Semi-Supervised Learning for Clusterable Graph Embeddings with NMF

Anasua Mitra[1], Priyesh Vijayan[2], Srinivasan Parthasarathy[3], Balaraman Ravindran[4]

[anasua.mitra@iitg.ac.in, priyesh@cse.iitm.ac.in, srini@cse.ohio-state.edu, ravi@cse.iitm.ac.in]

[1 - Indian Institute of Technology Guwahati] [2, 4 - Robert Bosch Centre for Data Science & Artificial Intelligence affiliated with Indian Institute of Technology Madras] [3 - Ohio State University]

Motivation & Objective :~

Encoding largely ignored **cluster assumption** of SSL to learn clusterable representations of nodes in a transductive graph based SSL framework.

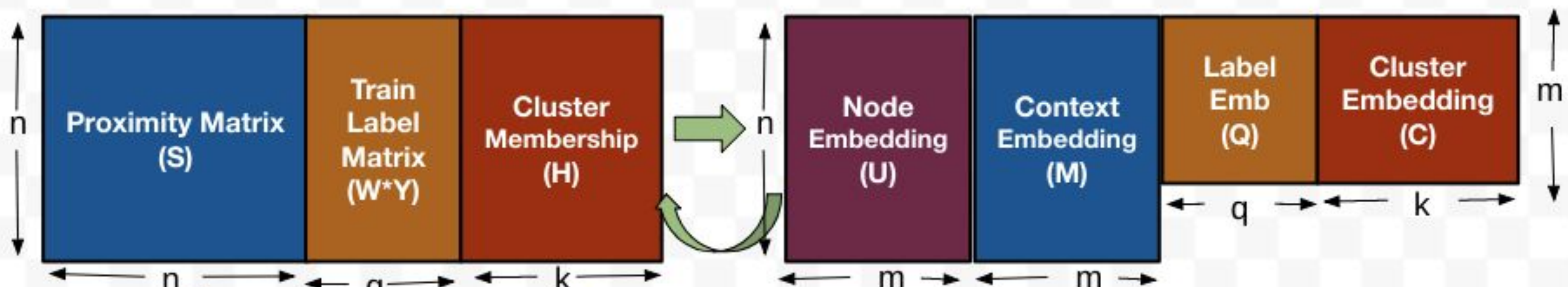
Contribution :~

- **Semi-Supervised Cluster Invariance Property** for nodes ~ clustering nodes with similar labels together.

Components of SSL:~ Well-separated classes, label smoothness assumption, clusterability, manifold assumption.

- **Cluster Assumption:~** If points are in the same cluster, they are likely to be of the same class.

SS-NMF Framework :~



$\|S - U^T M\|^2$

$L_{network}$: 1st + 2nd order proximity factorization

$\|W \odot (Y - QU)\|^2$

L_{label}: Train label matrix reconstruction

$\|H - CU\|^2$

L_{group}: Learn group assignments

$$L = \alpha L_{network} + \theta L_{label} + L_{group} + \lambda(L2_{reg})$$

Constraints:~

- Non-negative factor matrices
- Learn cluster assignment
- Semi-supervised cluster invariance

SS-Cluster Invariance to capture global structure : Nodes having same labels belong to the similar clusters.

$$\sum_{i,j} e'_{ij} \| h(u_i)-h(u_j) \|^2 = HLH^T$$

L = D - E', Graph Laplacian,
E' = $(W*Y)^T(W*Y)$, Cluster Proximity Matrix - CPM

Learn cluster assignment : Each node should be assigned to at-least one cluster. $\| HH^T - I \|^2$

Non-negativity : M, U, Q, C, H >= 0

$$L_{group} = \min_{H,C,U} \beta\|H - CU\|^2 + \phi Tr\{H\mathcal{L}(E)H^T\} + \zeta\|HH^T - I\|^2 : H \geq 0$$

Flow of supervision knowledge from labeled nodes to unlabeled nodes irrespective of their positions in graph !

Experiment Results & Analysis :~

Table 1: Node Classification Results | Micro-F1 Scores

	Non-negative Matrix Factorization models							Sampling
	Proposed	SOTA	Proposed Baseline Variants			SOTA	SOTA	SOTA
Datasets	SS-NMF	MMDW	MFDWL	MF-Plan	MNMFL	MNMF	MFDW	DW
Cora	**85.84**	83.92	83.69	84.38	84.55	82.66	80.37	80.15
Citeseer	**69.75**	67.25	68.62	69.52	69.00	63.57	59.71	57.27
Wiki	**67.02**	66.69	66.18	66.42	66.75	65.75	63.94	63.01
Washington	**66.09**	61.13	62.61	62.83	62.96	62.61	59.13	59.13
Wisconsin	**54.14**	50.67	50.38	52.13	52.76	51.13	49.02	48.12
Texas	**61.70**	56.38	58.51	59.36	57.45	57.45	56.38	58.51
Cornell	**52.04**	51.22	50.94	**52.04**	**52.04**	51.45	50.00	38.78
PPI	23.09	**23.58**	22.19	22.16	21.45	21.23	21.75	22.22
Blogcatalog	36.35	34.72	34.36	34.53	34.88	34.42	32.05	**40.59**
Rank	**1.33**	3.67	5.00	3.33	3.44	5.78	7.11	6.33
Penalty	**0.7267**	2.3144	2.6700	2.0156	2.1856	3.4711	5.4622	5.9700

Table 2: Node Clustering | (O)NMI Scores

	Non-negative Matrix Factorization models							Sampling
	Proposed	SOTA	Proposed Baseline Variants			SOTA	SOTA	SOTA
Dataset	**SS-NMF**	MMDW	MFDWL	MF-Plan	MNMFL	MNMF	MFDW	DW
Cora	**54.40**	36.44	51.38	51.80	53.21	39.29	34.40	34.28
Citeseer	**50.94**	22.61	28.94	48.19	41.19	29.96	17.71	19.04
Wiki	**52.60**	35.68	47.80	47.80	48.38	45.62	28.31	32.57
Washington	**40.27**	13.65	18.45	31.41	33.52	19.90	09.93	02.88
Wisconsin	**31.52**	07.79	06.81	28.38	17.89	11.20	06.09	05.04
Texas	**36.30**	07.63	10.61	28.72	15.14	09.00	02.85	02.70
Cornell	04.77	03.70	04.49	**04.89**	04.14	03.99	04.16	03.53
PPI	**09.76**	08.44	08.26	09.36	09.19	08.77	07.91	09.44
Blogcatalog	**14.31**	06.07	06.18	07.36	08.61	06.93	03.06	03.71
Rank	**1.11**	6.00	4.77	4.67	2.89	4.67	7.11	7.00
Penalty	**0.0133**	16.9977	12.4522	4.1200	7.0800	13.3700	20.0633	20.2000

Experiment Setup :~

- 5 fold cross-validation, 50% train-test split with random and stratified sampling.
- K-means & C-means clustering to detect (non)-overlapping clusters. Logistic Regression classifier for classification.
- **Penalty[model] = Avg(Best[Dataset - Performance[Model][Dataset])**

Stratified Sampling | Micro F1 scores in percentage	**Cora**	**Citeseer**	**Pubmed**
Planetoid	69.1	49.3	66.4
SS-NMF	78.8	50.6	79.6

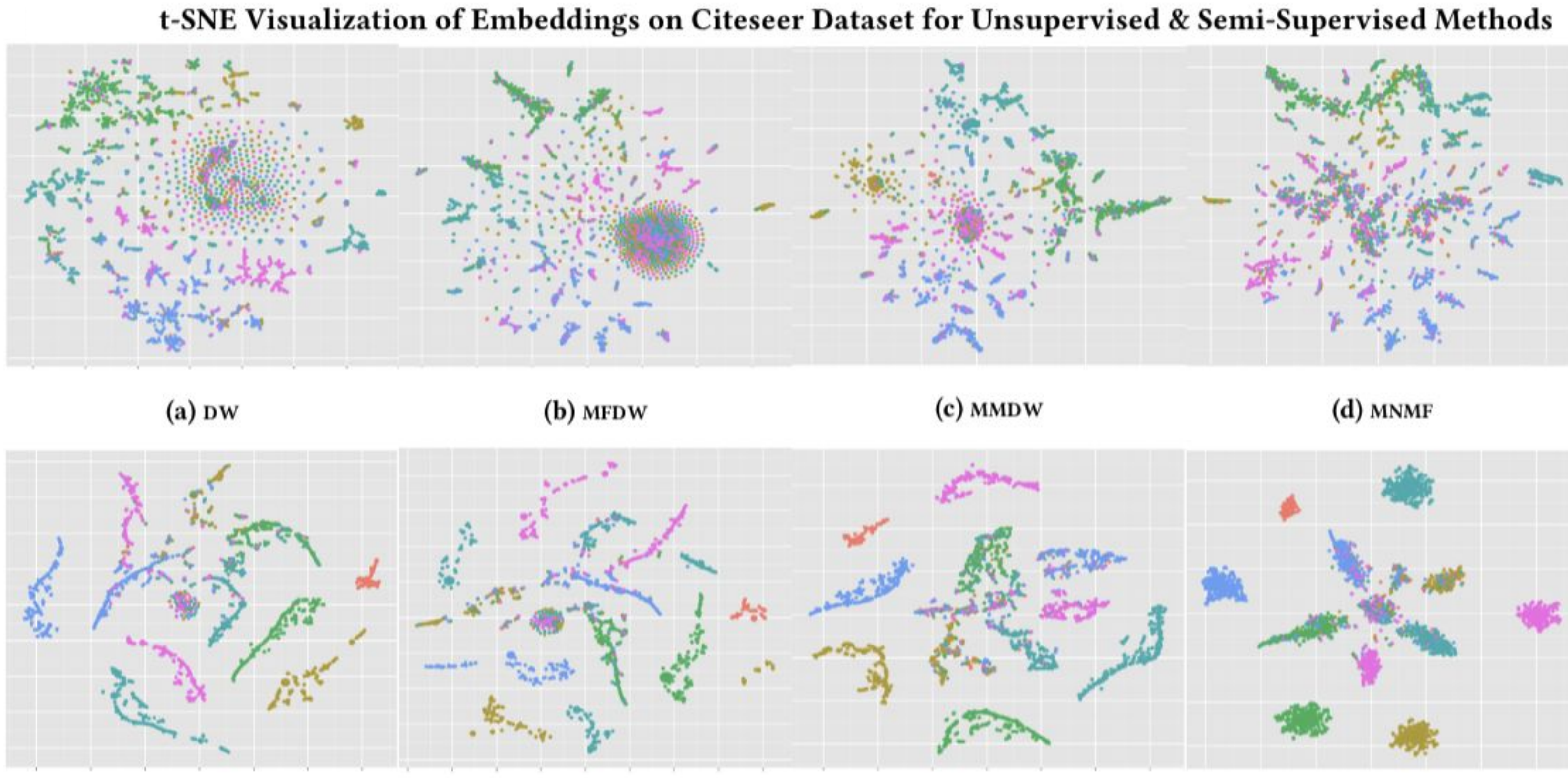
State-of-the-art Results :~

Robust performance (ranks first in 8/10 datasets and ranks second in rest 2/10) across 10 datasets in comparison with 8 baselines for node classification.

Performs outstandingly well in node clustering task with improvement upto 7% over the second best model **MNMFL**.

Well-separated and homophilous clusters !

t-SNE Visualization of Embeddings on Citeseer Dataset for Unsupervised & Semi-Supervised Methods



(a) DW (b) MFDW (c) MMDW (d) MNMF (e) MFDW+Y (f) MF-Plan (g) MNMF+Y (h) SSNMF

References :~

1. Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." *Proceedings of 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.
2. Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016.
3. Tu, Cunchao, et al. "Max-Margin DeepWalk: Discriminative Learning of Network Representation." *IJCAI*. 2016.
4. Wang, Xiao, et al. "Community Preserving Network Embedding." *AAAI*. 2017.
5. Yang, Zhilin, William W. Cohen, and Ruslan Salakhutdinov. "Revisiting semi-supervised learning with graph embeddings." *arXiv preprint arXiv:1603.08861* (2016).