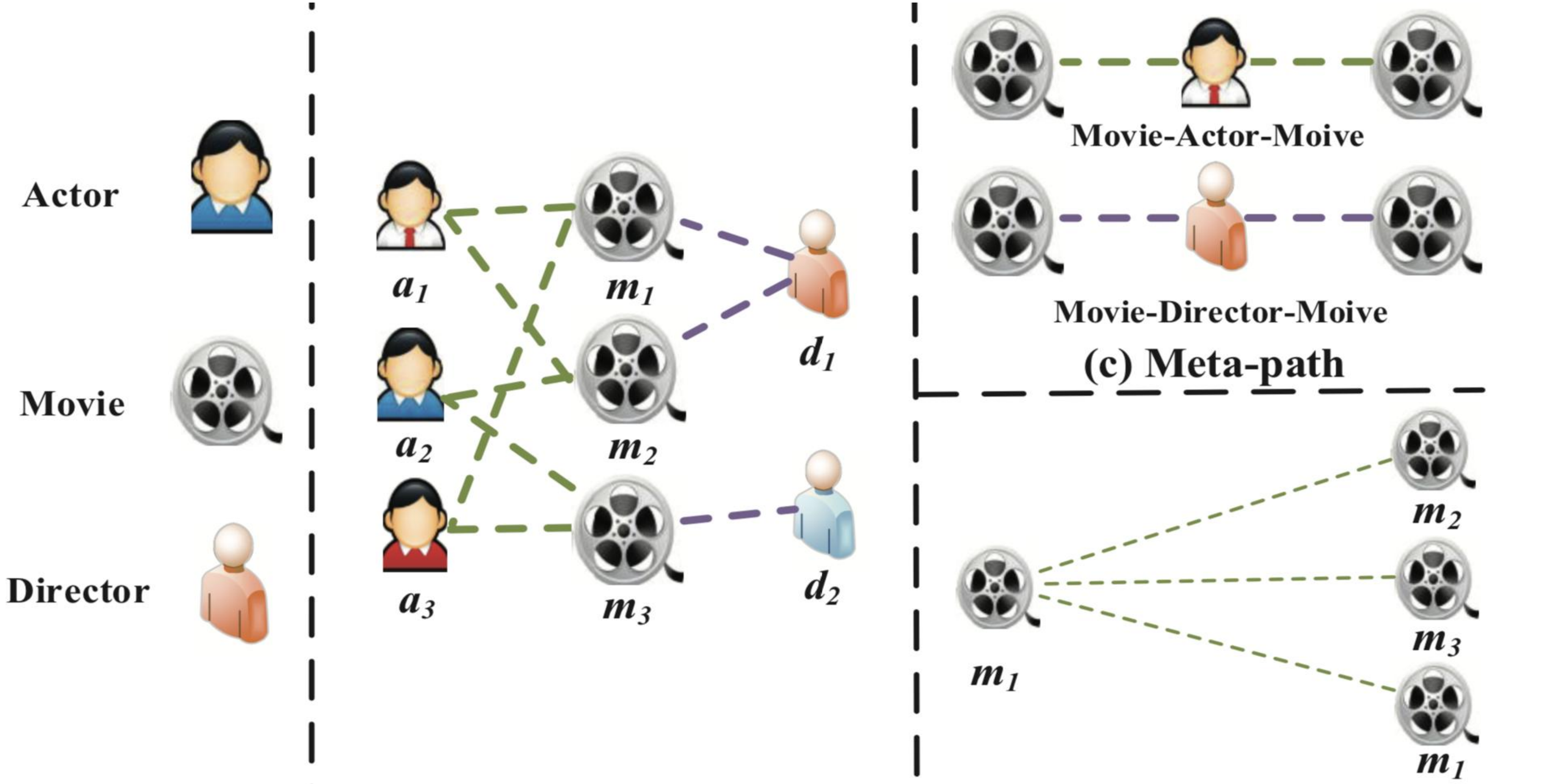
Long-range Meta-path Search on Large-scale Heterogeneous Graphs

Chao Li, Zijie Guo, Qiuting He, Kun He*

Hopcroft Center on Computing Science,
School of Computer Science and Technology,
Huazhong University of Science and Technology, China
(*Corresponding author)

Background

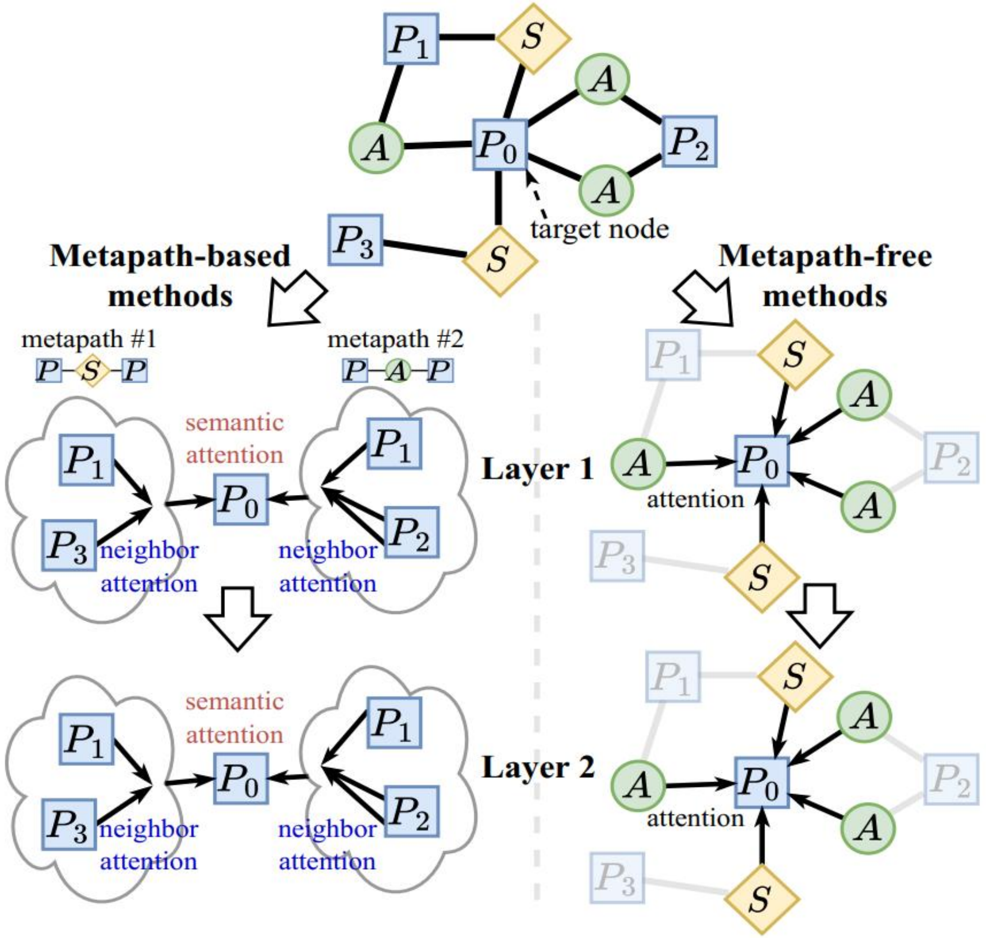
- **Heterogeneous Graph and Meta-path**



(a) Node Type (b) Heterogeneous Graph (c) Meta-path (d) Meta-path based Neighbors

Background

- **Classification of Heterogeneous Graph Neural Networks**



- **Metapath-based methods: First capture structural information from the same semantic relationships, then fuse different semantic vectors to generate the final output.**

- **Metapath-free methods: Capture structural and semantic information simultaneously.**

Motivation

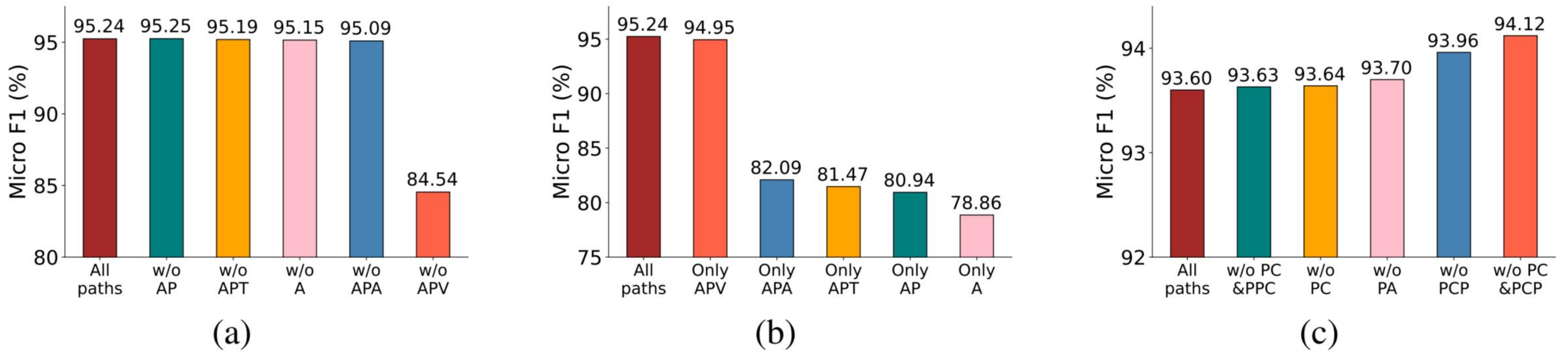


Figure 1: Analysis of the importance of different meta-paths. (a) illustrates the results after removing a single meta-path on DBLP; (b) shows the performance of utilizing a single meta-path on DBLP; (c) illustrates the performance after removing a part of meta-paths on ACM.

Two conclusions:

- **A small number of meta-paths provide major contributions.**
- **Certain meta-paths can have a negative impact for heterogeneous graphs.**

LMSPS

- **Key idea:**

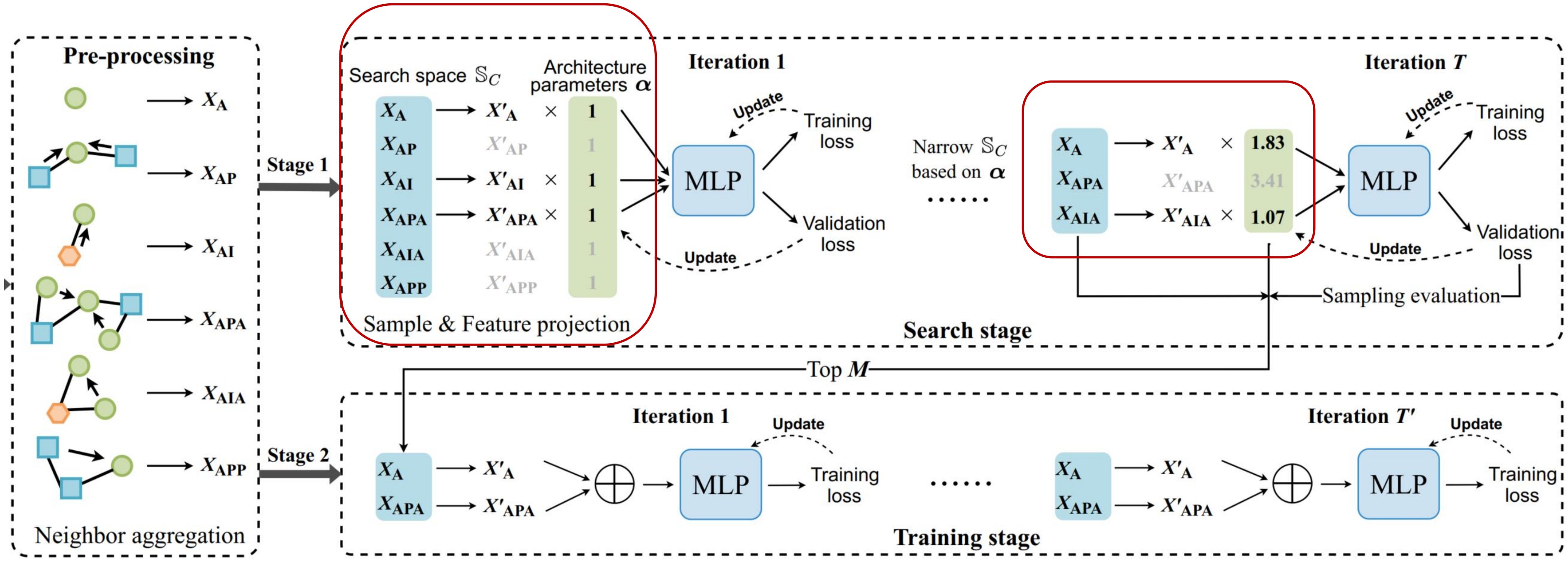
  Focusing on Long-range Dependency Issues in Large-scale Heterogeneous Graphs

- **Main Challenges:**
  - **Balancing Efficiency and Effectiveness: How to retain as much heterogeneous information as possible while reducing computational costs under the presence of heterogeneity.**
  - **Overcoming Over-smoothing Problem: Over-smoothing is a classic issue when graph neural networks utilize long-range dependencies. It is also necessary to consider how to overcome the over-smoothing problem in heterogeneous graphs.**
  - **Enhancing Generalization: How to make the discovered meta-paths effective on other HGNNs.**

- **Core Idea of LMSPS:**
  - **The proposed method falls into the category of meta-path-based approaches. It leverages meta-path search to identify effective meta-paths and eliminate ineffective or redundant ones.**

LMSPS

Balancing Efficiency and Effectiveness

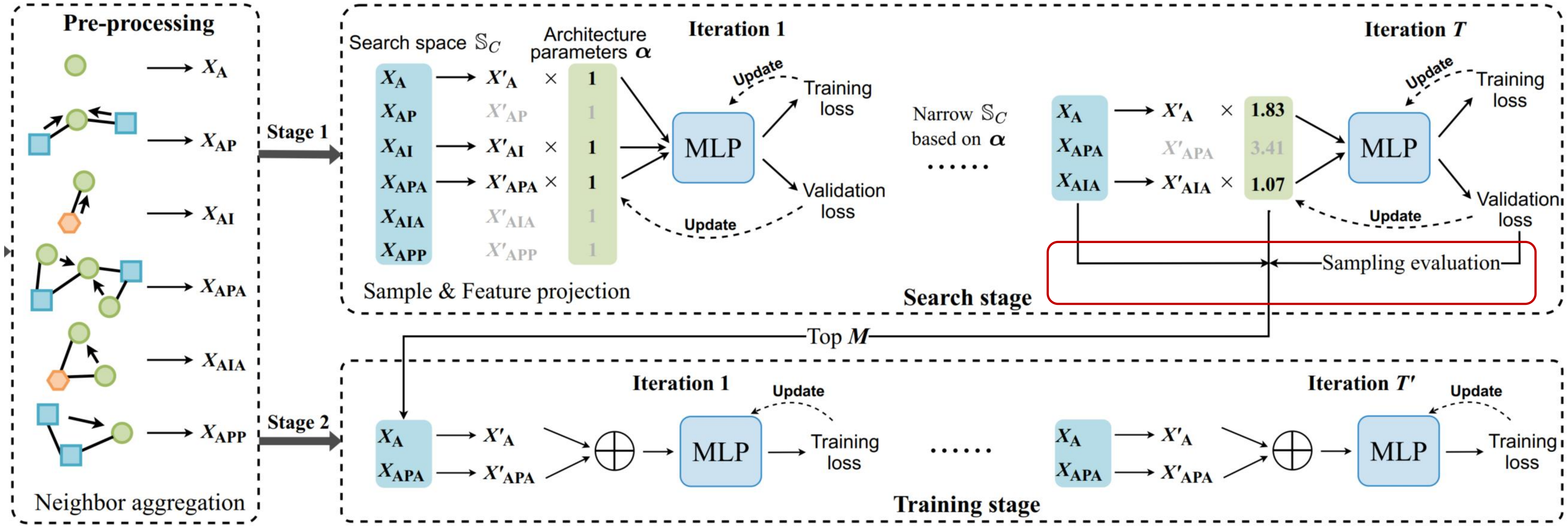


Strategies for Enhancing Efficiency:

- **Progressive Sampling Search**: To overcome efficiency challenges, only a portion of the meta-paths participate in updates during each iteration. Meanwhile, the search space gradually decreases throughout the search process.

LMSPS

Balancing Efficiency and Effectiveness

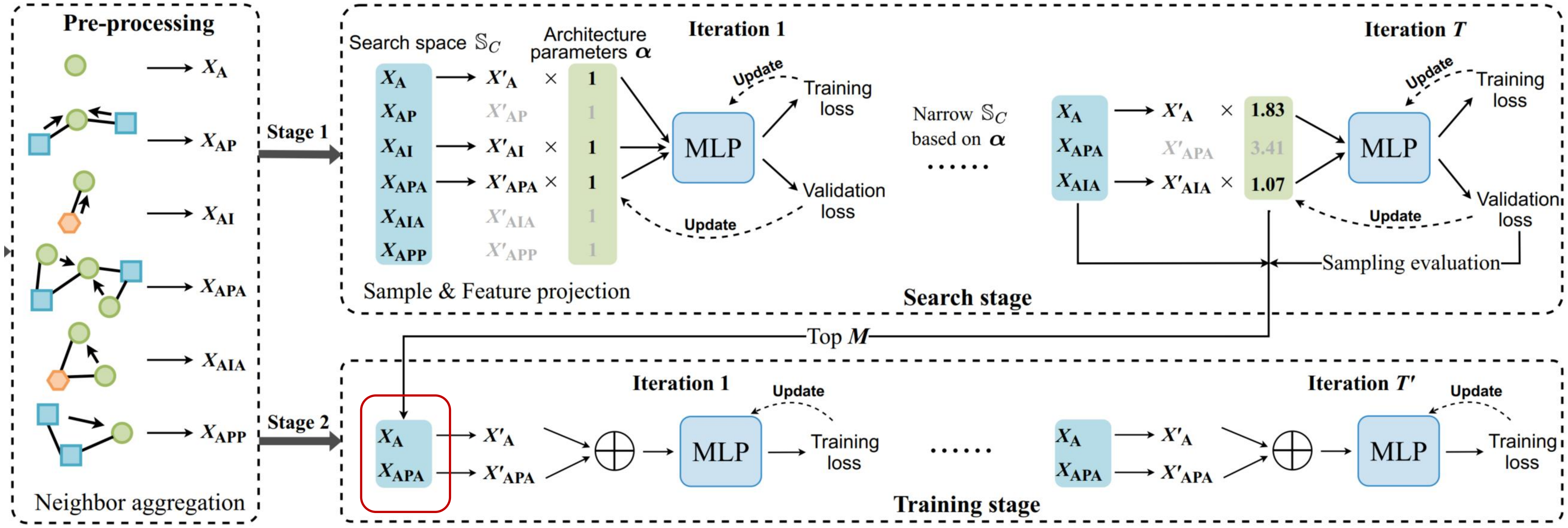


Strategies for Improving Effectiveness:

- **Sampling Evaluation: Within the compressed search space, a subset of meta-paths is sampled at each evaluation to assess performance. The subset of meta-paths with the minimum validation loss is selected as the final outcome.**

LMSPS

Overcoming the Over-smoothing Problem

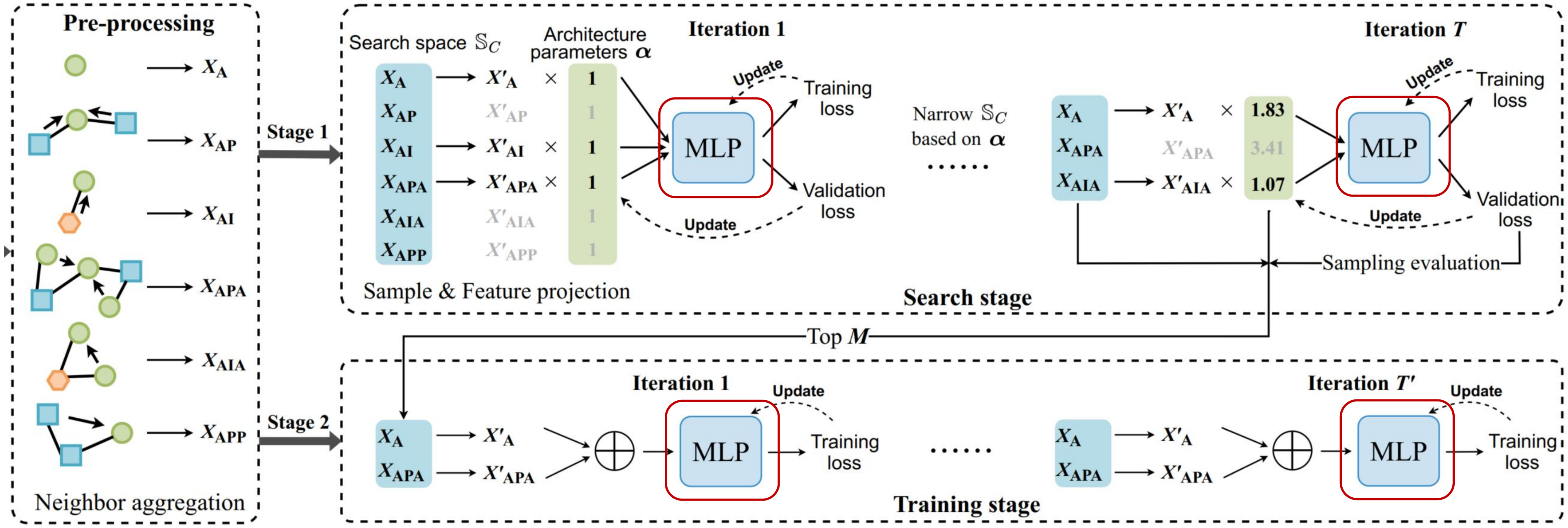


Over-smoothing Issue: As depth increases, node embeddings tend to become similar

- **Solution Approach: Each target node aggregates different neighboring nodes under the constraint of effective meta-path instances.**

LMSPS

Enhancing Generalization



Scheme for Enhancing Generalization: Utilizing Pure MLP Architecture

- **Compared to Transformers, MLPs involve less inductive bias, meaning there is less human intervention. This allows the search results to be unaffected by specific modules.**

Experiments

- ## Performance comparison

Table 1: Performance on small and large datasets. Best is in bold, and the runner-up is underlined.

Method	DBLP		IMDB		ACM		Freebase		OGBN-MAG*
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Test Acc.
MLP [19]	-	-	-	-	-	-	-	-	26.92 ± 0.26
GraphSAGE [14]	-	-	-	-	-	-	-	-	46.78 ± 0.67
RGCN [41]	91.52 ± 0.50	92.07 ± 0.50	58.85 ± 0.26	62.05 ± 0.15	91.55 ± 0.74	91.41 ± 0.75	46.78 ± 0.77	58.33 ± 1.57	47.37 ± 0.48
HAN [48]	91.67 ± 0.49	92.05 ± 0.62	57.74 ± 0.96	64.63 ± 0.58	90.89 ± 0.43	90.79 ± 0.43	21.31 ± 1.68	54.77 ± 1.40	*OOM*
GTN [56]	93.52 ± 0.55	93.97 ± 0.54	60.47 ± 0.98	65.14 ± 0.45	91.31 ± 0.70	91.20 ± 0.71	*OOM*	*OOM*	*OOM*
RSHN [63]	93.34 ± 0.58	93.81 ± 0.55	59.85 ± 3.21	64.22 ± 1.03	90.50 ± 1.51	90.32 ± 1.54	*OOM*	*OOM*	*OOM*
HetGNN [57]	91.76 ± 0.43	92.33 ± 0.41	48.25 ± 0.67	51.16 ± 0.65	85.91 ± 0.25	86.05 ± 0.25	*OOM*	*OOM*	*OOM*
MAGNN [11]	93.28 ± 0.51	93.76 ± 0.45	56.49 ± 3.20	64.67 ± 1.67	90.88 ± 0.64	90.77 ± 0.65	*OOM*	*OOM*	*OOM*
HetSANN [18]	78.55 ± 2.42	80.56 ± 1.50	49.47 ± 1.21	57.68 ± 0.44	90.02 ± 0.35	89.91 ± 0.37	*OOM*	*OOM*	*OOM*
GCN [26]	90.84 ± 0.32	91.47 ± 0.34	57.88 ± 1.18	64.82 ± 0.64	92.17 ± 0.24	92.12 ± 0.23	27.84 ± 3.13	60.23 ± 0.92	*OOM*
GAT [46]	93.83 ± 0.27	93.39 ± 0.30	58.94 ± 1.35	64.86 ± 0.43	92.26 ± 0.94	92.19 ± 0.93	40.74 ± 2.58	65.26 ± 0.80	*OOM*
Simple-HGN [34]	94.01 ± 0.24	94.46 ± 0.22	63.53 ± 1.36	67.36 ± 0.57	93.42 ± 0.44	93.35 ± 0.45	47.72 ± 1.48	66.29 ± 0.45	*OOM*
HGT [21]	93.01 ± 0.23	93.49 ± 0.25	63.00 ± 1.19	67.20 ± 0.57	91.12 ± 0.76	91.00 ± 0.76	29.28 ± 2.52	60.51 ± 1.16	46.78 ± 0.42
GraphMSE [31]	94.08 ± 0.14	94.44 ± 0.13	57.60 ± 2.13	62.37 ± 1.03	92.58 ± 0.50	92.54 ± 0.14	*OOM*	*OOM*	*OOM*
DiffMG [8]	94.01 ± 0.37	94.20 ± 0.36	58.09 ± 1.35	59.75 ± 1.23	88.16 ± 2.83	88.07 ± 3.04	*OOM*	*OOM*	*OOM*
Random	93.57 ± 0.64	93.84 ± 0.53	52.13 ± 0.74	53.83 ± 0.66	90.91 ± 1.02	90.82 ± 0.93	21.22 ± 2.58	37.54 ± 2.66	35.14 ± 3.78
NARS [55]	94.18 ± 0.47	94.61 ± 0.42	63.51 ± 0.68	66.18 ± 0.70	93.36 ± 0.32	93.31 ± 0.33	49.98 ± 1.77	63.26 ± 1.26	50.66 ± 0.22
space4HGNN [59]	94.24 ± 0.42	94.63 ± 0.40	61.57 ± 1.19	63.96 ± 0.43	92.50 ± 0.14	92.38 ± 0.10	41.37 ± 4.49	65.66 ± 4.94	*OOM*
PMMM [27]	94.82 ± 0.26	95.14 ± 0.22	65.81 ± 0.29	67.58 ± 0.22	93.78 ± 0.25	93.71 ± 0.17	*OOM*	*OOM*	*OOM*
HINormer [36]	94.57 ± 0.23	94.94 ± 0.21	64.65 ± 0.53	67.83 ± 0.34	93.91 ± 0.42	93.83 ± 0.45	52.18 ± 0.39	64.92 ± 0.43	*OOM*
SeHGNN [52]	94.86 ± 0.14	95.24 ± 0.13	66.63 ± 0.34	68.21 ± 0.32	93.95 ± 0.48	93.87 ± 0.50	50.71 ± 0.44	63.41 ± 0.47	51.45 ± 0.29
SlotGAT [62]	94.95 ± 0.20	95.31 ± 0.19	64.05 ± 0.60	68.64 ± 0.33	93.99 ± 0.23	94.06 ± 0.22	49.68 ± 1.97	**66.83 ± 0.30**	*OOM*
LMSPS (ours)	**95.35 ± 0.22**	**95.66 ± 0.20**	**66.99 ± 0.32**	**68.70 ± 0.26**	**94.73 ± 0.41**	**94.69 ± 0.36**	**53.26 ± 0.47**	66.09 ± 0.51	**54.83 ± 0.20**

* OGBN-MAG is a large dataset with nodes' numbers 10 to 175 times that of the other four datasets.

Experiments

- **Memory cost and training time comparison**

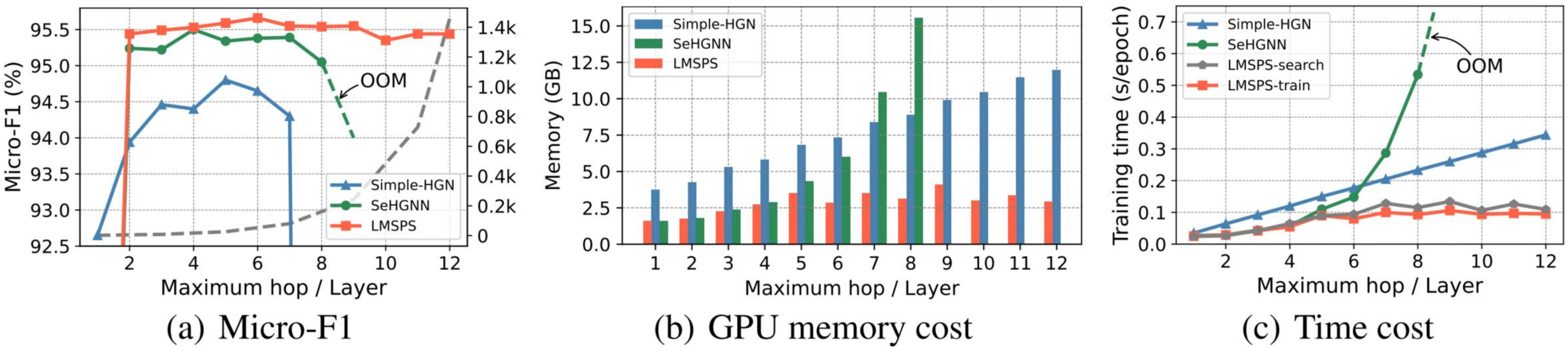


Figure 3: Illustration of (a) performance, (b) memory cost, (c) average training time of Simple-HGN, SeHGNN, and LMSPS relative to the maximum hop or layer on DBLP. The *gray dotted line* in (a) indicates the number of target-node-related meta-paths under different maximum hops, which is exponential.

Experiments

- **Study of maximum hop and search algorithm**

Table 2: Experiments on OGBN-MAG to analyze the performance of SeHGNN and LMSPS under different maximum hops. #MP is the number of meta-paths under different maximum hops.

Max hop	#MP	SeHGNN		LMSPS	
		Time	Test accuracy	Time	Test accuracy
1	4	4.35	47.18 ± 0.28	3.98	46.88 ± 0.10
2	10	6.44	51.79 ± 0.24	5.63	51.91 ± 0.13
3	23	11.28	52.44 ± 0.16	10.02	52.72 ± 0.24
4	50	OOM	OOM	14.34	53.43 ± 0.18
5	107	OOM	OOM	14.77	53.90 ± 0.19
6	226	OOM	OOM	14.71	$\mathbf{54.83 \pm 0.20}$

Table 3: Experiments to explore the effectiveness of our search algorithm. In our LMSPS, the meta-paths are replaced by those discovered by other methods.

Method	DBLP	IMDB	ACM	Freebase
HAN	95.44 ± 0.14	65.95 ± 0.31	90.66 ± 0.30	-
GTN	95.33 ± 0.05	65.99 ± 0.16	90.66 ± 0.30	-
DARTS	95.35 ± 0.17	66.23 ± 0.14	93.45 ± 0.13	63.25 ± 0.42
SPOS	95.41 ± 0.43	67.10 ± 0.29	93.64 ± 0.37	64.02 ± 0.62
DiffMG	95.45 ± 0.49	66.98 ± 0.37	93.61 ± 0.45	64.56 ± 0.78
PMMM	95.48 ± 0.27	67.49 ± 0.24	93.74 ± 0.22	64.83 ± 0.46
LMSPS	$\mathbf{95.66 \pm 0.20}$	$\mathbf{68.70 \pm 0.26}$	$\mathbf{94.69 \pm 0.36}$	$\mathbf{66.09 \pm 0.51}$

Experiments

- **Generalization and ablation study**

Table 4: Experiments on the generalization of the searched meta paths. * means using the meta-paths searched in LMSPS.

Method	DBLP	IMDB	ACM	Freebase
HAN	92.05 ± 0.62	64.63 ± 0.58	90.79 ± 0.43	54.77 ± 1.40
HAN*	93.54 ± 0.15	65.89 ± 0.52	92.28 ± 0.47	57.13 ± 0.72
SeHGNN	95.24 ± 0.13	68.21 ± 0.32	93.87 ± 0.50	63.41 ± 0.47
SeHGNN*	95.57 ± 0.23	68.59 ± 0.24	94.46 ± 0.18	65.37 ± 0.42

Table 5: Results of LMSPS and SeHGNN on the sparse large-scale heterogeneous graphs. ↑ means the improvements in test accuracy.

Dataset	SeHGNN	LMSPS	↑
OGBN-MAG-5	36.04 ± 0.64	40.82 ± 0.42	**4.78**
OGBN-MAG-10	38.27 ± 0.19	42.30 ± 0.23	**4.03**
OGBN-MAG-20	39.18 ± 0.09	42.65 ± 0.17	**3.47**
OGBN-MAG-50	39.50 ± 0.13	42.82 ± 0.16	**3.32**

Table 6: Experiments on small and large datasets to analyze the effects of different blocks in LMSPS. *PS* means progressive sampling strategy, and *SE* means sampling evaluation strategy. † means employing all meta-paths and replacing the concatenation operation with the transformer module.

Method	DBLP		IMDB		ACM		Freebase		OGBN-MAG
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Test Acc.
LMSPS w/o *PS*	94.71 ± 0.23	95.00 ± 0.19	64.85 ± 0.46	66.52 ± 0.37	93.19 ± 0.34	93.14 ± 0.41	48.89 ± 0.47	61.61 ± 0.51	47.66 ± 0.45
LMSPS w/o *SE*	95.15 ± 0.28	95.48 ± 0.24	65.46 ± 0.48	67.13 ± 0.47	94.20 ± 0.35	94.15 ± 0.31	52.08 ± 0.33	64.84 ± 0.38	52.94 ± 0.34
LMSPS †	95.06 ± 0.24	95.38 ± 0.21	66.85 ± 0.37	68.58 ± 0.34	94.60 ± 0.42	94.57 ± 0.39	*OOM*	*OOM*	*OOM*
LMSPS	**95.35 ± 0.22**	**95.66 ± 0.20**	**66.99 ± 0.32**	**68.70 ± 0.26**	**94.73 ± 0.41**	**94.69 ± 0.36**	**53.26 ± 0.47**	**66.09 ± 0.51**	**54.83 ± 0.20**

Conclusion

- We propose a novel meta-path search framework termed LMSPS, which to our knowledge is the first HGNNs to utilize long-range dependency in large-scale heterogeneous graphs.
- To search for effective meta-paths efficiently, we introduce a novel progressive sampling algorithmto reduce the search space dynamically and a sampling evaluation strategy for meta-path selection.
- Moreover, the searched meta-paths of LMSPS can be generalized to other HGNNs to boost their performance.
- We find that: 1) A minority of meta-paths provide the main contributions; 2) Certain meta-paths have negative impacts on heterogeneous graphs. These findings offer certain guidance on how to utilize meta-paths in heterogeneous graphs.

End

Thanks for your attention!