



北京大學
PEKING UNIVERSITY

面向子图结构的 图神经网络建模研究

报告人：龙晴晴

指导老师：宋国杰

2021年07月04日

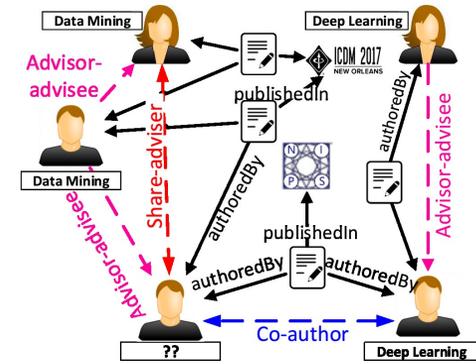
研究背景

□ 复杂图网络

- 由节点、边组成
- 在现实生活中大量存在



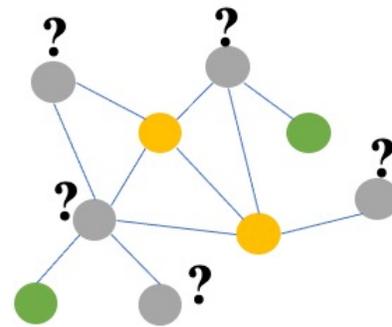
社交网络



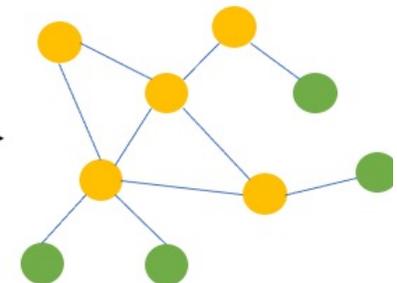
学术引用网络

□ 图机器学习/数据挖掘

- 节点分类：商品分类等
- 链接预测：点击率预测等
-



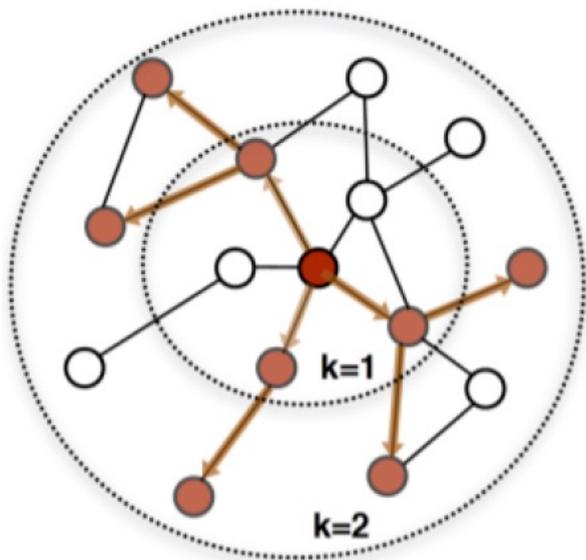
Machine Learning



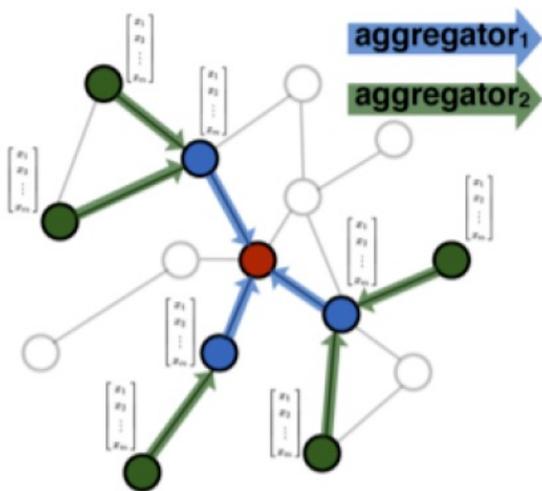
研究背景

□ 图神经网络 (GNN)

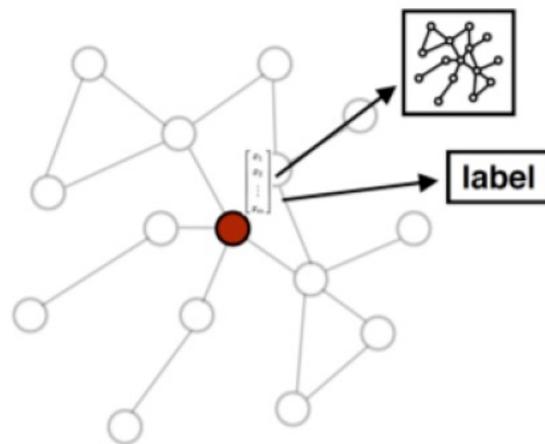
- 建模节点邻域结构特征，可解释性强、泛化性好、效果突出
- 是目前图网络应用广泛、重要的建模方法



(a) 邻居采样



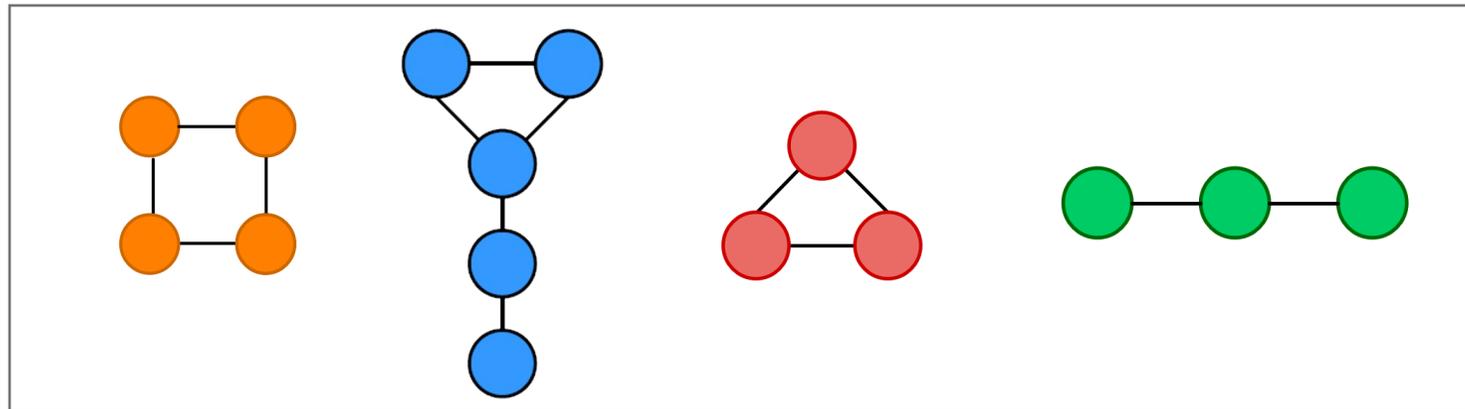
(b) 邻域节点特征聚合



(c) 利用标签信息进行参数更新

研究背景

- 图神经网络的**表达能力具有上界**^[1,2]
 - 1-Weisfeiler-Lehman Isomorphism Test (1-WL Test)
 - 但1-WL Test **无法建模**(区分/计数) 三角形、环形等子图结构
 - 图神经网络在**子图结构建模**方面具有**严重缺陷**

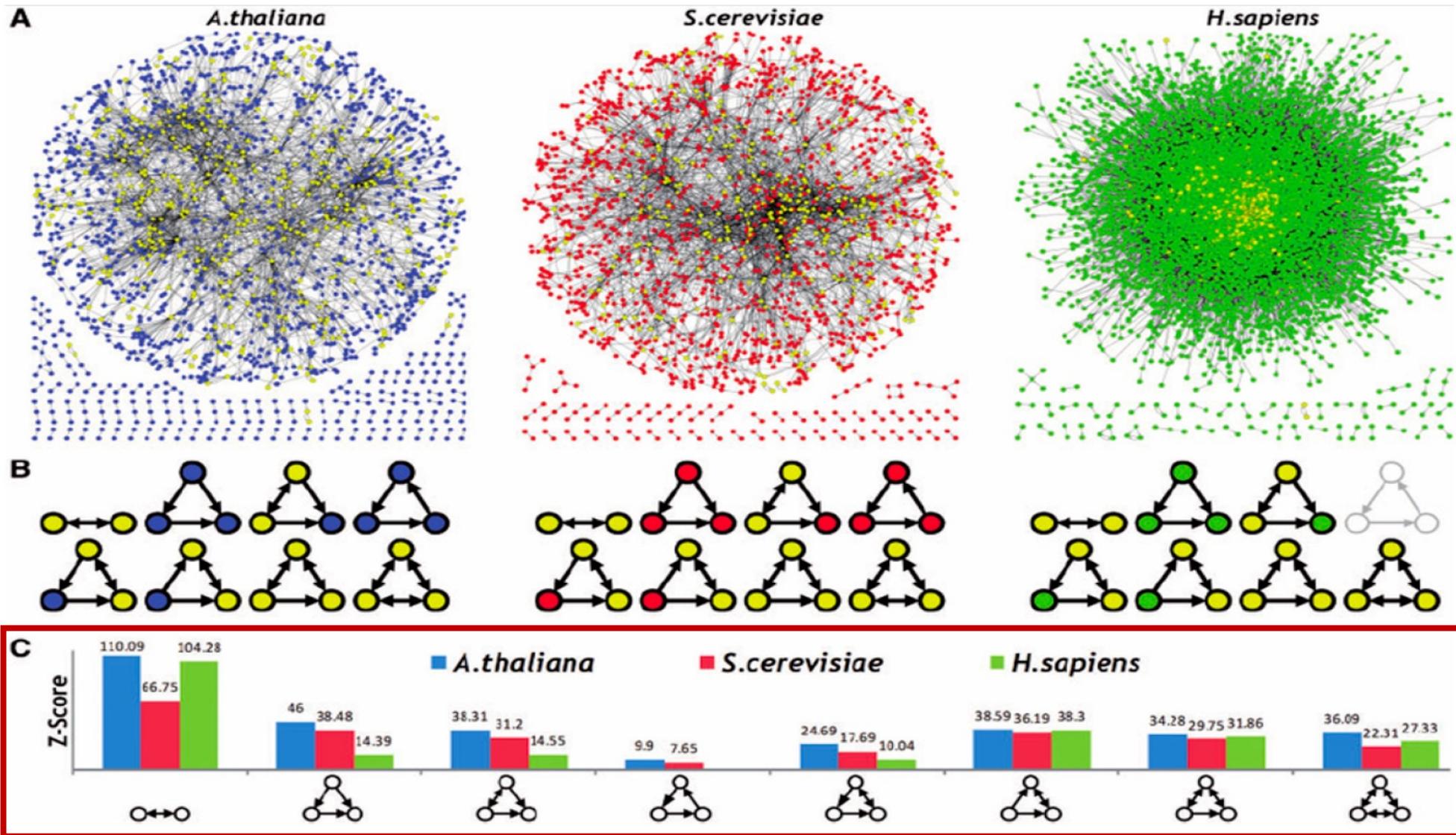


子图结构示意图

[1] Vikraman et al. On Weisfeiler-Leman invariance: subgraph counts and related graph properties.

[2] Chen et al. Can graph neural networks count substructures?

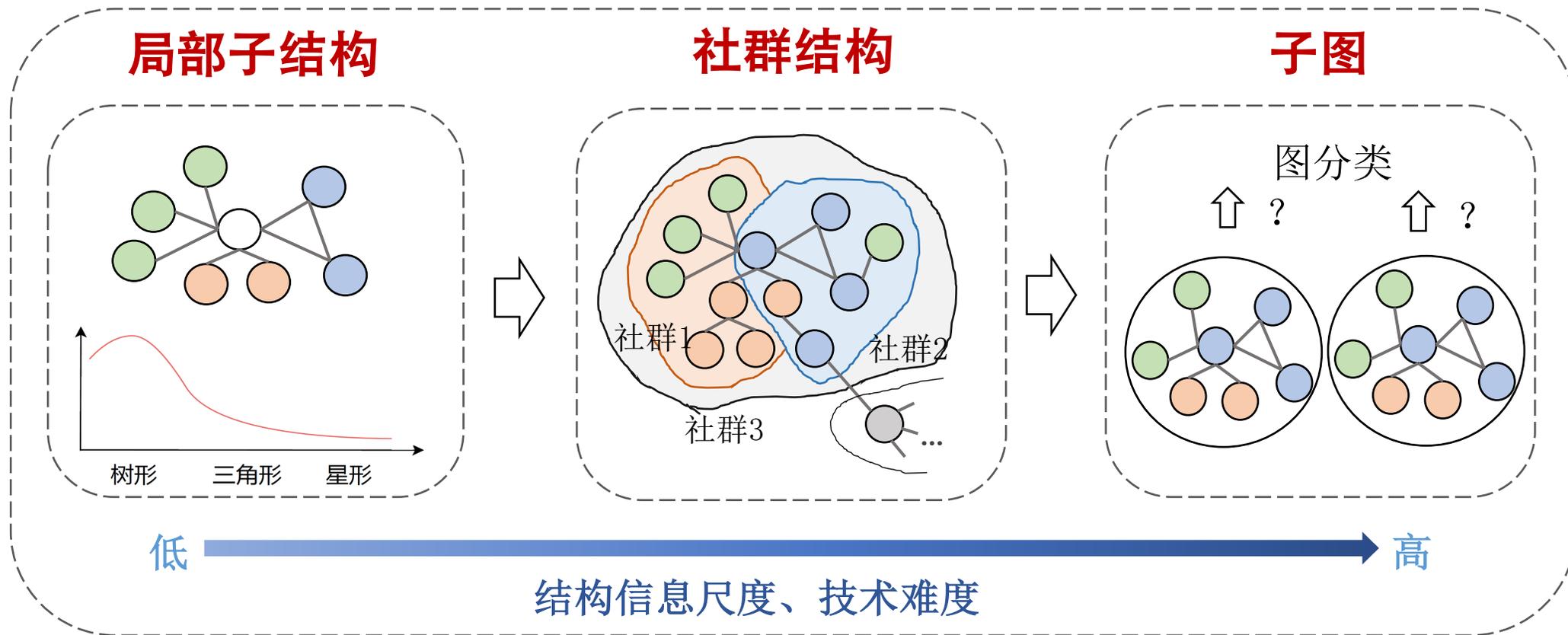
子图结构在图机器学习中扮演重要角色



子图结构
分布差异

研究问题：面向子图结构的图神经网络建模

研究内容



成果

第一作者发表 CCF-A 类长文 4 篇 (KDD、WWW)
B 类长文 1 篇 (CIKM 最佳论文第二名)

研究问题：面向子图结构的图神经网络建模

技术思路

局部子结构

基于子图结构的共现关系，进行矩阵分解

社群结构

子图结构投影至线形子空间，建模社群相似性

子图

投影至高维非线性隐空间、利用节点特征、核函数度量

研究贡献

率先提出图主题模型，提升GNN建模**局部子结构**能力

KDD

率先将子空间理论引入表示学习，**弥补**其建模**多层级社群**的**空白**

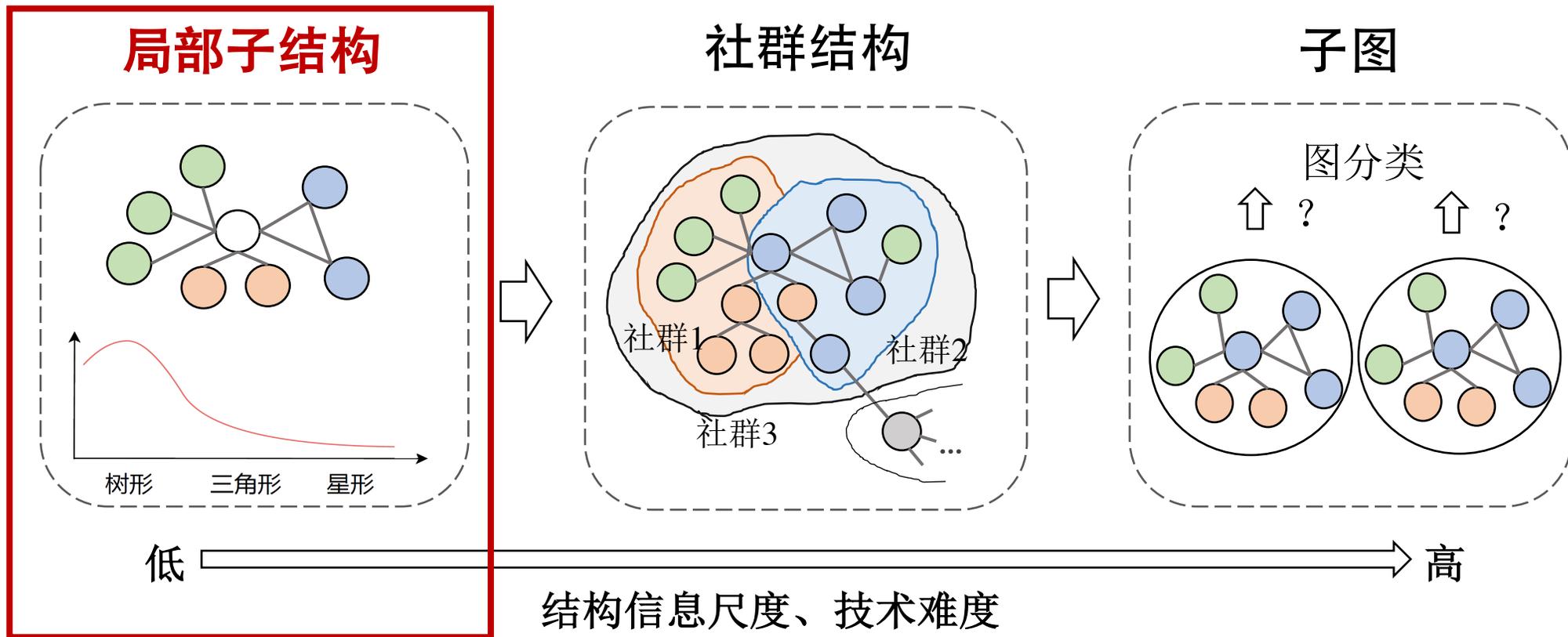
CIKM、KDD

理论上突破了现有GNN**表达能力上界**

WWW、KDD

研究问题：面向子图结构的图神经网络建模

研究思路

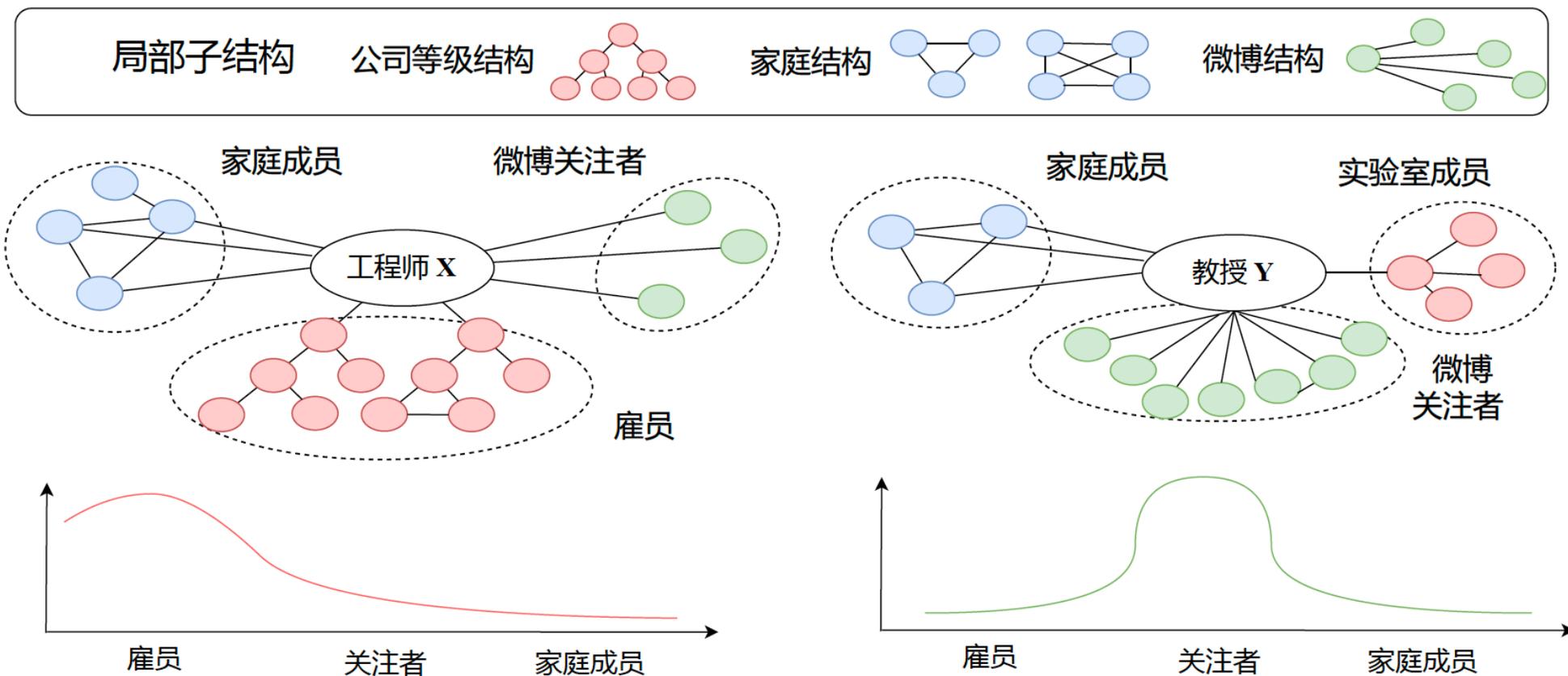


KDD 2020

研究背景



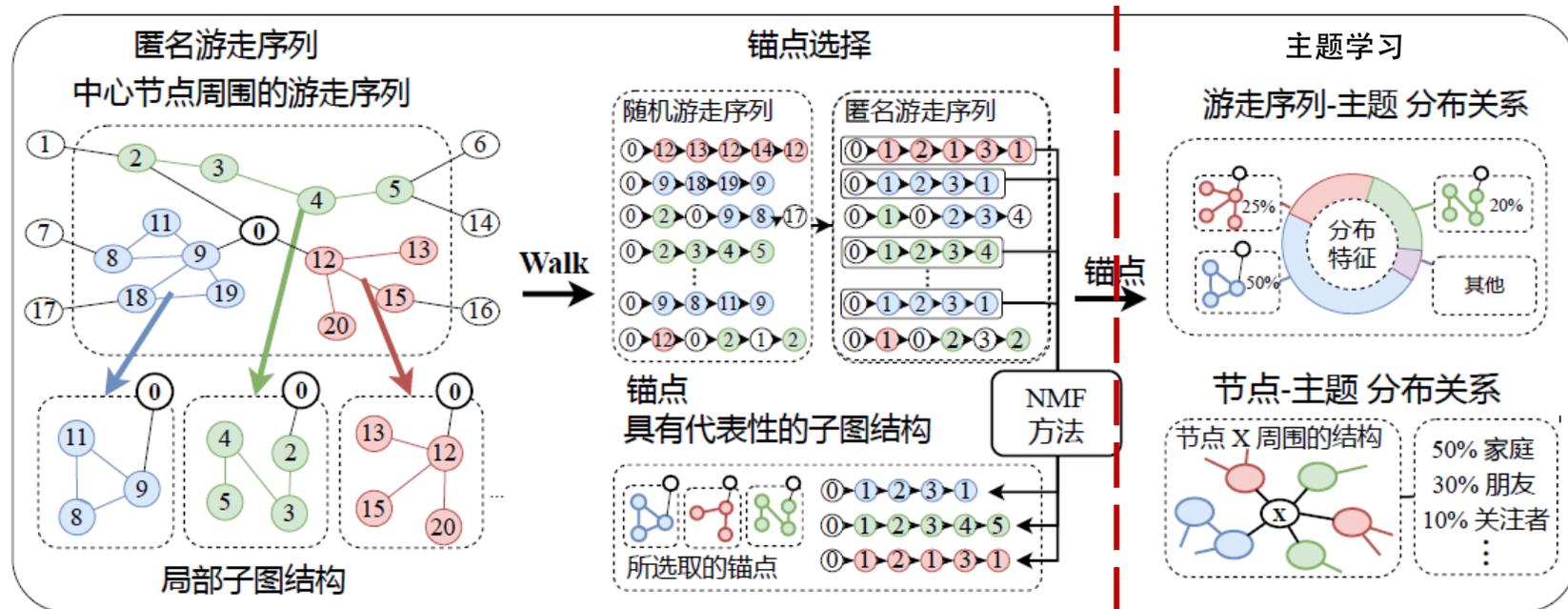
□ 局部子图结构分布



模型框图: GraphSTONE



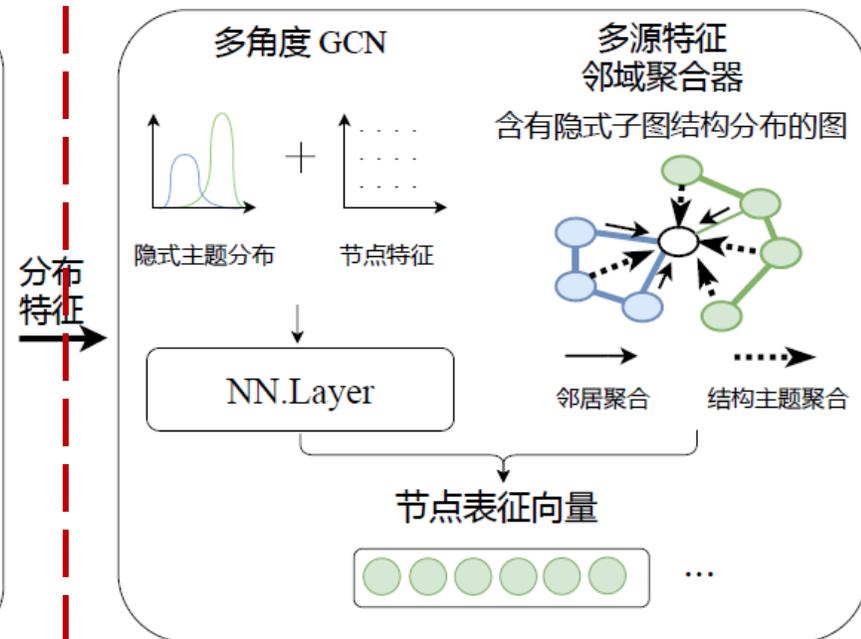
主题模型: 基于结构锚点的子图结构分布模型



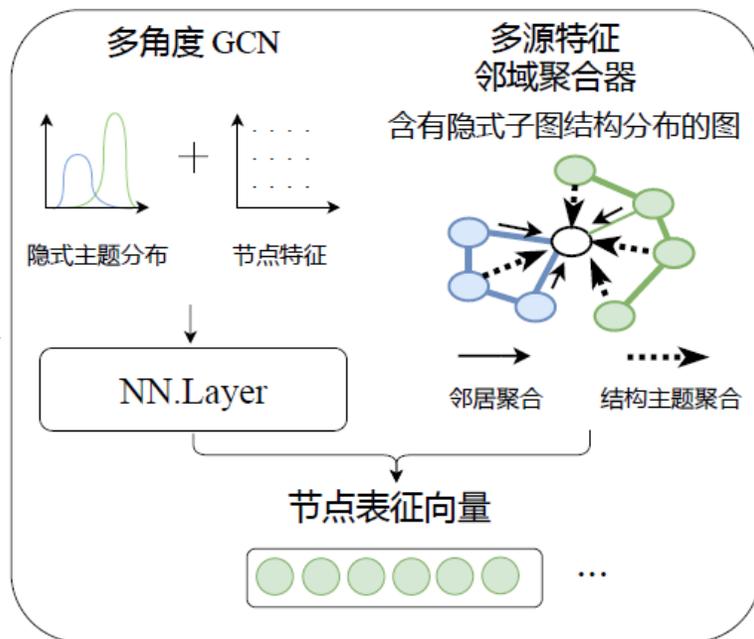
1 自适应提取子图结构

2 学习结构分布

融合多源信息的图神经网络



3 改进GCN卷积方式





1 自适应提取子图结构

匿名游走

➤ 节点被编号为其在序列中第一次出现位置

➤ 例子

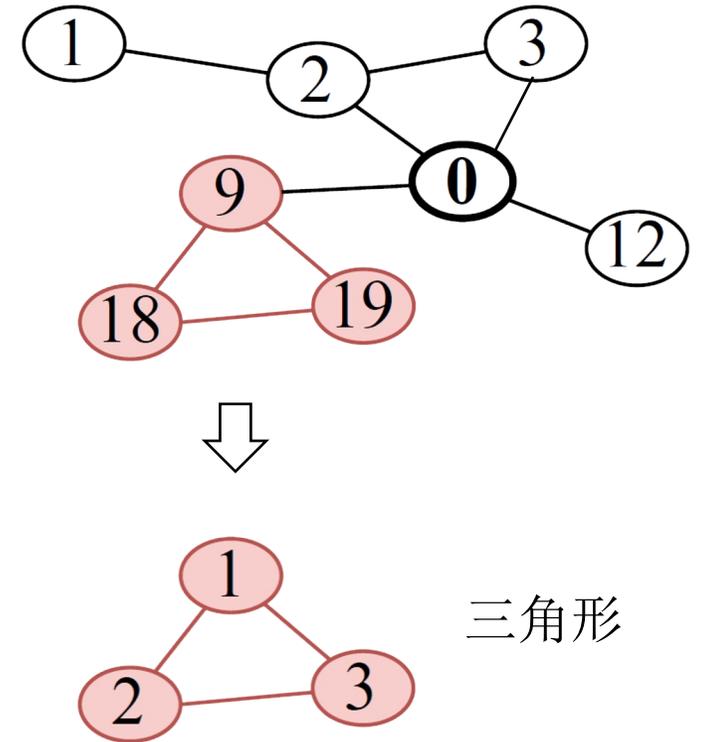
✓ 随机游走序列: $(9, 18, 19, 9)$ 、 $(0, 2, 3, 0)$

✓ 匿名游走序列: $(1, 2, 3, 1)$ 、 $(1, 2, 3, 1)$

✓ 是三角形结构

➤ 可以将相似结构映射为同样的编码^[1]

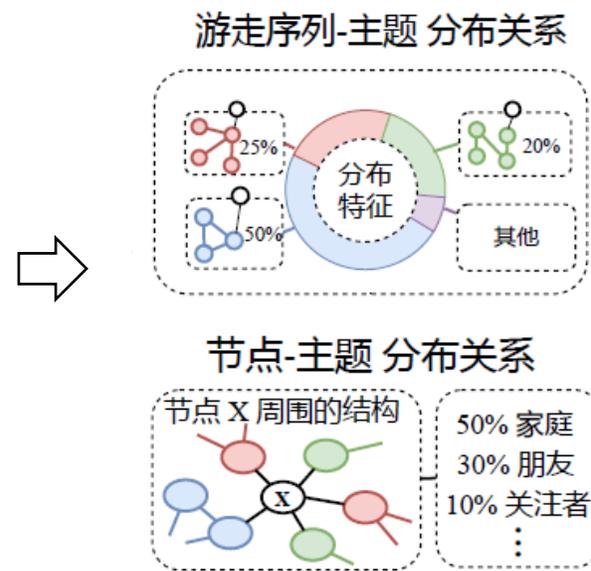
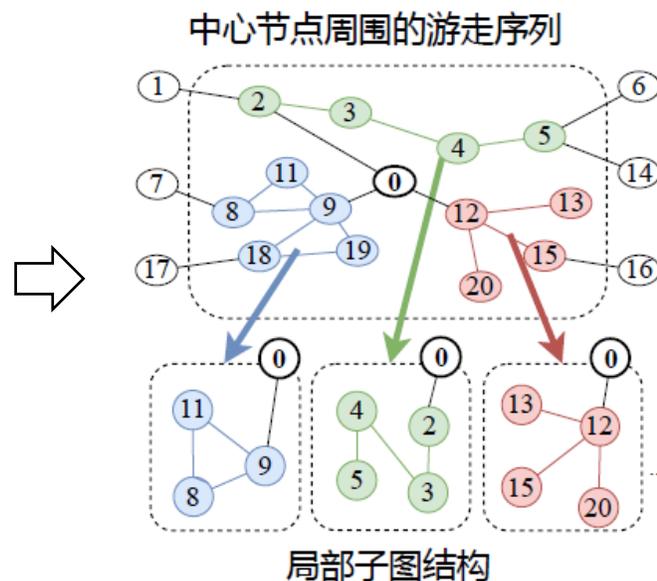
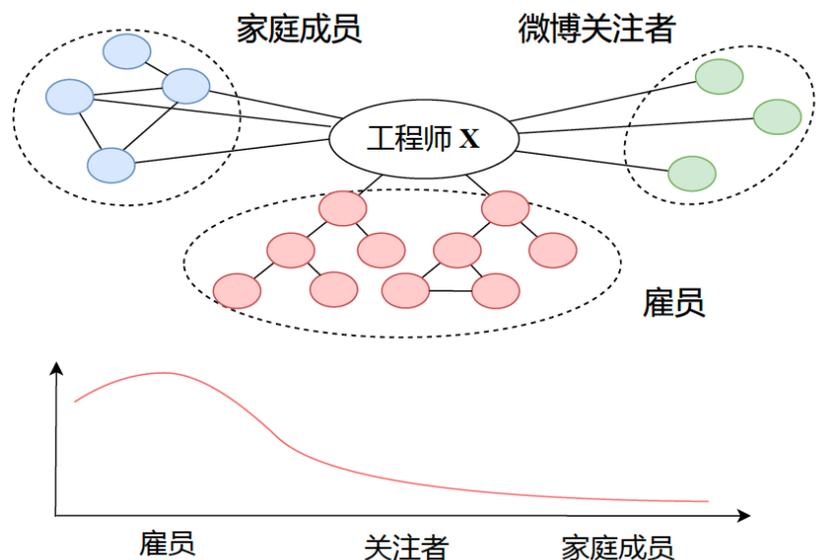
✓ (完备性): 通过匿名游走可重建原始子图



[1] Micali and Zhu. Reconstructing markov processes from independent and anonymous experiments. In Applied Discrete Maths, 2016.

2 学习结构分布——图网络上的主题模型

- 首次在图网络中提出主题模型 (与NLP中的主题模型进行类比)
 - 子图结构 (匿名游走) \Leftrightarrow 单词
 - 中心节点周围的子图结构集合 \Leftrightarrow 文档



3 考虑子图结构分布的动态卷积

□ 多角度GNN

$$h_i^{(L)} = (\mathbf{W} \cdot \text{ReLU}([h_{i,n}^{(L)} \otimes h_{i,s}^{(L)}]) + \mathbf{b})$$

□ 考虑结构主题信息的卷积方式改进

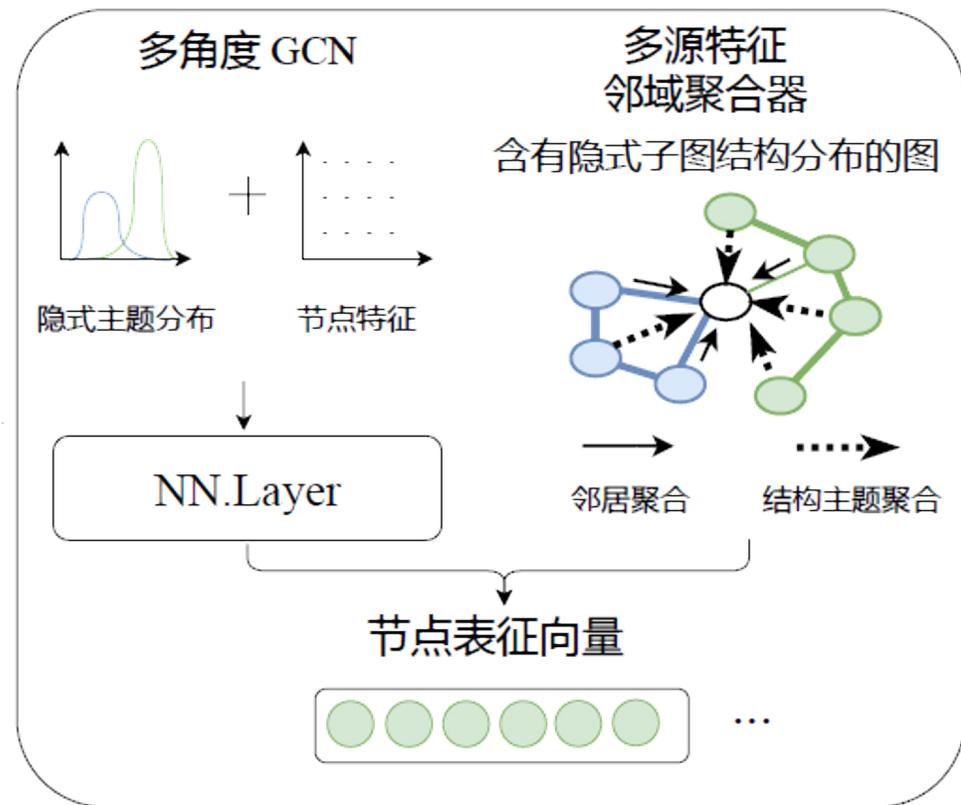
$$h_i^{(k)} = \text{AGGREGATE} \left(\left\{ \frac{R_i^T R_j}{\sum_j R_i^T R_j} h_j^{(k-1)}, v_j \in N(v_i) \right\} \right)$$

□ 无监督优化目标

➤ 类似于 GraphSAGE^[1]

$$\mathcal{L} = -\log[\sigma(h_i^{(L)T} h_j^{(L)})] - q \cdot \mathbb{E}_{v_n \sim P_n(v)} \log[\sigma(h_i^{(L)T} h_n^{(L)})]$$

融合多源信息的图神经网络



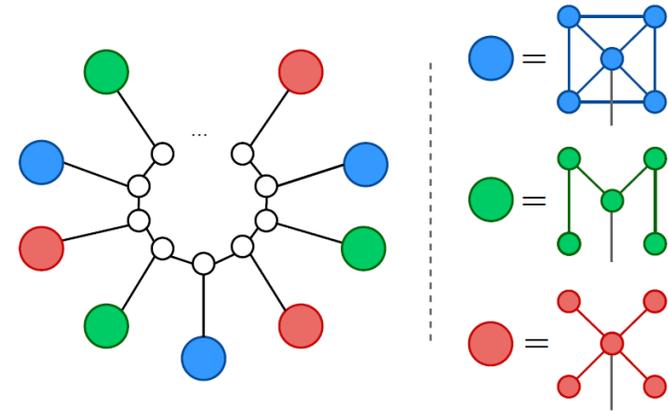
概念性验证

□ 仿真数据

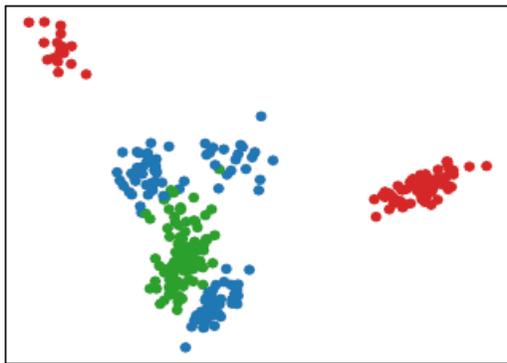
- 包含三种不同子结构的仿真数据

□ 结论

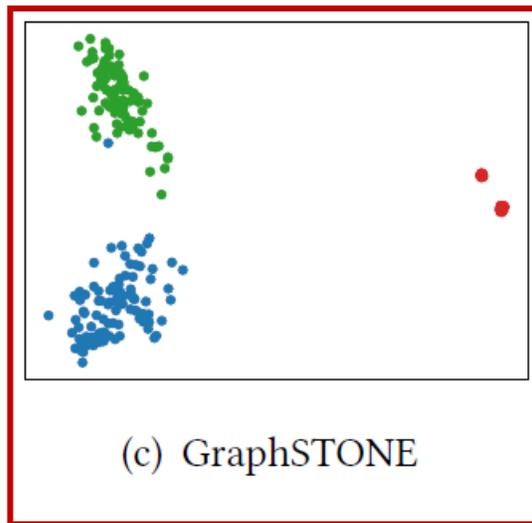
- GraphSTONE 可以有效 **区分** 不同子图结构



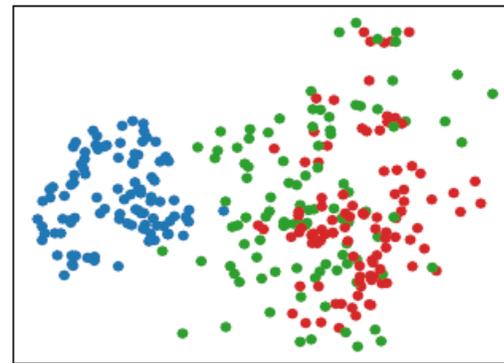
(a) Illustration of $G(n)$



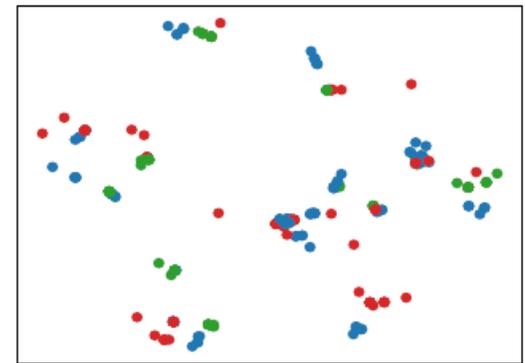
(b) Graph Anchor LDA



(c) GraphSTONE



(d) GraLSP

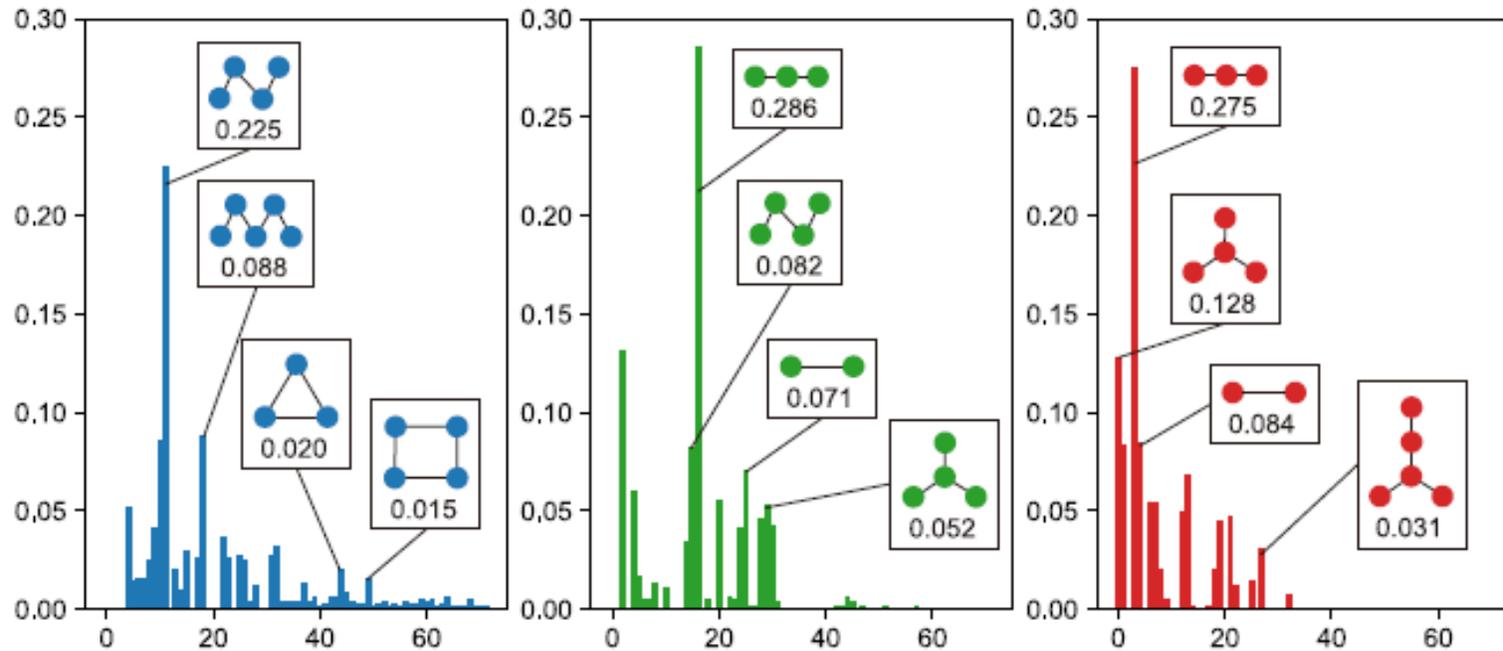


(e) MNMF

概念性验证

□ 所学习的主题分布

➤ 子图结构 在不同主题下 分布差异明显



(a) Walk-topic distribution by Graph Anchor LDA

实验与分析



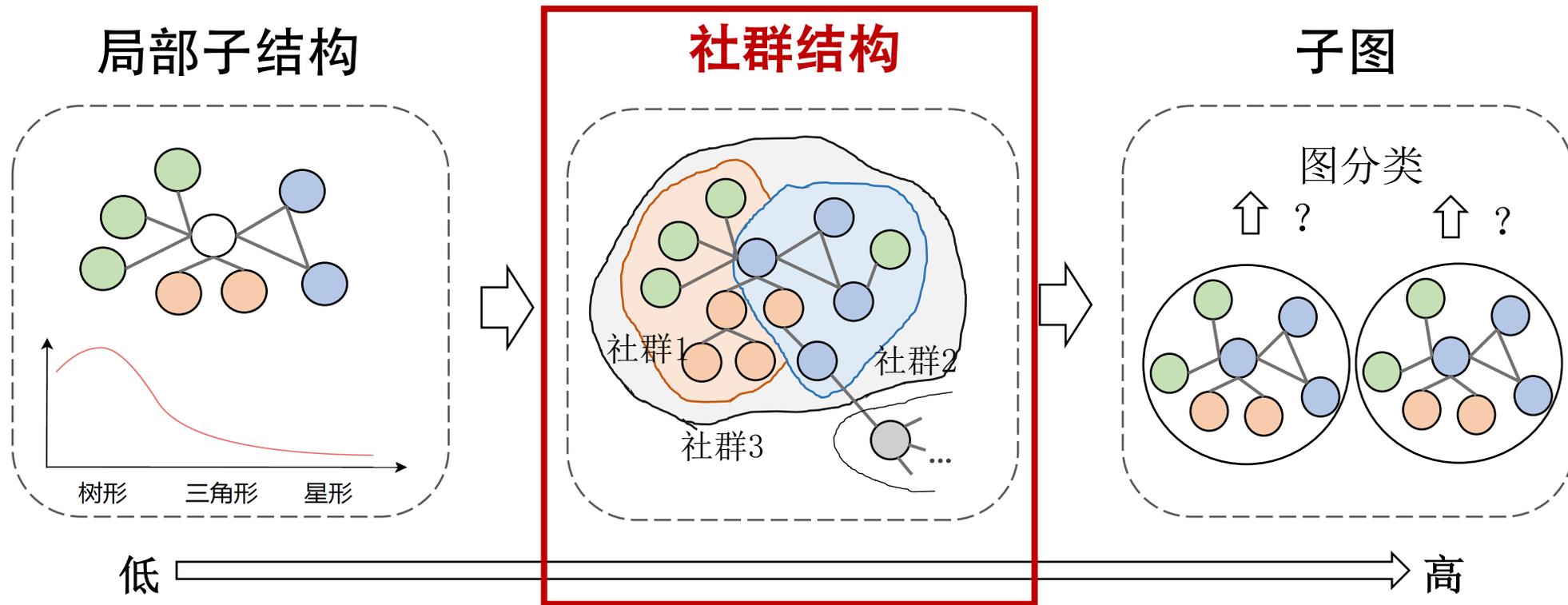
公开数据集（节点分类）

Input	Model	Cora				AMiner				Pubmed			
		Macro-f1		Micro-f1		Macro-f1		Micro-f1		Macro-f1		Micro-f1	
		30%	70%	30%	70%	30%	70%	30%	70%	30%	70%	30%	70%
No features	Struc2Vec	17.55	18.92	29.07	31.34	23.17	21.80	36.11	38.44	31.29	31.31	41.50	41.49
	GraLSP	58.86	63.62	60.88	64.45	43.19	43.03	45.85	45.92	38.89	38.84	45.88	46.01
	GCN	11.65	11.94	32.30	32.83	14.86	16.81	41.24	42.51	35.07	36.51	46.56	47.83
	GraphSTONE (nf)	70.75	71.83	72.73	73.52	57.11	56.70	58.21	58.91	56.87	58.88	60.47	60.69
	GCN	79.84	81.09	80.97	81.94	65.02	67.33	64.89	66.72	76.93	77.21	76.42	77.49
Features	GAT	79.33	82.08	80.41	83.43	68.76	69.10	67.92	68.16	76.94	76.92	77.64	77.82
	GraLSP	82.43	83.27	83.67	84.31	68.82	70.15	69.12	69.73	81.21	81.38	81.43	81.52
	GraphSAGE	80.52	81.90	82.13	83.17	67.40	68.32	66.59	67.54	76.61	77.24	77.36	77.84
	GraphSTONE	82.78	83.54	83.88	84.73	69.37	71.16	69.51	69.93	78.61	78.87	79.53	81.03

- GraphSTONE 效果优于其他所有baseline
- 节点无特征的情况下，效果相比最好baseline提升**15%-25%**

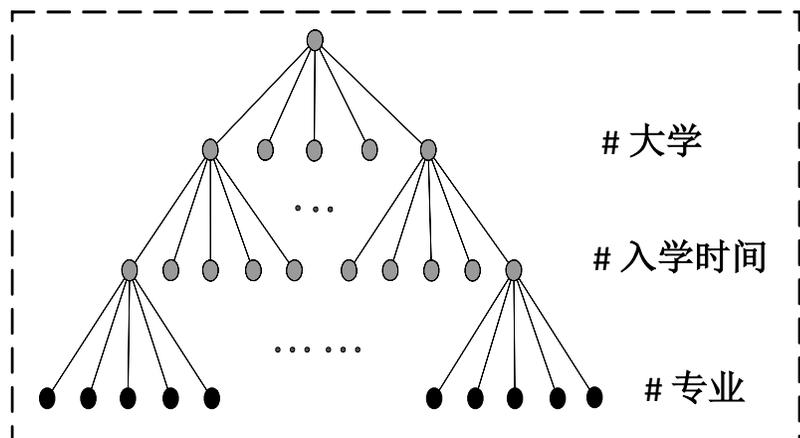
研究问题：面向子图结构的图神经网络建模

研究内容

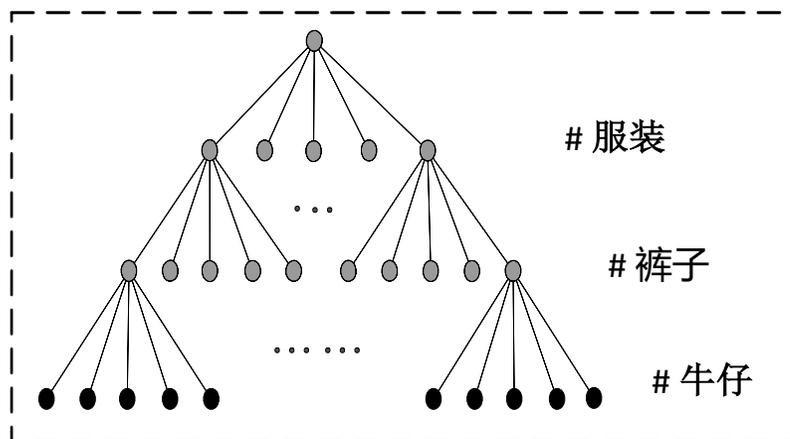


CIKM 最佳论文第二名

基于子空间理论的多层级社群表征模型

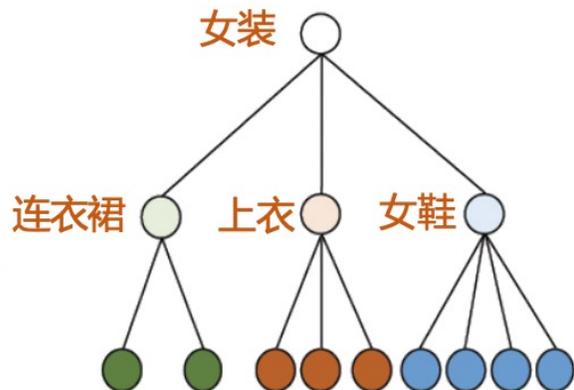
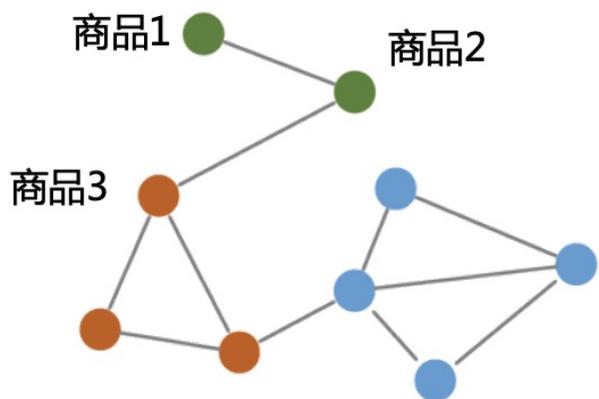


Facebook 社交网络



阿里电商网络

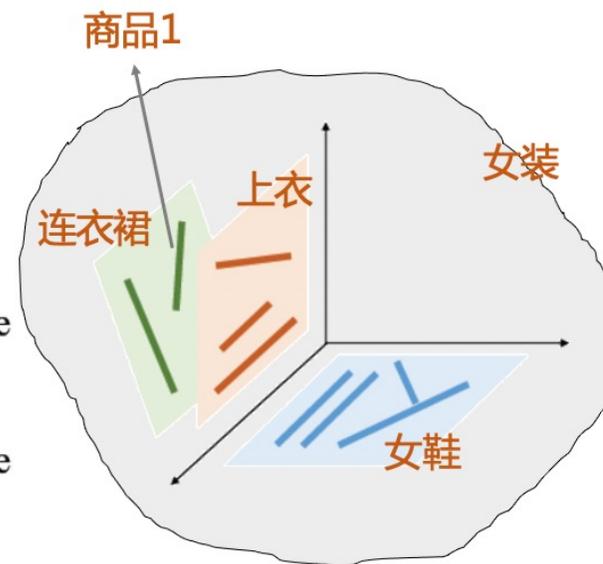
贡献: **弥补了**
表示学习建模
多层次社群的
空白



d-dimensional subspace

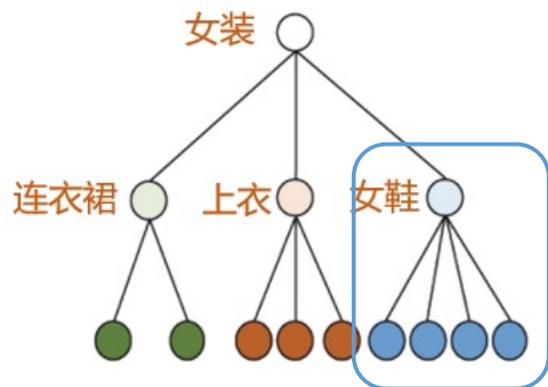
(d-1)-dimensional subspace

(d-2)-dimensional subspace

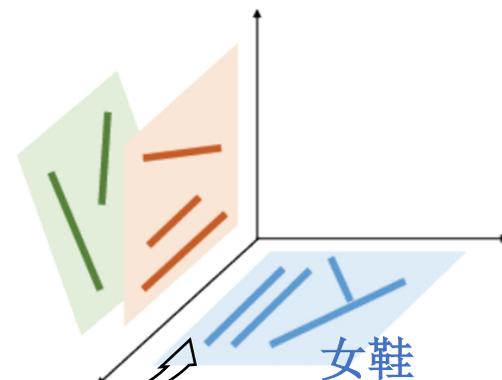


基于子空间理论的多层级社群表征模型

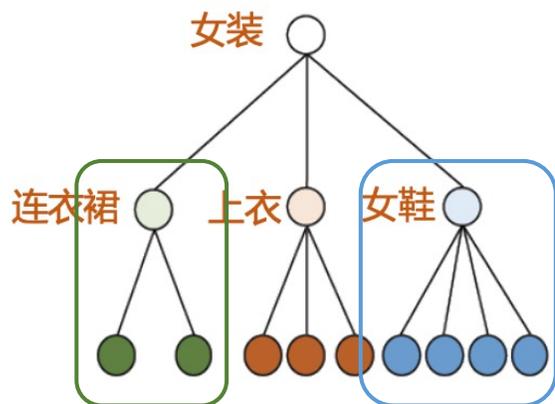
□ 多层级信息建模



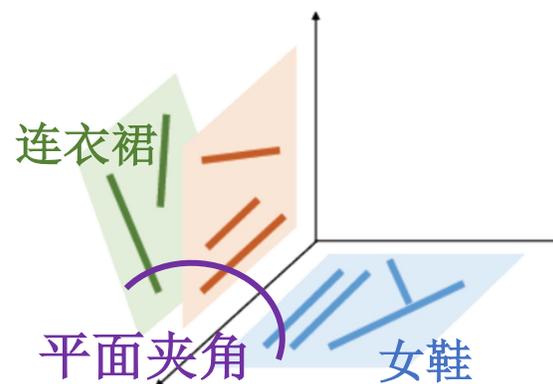
投影矩阵 S_i^l
(要学习的参数)



□ 社群间关系保持



投影矩阵 S_j^l, S_i^l
(要学习的参数)

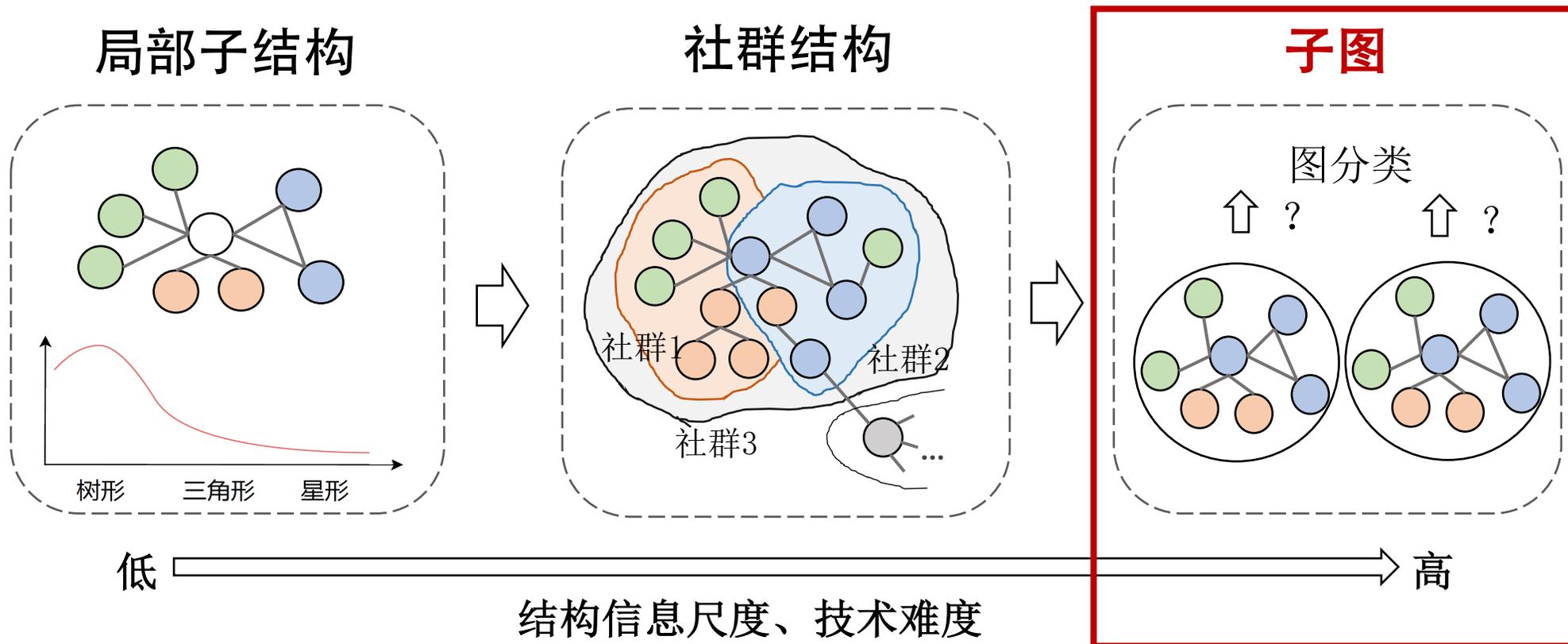


社群相似度

投影前后，社群间相似度保持一致

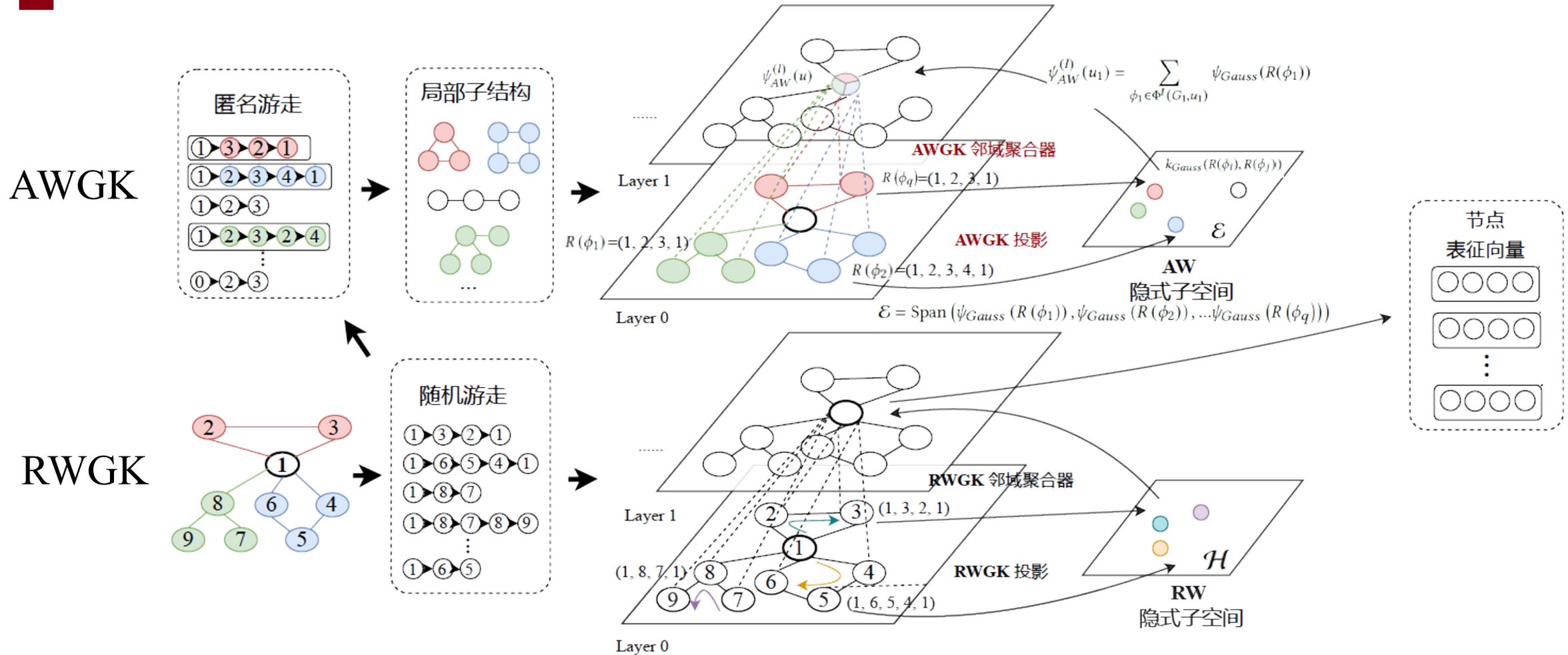
研究问题：面向子图结构的图神经网络建模

研究内容



WWW 2021、KDD 2021

基于子图核函数的图神经网络建模



- 融合随机游走图核函数 (RWGK) 与匿名游走图核函数 (AWGK)
- 可以与图神经网络进行融合

理论分析



Lemma 1.

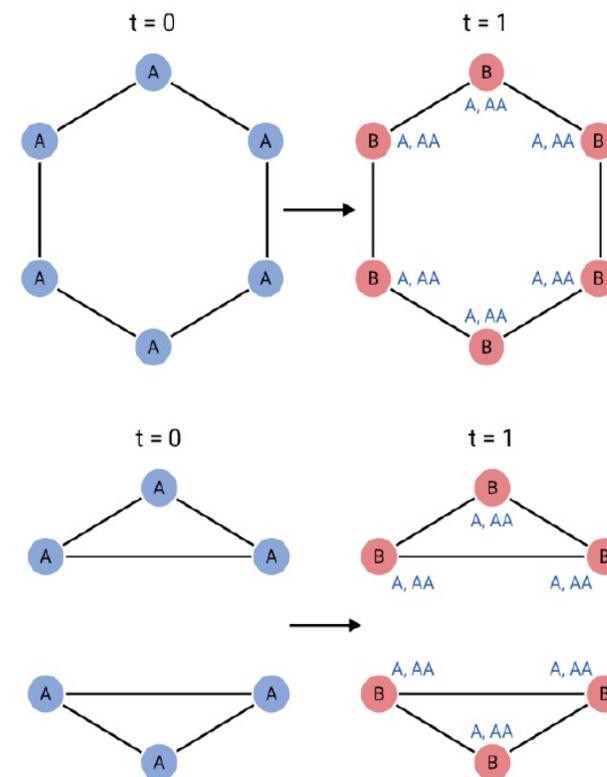
GSKN 的表达能能力大于等于1-WL test, 即表达能力最强的MPGNN (Message Passing GNNs)。

Lemma 2.

存在图G 满足 $\Psi_{AR,L}^{(l)}$ 的表达能能力严格大于1-WL Test。



贡献: 突破了现有GNN表达能力上界



Q & A



请各位老师提出宝贵意见