GraphSHA: Synthesizing Harder Samples for Class-Imbalanced Node Classification

Wen-Zhi Li[1,2], Chang-Dong Wang[1], Hui Xiong[2,3], Jian-Huang Lai[1]

[1]CSE, Sun Yat-sen University

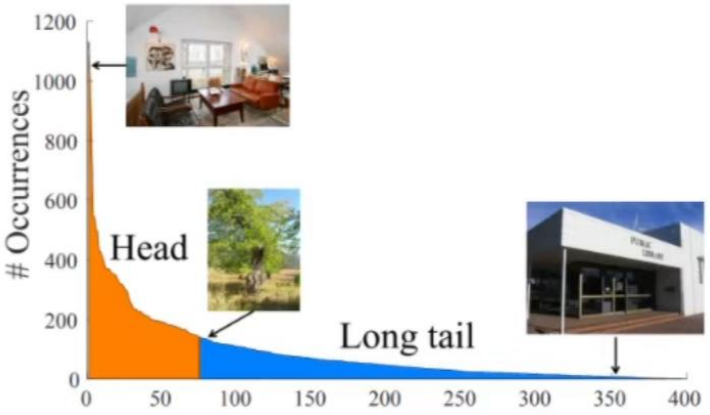
[2]AI Thrust, HKUST(GZ)

[3]CSE, HKUST

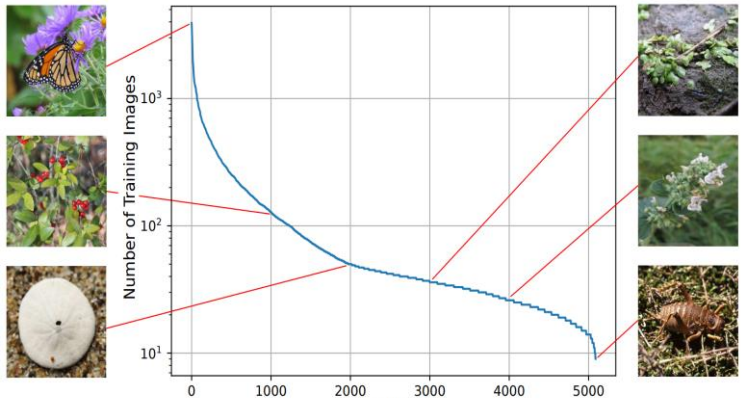
Project Page: https://wenzhilics.github.io/GraphSHA.html

Contact: liwzh63@mail2.sysu.edu.cn

Class imbalanced data in-the-wild

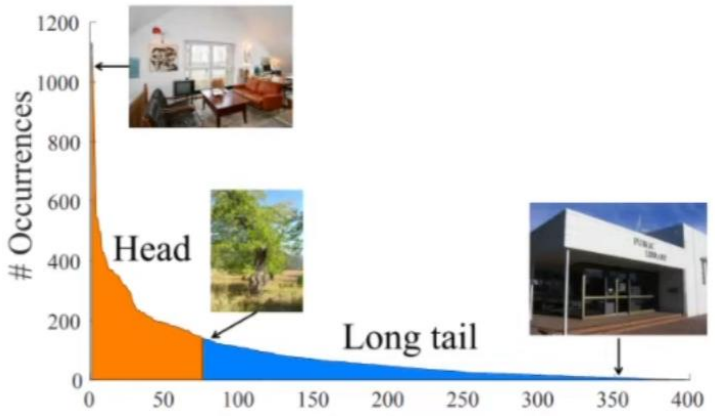


Places

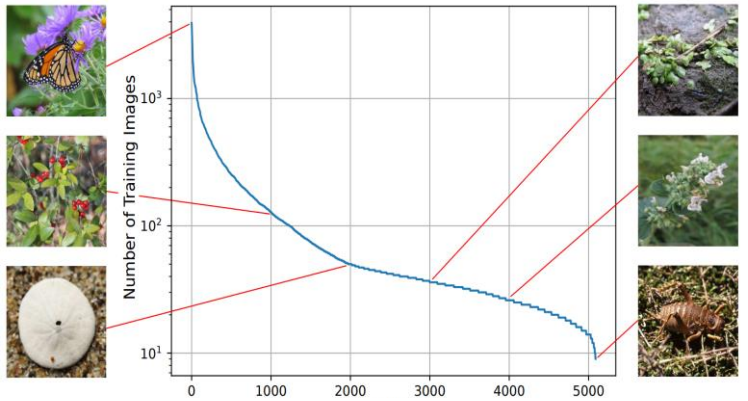


Species

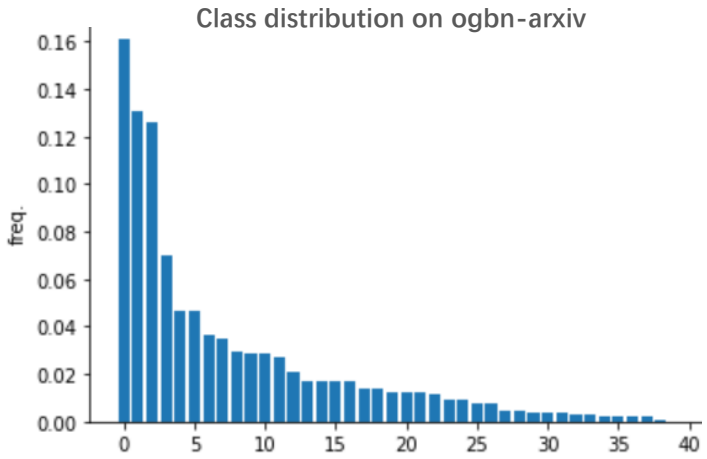
Class imbalanced data in-the-wild



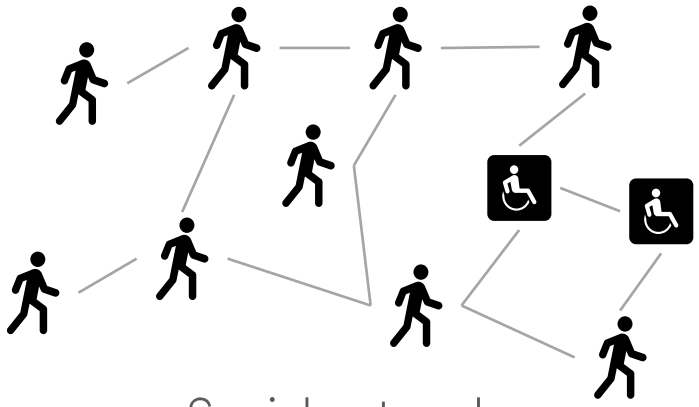
Places



Species

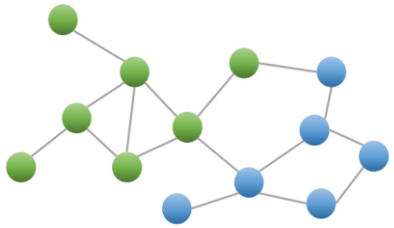


Citation networks

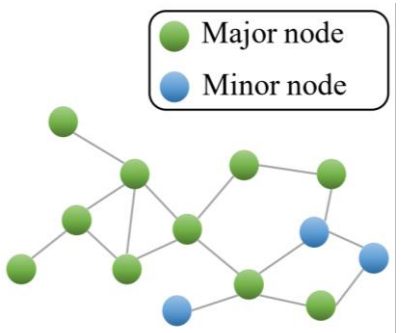


Social networks

Class imbalanced graph



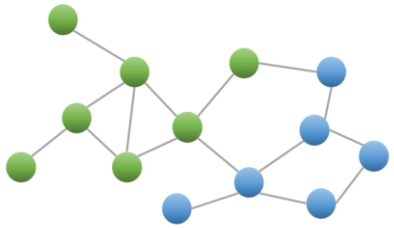
(a) Class-balanced graph



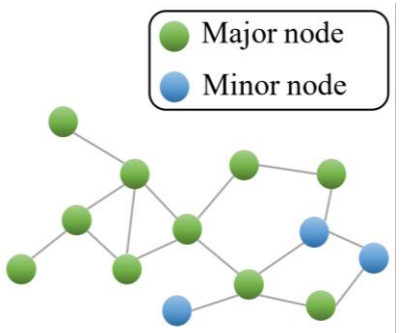
(b) Class-imbalanced graph

- **GNNs** — class balance assumption.
- **Under-represent** minor classes.

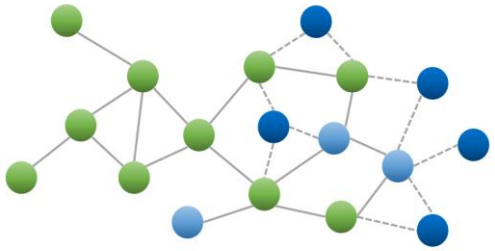
Class imbalanced graph



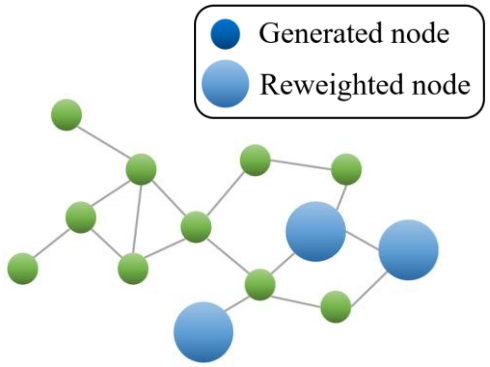
(a) Class-balanced graph



(b) Class-imbalanced graph



(c) Generative approach

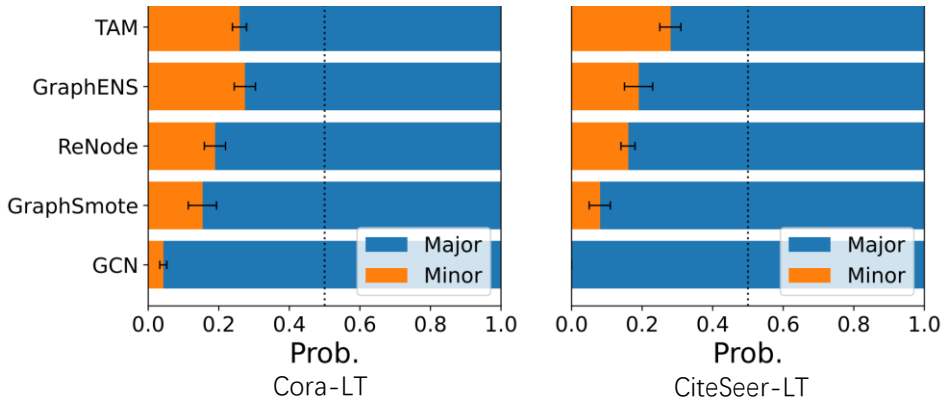


(d) Loss-modifying approach

- **GNNs** — class balance assumption.
- **Under-represent** minor classes.

- **Generative approaches**
  Augmenting the original class-imbalanced graph by **synthesizing** plausible minor nodes.
- **Loss-modifying approaches**
  Adjusting the **objective function** to pay more attention to minor class samples.

Empirical study: squeezed minority problem

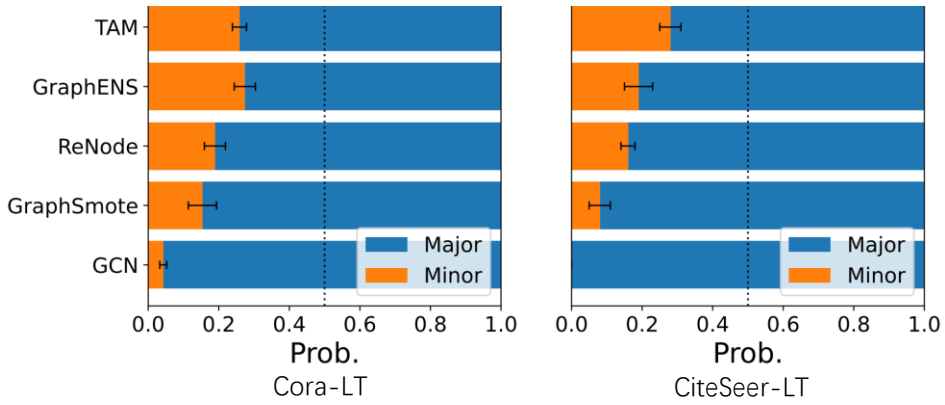


Probability distribution of misclassified samples.

- Minor classes are **squeezed** by major ones.
- Baselines cannot tackle the problem.

  → Enlarging the minor decision boundaries!

Empirical study: squeezed minority problem

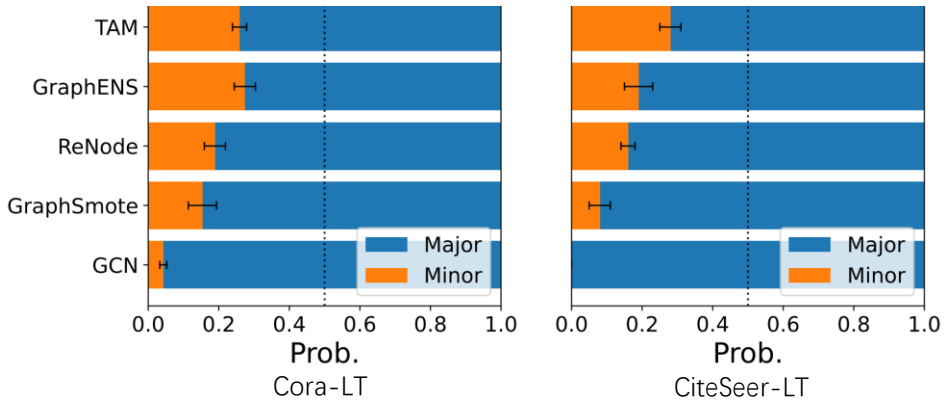


Probability distribution of misclassified samples.

- Minor classes are **squeezed** by major ones.
- Baselines cannot tackle the problem.

  → Enlarging the minor decision boundaries!

Empirical study: squeezed minority problem



Probability distribution of misclassified samples.

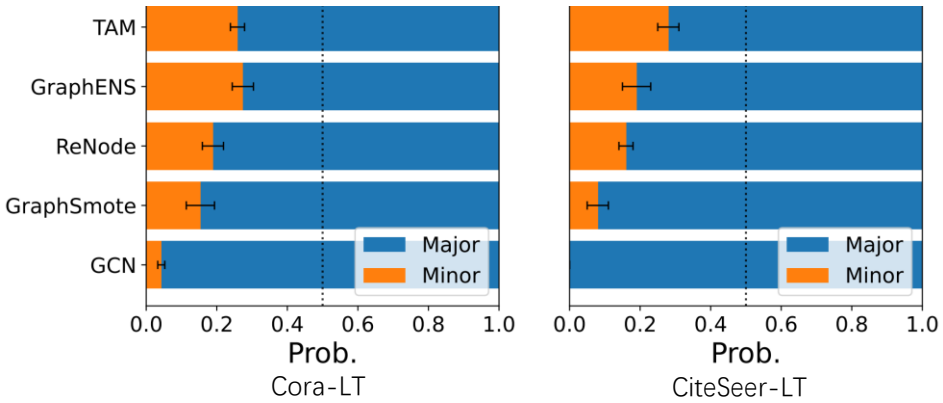
- Minor classes are **squeezed** by major ones.
- Baselines cannot tackle the problem.

→ Enlarging the minor decision boundaries!

Challenges

- The decision boundary is **shared** by a minor class and its neighbor class.
- **Enlarging the subspaces of minor classes while avoiding deteriorating those of the neighbor ones.**

Empirical study: squeezed minority problem



Probability distribution of misclassified samples.

- Minor classes are **squeezed** by major ones.
- Baselines cannot tackle the problem.

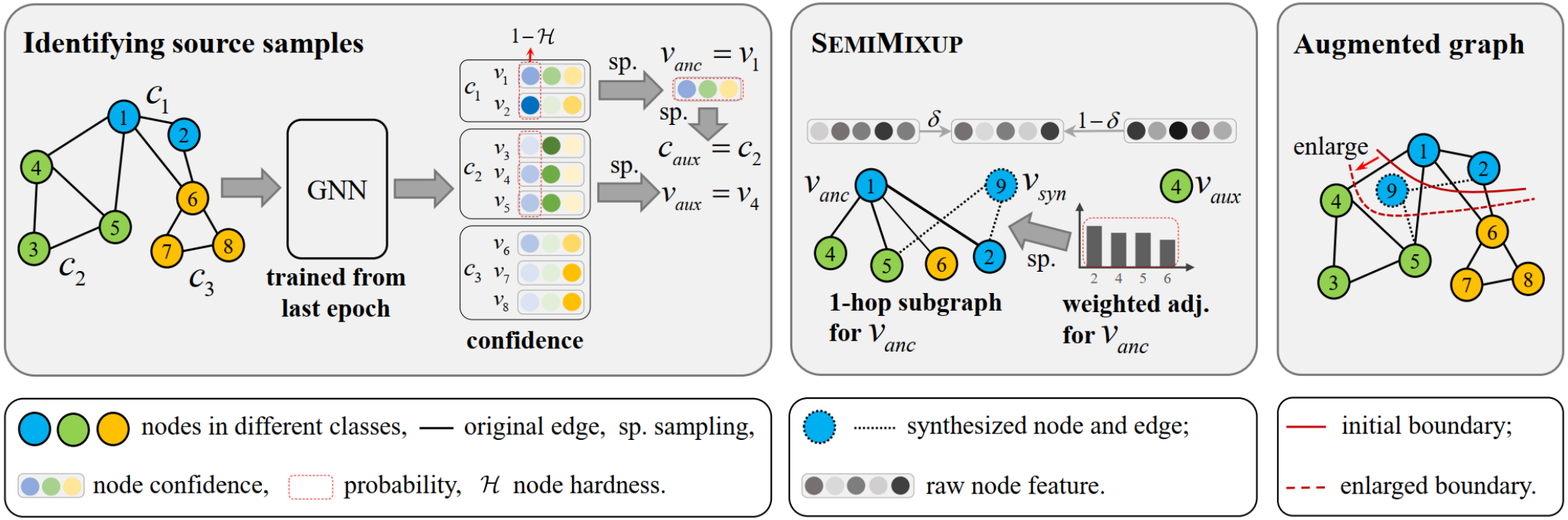
→ Enlarging the minor decision boundaries!

Challenges

- The decision boundary is **shared** by a minor class and its neighbor class.
- **Enlarging the subspaces of minor classes while avoiding deteriorating those of the neighbor ones.**

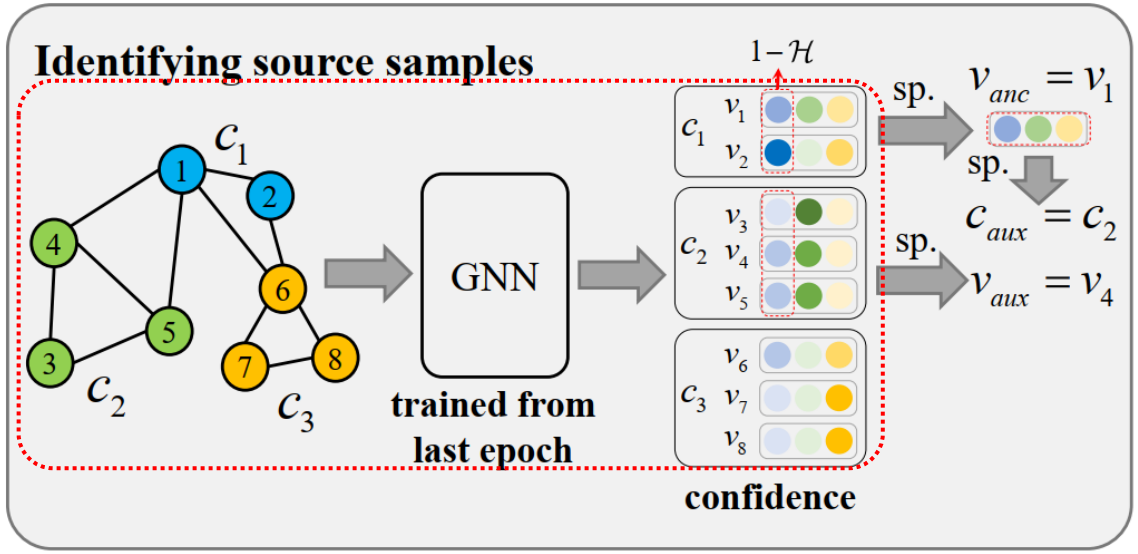
Our solution: **GraphSHA** for **S**ynthesizing **HA**rder minor samples.

GraphSHA: Synthesizing harder minor samples



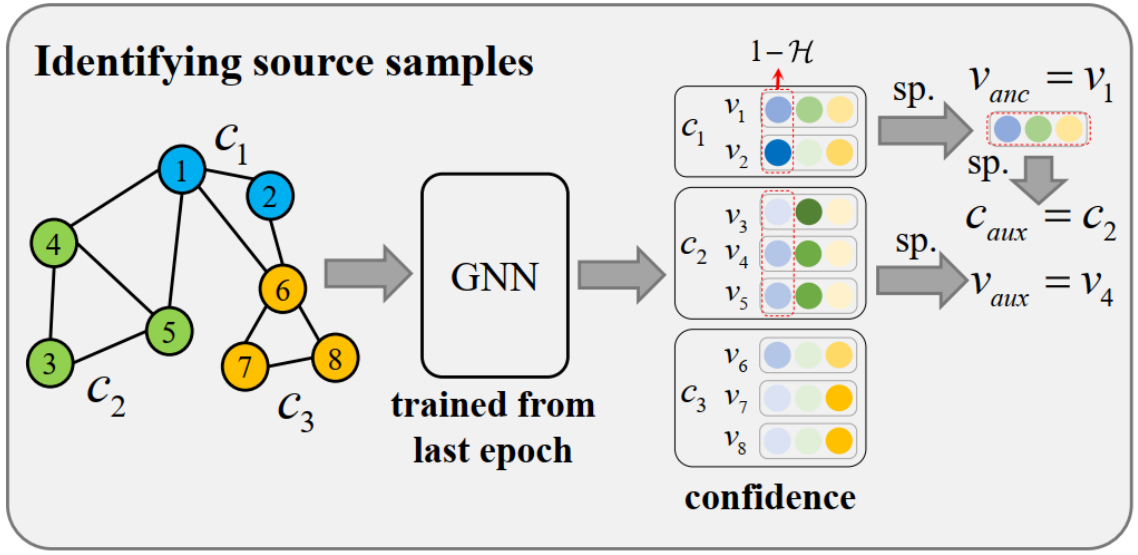
- **Left**: Identifying two source nodes near the boundary via three samplings.
- **Middle**: Enlarging the minor decision boundary via the SEMIMIXUP module.
- **Right**: The enlarged minor decision boundary.

Identifying source samples



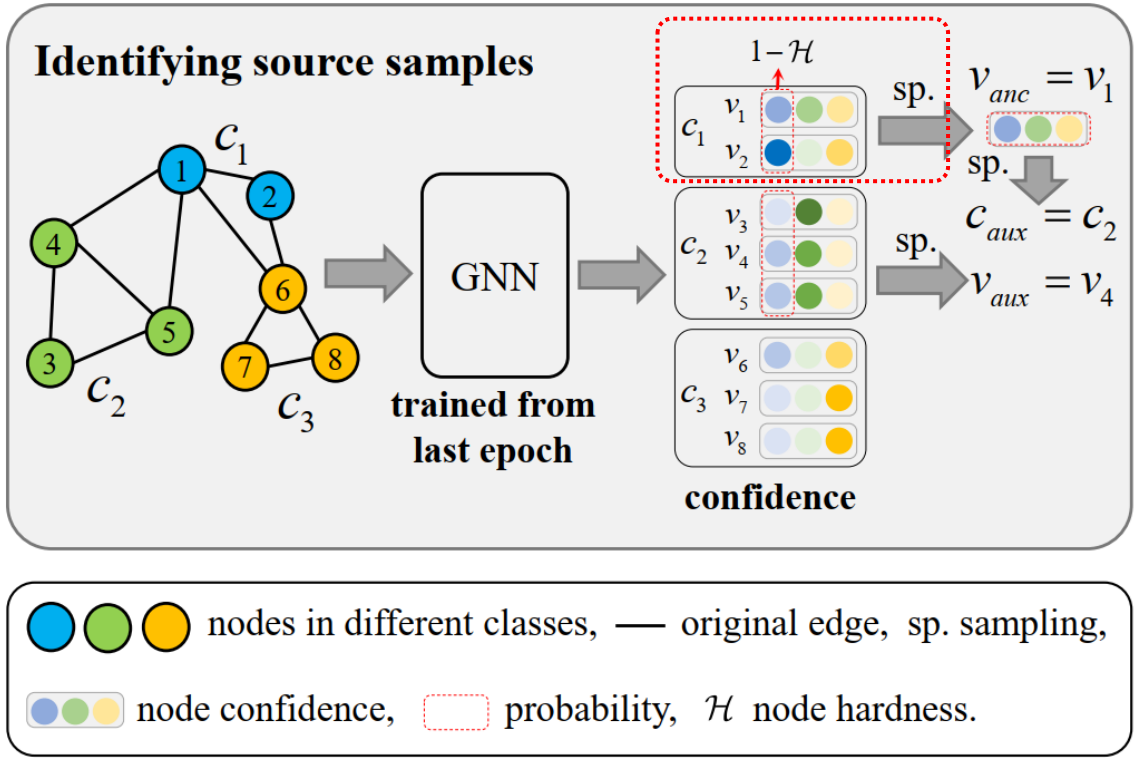
Identifying source samples

Def. (node hardness): $\mathcal{H}_i = 1 - \text{softmax}\left(\boldsymbol{Z}_{i,\boldsymbol{Y}(v_i)}\right)$, where $\boldsymbol{Z}_i = f_\theta(v_i) \in \mathbb{R}^C$.



Identifying source samples

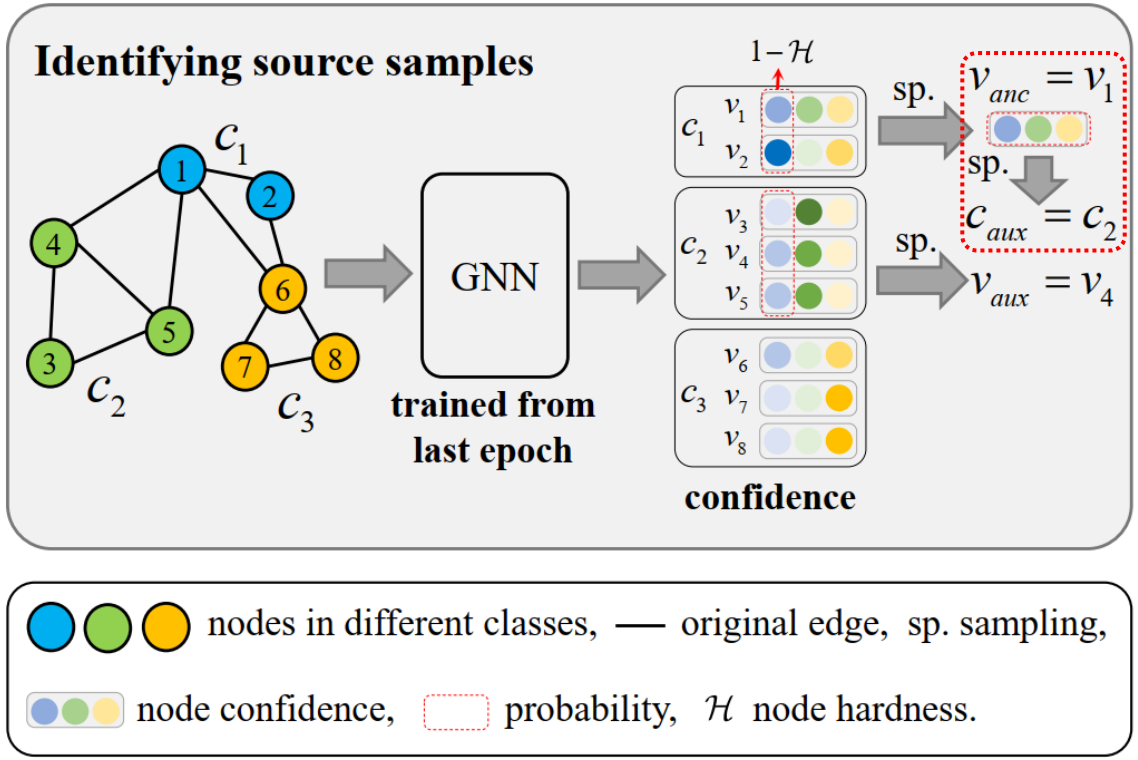
Def. (node hardness): $\mathcal{H}_i = 1 - \text{softmax}\left(\boldsymbol{Z}_{i,\boldsymbol{Y}(v_i)}\right)$, where $\boldsymbol{Z}_i = f_\theta(v_i) \in \mathbb{R}^C$.



- Sampling from minor nodes in c_1 according to $\mathcal{H}$ to get v_{anc}.

Identifying source samples

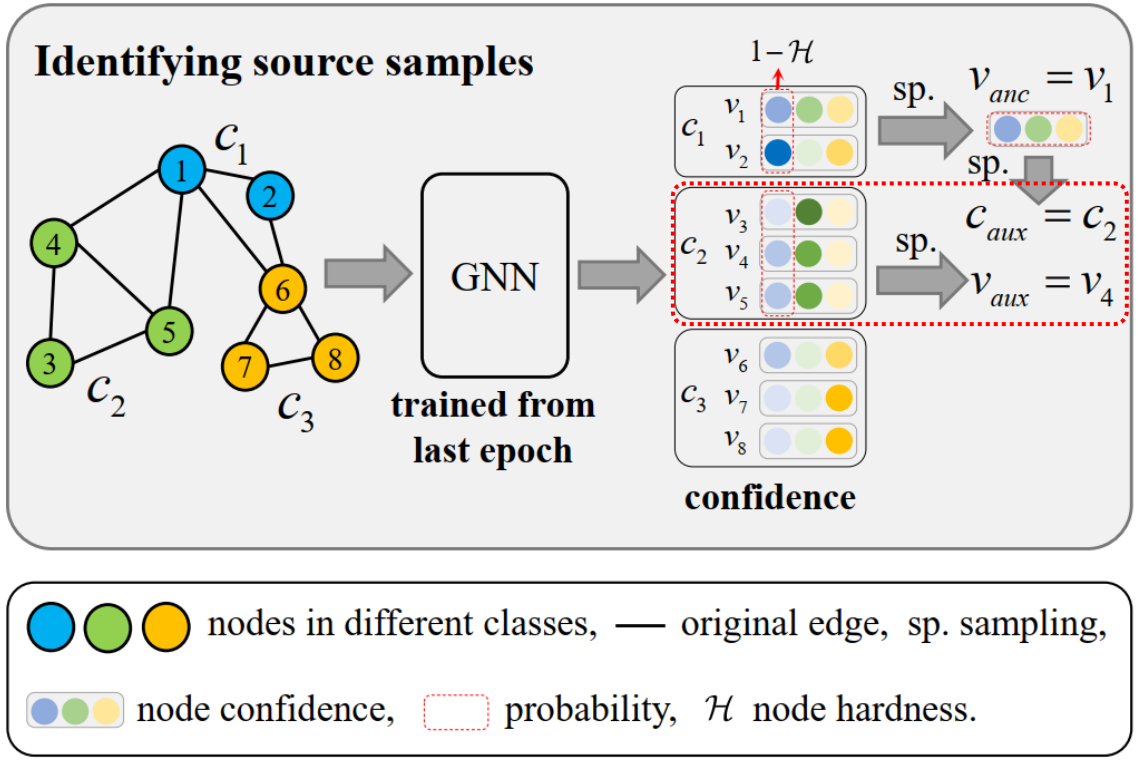
Def. (node hardness): $\mathcal{H}_i = 1 - \text{softmax}\left(\boldsymbol{Z}_{i,\boldsymbol{Y}(v_i)}\right)$, where $\boldsymbol{Z}_i = f_\theta(v_i) \in \mathbb{R}^C$.



- Sampling from minor nodes in c_1 according to $\mathcal{H}$ to get v_{anc}.
- Sampling from major classes c_2, c_3 according to v_{anc}'s confidence on them to get neighbor class c_{aux}.

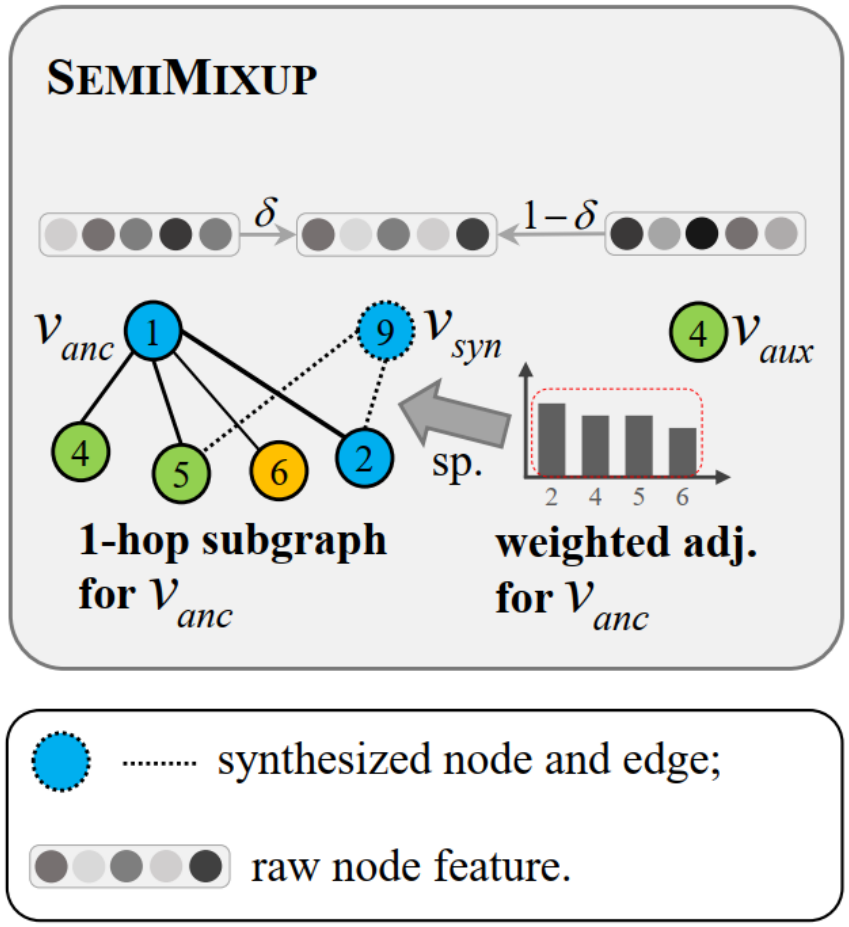
Identifying source samples

Def. (node hardness): $\mathcal{H}_i = 1 - \text{softmax}\left(\boldsymbol{Z}_{i,\boldsymbol{Y}(v_i)}\right)$, where $\boldsymbol{Z}_i = f_\theta(v_i) \in \mathbb{R}^C$.

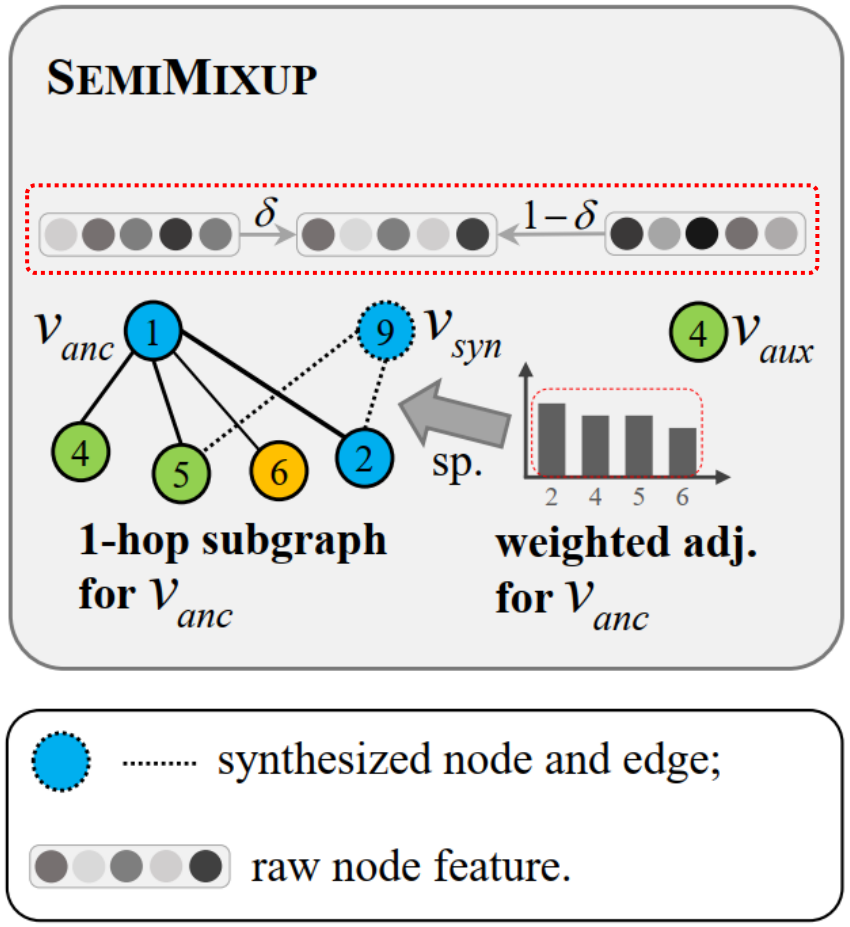


- Sampling from minor nodes in c_1 according to $\mathcal{H}$ to get v_{anc}.
- Sampling from major classes c_2, c_3 according to v_{anc}'s confidence on them to get neighbor class c_{aux}.
- Sampling from nodes in neighbor class c_{aux} according to their confidences on minor class c_1 to get v_{aux}.

SEMIMIXUP Module



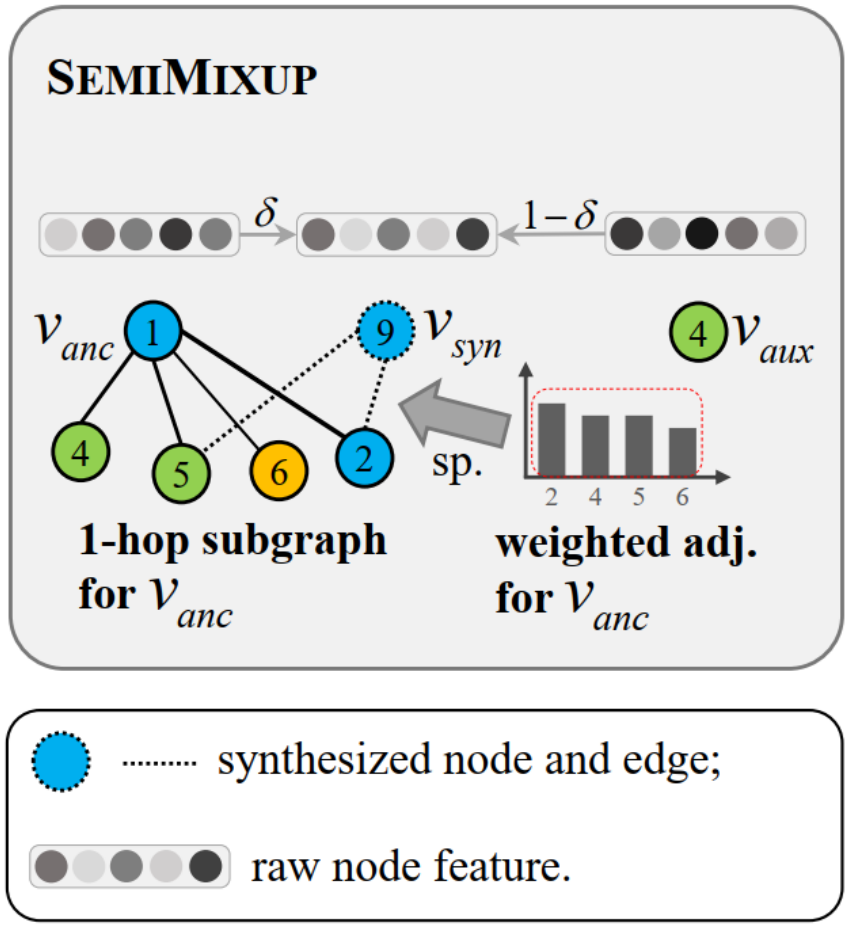
SEMIMIXUP Module



- **Node feature synthesis**

$$X_{syn} = \delta X_{anc} + (1-\delta)X_{aux},\ \delta \in [0,1].$$

SEMIMIXUP Module

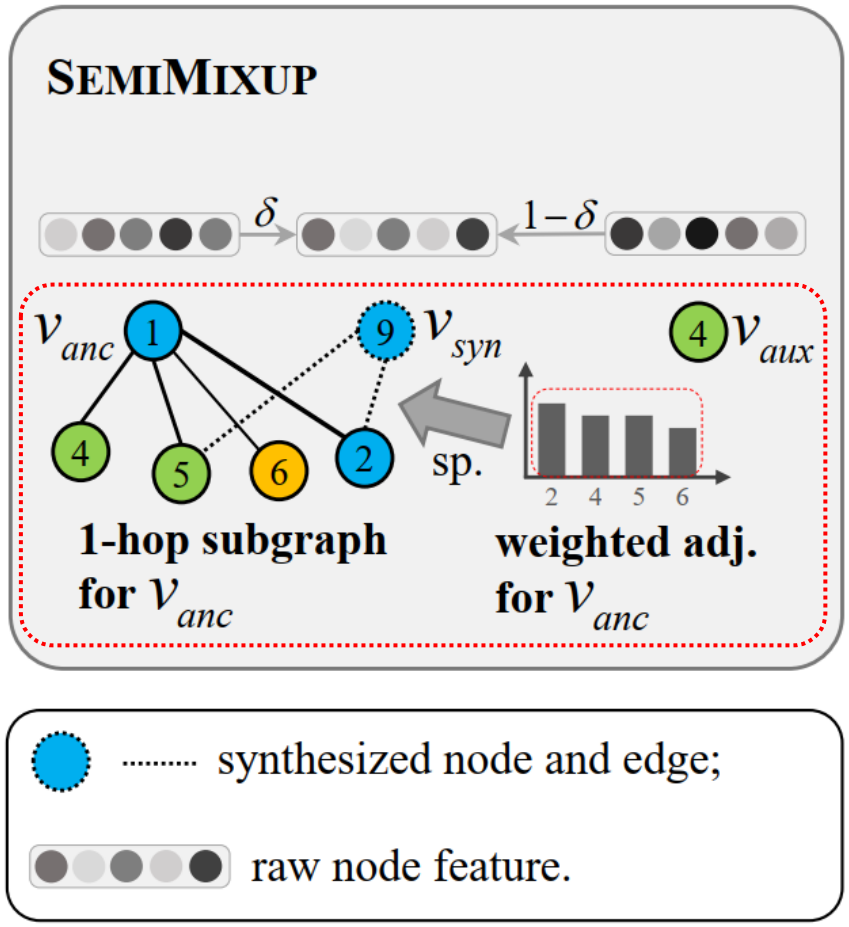


- **Node feature synthesis**

$$X_{syn} = \delta X_{anc} + (1-\delta)X_{aux},\ \delta \in [0,1].$$

- **Edge synthesis**
  - Enabling enlarged boundary propagating to the interior of minor class;
  - Blocking propagation from minor class to neighbor class.

SEMIMIXUP Module



- **Node feature synthesis**

$$X_{syn} = \delta X_{anc} + (1-\delta)X_{aux},\ \delta \in [0,1].$$

- **Edge synthesis**
  - Enabling enlarged boundary propagating to the interior of minor class;
  - Blocking propagation from minor class to neighbor class.

  - Leveraging GDC to build weighted adjacency matrix $\tilde{S}$.
  - Sampling according to $\tilde{S}_{anc}$ to get the neighbor set.

Experiments

How does GraphSHA perform in real-world class-imbalanced node classification problems?

Can the SEMIMIXUP module avoid deteriorating the subspaces of major classes in the latent space?

Does GraphSHA really solve the squeezed minority problem?

Evaluation results

- Long-tailed setting
- Step setting

Evaluation results

- Long-tailed setting

	Dataset
	ρ=100
GCN	Vanilla
	Reweight
	PC Softmax
	CB Loss
	Focal Loss
	ReNode
	Upsample
	GraphSmote
	GraphENS
	TAM (G-ENS)
	GraphSHA

GNN

Loss-modifying approaches

Generative approaches

Evaluation results

- Long-tailed setting

	Dataset	Cora-LT			CiteSeer-LT			PubMed-LT		
	ρ=100	Acc.	bAcc.	F1	Acc.	bAcc.	F1	Acc.	bAcc.	F1
GCN	Vanilla	72.02±0.50	59.42±0.74	59.23±1.02	51.40±0.44	44.64±0.42	37.82±0.67	51.58±0.60	42.11±0.48	34.73±0.71
	Reweight	78.42±0.10	72.66±0.17	73.75±0.15	63.61±0.22	56.80±0.20	55.18±0.18	77.02±0.14	72.45±0.17	72.12±0.15
	PC Softmax	77.30±0.13	72.08±0.30	71.65±0.34	62.15±0.45	**59.08**±0.28	58.13±0.31	74.36±0.62	72.59±0.34	71.79±0.50
	CB Loss	77.97±0.19	72.70±0.28	73.17±0.22	61.47±0.51	55.18±0.52	53.47±0.65	76.57±0.19	72.16±0.18	72.84±0.19
	Focal Loss	78.43±0.19	73.17±0.23	73.76±0.20	59.66±0.38	53.39±0.33	51.80±0.39	75.67±0.20	71.34±0.24	72.03±0.21
	ReNode	78.93±0.13	73.13±0.17	74.46±0.16	62.39±0.31	55.62±0.27	54.05±0.24	76.00±0.16	70.68±0.15	71.41±0.15
	Upsample	75.52±0.11	66.68±0.14	68.35±0.15	55.05±0.11	48.41±0.11	45.22±0.14	71.58±0.06	63.79±0.06	64.62±0.07
	GraphSmote	75.44±0.43	68.99±0.51	70.41±0.52	56.58±0.29	50.39±0.28	47.96±0.33	74.62±0.08	69.53±0.10	71.18±0.09
	GraphENS	76.15±0.24	71.16±0.40	70.85±0.49	63.14±0.35	56.92±0.37	55.54±0.41	77.11±0.11	71.89±0.15	72.71±0.14
	TAM (G-ENS)	77.30±0.23	72.10±0.29	72.25±0.29	63.40±0.34	57.15±0.35	55.68±0.40	78.07±0.15	72.63±0.23	72.96±0.22
	GraphSHA	**79.90**±0.29	**74.62**±0.35	**75.74**±0.32	**64.50**±0.41	59.04±0.34	**59.16**±0.21	**79.20**±0.13	**74.46**±0.17	**75.24**±0.27

Evaluation results

- Step setting

	Dataset	Photo-ST			Computer-ST			CS-ST		
	ρ=20	Acc.	bAcc.	F1	Acc.	bAcc.	F1	Acc.	bAcc.	F1
SAGE	Vanilla	59.19±2.32	59.98±1.82	47.11±2.95	63.88±0.05	46.96±0.04	30.08±0.11	74.81±0.35	79.69±0.19	64.68±0.52
	Reweight	84.85±0.23	87.30±0.24	82.89±0.16	83.59±0.31	87.91±0.10	77.59±0.41	91.02±0.33	90.87±0.30	75.50±0.35
	PC Softmax	86.16±0.13	86.93±0.15	83.55±0.09	81.38±0.17	80.50±0.76	72.30±0.57	92.58±0.23	92.11±0.37	78.00±0.53
	CB Loss	83.02±0.29	85.79±0.21	80.48±0.31	83.75±0.21	87.38±0.10	77.08±0.20	90.85±0.14	90.77±0.10	79.04±0.85
	Focal Loss	82.58±0.39	85.42±0.32	79.28±0.35	82.56±0.22	87.38±0.08	76.53±0.15	90.08±0.19	90.01±0.16	79.56±0.26
	ReNode	84.83±0.15	86.43±0.20	81.85±0.22	81.29±0.34	87.33±0.17	76.60±0.28	90.98±0.31	91.17±0.35	**81.22**±0.43
	Upsample	82.20±0.34	84.86±0.08	79.38±0.24	82.99±0.24	87.02±0.09	77.10±0.33	87.23±0.18	87.99±0.11	76.38±0.21
	GraphSmote	80.21±0.27	84.68±0.31	79.05±0.38	83.62±0.25	88.15±0.21	76.02±0.30	86.30±0.12	85.66±0.09	69.19±0.14
	GraphENS	88.02±0.09	90.55±0.11	86.70±0.10	83.28±0.38	88.54±0.10	76.77±0.52	92.13±0.16	92.53±0.22	78.23±0.20
	TAM (G-ENS)	87.61±0.13	89.17±0.17	85.74±0.16	80.31±0.52	86.74±0.22	76.96±0.59	92.60±0.22	92.39±0.19	78.52±0.23
	GraphSHA	**89.14**±0.22	**90.60**±0.10	**87.25**±0.18	**84.21**±0.50	**89.49**±0.10	**77.93**±0.91	**92.93**±0.05	**92.78**±0.13	79.68±0.16

Evaluation results

	Dataset	Cora-LT			CiteSeer-LT			PubMed-LT		
	ρ=100	Acc.	bAcc.	F1	Acc.	bAcc.	F1	Acc.	bAcc.	F1
GAT	Vanilla	67.52±0.58	54.20±0.79	55.34±0.74	49.16±0.19	42.58±0.18	35.75±0.29	47.83±1.57	39.09±1.27	29.62±2.15
	Reweight	77.77±0.28	72.03±0.58	72.79±0.58	61.95±0.57	55.40±0.63	53.71±0.72	74.08±0.39	69.35±0.79	69.52±0.70
	PC Softmax	68.75±0.77	64.07±0.86	64.17±0.74	56.70±1.61	56.31±1.30	55.31±1.53	76.70±0.32	73.22±0.15	73.25±0.28
	CB Loss	77.29±0.36	72.07±0.63	72.79±0.43	61.44±0.52	55.17±0.52	53.63±0.54	74.81±0.25	69.54±0.64	70.55±0.59
	Focal Loss	77.97±0.11	72.47±0.21	73.15±0.21	59.75±0.36	53.44±0.34	52.12±0.29	74.23±0.27	70.36±0.34	70.63±0.20
	ReNode	78.09±0.24	71.78±0.34	73.41±0.34	60.87±0.37	54.01±0.37	51.98±0.44	74.09±0.28	69.02±0.34	69.55±0.32
	Upsample	72.62±0.31	62.39±0.37	65.08±0.28	53.41±0.22	46.89±0.22	43.10±0.44	67.61±0.95	57.29±0.64	54.99±1.02
	GraphSmote	74.65±0.29	67.71±0.37	69.10±0.39	57.45±0.26	51.33±0.31	49.38±0.54	74.04±0.38	69.04±0.35	70.62±0.42
	GraphENS	77.08±0.26	72.07±0.38	72.09±0.48	61.91±0.34	55.88±0.32	54.38±0.41	76.65±0.11	70.43±0.20	71.25±0.20
	TAM (G-ENS)	77.69±0.21	72.87±0.30	72.99±0.31	**64.06**±0.34	57.77±0.31	56.38±0.32	77.94±0.18	71.98±0.29	73.07±0.27
	GraphSHA	**79.07**±0.18	**74.08**±0.26	**75.02**±0.18	63.94±0.44	**58.14**±0.35	**57.71**±0.40	**78.40**±0.20	**73.82**±0.17	**74.66**±0.21
SAGE	Vanilla	73.30±0.09	61.83±0.12	63.25±0.13	47.90±0.24	41.80±0.22	36.96±0.31	58.78±0.08	47.92±0.06	42.34±0.07
	Reweight	76.81±0.15	68.74±0.31	70.22±0.37	57.30±0.53	50.90±0.46	49.15±0.49	65.94±0.53	59.83±1.24	58.89±1.14
	PC Softmax	76.92±0.22	73.25±0.28	73.54±0.26	58.35±0.25	56.06±0.18	56.65±0.18	71.60±0.13	73.83±0.15	70.28±0.12
	CB Loss	77.04±0.30	70.25±0.37	71.26±0.30	57.63±0.34	51.19±0.32	48.70±0.35	67.78±0.36	60.67±0.46	61.46±0.52
	Focal Loss	77.17±0.16	69.78±0.27	70.76±0.25	57.02±0.72	50.77±0.66	48.42±0.79	70.59±0.35	65.69±0.45	66.25±0.44
	ReNode	77.26±0.15	69.22±0.21	71.13±0.22	57.82±0.50	51.27±0.49	49.04±0.45	67.60±0.51	60.65±0.82	60.78±0.78
	Upsample	73.80±0.12	63.45±0.20	65.83±0.16	50.32±0.11	44.24±0.11	41.46±0.17	64.08±0.06	54.64±0.07	53.39±0.10
	GraphSmote	74.24±0.19	66.15±0.38	67.89±0.41	52.85±0.64	46.99±0.63	44.20±0.74	65.10±0.42	56.82±0.49	56.85±0.54
	GraphENS	76.69±0.20	70.07±0.25	70.37±0.30	62.63±0.34	56.14±0.37	54.13±0.39	77.62±0.14	72.54±0.23	73.21±0.18
	TAM (G-ENS)	77.31±0.30	71.02±0.34	71.14±0.36	62.93±0.21	56.44±0.19	54.50±0.21	78.12±0.31	72.80±0.76	73.69±0.67
	GraphSHA	**78.80**±0.24	**73.56**±0.35	**74.27**±0.30	**63.76**±0.38	**58.25**±0.37	**58.04**±0.45	**78.20**±0.19	**74.07**±0.23	**74.93**±0.23

	Dataset	Photo-ST			Computer-ST			CS-ST		
	ρ=20	Acc.	bAcc.	F1	Acc.	bAcc.	F1	Acc.	bAcc.	F1
GCN	Vanilla	37.79±0.22	46.77±0.11	27.15±0.43	56.12±1.41	41.49±0.63	27.76±0.53	37.36±0.97	54.35±0.72	30.47±1.19
	Reweight	85.81±0.13	88.62±0.06	83.30±0.14	78.77±0.25	85.30±0.10	74.31±0.23	91.86±0.06	91.62±0.06	82.46±0.15
	PC Softmax	64.66±1.73	71.56±1.16	61.31±1.25	73.33±1.22	60.07±1.82	55.09±2.27	87.38±0.49	87.46±0.39	74.24±0.77
	CB Loss	86.85±0.05	88.69±0.05	84.78±0.12	**82.22**±0.13	86.71±0.05	75.80±0.13	91.43±0.05	91.25±0.07	77.72±0.85
	Focal Loss	86.14±0.17	88.44±0.11	84.12±0.23	81.01±0.19	**86.89**±0.07	75.50±0.17	91.01±0.08	90.72±0.04	79.80±0.77
	ReNode	86.08±0.18	87.34±0.34	82.51±0.29	72.92±0.97	78.12±0.84	67.04±1.13	92.02±0.21	91.08±0.19	82.87±0.97
	Upsample	85.40±0.18	87.32±0.15	82.79±0.22	80.07±0.31	85.10±0.11	74.85±0.21	86.11±0.14	86.82±0.10	75.55±0.13
	GraphSmote	83.99±0.20	86.53±0.19	81.86±0.21	76.76±0.18	84.10±0.17	69.40±0.19	86.20±0.17	85.44±0.15	69.04±0.64
	GraphENS	87.00±0.07	**89.19**±0.06	84.66±0.09	79.71±0.08	86.50±0.08	74.55±0.10	92.17±0.10	91.94±0.11	82.90±0.43
	TAM (G-ENS)	84.37±0.11	86.41±0.09	81.91±0.10	76.26±0.23	83.38±0.26	73.85±0.22	92.15±0.22	91.92±0.24	83.13±0.53
	GraphSHA	**87.40**±0.09	88.92±0.09	**85.18**±0.11	81.75±0.14	86.75±0.09	**76.86**±0.30	**92.38**±0.09	**92.01**±0.06	**83.33**±0.45
GAT	Vanilla	37.54±0.34	45.95±0.32	28.87±0.49	58.00±0.69	42.82±0.39	26.79±0.19	34.48±0.42	50.08±0.65	24.92±1.00
	Reweight	80.34±1.02	83.08±0.50	76.64±0.92	72.65±0.40	76.81±0.37	64.00±0.35	88.31±0.38	87.33±0.39	71.67±0.59
	PC Softmax	51.74±3.22	61.48±1.90	51.17±2.44	31.56±2.89	51.83±2.52	37.70±2.25	78.84±0.58	77.80±0.61	65.46±0.64
	CB Loss	82.82±0.79	86.44±0.58	79.57±0.93	79.60±0.71	84.78±0.23	74.11±0.66	89.74±0.26	89.68±0.22	75.00±0.83
	Focal Loss	83.03±0.58	85.86±0.43	79.39±0.73	79.49±0.45	85.04±0.23	74.10±0.48	88.73±0.20	88.03±0.22	73.08±0.73
	ReNode	76.49±1.00	81.35±0.95	73.33±0.86	71.71±0.65	75.20±0.36	63.94±0.77	87.86±0.29	85.55±0.34	69.80±0.47
	Upsample	77.89±0.83	81.16±0.33	73.91±0.59	74.86±0.69	78.18±0.45	66.28±0.81	82.23±0.18	82.70±0.10	65.74±0.17
	GraphSmote	80.71±0.33	81.48±0.38	76.96±0.30	79.38±0.25	84.66±0.20	70.75±0.27	83.46±0.18	82.75±0.18	67.02±0.22
	GraphENS	**84.22**±0.36	86.45±0.19	80.02±0.30	**80.78**±0.18	84.82±0.19	**75.13**±0.43	89.93±0.30	89.71±0.29	79.66±0.38
	TAM (G-ENS)	80.94±0.42	83.09±0.36	78.89±0.45	77.68±0.24	82.97±0.18	74.22±0.39	**91.86**±0.36	90.96±0.33	**80.41**±0.35
	GraphSHA	84.09±0.90	**86.61**±0.81	**80.85**±0.72	80.01±0.42	**84.89**±0.27	71.64±0.55	91.79±0.13	**91.46**±0.10	76.66±0.18

Method	Val Acc.	Test Acc.	Test bAcc.	Test F1
Vanilla (GCN)	73.02±0.14	71.81±0.26	50.96±0.21	50.42±0.18
Reweight	67.49±0.32	66.07±0.55	53.34±0.30	48.07±0.77
PC Softmax	72.19±0.11	71.49±0.25	48.14±0.14	50.59±0.13
CB Loss	65.75±0.23	64.73±0.86	52.66±0.72	47.24±1.25
Focal Loss	67.36±0.24	65.93±0.58	53.06±0.21	48.89±0.72
ReNode	66.44±0.51	65.91±0.20	53.39±0.40	48.18±0.52
TAM (ReNode)	67.91±0.27	66.63±0.66	53.40±0.24	48.71±0.49
Upsample	70.53±0.08	69.55±0.37	46.82±0.07	45.49±0.20
GraphSmote	OOM	OOM	OOM	OOM
GraphENS	OOM	OOM	OOM	OOM
GraphSHA	**73.04**±0.11	**72.14**±0.28	**53.75**±0.16	**53.13**±0.20

ρ=100	Acc./bAcc.	F1
Vanilla (SAGE)	56.40±0.71	51.84±0.78
Reweight	68.17±0.96	67.14±0.96
PC Softmax	66.06±0.38	64.74±0.33
CB Loss	69.89±1.11	69.18±1.21
Focal Loss	67.83±1.20	66.54±1.33
ReNode	59.52±0.58	57.94±0.52
Upsample	63.14±0.77	61.70±0.72
GraphSmote	58.86±0.59	56.95±0.66
GraphENS	63.14±0.92	62.28±0.88
TAM (G-ENS)	64.86±0.78	63.73±0.72
GraphSHA	**73.43**±0.42	**72.50**±0.43

Evaluation results

	Dataset	Cora-LT			CiteSeer-LT			PubMed-LT		
	ρ=100	Acc.	bAcc.	F1	Acc.	bAcc.	F1	Acc.	bAcc.	F1
GAT	Vanilla	67.52±0.58	54.20±0.79	55.34±0.74	49.16±0.19	42.58±0.18	35.75±0.29	47.83±1.57	39.09±1.27	29.62±2.15
	Reweight	77.77±0.28	72.03±0.58	72.79±0.58	61.95±0.57	55.40±0.63	53.71±0.72	74.08±0.39	69.35±0.79	69.52±0.70
	PC Softmax	68.75±0.77	64.07±0.86	64.17±0.74	56.70±1.61	56.31±1.30	55.31±1.53	76.70±0.32	73.22±0.15	73.25±0.28
	CB Loss	77.29±0.36	72.07±0.63	72.79±0.43	61.44±0.52	55.17±0.52	53.63±0.54	74.81±0.25	69.54±0.64	70.55±0.59
	Focal Loss	77.97±0.11	72.47±0.21	73.15±0.21	59.75±0.36	53.44±0.34	52.12±0.29	74.23±0.27	70.36±0.34	70.63±0.20
	ReNode	78.09±0.24	71.78±0.34	73.41±0.34	60.87±0.37	54.01±0.37	51.98±0.41	74.09±0.28	69.02±0.31	69.55±0.32
	GraphSHA	78.80±0.24	73.56±0.35	74.27±0.30	63.76±0.38	58.25±0.37	58.04±0.45	78.20±0.19	74.07±0.23	74.93±0.23

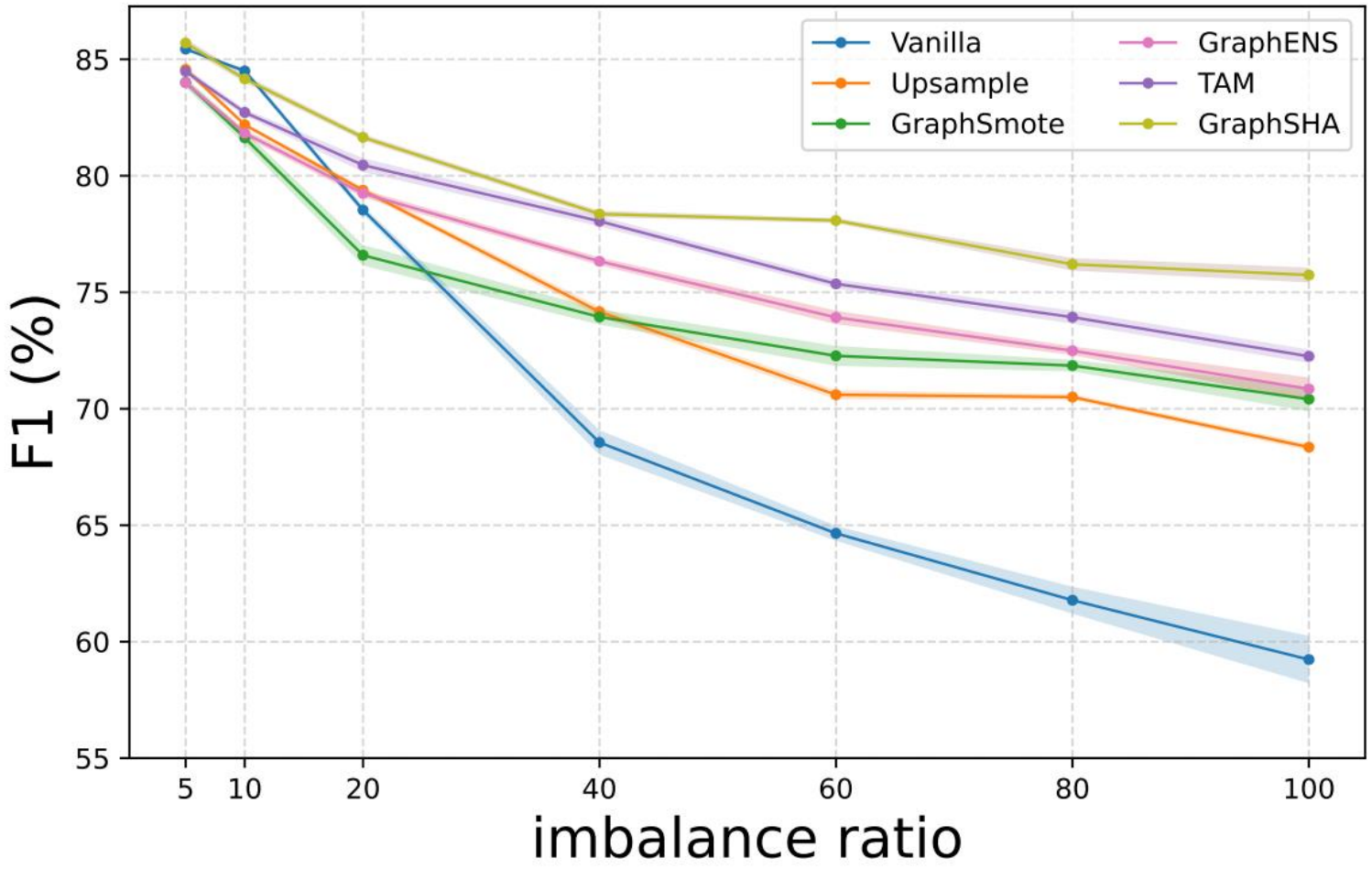
	Dataset	Photo-ST			Computer-ST			CS-ST		
	ρ=20	Acc.	bAcc.	F1	Acc.	bAcc.	F1	Acc.	bAcc.	F1
GCN	Vanilla	37.79±0.22	46.77±0.11	27.15±0.43	56.12±1.41	41.49±0.63	27.76±0.53	37.36±0.97	54.35±0.72	30.47±1.19
	Reweight	85.81±0.13	88.62±0.06	83.30±0.14	78.77±0.25	85.30±0.10	74.31±0.23	91.86±0.06	91.62±0.06	82.46±0.15
	PC Softmax	64.66±1.73	71.56±1.16	61.31±1.25	73.33±1.22	60.07±1.82	55.09±2.27	87.38±0.49	87.46±0.39	74.24±0.77
	CB Loss	86.85±0.05	88.69±0.05	84.78±0.12	**82.22**±0.13	86.71±0.05	75.80±0.13	91.43±0.05	91.25±0.07	77.72±0.85
	Focal Loss	86.14±0.17	88.44±0.11	84.12±0.23	81.01±0.19	**86.89**±0.07	75.50±0.17	91.01±0.08	90.72±0.04	79.80±0.77
	ReNode	86.08±0.18	87.34±0.34	82.51±0.29	72.92±0.97	78.12±0.84	67.04±1.13	92.02±0.21	91.08±0.19	82.87±0.97
	GraphSHA	84.09±0.90	**86.61**±0.81	**80.85**±0.72	80.01±0.42	**84.89**±0.27	71.64±0.55	91.79±0.13	**91.46**±0.10	76.66±0.18

GraphSHA achieves remarkable performance over all tested regions!
(complete results in paper)

Method	Val Acc.	Test Acc.	Test bAcc.	Test F1
Vanilla (GCN)	73.02±0.14	71.81±0.26	50.96±0.21	50.42±0.18
Reweight	67.49±0.32	66.07±0.55	53.34±0.30	48.07±0.77
PC Softmax	72.19±0.11	71.49±0.25	48.14±0.14	50.59±0.13
CB Loss	65.75±0.23	64.73±0.86	52.66±0.72	47.24±1.25
Focal Loss	67.36±0.24	65.93±0.58	53.06±0.21	48.89±0.72
ReNode	66.44±0.51	65.91±0.20	53.39±0.40	48.18±0.52
TAM (ReNode)	67.91±0.27	66.63±0.66	53.40±0.24	48.71±0.49
Upsample	70.53±0.08	69.55±0.37	46.82±0.07	45.49±0.20
GraphSmote	OOM	OOM	OOM	OOM
GraphENS	OOM	OOM	OOM	OOM
GraphSHA	**73.04**±0.11	**72.14**±0.28	**53.75**±0.16	**53.13**±0.20

ρ=100	Acc./bAcc.	F1
Vanilla (SAGE)	56.40±0.71	51.84±0.78
Reweight	68.17±0.96	67.14±0.96
PC Softmax	66.06±0.38	64.74±0.33
CB Loss	69.89±1.11	69.18±1.21
Focal Loss	67.83±1.20	66.54±1.33
ReNode	59.52±0.58	57.94±0.52
Upsample	63.14±0.77	61.70±0.72
GraphSmote	58.86±0.59	56.95±0.66
GraphENS	63.14±0.92	62.28±0.88
TAM (G-ENS)	64.86±0.78	63.73±0.72
GraphSHA	**73.43**±0.42	**72.50**±0.43

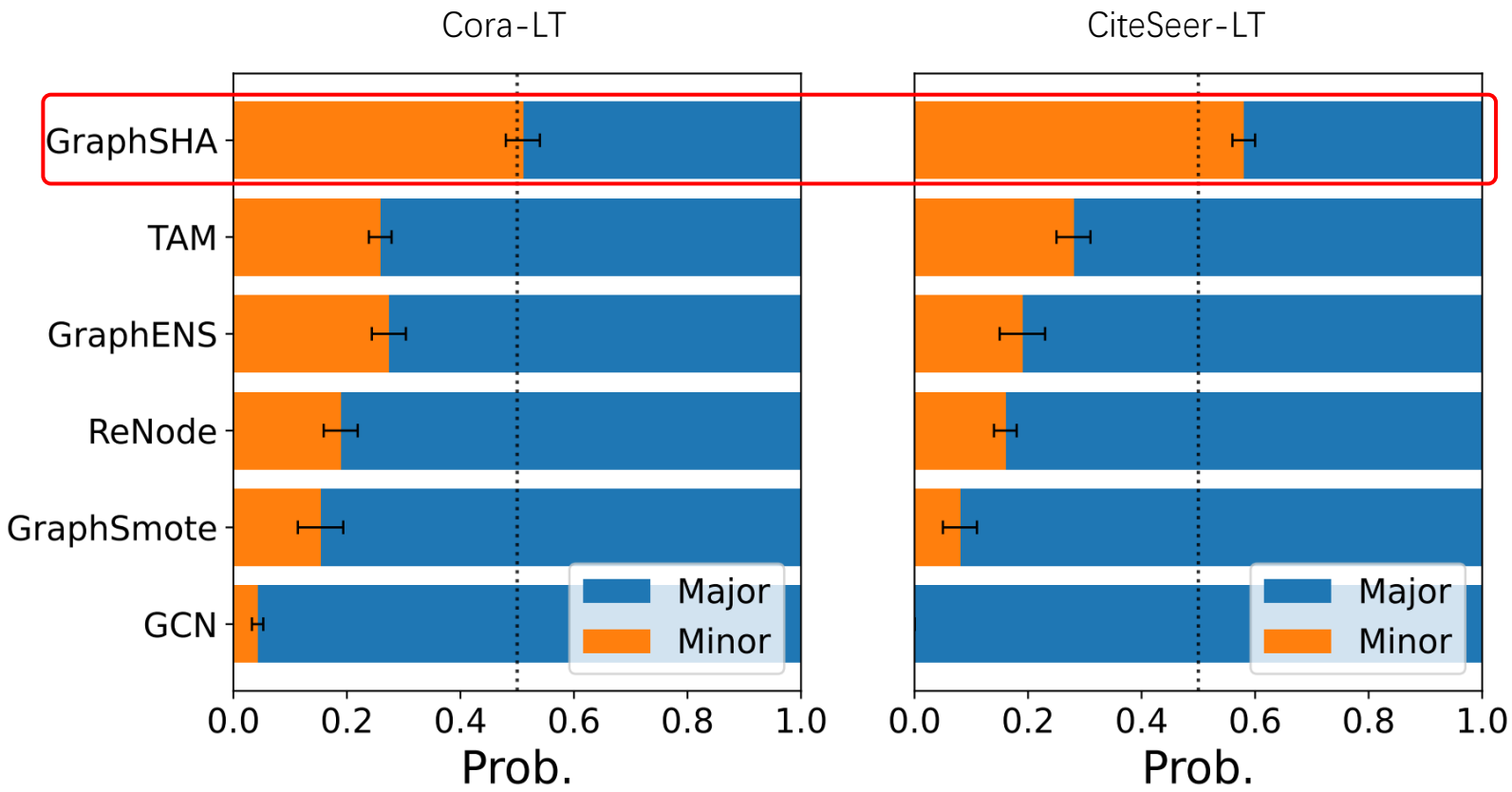
Influence of imbalance ratio



Ablation study

Method	Acc.	bAcc.	F1	C0 (0.5%)	C1 (1.1%)	C2 (2.4%)	C3 (5.4%)	C4 (11.6%)	C5 (25.0%)	C6 (54.0%)
GCN	72.02±0.50	59.42±0.74	59.23±1.02	0.0	28.6	67.0	60.0	81.2	93.8	93.1
+easy samples	76.90±0.19	69.55±0.21	71.28±0.25	21.1	69.4	67.9	63.1	73.1	95.1	94.7
+harder samples w/o SemiMixup	75.84±0.38	71.38±0.58	71.44±0.59	54.7	71.7	63.1	58.4	74.2	92.5	82.6
+harder samples w/ SemiMixup	79.16±0.25	72.89±0.32	74.62±0.27	42.2	74.3	71.8	62.3	72.5	94.4	93.4
+harder samples w/ SemiMixup (HK)	79.60±0.17	74.37±0.18	75.17±0.15	48.4	75.8	68.3	63.2	77.8	93.5	92.8
+harder samples w/ SemiMixup (PPR)	79.90±0.29	74.62±0.35	75.74±0.32	51.6	76.9	66.0	65.4	76.5	93.8	92.1

Analysis of the squeezed minority problem

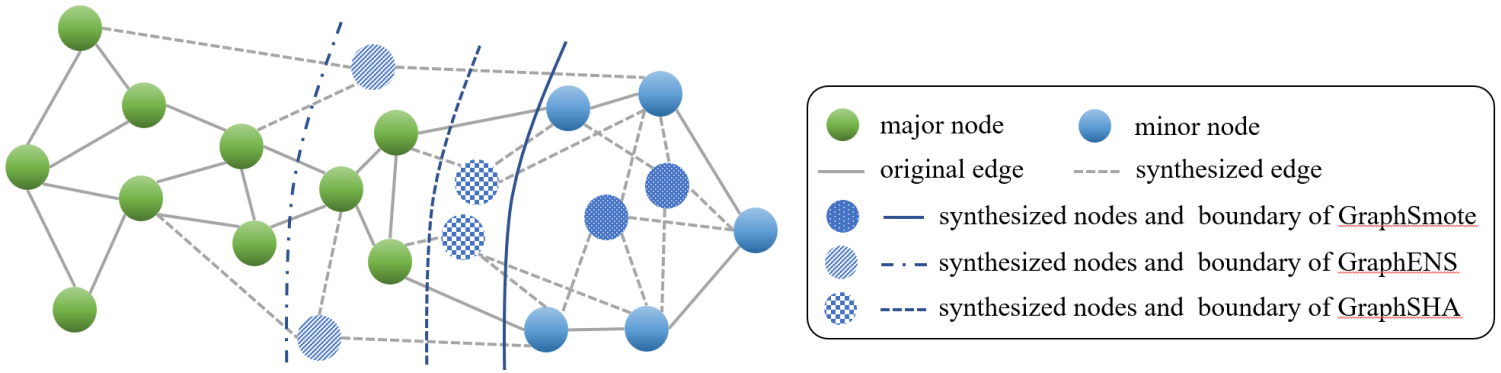


Case study on per-class accuracy

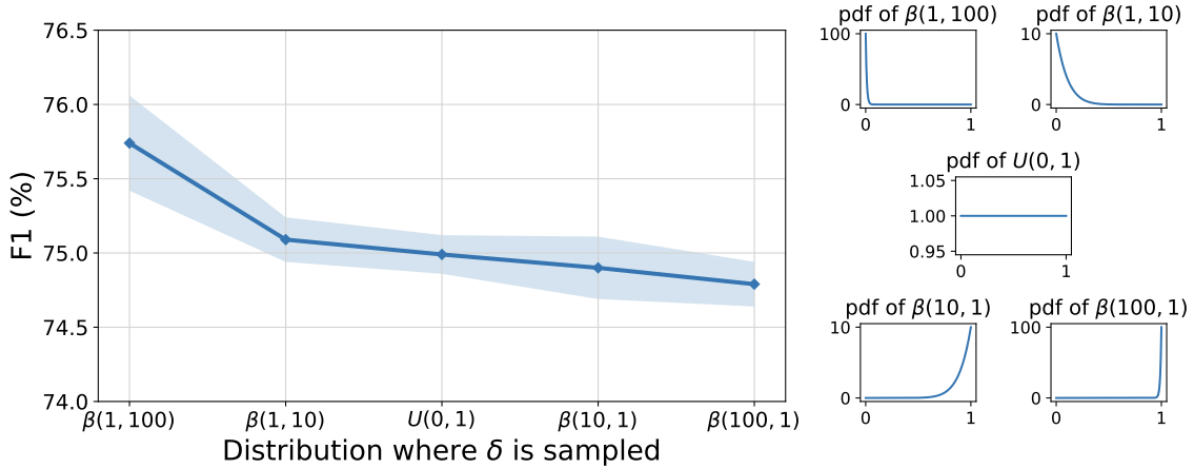
Class Distribution	C_0 (0.5%)	C_1 (1.1%)	C_2 (2.4%)	C_3 (5.4%)	C_4 (11.6%)	C_5 (25.0%)	C_6 (54.0%)
Vanilla (GCN)	0.0	28.6	67.0	60.0	81.2	93.8	93.1
Upsample	12.5	58.2	65.1	67.5	76.5	92.4	89.7
GraphSmote	22.2	66.4	68.9	62.1	79.4	93.4	89.4
GraphENS	37.7	72.2	73.2	63.5	74.6	94.7	82.3
GraphSHA	51.6	76.9	66.0	65.4	76.5	93.8	92.1

Case study on per-class accuracy

Class Distribution	C_0 (0.5%)	C_1 (1.1%)	C_2 (2.4%)	C_3 (5.4%)	C_4 (11.6%)	C_5 (25.0%)	C_6 (54.0%)
Vanilla (GCN)	0.0	28.6	67.0	60.0	81.2	93.8	93.1
Upsample	12.5	58.2	65.1	67.5	76.5	92.4	89.7
GraphSmote	22.2	66.4	68.9	62.1	79.4	93.4	89.4
GraphENS	37.7	72.2	73.2	63.5	74.6	94.7	82.3
GraphSHA	51.6	76.9	66.0	65.4	76.5	93.8	92.1

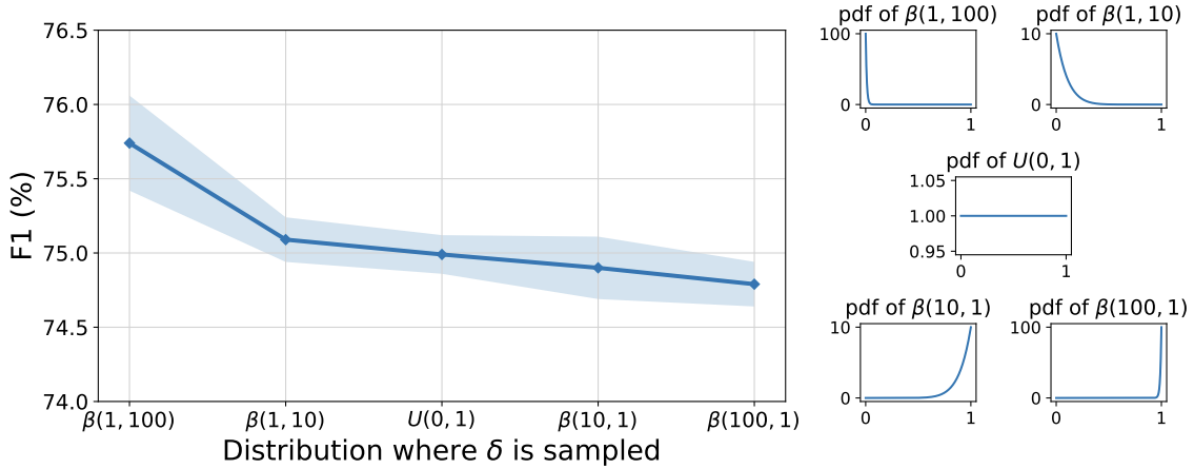


Hyper-parameter analysis

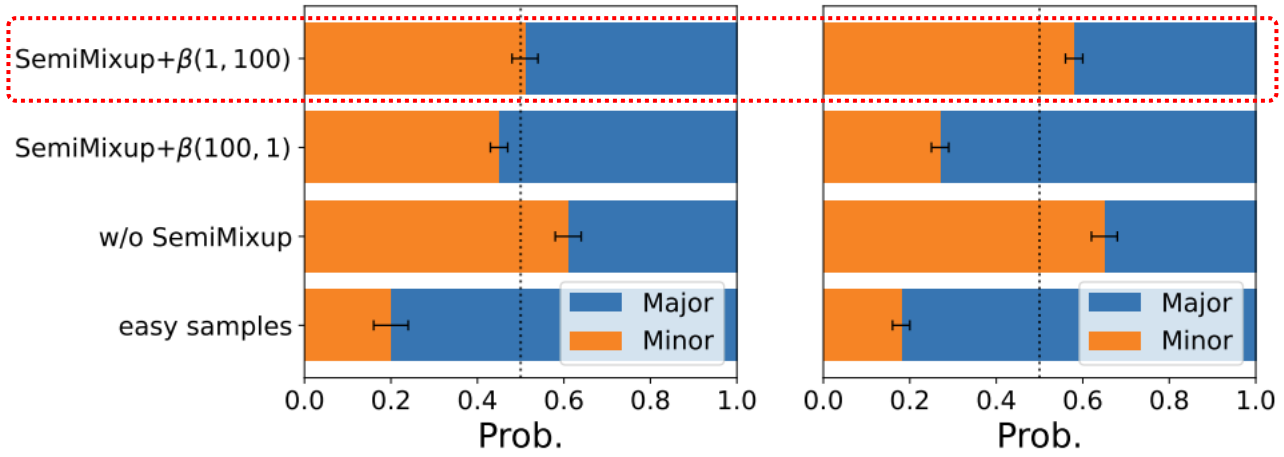


$$X_{syn} = \delta X_{anc} + (1 - \delta) X_{aux}, \ \delta \in [0, 1].$$

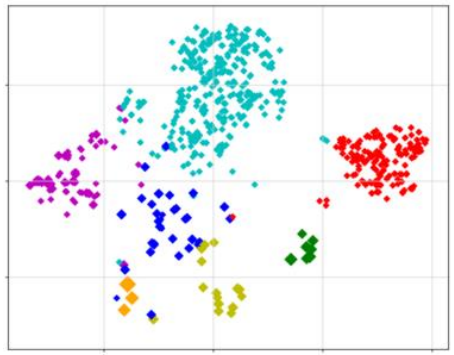
Hyper-parameter analysis



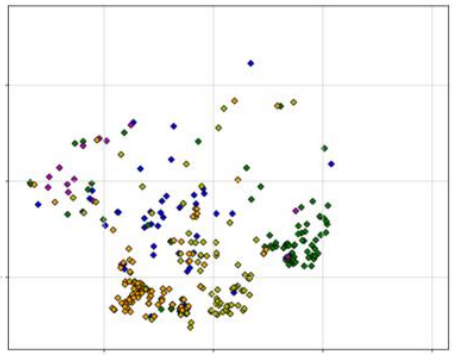
$$X_{syn} = \delta X_{anc} + (1 - \delta) X_{aux}, \ \delta \in [0, 1].$$



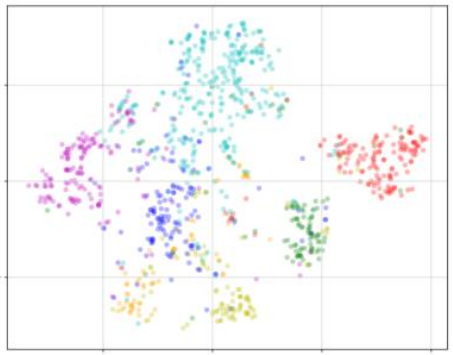
Visualization



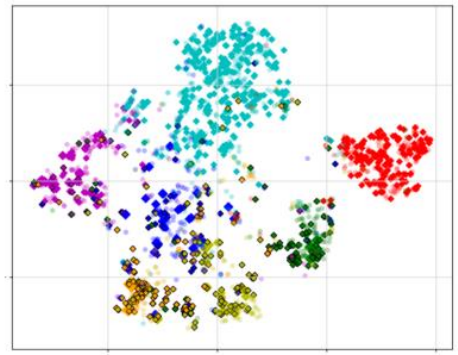
(a) Training samples



(b) Synthesized samples



(c) Test samples



(d) All samples from above

Summary

- **New entry point**: squeezed minority problem.
- **New technique**: GraphSHA for Synthesizing HArder minor samples.
- **Experiments**: Effectiveness of GraphSHA empirically.

Summary

- **New entry point**: squeezed minority problem.
- **New technique**: GraphSHA for Synthesizing HArder minor samples.
- **Experiments**: Effectiveness of GraphSHA empirically.

Check out our paper and code at...

- **Paper**: https://arxiv.org/abs/2306.09612
- **Project Page**: https://wenzhilics.github.io/GraphSHA.html
- **Code**: https://github.com/wenzhilics/GraphSHA

THANKS

Q&A

Wen-Zhi Li, Sun Yat-sen University & HKUST(GZ)
E-mail: liwzh63@mail2.sysu.edu.cn