.2	46.7	31.2
1.0	0.0	45.4	30.4

local proximity will impair the quality of learned embeddings in real-world tasks. Therefore, the structural rate α is crucial for enhancing the quality of learned embeddings.

4.2 Effects of λ_1 and λ_2

To investigate the influence of ratios λ_1 and λ_2, we vary the value of λ_1 from 0.2 to 1.0 and the value of λ_2 from 0.2 to 0.6, both with an interval of 0.2. Results of node classification on subPPI are reported in Table 2, from which we can conclude that the interaction between the generator and $\boldsymbol{W}_S$ successfully enhances the structure information carried in the generator as expected, as $\lambda_1 = 1.0$ (no interaction) achieves the worst performance.

5 Case Study

To intuitively analyse the impact of jointly considering local proximity and structural identity on classification performance, we conduct case studies on the Cora dataset, in the transductive setting. Specifically, we select several representative case from nodes that GraphSAGE and node2vec both misclassify. From our observation, the integration of local proximity and structural identity can significantly benefit the classification performance on two types of nodes: nodes with few or lots of neighbors.

We first illustrate cases of nodes with few neighbors in Figure 2. These nodes are similar to the ordinary users we discussed in Introduction. (blue nodes in Figure 1), which are also the majority type of nodes in real-world network dataset. Therefore the classification performance on these nodes has a heavy impact on the overall results. However, normally these nodes have few neighbors, which implies limited local proximity to utilize, inhibiting the methods that only consider local proximity. In Cora, there are 38.3% nodes with no more than 2 neighbors, and this percentage rises to 50.7% among the nodes misclassified by node2vec or GraphSAGE, indicating it is harder for local proximity methods to deal with ordinary nodes than nodes with more neighbors.

Dataset	Pearson (p-value)	Spearman (p-value)
Citeseer	0.72 (0.0)	0.74 (0.0)
Cora	0.77 (0.0)	0.79 (0.0)
Pubmed	0.78 (0.0)	0.84 (0.0)

Table 3: Pearson and Spearman coefficients between structural distance and Euclidean distance for connected node pairs on citation networks.

SPINE handles this issue by introducing structural identities. As illustrated in Figure 2, we list the local structure of top 3 or 5 nodes with highest structural similarity (computed by line 4, Algorithm 3) to the current node. Obviously they have similar local structures, e.g., both have similar numbers of 1-hop and 2-hop neighbors, and more importantly, most of these nodes have the same label with the current node. SPINE successfully detects these nodes as expected. Thus, instead of singly leveraging the current node's local information, we are able to leverage much more structural and local information around these selected nodes. Finally, embeddings with higher quality are learned and better classification performance are achieved.

Another case of nodes that SPINE can better deal with is those with lots of neighbors but in different classes. These nodes act as bridges between different academic fields or communities, also know as the structural hole spanners as discussed in Introduction (red nodes in Figure 1). Therefore, these nodes tend to have many neighbors from different classes, making it hard to judge their class only depending on the local information, as illustrated in Figure 3.

Again, we alleviate this problem by jointly considering local proximity and structural identity. We also list the 3 most similar nodes for each case. From Figure 3, we can observe that the nodes we select not only have similar local structures to the current node, but also have many neighbors in the same class to the current node at the same time. As a benefit of the extra local structure information, better performance is achieved in classifying this kind of nodes, comparing to the local proximity methods.

6 Experiments on Structural Identity

6.1 Real World Tasks

We verify that embeddings learned by SPINE also preserve structural identities in real-world tasks. We compute the correlation between the structural distance (or similarity) defined in line 4, Algorithm 3 and the Euclidean distance in the embedding space for all the connected node pairs. The values of correlation measured by Pearson and Spearman coefficients are listed in Table 3, which indicates that there indeed exists a strong correlation between the two distances, validating that SPINE successfully preserves the defined structural similarity in the embedding space.

6.2 Inductive Setting

To verify whether SPINE can capture the structural identity across networks, we generate four new networks G_1, G_2, G_3

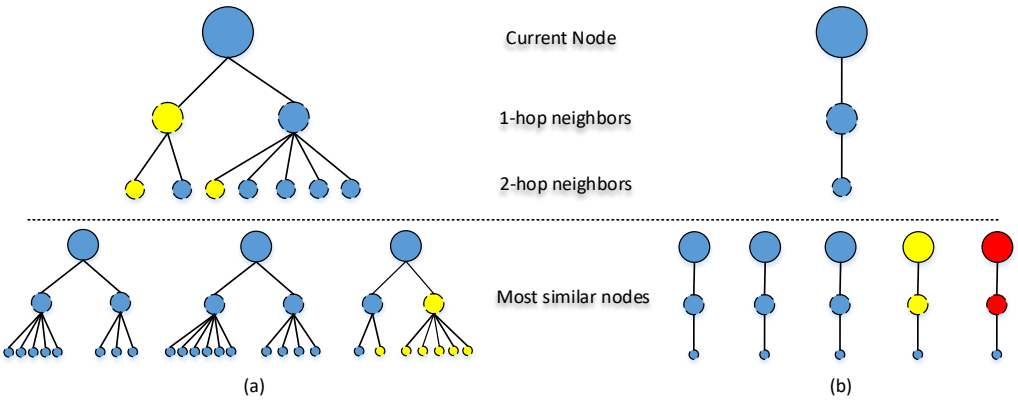

Figure 2: Case study for nodes with few neighbors. The top half are the current nodes and their local structures, and the bottom half are the nodes with highest structural similarities to the current node, as well as their corresponding local structures. Different colors indicate different classes. Both GraphSAGE and node2vec misclassify the current nodes. In (a) we present the current node with two neighbors, while the detected nodes have similar local and class information. Same situation occurs in (b).

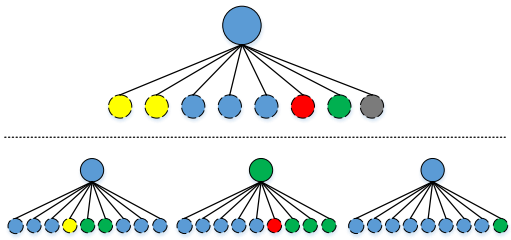

Figure 3: Case study for nodes with many neighbors. Following same settings in Figure 2.

and G_4, from the original FB-686 network with $s = 0.2$. The combination of G_1 and G_2 is considered as training networks, while G_3 and G_4 constitute the test data. After training, the embeddings of nodes in G_3 and G_4 are inferred, on which we compute the distance distributions between embeddings of connected node pairs and mirrored node pairs. Intuitively, as G_1 and G_2 are separated, baselines which only consider local proximities are not able to capture the structural similarity between networks. Results are shown in Figure 4. Although the structural correlation between training and testing networks is small given $s = 0.2$, the two distance distributions learned by SPINE are still strikingly different, indicating that SPINE can learn high-level representations of structural identities from training networks rather than just storing them, which leads to the generalization ability to identify similar structural identities in unseen networks. The two distributions learned from Graphsage are practically identical, justifying our intuition and the necessity of preserving structural identities.

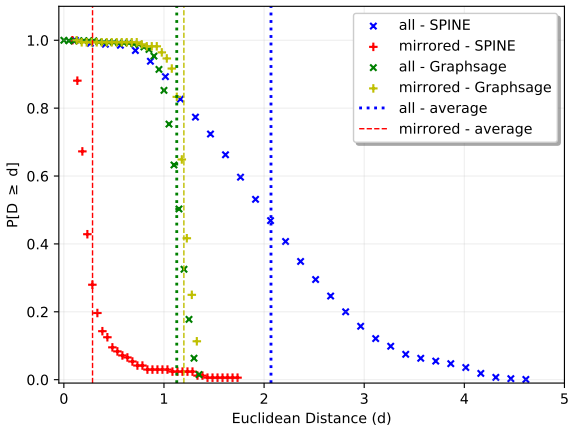

Figure 4: Euclidean distance distribution between inductively trained embeddings of mirrored node pairs and connected node pairs on the perturbed FB-686 with $s = 0.2$.