Appendix

A. More on experiments

A.1. Dataset information

Table 5. Multi-class node classification Dataset statistics as reported in Kipf & Welling (2016).

Dataset	# Nodes	# Edges	# Classes	# Features	Label rate
Citeseer	3,327	4,732	6	3,703	3.6%
Cora	2,708	5,429	7	1,433	5.2%
Pubmed	19,717	44,338	3	500	0.3%

Table 6. Multi-label node classification Dataset statistics

Dataset	# Nodes	# Edges	# Labels	Label type	Graph type
BlogCatalog	10,312	333,983	39	membership	social network
PPI(transductive)	3,890	76,584	50	Bio-states	protein
Wikipedia	4,777	184,812	40	POS-tag	word-net
Amazon	334,863	925,872	58	product type	co-purchasing
PPI(inductive)	56,944	818,716	121	Bio-states	protein

The real-world dataset used are shown in Table 5 and Table 6. The multi-class classification datasets are from Kipf & Welling (2016), where the multi-label classification datasets are from Grover & Leskovec (2016), Hamilton et al. (2017a) and SNAP website. Datasets in Table 5 and also the inductive PPI dataset have extra node features. When available, we use the same train/valid/test split as in original paper.

A.2. Experiments on small graphs

In this section, we compare with baseline algorithms on small benchmark datasets. We show that in the graphs where the diameter is small, existing algorithms can do pretty good, since local information is almost equivalent to global-range information. Nonetheless, our SSE can still achieve comparable performance in this scenario.

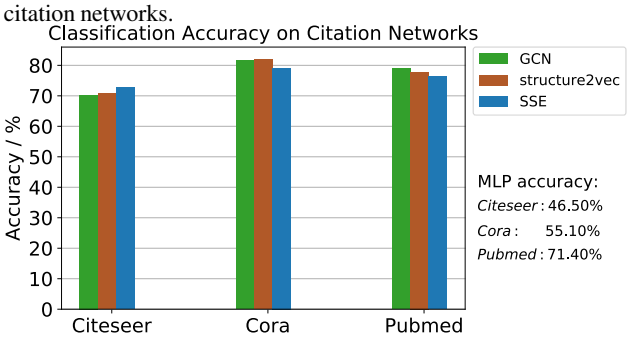
A.2.1. Multi-label classification

Here we compare our proposed method on the multi-label benchmark datasets. We include all the datasets used in Grover & Leskovec (2016) (namely, BlogCatalog (Zafarani & Liu, 2009), Protein-Protein Interactions (PPI) (Breitkreutz et al., 2007) and Wikipedia (Mahoney, 2011)). All the statistics of the dataset can be found in Table 6.

The evaluation metric we used here is Micro-F1 and Macro-F1 score. We tuned the hyperparameters for all the algorithms on 10% of training nodes, and then trained the model on full training set. The dimension of the embedding is set to 128. The results are shown in Table 7. We achieve the best results in Wikipedia, while getting comparable performances on the other two. In dataset like Blogcatalog, a small local neighborhood would be enough to infer the group membership of users in this friendship network, thus our approach would not benefit from taking global-range of information into account. However, in the Wikipedia dataset where we achieves the best Micro-F1 and Macro-F1 scores, it is important to know long range information to get a consensus among POS-tag labeling.

A.2.2. Document Classification

Figure 5. The document classification accuracy on benchmark citation networks.



In this section, we evaluate the performances on several benchmark citation graphs, namely Citeseer, Cora an Pubmed (Sen et al., 2008). The task is to do document classification, where each node in the citation graph represents the corresponding document. Different from the experiment in Section A.2.1, here the documents have auxiliary bag-of-words features. Since the document classes are mutually exclusive, we train all the models with Cross Entropy loss.

The statistics of the datasets are shown in Table 5. The number of features corresponds to the vocabulary size in each dataset. The edges (undirected) are formed by the citation relationship between articles. We use the same training/validation/test splits as in Kipf & Welling (2016). During training, only 20 instances per class are provided with corresponding labels.

We report the test classification accuracy in Figure 5. When possible, we include the baselines' performances directly from previously published results (Kipf & Welling, 2016). From the figure we can see, the proposed SSE performs the best in Citeseer dataset, while being slightly worse than other GNN models in Cora and Pubmed dataset, respectively. We've also include the results that using the node feature with multi-layer perceptron (MLP). The MLP doesn't consider any graph structure into consideration, which serves a sanity check.

Table 7. Multi-label classification in small datasets. We report both Micro-F1 and Macro-F1 on held-out test set.

Blogcatalog	Micro-F1/%									Macro-F1/%								
Methods	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
structure2vec	**35.05**	**36.65**	**38.43**	**39.35**	**40.48**	**40.89**	**42.56**	**42.58**	**42.61**	19.78	22.39	**23**	**25.16**	**25.89**	**26.96**	**26.86**	**27.46**	**27.69**
GCN	36.80	38.42	39.47	40.88	40.88	41.69	42.06	42.43	42.50	19.31	20.96	20.43	22.3	21.86	22.14	23.06	23.2	23.43
SSE	33.90	36.42	36.80	37.39	37.91	37.92	38.58	39.10	40.28	**19.88**	**22.68**	22.88	23.89	23.89	24.08	24.38	25.12	24.99
PPI	Micro-F1/%									Macro-F1/%								
Methods	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
structure2vec	**19.86**	**23.19**	24.73	25.46	25.29	**27.79**	27.75	**28.32**	**28.99**	15.14	15.94	18.32	18.41	19.04	20.41	20.56	22.01	23.83
GCN	18.85	22.52	**25.40**	**26.36**	**26.52**	**27.80**	**27.96**	**28.28**	28.44	**16.03**	17.09	**19.01**	**20.45**	**21.01**	**21.62**	23.50	**23.29**	**24.13**
SSE	19.17	22.04	23.64	23.64	25.24	24.44	26.36	26.20	27.16	15.58	**17.79**	18.36	19.30	**20.99**	20.16	22.64	22.80	22.63
Wikipedia	Micro-F1/%									Macro-F1/%								
Methods	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%
structure2vec	**47.94**	**50.24**	49.98	50.76	**52.45**	**53.54**	53.21	54.07	54.95	11.53	11.63	12.38	13.05	14.12	**16.65**	16.80	**17.37**	17.27
GCN	46.94	49.14	49.61	48.82	49.61	49.92	49.61	51.02	50.55	11.30	11.64	12.41	12.32	13.11	12.98	13.47	13.87	14.34
SSE	46.94	49.76	**51.33**	**51.44**	51.18	52.91	**54.32**	**54.33**	**55.26**	**13.63**	**13.70**	**16.00**	**16.26**	**16.33**	16.41	**17.00**	**17.33**	**17.42**