RoSA: A Robust Self-Aligned Framework for Node-Node Graph Contrastive Learning

Yun Zhu[*] , Jianhao Guo[*] , Fei Wu and Siliang Tang[†]
Zhejiang University
{zhuyun_dcd,guojianhao,wufei,siliang}@zju.edu.cn

Code https://github.com/ZhuYun97/RoSA

IJCAI-2022

Reported by Dongdong Hu

Introduction

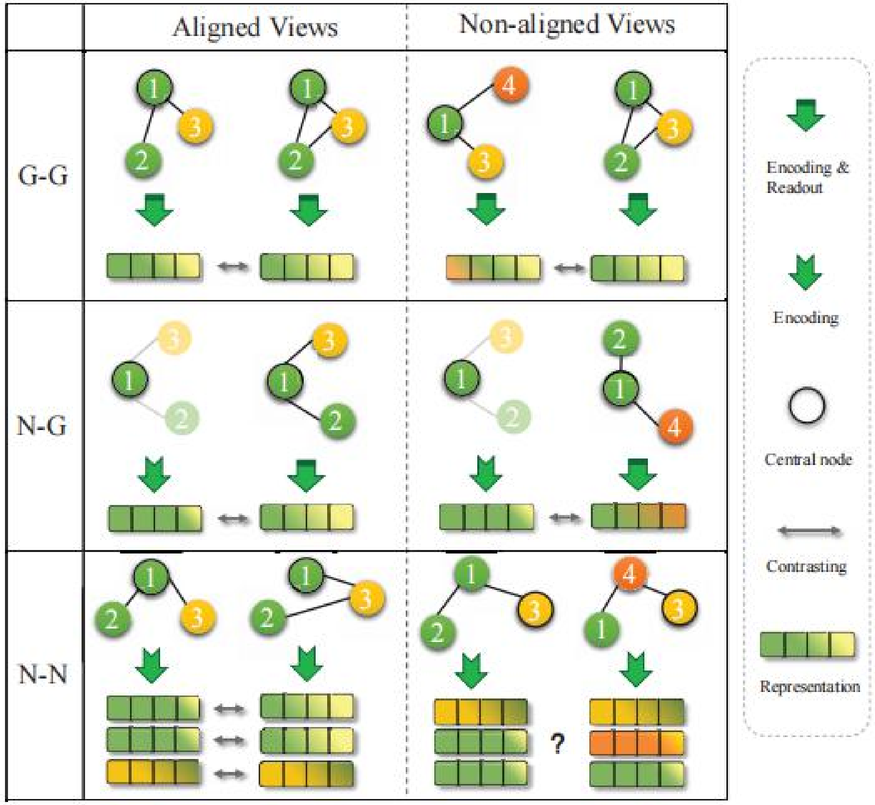


Figure 1: An illustration of different levels of contrasting methods, where G-G means graph-graph, N-G means node-graph and N-N means node-node contrasting level. We only show how a positive pair looks like, where the central node of subgraph is surrounded by a black circle. The number on nodes corresponds to their indices in the original full graph, and the color represents their labels.

Method

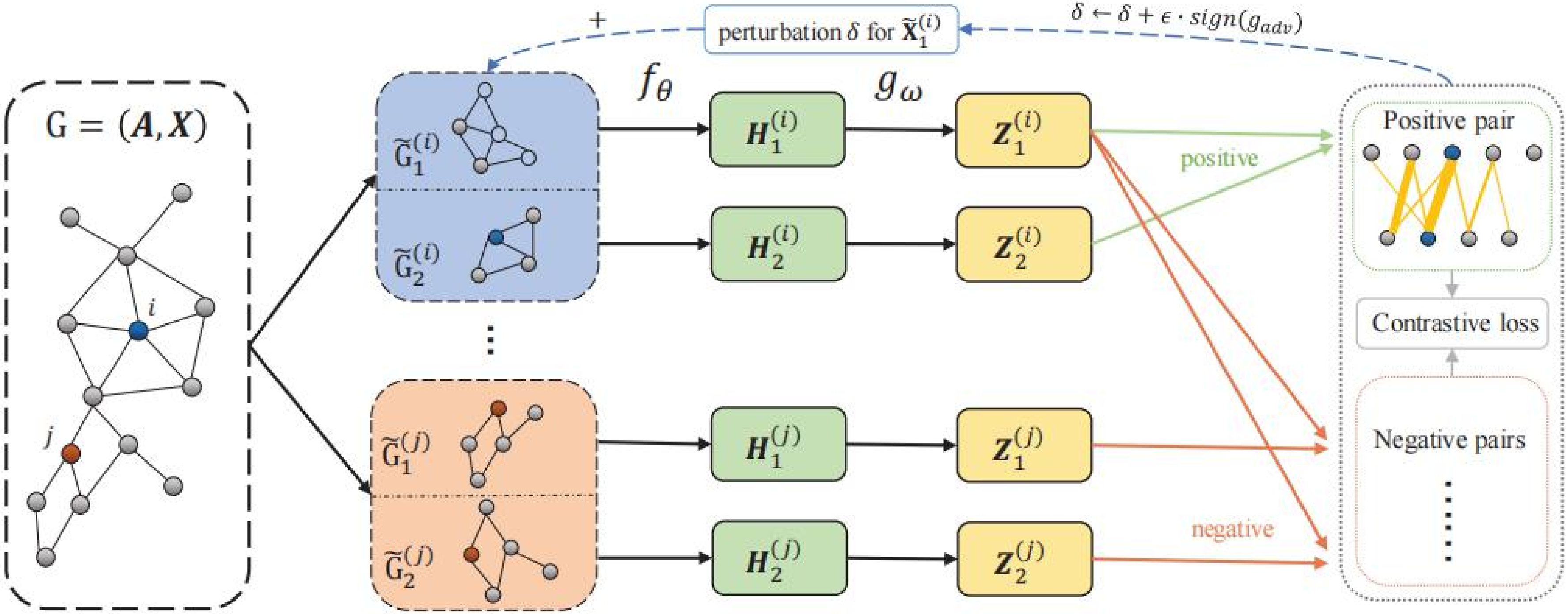
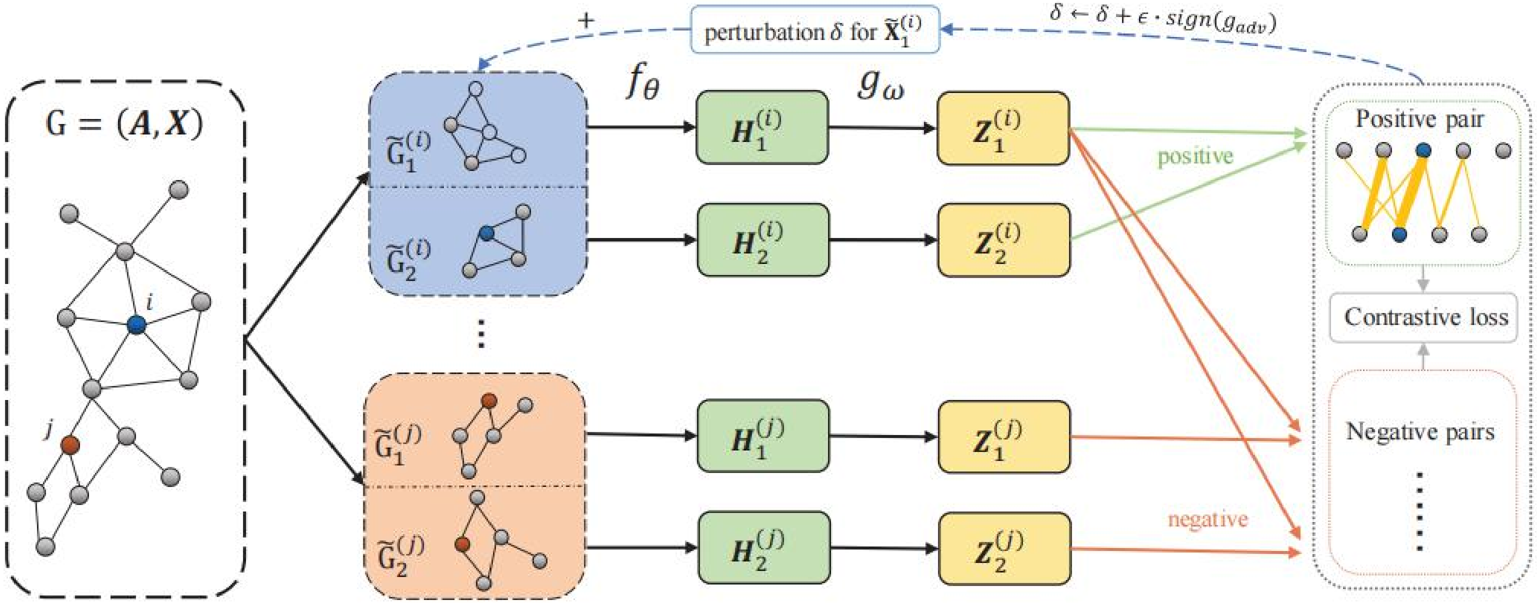


Figure 2: The overview of our proposed method: RoSA. The input is a series of subgraphs sampled from a full graph, where different random walk views from the same central node are recognized as positive pairs and views from different central nodes are treated as negative pairs. Then the subgraphs are fed into the encoder and projector to obtain node embeddings for contrasting. The self-aligned EMD-based contrastive loss will maximize the mutual information (MI) between positive pairs and minimize MI between negative pairs, guiding the model to learn rich representations. Besides, introducing adversarial training into this workflow enhances the robustness of the model.

Method



$$\mathcal{G} = (\mathbf{X}, \mathbf{A})$$

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in R^{N\times d}$$

$$\mathbf{x}_i \in \mathbb{R}^d,\ \mathbf{A} \in \mathbb{R}^{N\times N}$$

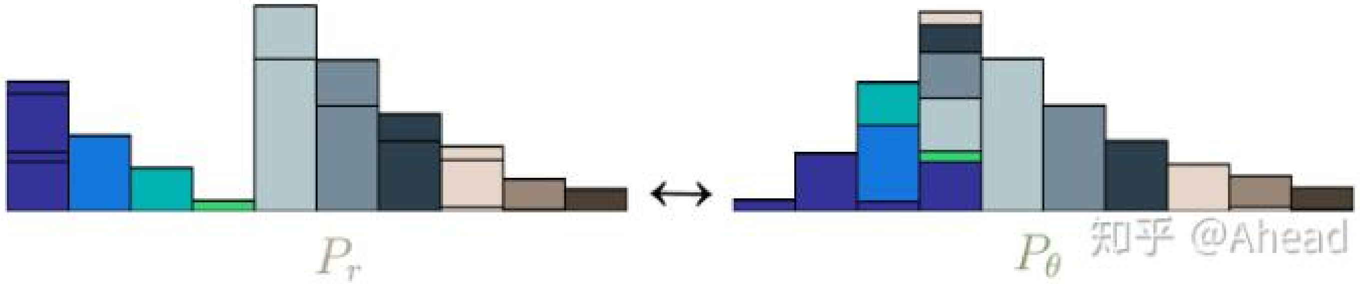
$$\mathcal{G}^{(i)} \qquad \tilde{\mathcal{G}}_k^{(i)}$$

encoder f_θ

embeddings $\mathbf{H}_1^{(i)}$ and $\mathbf{H}_2^{(i)}$

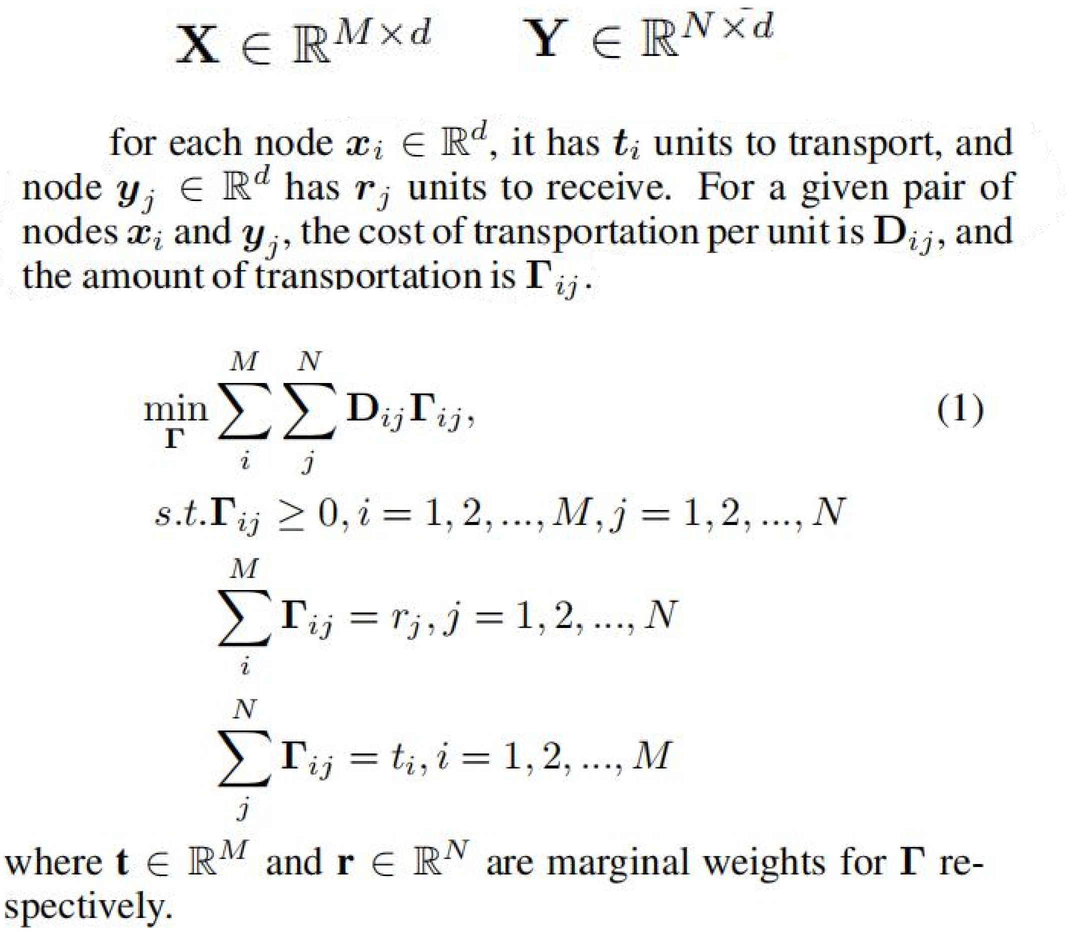
linear projector g_ω

Method



注意从左到右的颜色小块的变动

g-EMD:



$$\mathbf{X} \in \mathbb{R}^{M\times d} \qquad \mathbf{Y} \in \mathbb{R}^{N\times d}$$

for each node $\boldsymbol{x}_i \in \mathbb{R}^d$, it has $\boldsymbol{t}_i$ units to transport, and node $\boldsymbol{y}_j \in \mathbb{R}^d$ has $\boldsymbol{r}_j$ units to receive. For a given pair of nodes $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, the cost of transportation per unit is $\mathbf{D}_{ij}$, and the amount of transportation is $\mathbf{\Gamma}_{ij}$.

$$\min_{\mathbf{\Gamma}} \sum_i^M \sum_j^N \mathbf{D}_{ij}\mathbf{\Gamma}_{ij}, \tag{1}$$

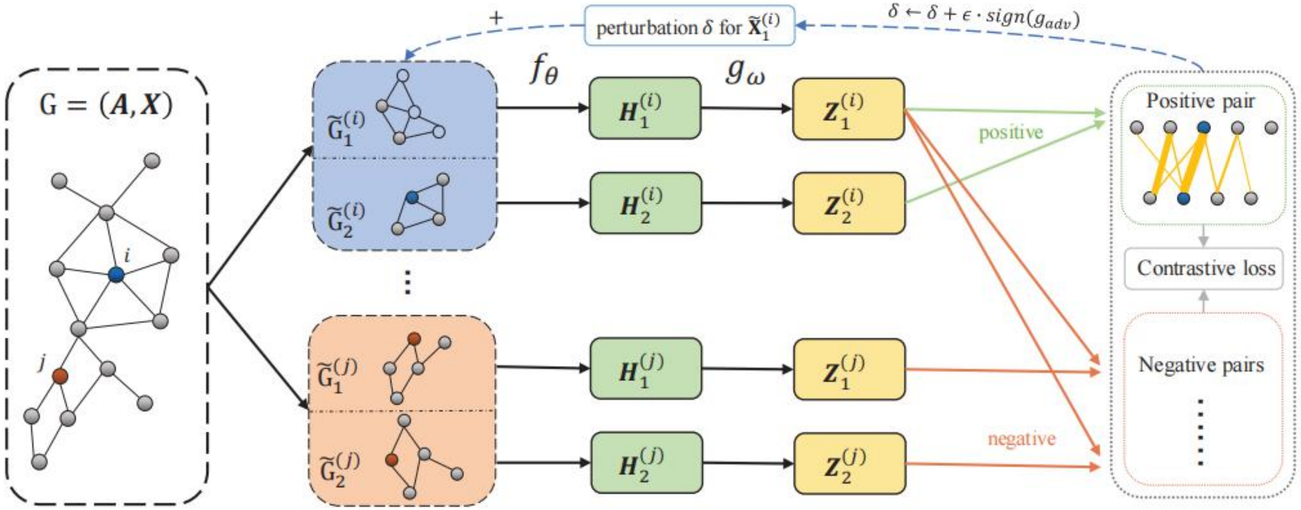
$$s.t.\mathbf{\Gamma}_{ij} \geq 0, i = 1, 2, ..., M, j = 1, 2, ..., N$$

$$\sum_i^M \mathbf{\Gamma}_{ij} = r_j, j = 1, 2, ..., N$$

$$\sum_j^N \mathbf{\Gamma}_{ij} = t_i, i = 1, 2, ..., M$$

where $\mathbf{t} \in \mathbb{R}^M$ and $\mathbf{r} \in \mathbb{R}^N$ are marginal weights for $\mathbf{\Gamma}$ respectively.

Method



The set of all possible transportation matrices Γ can be defined as

$$\Pi(\mathbf{t}, \mathbf{r}) = \{\mathbf{\Gamma} \in \mathbb{R}^{M \times N} | \mathbf{\Gamma} \mathbf{1}_M = \mathbf{t}, \mathbf{\Gamma}^T \mathbf{1}_N = \mathbf{r}\}, \quad (2)$$

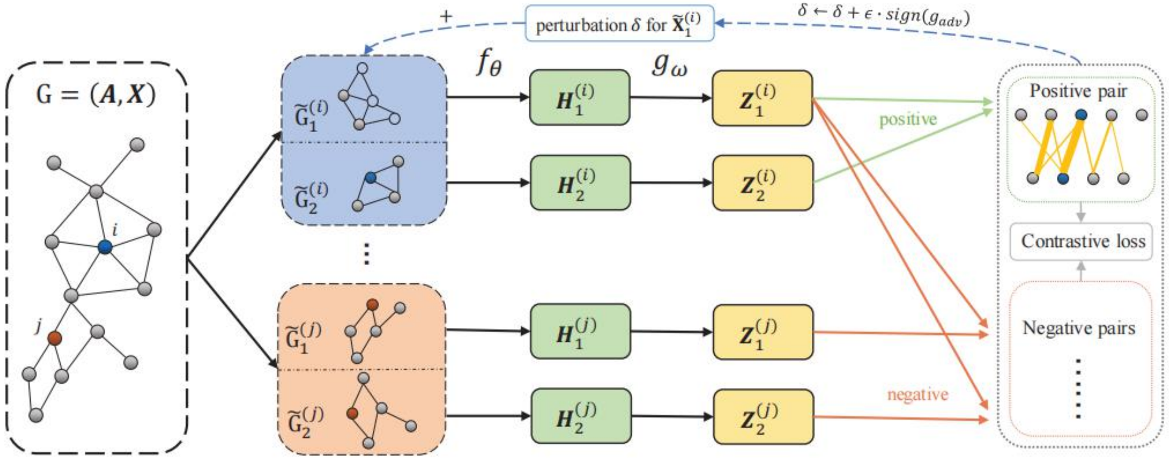
$$\mathbf{D}_{ij} = 1 - \frac{\boldsymbol{x}_i^T \boldsymbol{y}_j}{||\boldsymbol{x}_i|| ||\boldsymbol{y}_j||},$$

$$\mathbf{S}_{i,j} = S(\Psi_{i,j}) = \frac{1}{1 + e^{-\Psi_{i,j}/\tau}}, \quad (4)$$

$$\mathbf{D} = \mathbf{D} \circ \mathbf{S}, \quad (5)$$

$$\text{g-EMD}(\mathbf{X}, \mathbf{Y}, \mathbf{S}) = \inf_{\mathbf{\Gamma} \in \Pi} \langle \mathbf{\Gamma}, \mathbf{D} \rangle_{\mathrm{F}} + \underbrace{\frac{1}{\lambda} \mathbf{\Gamma}(\log \mathbf{\Gamma} - 1)}_{\text{regularization term}}, \quad (6)$$

Method



$$\tilde{\Gamma} = diag(\mathbf{v})\mathbf{P}diag(\mathbf{u}), \qquad (7)$$

where $\mathbf{P} = e^{-\lambda\mathbf{D}}$, and $\mathbf{v}$, $\mathbf{u}$ are two coefficient vectors whose values can be iteratively updated as

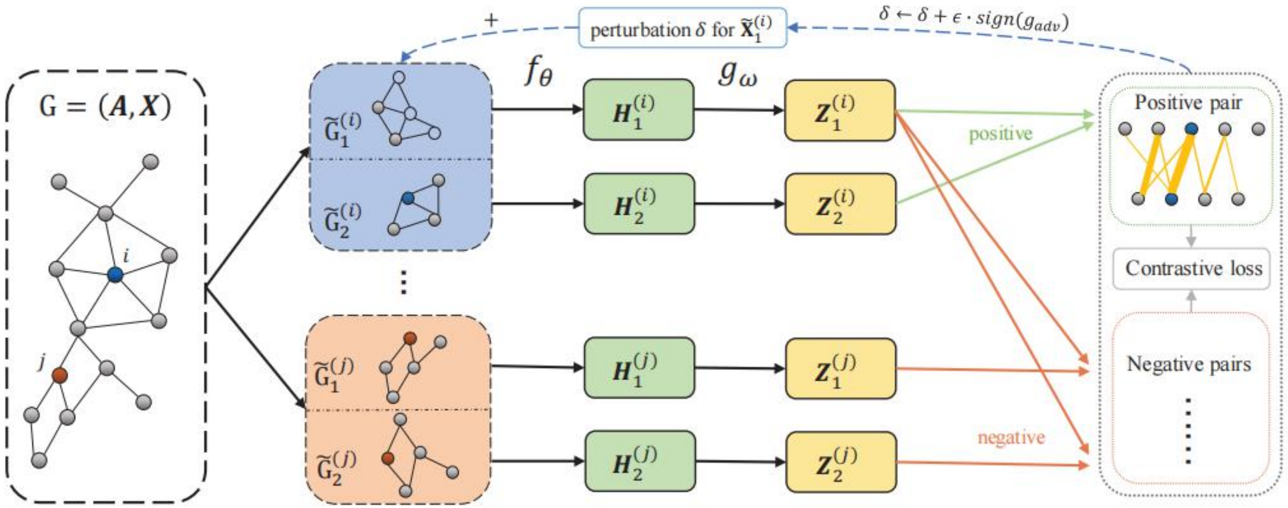
$$\boldsymbol{v}_i^{t+1} = \frac{\boldsymbol{t}_i}{\sum_{j=1}^{N} \mathbf{P}_{ij}\boldsymbol{u}_j^t}, \qquad (8)$$
$$\boldsymbol{u}_j^{t+1} = \frac{\boldsymbol{r}_j}{\sum_{i=1}^{M} \mathbf{P}_{ij}\boldsymbol{v}_i^{t+1}}.$$

The weight represents a node's contribution in comparison of two views, where a node should have larger weight if its semantic meaning is close to the other view.

$$\boldsymbol{t}_i = \max\{\boldsymbol{x}_i^T \cdot \frac{\sum_{j=1}^{N} \boldsymbol{y}_j}{N}, 0\}, \qquad (9)$$
$$\boldsymbol{r}_j = \max\{\boldsymbol{y}_j^T \cdot \frac{\sum_{i=1}^{M} \boldsymbol{x}_i}{M}, 0\}.$$

$$\text{g-EMD}(\mathbf{X}, \mathbf{Y}, \mathbf{S}) = \langle\tilde{\Gamma}, \mathbf{D}\rangle_{\text{F}}. \qquad (10)$$

Method



$$\ell(\mathbf{Z}_1^{(i)}, \mathbf{Z}_2^{(i)}) =$$

$$-\log\left(\frac{e^{\mathrm{s}(\mathbf{Z}_1^{(i)}, \mathbf{Z}_2^{(i)}))/\tau}}{\sum_{k=1}^{N} e^{\mathrm{s}(\mathbf{Z}_1^{(i)}, \mathbf{Z}_2^{(k)}))/\tau} + \sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} e^{\mathrm{s}(\mathbf{Z}_1^{(i)}, \mathbf{Z}_1^{(k)}))/\tau}}\right), \tag{11}$$

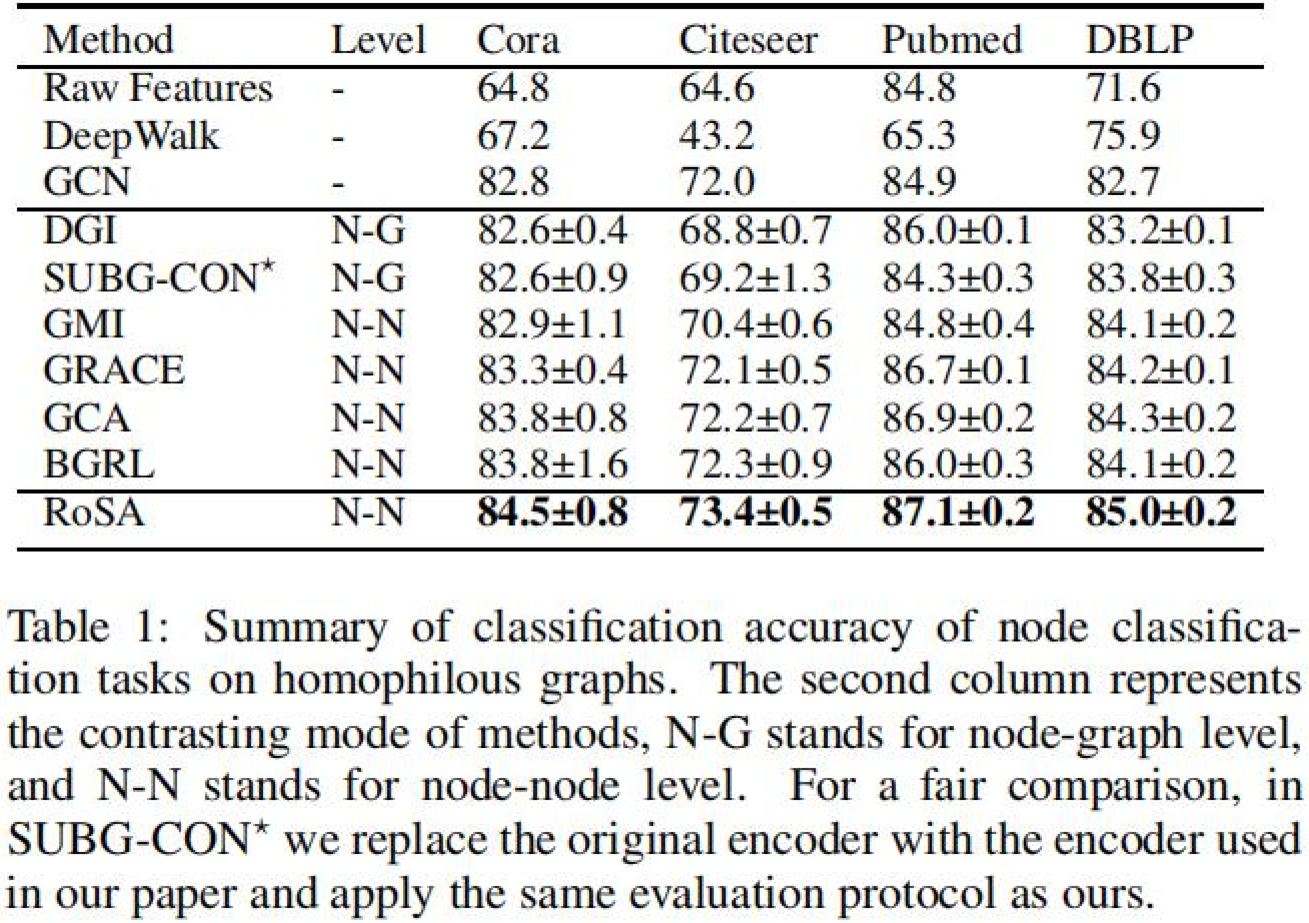
where $\mathrm{s}(\boldsymbol{x}, \boldsymbol{y})$ is a function that calculates the similarity between $\boldsymbol{x}$ and $\boldsymbol{y}$, here we use $1 - \mathrm{EMD}(\boldsymbol{x}, \boldsymbol{y})$ to replace $\mathrm{s}(\boldsymbol{x}, \boldsymbol{y})$; $\mathbf{1}$ is an indicator function which returns 1 if $i \neq k$ otherwise returns 0; and τ is temperature parameter. Adding all nodes in $\mathcal{N}$, the overall contrastive loss is given by:

$$\mathcal{J} = \frac{1}{2N} \sum_{i=1}^{N} \left[\ell\left(\mathbf{Z}_1^{(i)}, \mathbf{Z}_2^{(i)}\right) + \ell\left(\mathbf{Z}_2^{(i)}, \mathbf{Z}_1^{(i)}\right) \right]. \tag{12}$$

$$\min_{\theta,\omega} \mathbb{E}_{(\mathbf{X}_1^{(i)}, \mathbf{X}_2^{(i)}) \sim \mathbb{D}} \left[\frac{1}{M} \sum_{t=0}^{M-1} \max_{\boldsymbol{\delta}_t \in \mathcal{I}_t} \mathcal{J}\left(\mathbf{X}_1^{(i)} + \boldsymbol{\delta}_t, \mathbf{X}_2^{(i)}\right) \right], \tag{13}$$

where θ, ω are the parameters of encoder and projector, $\mathbb{D}$ is data distribution, $\mathcal{I}_t = \mathcal{B}_{\mathbf{X}+\boldsymbol{\delta}_0}(\alpha t) \cap \mathcal{B}_{\mathbf{X}}(\epsilon)$ where ϵ is the perturbation budget. For efficiency, the inner loop runs M times, the gradient of δ, θ_{t-1} and ω_{t-1} will be accumulated in each time, and the accumulated gradients will be used for updating θ_{t-1} and ω_{t-1} during outer update.

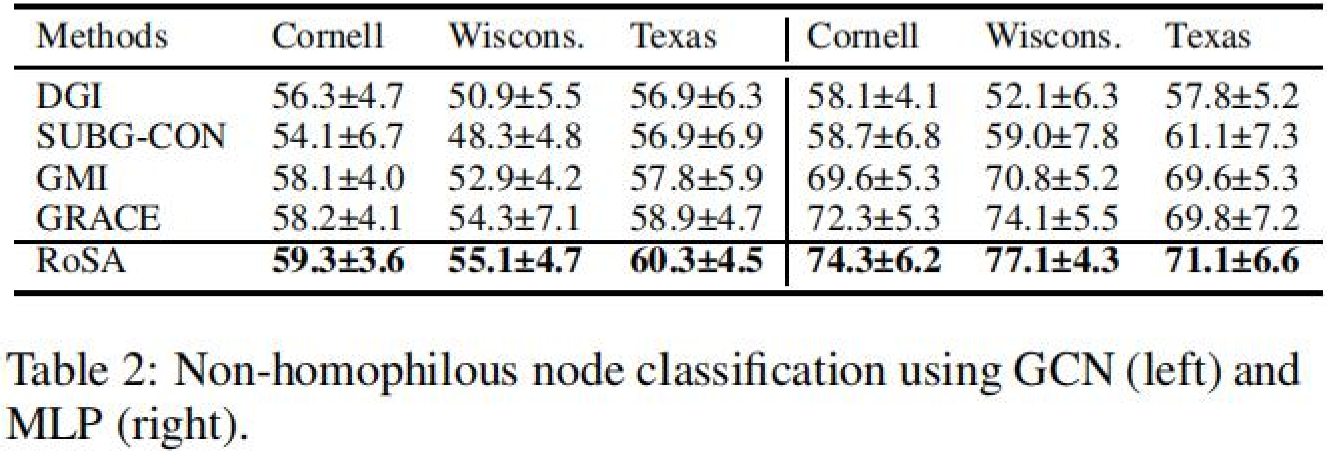
Experiments



Method	Level	Cora	Citeseer	Pubmed	DBLP
Raw Features	-	64.8	64.6	84.8	71.6
DeepWalk	-	67.2	43.2	65.3	75.9
GCN	-	82.8	72.0	84.9	82.7
DGI	N-G	82.6±0.4	68.8±0.7	86.0±0.1	83.2±0.1
SUBG-CON*	N-G	82.6±0.9	69.2±1.3	84.3±0.3	83.8±0.3
GMI	N-N	82.9±1.1	70.4±0.6	84.8±0.4	84.1±0.2
GRACE	N-N	83.3±0.4	72.1±0.5	86.7±0.1	84.2±0.1
GCA	N-N	83.8±0.8	72.2±0.7	86.9±0.2	84.3±0.2
BGRL	N-N	83.8±1.6	72.3±0.9	86.0±0.3	84.1±0.2
RoSA	N-N	**84.5±0.8**	**73.4±0.5**	**87.1±0.2**	**85.0±0.2**

Table 1: Summary of classification accuracy of node classification tasks on homophilous graphs. The second column represents the contrasting mode of methods, N-G stands for node-graph level, and N-N stands for node-node level. For a fair comparison, in SUBG-CON* we replace the original encoder with the encoder used in our paper and apply the same evaluation protocol as ours.

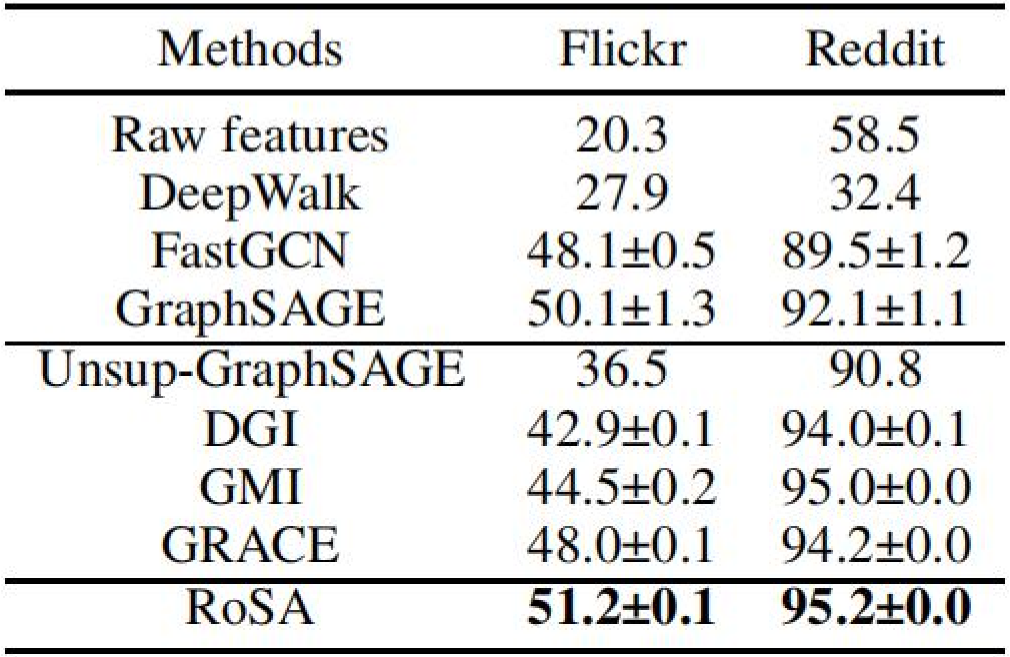
Method



Methods	Cornell	Wiscons.	Texas	Cornell	Wiscons.	Texas
DGI	56.3±4.7	50.9±5.5	56.9±6.3	58.1±4.1	52.1±6.3	57.8±5.2
SUBG-CON	54.1±6.7	48.3±4.8	56.9±6.9	58.7±6.8	59.0±7.8	61.1±7.3
GMI	58.1±4.0	52.9±4.2	57.8±5.9	69.6±5.3	70.8±5.2	69.6±5.3
GRACE	58.2±4.1	54.3±7.1	58.9±4.7	72.3±5.3	74.1±5.5	69.8±7.2
RoSA	**59.3±3.6**	**55.1±4.7**	**60.3±4.5**	**74.3±6.2**	**77.1±4.3**	**71.1±6.6**

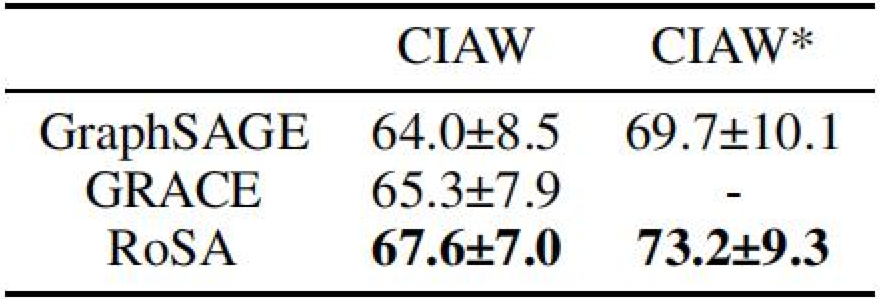
Table 2: Non-homophilous node classification using GCN (left) and MLP (right).

Experiments



Methods	Flickr	Reddit
Raw features	20.3	58.5
DeepWalk	27.9	32.4
FastGCN	48.1±0.5	89.5±1.2
GraphSAGE	50.1±1.3	92.1±1.1
Unsup-GraphSAGE	36.5	90.8
DGI	42.9±0.1	94.0±0.1
GMI	44.5±0.2	95.0±0.0
GRACE	48.0±0.1	94.2±0.0
RoSA	**51.2±0.1**	**95.2±0.0**

Table 3: Result for inductive learning on large-scale datasets.



	CIAW	CIAW*
GraphSAGE	64.0±8.5	69.7±10.1
GRACE	65.3±7.9	-
RoSA	**67.6±7.0**	**73.2±9.3**

Table 4: Node classification using GraphSAGE on dynamic graphs.

Experiments

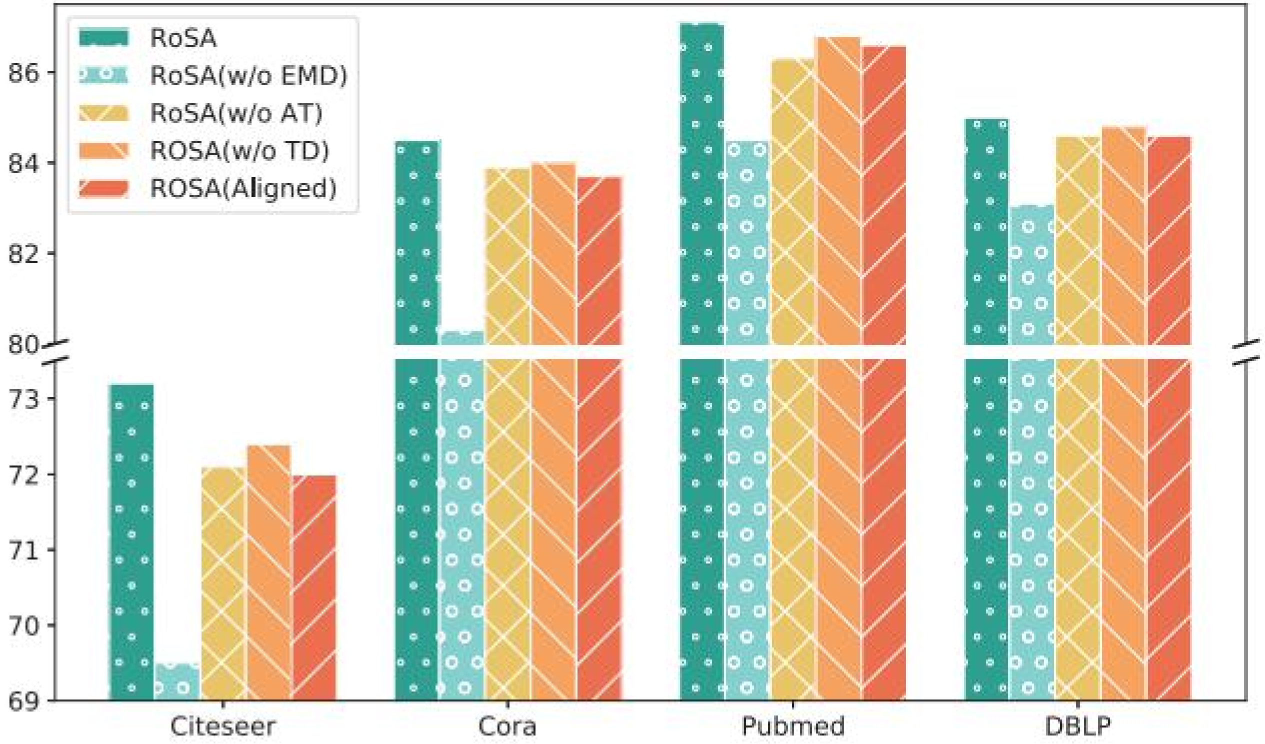


Figure 3: Abalation study on RoSA

Thanks