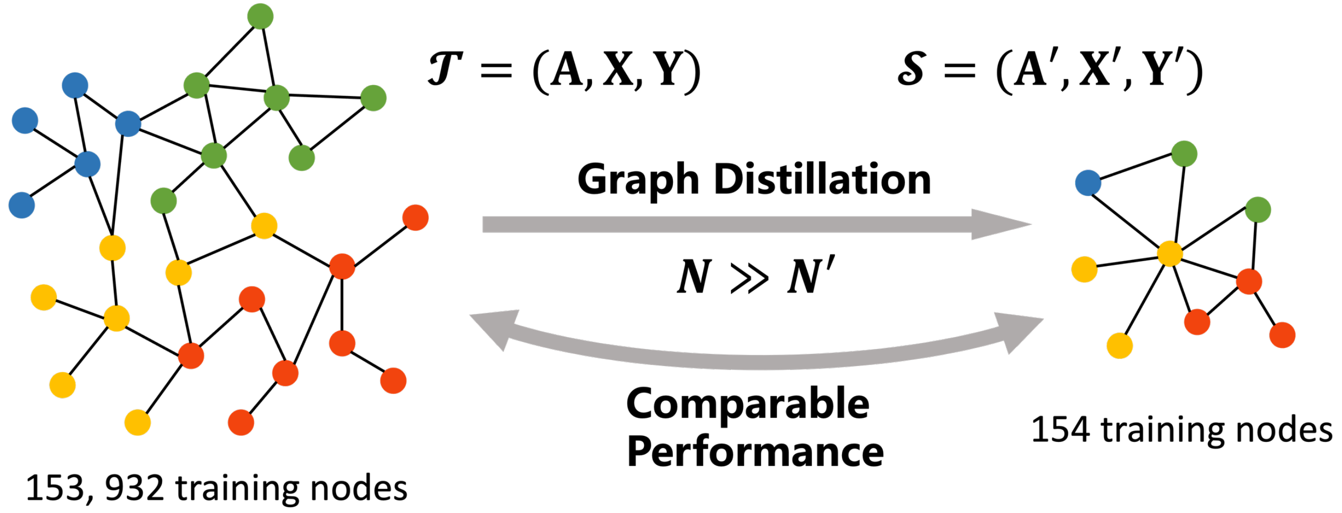
Graph Distillation with Eigenbasis Matching

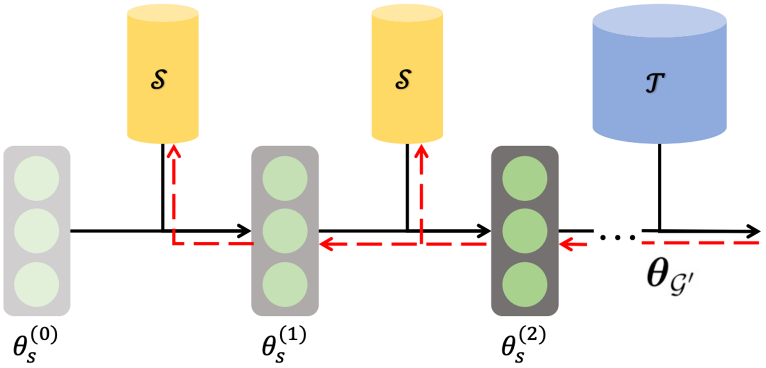
Yang Liu[1], Deyu Bo[1], Chuan Shi[1†]

[1] Beijing University of Posts and Telecommunication.

Graph Distillation

Distilling the knowledge of a large-scale real graph $\mathcal{T}$ into a small synthetic graph $\mathcal{S}$, thus accelerating the training of GNNs.



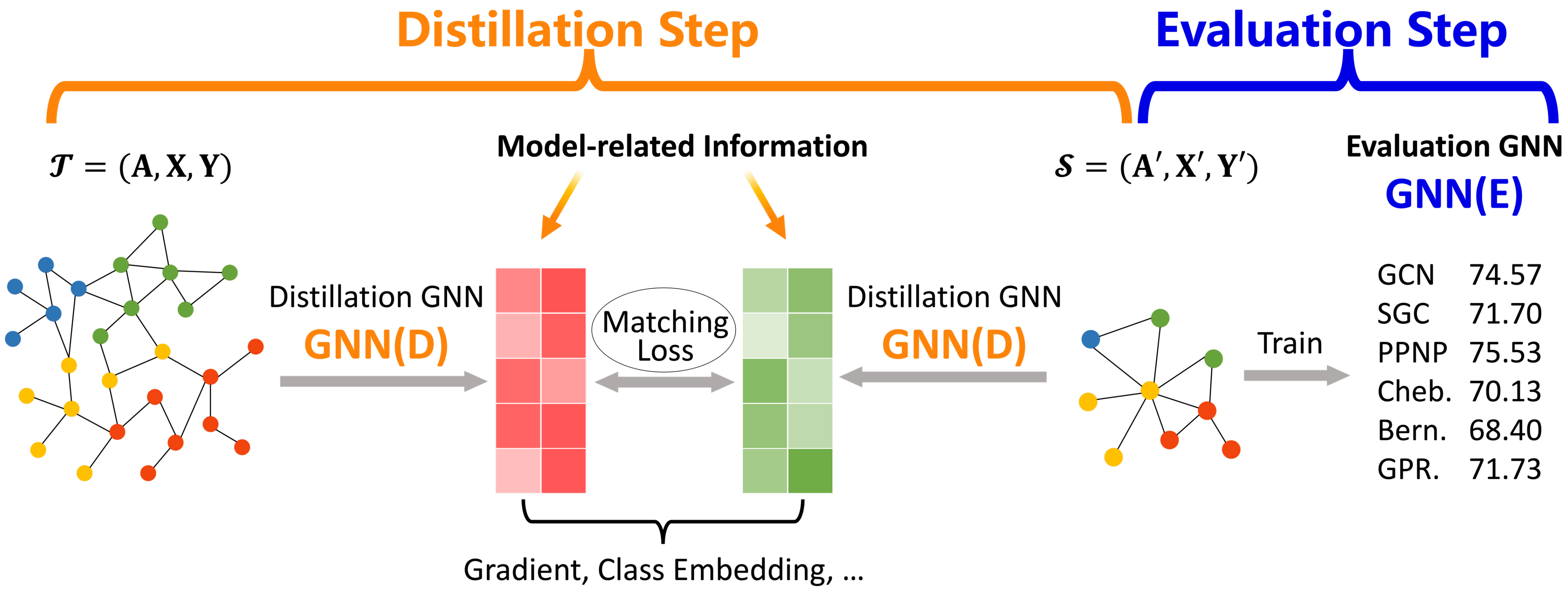


$$\min_{\mathcal{G}'} \mathcal{L}\left(\mathrm{GNN}_{\boldsymbol{\theta}_{\mathcal{G}'}}(\mathbf{A}, \mathbf{X}), \mathbf{Y}\right)$$

$$\text{s.t} \quad \boldsymbol{\theta}_{\mathcal{G}'} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathrm{GNN}_{\boldsymbol{\theta}}(\mathbf{A}', \mathbf{X}'), \mathbf{Y}')$$

Bilevel-optimization

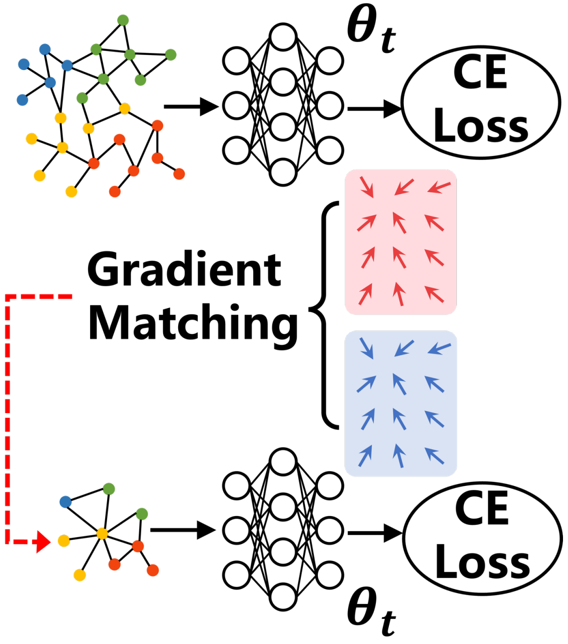
Pipeline of Graph Distillation



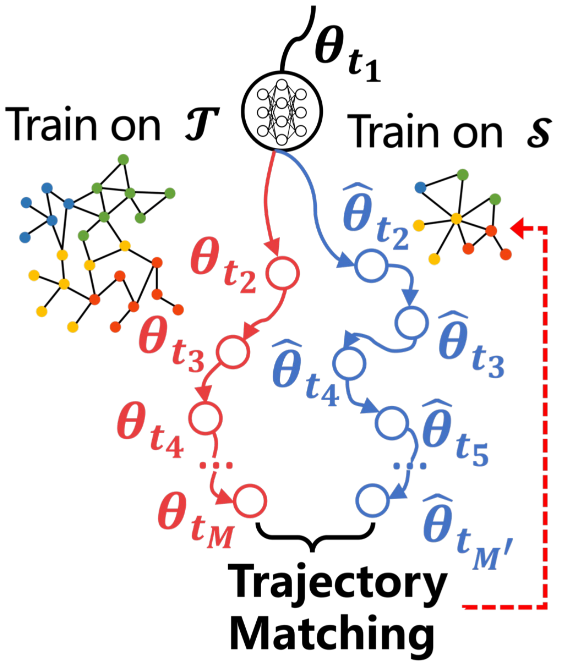
1 Background

Existing Methods

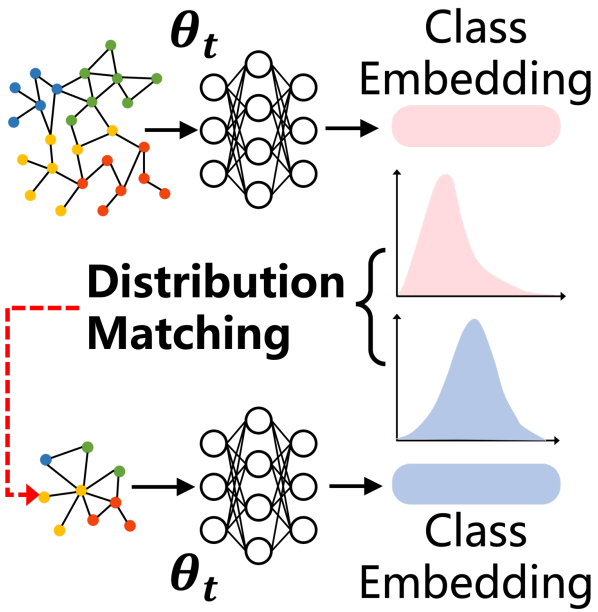
Gradient Matching (GM)



Trajectory Matching (TM)



Distribution Matching (DM)



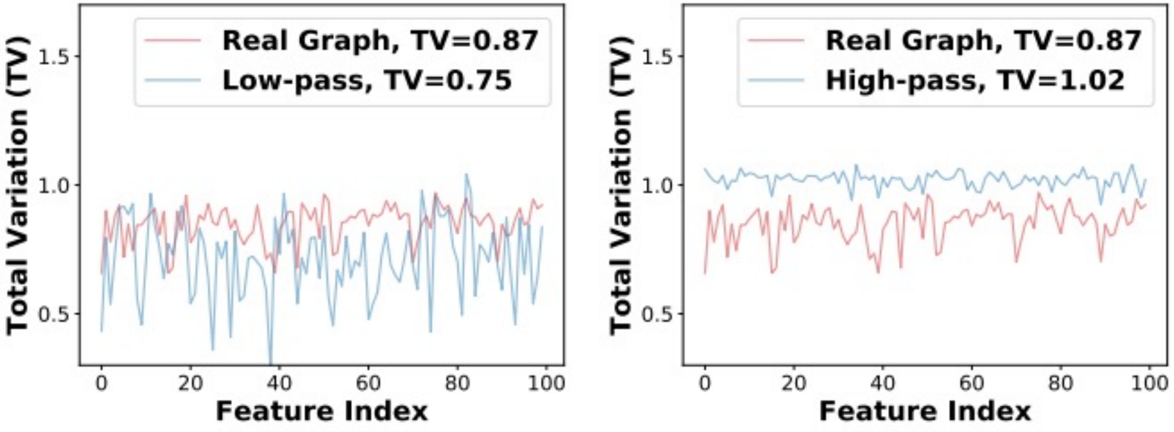
GM: $\min\limits_{\mathbf{A}',\mathbf{X}'} \mathop{\mathbb{E}}\limits_{\theta\sim P_\theta} \left[D\left(\nabla_\theta \mathcal{L}\left(\Phi_\theta\left(\mathbf{A}',\mathbf{X}'\right),\mathbf{Y}'\right),\nabla_\theta \mathcal{L}\left(\Phi_\theta\left(\mathbf{A},\mathbf{X}\right),\mathbf{Y}\right)\right)\right]$

TM: $\min\limits_{\mathbf{A}',\mathbf{X}'} \mathop{\mathbb{E}}\limits_{\theta\sim P_\theta} \left[\sum_{t=1}^{L} D\left(\Phi_\theta^t\left(\mathbf{A}',\mathbf{X}'\right),\Phi_\theta^t\left(\mathbf{A},\mathbf{X}\right)\right)\right]$

DM: $\min\limits_{\mathbf{A}',\mathbf{X}'} \mathop{\mathbb{E}}\limits_{\theta_t^{*,i}\sim P_{\Theta\mathcal{T}}} \left[\mathcal{L}_{\text{meta-tt}}\left(\theta_t^*|_{t=t_0}^{p},\tilde{\theta}_t|_{t=t_0}^{q}\right)\right]$

2 Motivation

- Existing methods are **model-specific**, which have two limitations:
  - Introduce distribution bias in the synthetic graph.
  - Need to traverse various distillation GNNs for optimal performance.



D \ E	GCN	SGC	PPNP	Cheb.	Bern.	GPR.
GCN	74.57	71.70	75.53	70.13	68.40	71.73
SGC	**77.72**	**77.60**	77.34	76.03	74.42	76.52
PPNP	72.70	70.40	77.46	73.38	70.56	74.02
Cheb.	73.60	70.62	75.10	**77.30**	77.62	78.10
Bern.	67.68	73.76	74.30	77.20	**78.12**	**78.28**
GPR.	76.04	72.20	**77.94**	75.92	77.12	77.96

Left: Low-pass filter (AXW).
Right: High-pass filter (LXW).

$$\mathrm{TV} = \mathbf{x}^\top \mathbf{L} \mathbf{x} = \sum_{(i,j)\in\mathcal{E}} (x_i - x_j)^2$$

Table 1. Cross-architecture performance of GCOND with various distillation (D) and evaluation (E) GNNs.

- **Goal:** Distilling graphs without being affected by different GNNs.

Upper-bound of Gradient Matching

The objective of distillation GNNs can be simplified into a MSE loss:

$$\mathcal{L} = \|g(\mathbf{L})\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2$$

The gradients on the real and synthetic graphs are:

$$\nabla_{\mathbf{W}} = (g(\mathbf{L})\mathbf{X})^T (g(\mathbf{L})\mathbf{X}\mathbf{W} - \mathbf{Y}),$$

$$\nabla'_{\mathbf{W}} = (g(\mathbf{L}')\mathbf{X}')^T (g(\mathbf{L}')\mathbf{X}'\mathbf{W} - \mathbf{Y}')$$

The upper-bound of MSE loss between two gradients:

Target Distribution

$$\begin{aligned} \mathcal{L}_{GM} &= \|\nabla_{\mathrm{W}} - \nabla'_{\mathrm{w}}\|_F^2 \\ \textbf{Unsupervised Loss} \quad &\leq \|\mathbf{W}\|_F^2 \left\| \boxed{\mathbf{X}^\top g(\mathbf{L})^{2t}\mathbf{X}} - \mathbf{X}'^\top g(\mathbf{L}')^{2t}\mathbf{X}' \right\|_F^2 \\ \textbf{Supervised Loss} \quad &+ \left\| \boxed{\mathbf{X}^\top g(\mathbf{L})^{t}\mathbf{Y}} - \mathbf{X}'^\top g(\mathbf{L}')^{t}\mathbf{Y}' \right\|_F^2, \end{aligned}$$

Spectrum Bias in Synthetic Graph

The target distribution can be formulated as:

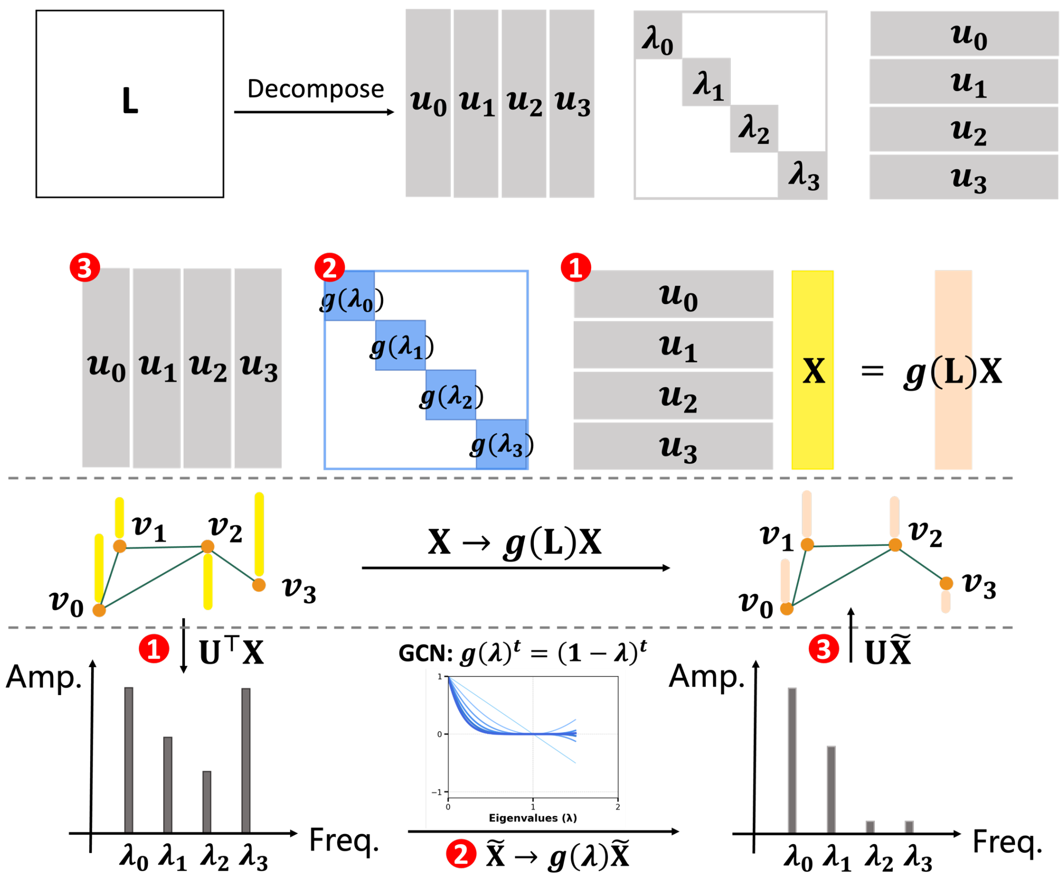
$$\mathbf{X}^\top g(\mathbf{L})^{2t}\mathbf{X} = \sum_{i=1}^{N} g(\lambda_i)^{2t}\boxed{\mathbf{X}^\top \mathbf{u}_i\mathbf{u}_i^\top \mathbf{X}}$$

$$\mathbf{X}^\top g(\mathbf{L})^{t}\mathbf{Y} = \sum_{i=1}^{N} g(\lambda_i)^{t}\boxed{\mathbf{X}^\top \mathbf{u}_i\mathbf{u}_i^\top \mathbf{Y}}$$

invariant

Lemma. The target distribution is dominated by the eigenvalues whose filtered values are greater than 1, i.e., $g(\lambda_i)\geq 1$.

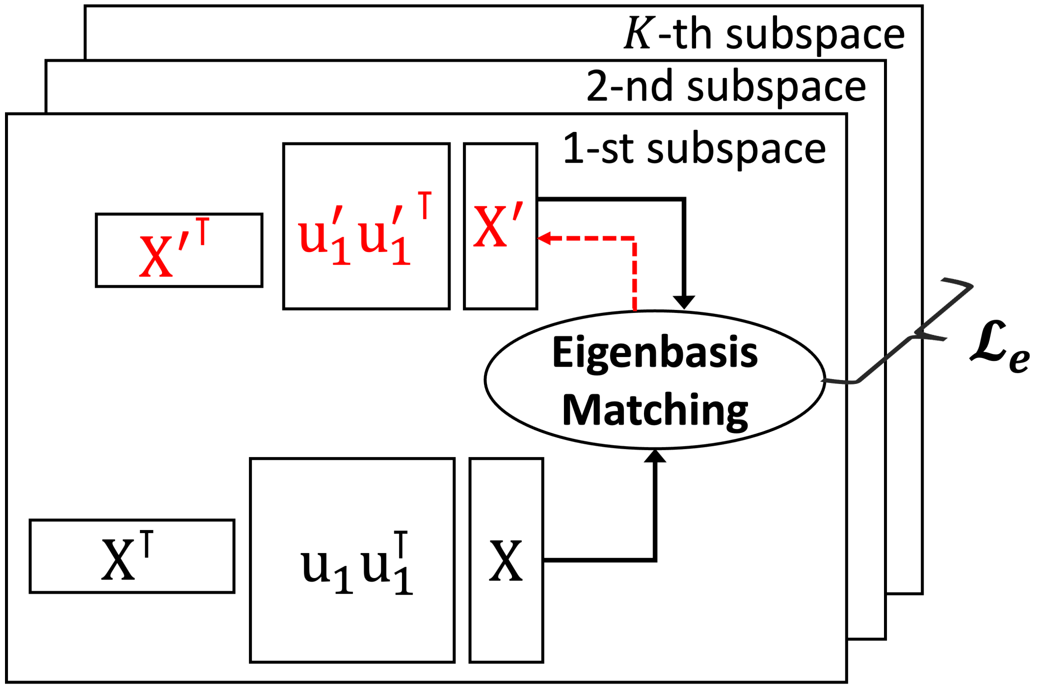
When distillation GNN is a GCN and t goes to infinity, the target distribution will be dominated by $\mathbf{X}^\top \mathbf{u}_0\mathbf{u}_0^\top \mathbf{X}$.



Solution: Match *invariant information* of the graph and design a model-agnostic graph distillation.

4 GDEM Eigenbasis Matching

- Matching the eigenbasis and node features between the real and synthetic graphs.



➢ Matching eigenvectors with the K_1 smallest and the K_2 largest eigenvalues, $K_1 + K_2 = K \leq N'$.

$$\mathcal{L}_e = \sum_{k=1}^{K} \left\| \mathbf{X}^{\top} \mathbf{u}_k \mathbf{u}_k^{\top} \mathbf{X} - \mathbf{X}'^{\top} \mathbf{u}'_k \mathbf{u}'^{\top}_k \mathbf{X}' \right\|_F^2$$

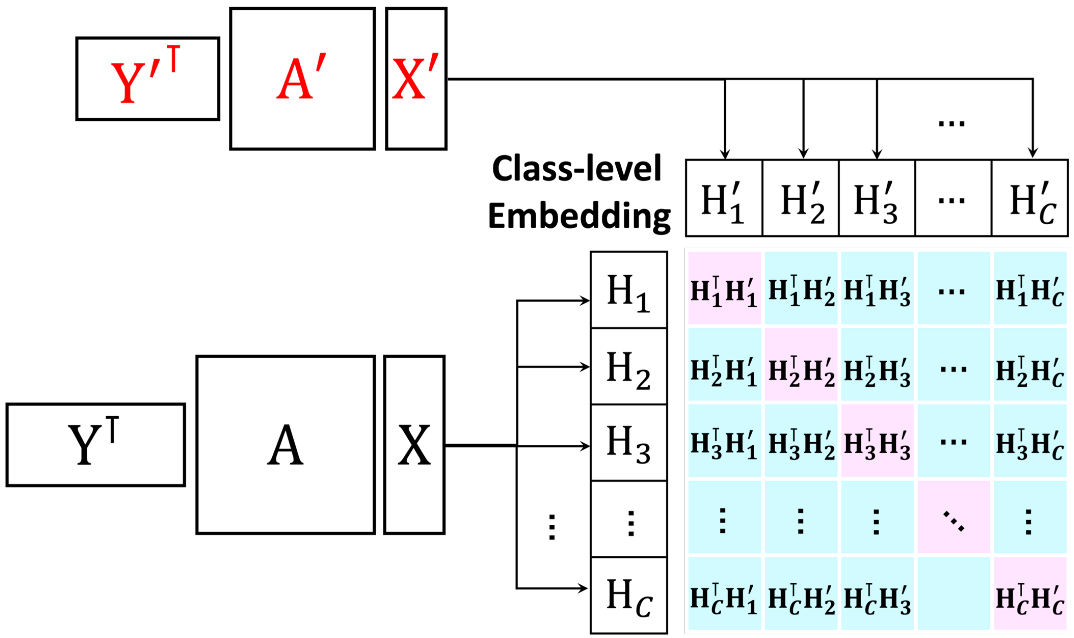
➢ Orthonormal regularization

$$\mathcal{L}_o = \left\| \mathbf{U}'^{\top}_K \mathbf{U}'_K - \mathbf{I}_K \right\|_F^2$$

Limitations of Eigenbasis Matching

Eigenbasis matching improves the cross-architecture generalization but **contributes less to the performance of node classification**:

- Only preserves the distribution of $\mathbf{X}^\top \mathbf{u}\mathbf{u}^\top \mathbf{X}$.
- Neglect the information of downstream tasks.



Design of Discrimination Constraint

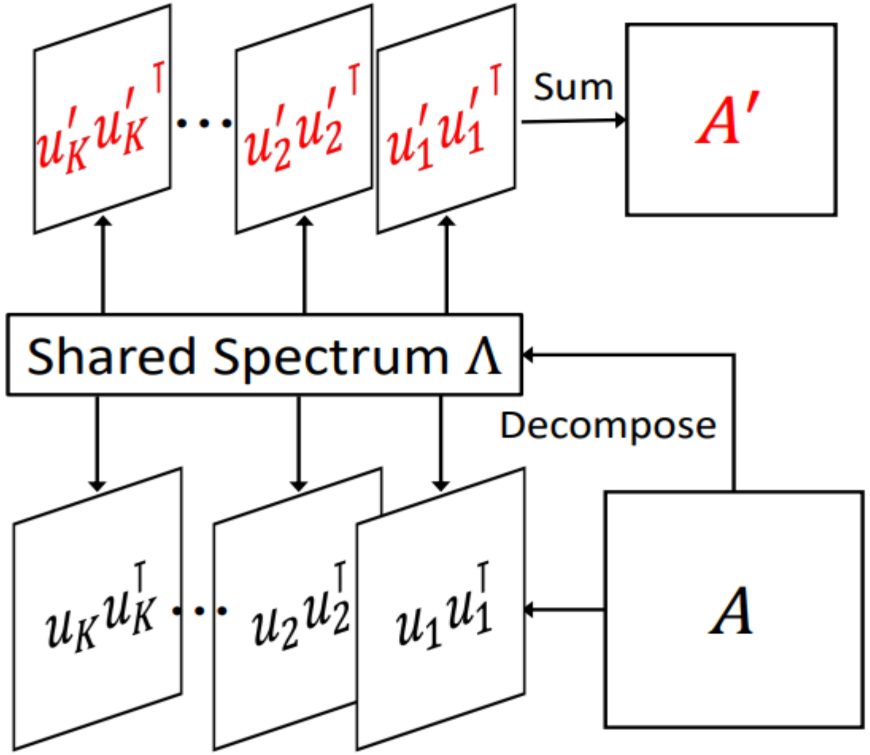
Improve classification performance by constraining the representation between real and synthetic graphs:

$$\mathbf{H} = \mathbf{Y}^\top \mathbf{A}\mathbf{X}, \quad \mathbf{H}' = \mathbf{Y}'^\top \sum_{k=1}^{K} (1-\lambda_k)\mathbf{u}'_k \mathbf{u}'^\top_k \mathbf{X}',$$

$$\mathcal{L}_d = \sum_{i=1}^{C} \underbrace{\left(1 - \frac{\mathbf{H}_i^\top \cdot \mathbf{H}'_i}{||\mathbf{H}_i||\,||\mathbf{H}'_i||}\right)}_{\text{Intra-class}} + \sum_{\substack{i,j=1 \\ i\neq j}}^{C} \underbrace{\frac{\mathbf{H}_i^\top \cdot \mathbf{H}'_j}{||\mathbf{H}_i||\,||\mathbf{H}'_j||}}_{\text{Inter-class}}.$$

Synthetic Graph Constructing

- Constructing the synthetic graph by using the synthesized eigenbasis and replicating the spectrum of the real graph.



$$\mathbf{L}' = \sum_{k=1}^{K} \lambda_k \mathbf{u}'_k \mathbf{u}'^{\top}_k$$

$$\mathbf{A}' = \sum_{k=1}^{K} (1 - \lambda_k) \mathbf{u}'_k \mathbf{u}'^{\top}_k$$

Algorithm of GDEM

Algorithm 1 GDEM for Graph Distillation

Input: Real graph $\mathcal{G} = (\mathbf{A}, \mathbf{X}, \mathbf{Y})$ with eigenvalues $\{\lambda_i\}_{i=1}^{K}$ and eigenbasis $\mathbf{U}_K$

Init: Synthetic graph $\mathcal{G}'$ with eigenbasis $\mathbf{U}'_K$, node features $\mathbf{X}'$, and labels $\mathbf{Y}'$

for $t = 1$ **to** T **do**

  Compute $\mathcal{L}_e$, $\mathcal{L}_o$, and $\mathcal{L}_d$ via Eqs. 5, 6, and 8

  Compute $\mathcal{L}_{total} = \alpha\mathcal{L}_e + \beta\mathcal{L}_d + \gamma\mathcal{L}_o$

  if $t\%(\tau_1 + \tau_2) < \tau_1$ **then**

    Update $\mathbf{U}'_K \leftarrow \mathbf{U}'_K - \eta_1 \nabla_{\mathbf{U}'_K} \mathcal{L}_{total}$

  else

    Update $\mathbf{X}' \leftarrow \mathbf{X}' - \eta_2 \nabla_{\mathbf{X}'} \mathcal{L}_{total}$

  end if

end for

Compute $\mathbf{A}' = \sum_{k=1}^{K} (1 - \lambda_k) \mathbf{u}'_k {\mathbf{u}'_k}^{\top}$

Return: $\mathbf{A}'$, $\mathbf{X}'$

Performance of Node Classification

Dataset	Ratio (r)	Traditional Methods				Graph Distillation Methods				Whole Dataset
		Random ($\mathbf{A}', \mathbf{X}'$)	Coarsening ($\mathbf{A}', \mathbf{X}'$)	Herding ($\mathbf{A}', \mathbf{X}'$)	K-Center ($\mathbf{A}', \mathbf{X}'$)	GCOND ($\mathbf{A}', \mathbf{X}'$)	SFGC ($\mathbf{X}'$)	SGDD ($\mathbf{A}', \mathbf{X}'$)	GDEM ($\mathbf{U}', \mathbf{X}'$)	
Citeseer	0.90%	54.4±4.4	52.2±0.4	57.1±1.5	52.4±2.8	70.5±1.2	71.4±0.5	69.5±0.4	**72.3±0.3**	71.7±0.1
	1.80%	64.2±1.7	59.0±0.5	66.7±1.0	64.3±1.0	70.6±0.9	72.4±0.4	70.2±0.8	**72.6±0.6**	
	3.60%	69.1±0.1	65.3±0.5	69.0±0.1	69.1±0.1	69.8±1.4	70.6±0.7	70.3±1.7	**72.6±0.5**	
Pubmed	0.08%	69.4±0.2	18.1±0.1	76.7±0.7	64.5±2.7	76.5±0.2	76.4±1.2	77.1±0.5	**77.7±0.7**	79.3±0.2
	0.15%	73.3±0.7	28.7±4.1	76.2±0.5	69.4±0.7	77.1±0.5	77.5±0.4	78.0±0.3	**78.4±1.8**	
	0.30%	77.8±0.3	42.8±4.1	78.0±0.5	**78.2±0.4**	77.9±0.4	77.9±0.3	77.5±0.5	**78.2±0.8**	
Ogbn-arxiv	0.05%	47.1±3.9	35.4±0.3	52.4±1.8	47.2±3.0	59.2±1.1	**65.5±0.7**	60.8±1.3	63.7±0.8	71.4±0.1
	0.25%	57.3±1.1	43.5±0.2	58.6±1.2	56.8±0.8	63.2±0.3	**66.1±0.4**	65.8±1.2	63.8±0.6	
	0.50%	60.0±0.9	50.4±0.1	60.4±0.8	60.3±0.4	64.0±0.4	**66.8±0.4**	66.3±0.7	64.1±0.3	
Flickr	0.10%	41.8±2.0	41.9±0.2	42.5±1.8	42.0±0.7	46.5±0.4	46.6±0.2	46.9±0.1	**49.9±0.8**	47.2±0.1
	0.50%	44.0±0.4	44.5±0.1	43.9±0.9	43.2±0.1	47.1±0.1	47.0±0.1	47.1±0.3	**49.4±1.3**	
	1.00%	44.6±0.2	44.6±0.1	44.4±0.6	44.1±0.4	47.1±0.1	47.1±0.1	47.1±0.1	**49.9±0.6**	
Reddit	0.05%	46.1±4.4	40.9±0.5	53.1±2.5	46.6±2.3	88.0±1.8	89.7±0.2	91.8±1.9	**92.9±0.3**	93.9±0.0
	0.10%	58.0±2.2	42.8±0.8	62.7±1.0	53.0±3.3	89.6±0.7	90.0±0.3	91.0±1.6	**93.1±0.2**	
	0.50%	66.3±1.9	47.4±0.9	71.0±1.6	58.5±2.1	90.1±0.5	89.9±0.4	91.6±1.8	**93.2±0.4**	
Squirrel	0.60%	22.4±1.6	20.9±1.1	21.3±1.1	21.8±0.3	27.0±1.3	24.0±0.4	24.1±2.3	**28.4±2.0**	33.0±0.4
	1.20%	25.0±0.2	21.1±0.4	21.4±2.1	22.8±0.9	25.7±2.3	26.9±2.5	24.7±2.5	**28.2±2.4**	
	2.50%	26.9±1.4	21.5±0.3	22.4±1.6	22.9±1.7	25.3±0.8	26.1±0.8	25.8±1.8	**27.8±1.6**	
Gamers	0.05%	56.6±1.8	56.1±0.1	56.7±1.7	52.5±4.2	58.5±1.5	58.2±1.1	57.5±1.8	**59.3±1.9**	62.6±0.0
	0.25%	60.5±1.0	56.9±3.0	57.5±2.0	57.2±2.3	58.9±1.8	58.8±0.5	57.7±1.0	**60.8±0.4**	
	0.50%	60.0±0.5	57.1±0.4	58.6±1.3	57.8±1.7	58.5±1.9	59.9±0.3	58.4±1.7	**61.2±0.3**	

Table 2. Node classification performance of different distillation methods.

Cross-architecture Generalization

Datasets (Ratio)	Methods	Spatial GNNs			Spectral GNNs			Avg. (↑)	Std. (↓)	Impro. (↑)
		GCN	SGC	PPNP	ChebyNet	BernNet	GPR-GNN			
Ogbn-arxiv ($r = 0.25\%$)	GCOND	63.2	63.7	63.4	54.9	55.0	60.5	60.12	3.80	(+) 2.90
	SFGC	65.1	64.8	63.9	60.7	63.8	64.9	**63.87**	1.50	(-) 0.85
	SGDD	65.8	64.0	63.6	56.4	62.0	64.0	62.63	3.00	(+) 0.39
	GDEM	63.8	62.9	63.5	62.4	61.9	63.6	63.02	**0.69**	-
Flickr ($r = 0.50\%$)	GCOND	47.1	46.1	45.9	42.8	44.3	46.4	45.43	1.45	(+) 3.90
	SFGC	47.1	42.5	40.7	45.4	45.7	46.4	44.63	2.27	(+) 4.70
	SGDD	47.1	46.5	44.3	45.3	46.0	46.8	46.00	0.96	(+) 3.33
	GDEM	49.4	50.3	49.4	48.3	49.6	49.0	**49.33**	**0.60**	-
Reddit ($r = 0.10\%$)	GCOND	89.4	89.6	87.8	75.5	67.1	78.8	81.37	8.35	(+) 10.10
	SFGC	89.7	89.5	88.3	82.8	87.8	85.4	87.25	2.44	(+) 4.22
	SGDD	91.0	89.4	89.2	78.4	72.4	81.4	83.63	6.80	(+) 7.84
	GDEM	93.1	90.0	92.6	90.0	92.7	90.4	**91.47**	**1.35**	-
Gamers ($r = 0.25\%$)	GCOND	58.9	54.2	60.1	60.3	59.1	59.3	58.65	2.05	(+) 1.57
	SFGC	58.8	55.0	56.3	57.2	57.5	59.8	57.43	1.57	(+) 2.79
	SGDD	57.7	54.6	56.0	57.3	58.8	58.6	57.17	1.47	(+) 3.05
	GDEM	60.8	59.5	61.0	59.9	59.8	60.3	**60.22**	**0.54**	-

Table 3. Generalization of different distillation methods across GNNs.

Optimal Performance and Time Overhead

- Optimal performance of GDEM and baseline

Evaluation	GCN	SGC	PPNP	Cheb.	Bern.	GPR.
GCOND	77.7	**77.6**	77.9	77.3	**78.2**	78.3
SGDD	78.0	76.6	**78.7**	77.5	78.0	78.3
GDEM	**78.4**	76.1	78.1	**78.1**	**78.2**	**78.6**

Table 4. Optimal performance of different methods.

- Time overhead

Distillation	GCN	SGC	PPNP	Cheb.	Bern.	GPR.	Overall
GCOND	1.99	1.36	1.52	3.89	56.94	3.05	68.75
SGDD	2.95	2.18	2.33	4.95	58.07	4.28	74.76
GDEM	-	-	-	-	-	-	**1.79**

Table 5. Time overhead (s) of different methods.

Ablation Study

- Ablation studies

Pubmed	GCN (↑)	GPR. (↑)	Avg. (↑)	Var. (↓)
GDEM	78.4 / 60.8	78.6 / 60.3	77.92 / 60.22	0.69 / 0.29
w/o $\mathcal{L}_e$	76.1 / 56.5	76.9 / 59.8	76.13 / 58.93	1.18 / 2.39
w/o $\mathcal{L}_o$	77.9 / 59.0	76.4 / 58.9	77.07 / 58.85	2.15 / 2.34
w/o $\mathcal{L}_d$	76.7 / 59.9	77.2 / 60.3	76.77 / 59.78	0.21 / 0.13

Table 6. Ablation studies of on Pubmed / Gamers

- Parameters analysis

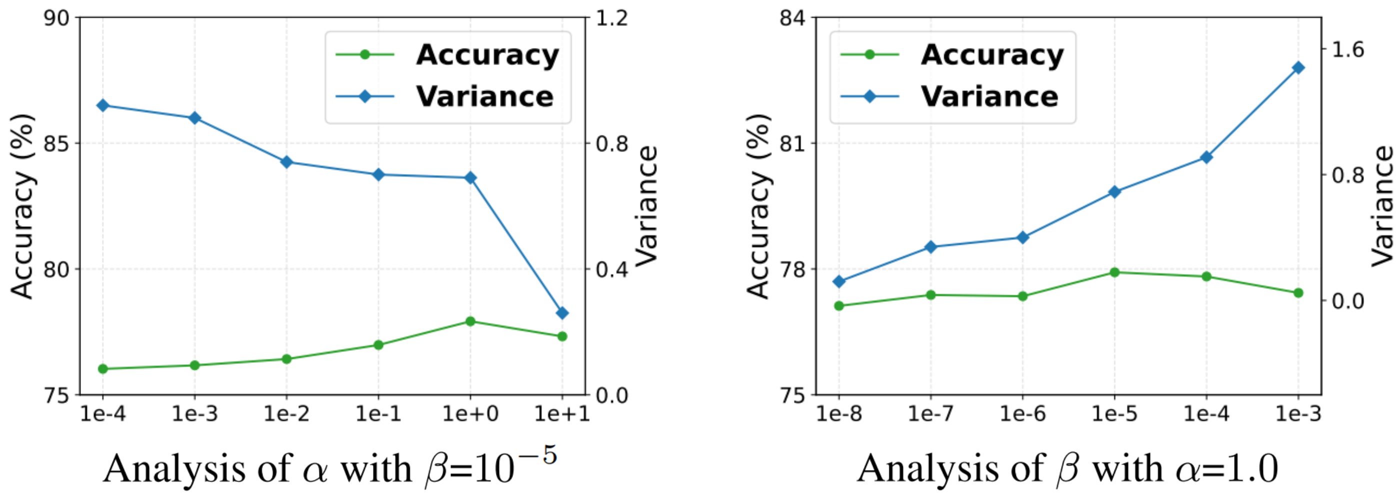
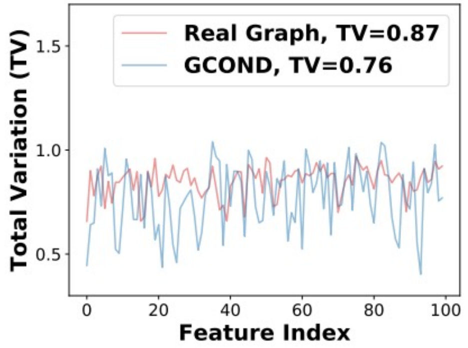


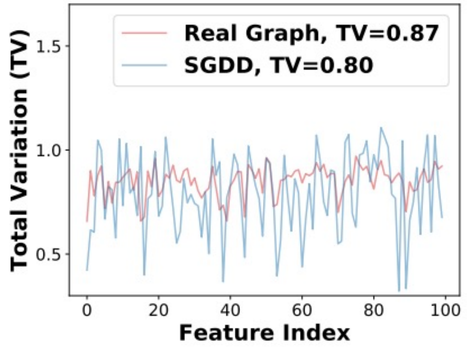
Figure 3. Parameters analysis of α and β.

3 Experiments Visualization

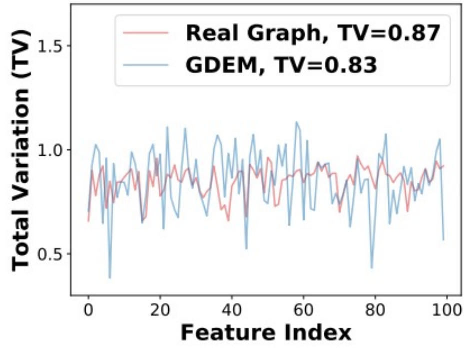
- TVs of synthetic graphs distilled by different methods.



(a) GCOND



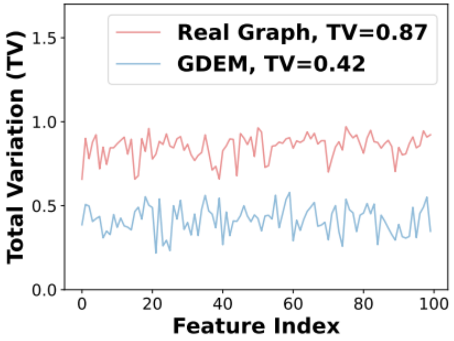
(b) SGDD



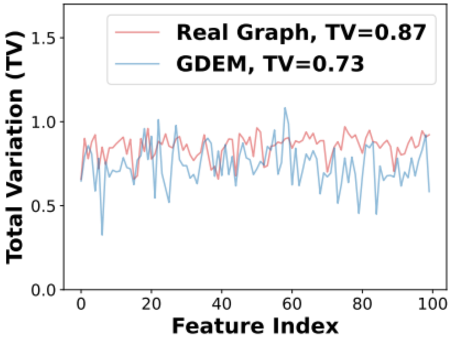
(c) GDEM

Figure 4. TVs of synthetic graphs distilled by different methods.

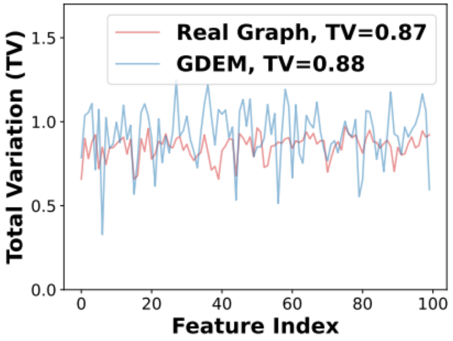
- TVs of synthetic graphs constructed by GDEM at different epochs.



(a) Epoch: 50



(b) Epoch: 100



(c) Epoch: 200

Figure 5. TVs of synthetic graphs at different epochs (GDEM).

4 Conclusion

- **Systematically analysis:** We systematically analyze the limitations of existing distillation methods, including spectrum bias and traversal requirement.

- **Novel framework:** We propose GDEM, a novel graph distillation framework, which mitigates the dependence on GNNs by matching the eigenbasis instead of the entire graph structure.

- **SOTA performance:** Extensive experiments on seven graph datasets validate the superiority of GDEM over state-of-the-art GD methods in terms of effectiveness, generalization, and efficiency.

Paper

Code

Thanks

Q&A

千

扫一扫上面的二维码图案，加我为好友。

liuyangjanet@bupt.edu.cn