2020年智能系博士生答辩

大规模图的解构及其在图挖掘任务中的应用

答辩人： 吕天舒

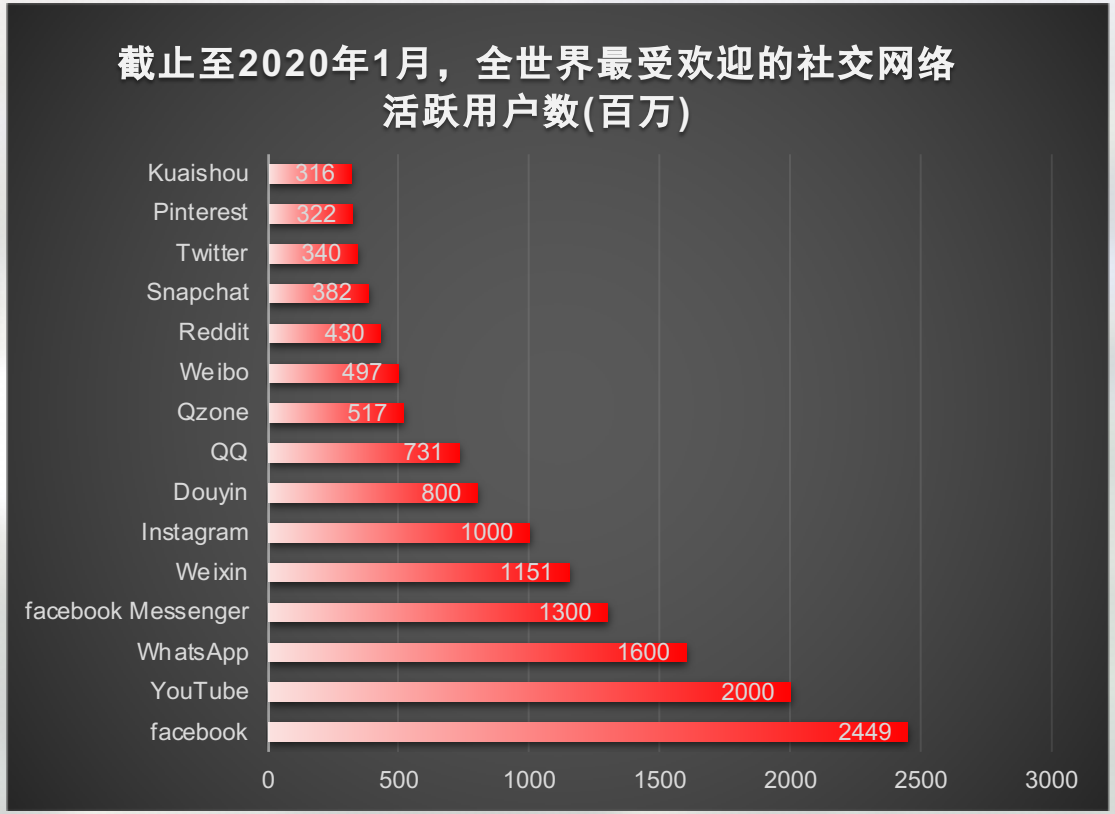
导师： 张岩 教授

2020.6.9

引言

关系型数据与大规模图

- 多领域、大规模
  - 线上社交网络、金融、生物、交通等
  - 百万级
- 异构信息图
  - 文本、标签、属性、图像、视频等
- 应用场景丰富
  - 点分类：e. g. 用户画像
  - 边预测：e. g. 好友推荐
  - 图分类：e. g. 异常检测



引言 — 项目支持

- 国家重点基础研究计划(973)项目“**网络大数据计算的基础理论及其应用研究**” 之课题“网络大数据模式发现与效应分析方法研究”(课题号:2014CB340405)
- 国家自然科学基金重点项目，面向课程的大规模在线教育资源组织与持续优化 的理论与方法(课题号:61532001)
- 教育部---中国移动科研基金项目，慕课教学效果与慕课的教育资源质量评价体 系及应用研究(课题号:MCM20170503)

引言 — 研究背景

研究课题

图挖掘

图挖掘

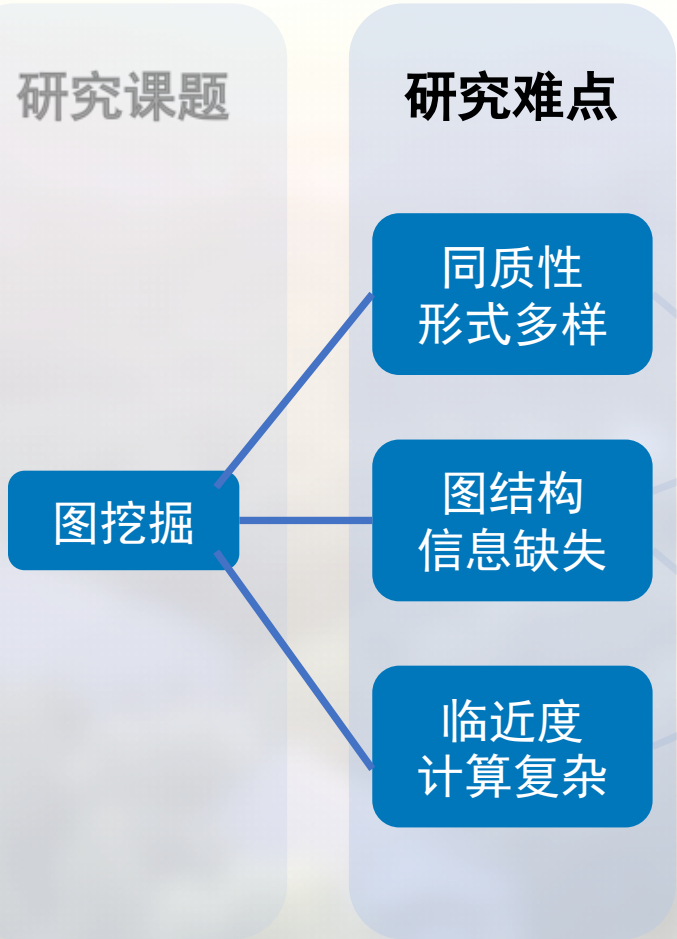
图的类型

1. 数学意义上的图
   (点集+边集)
2. 现实中的关系型数据
   (点集+边集+属性信息集)

挖掘方法

1. 图的表示方法+机器学习模型
2. 端到端训练

引言 — 研究难点



研究难点

- **同质性形式多样**
  相似年龄，相似地位，相似地域 ……
- **图结构信息缺失**
  点、边、属性信息不完整
- **临近度计算复杂**
  图上最优路径计算复杂度高

引言 — 研究问题

- **解构策略**

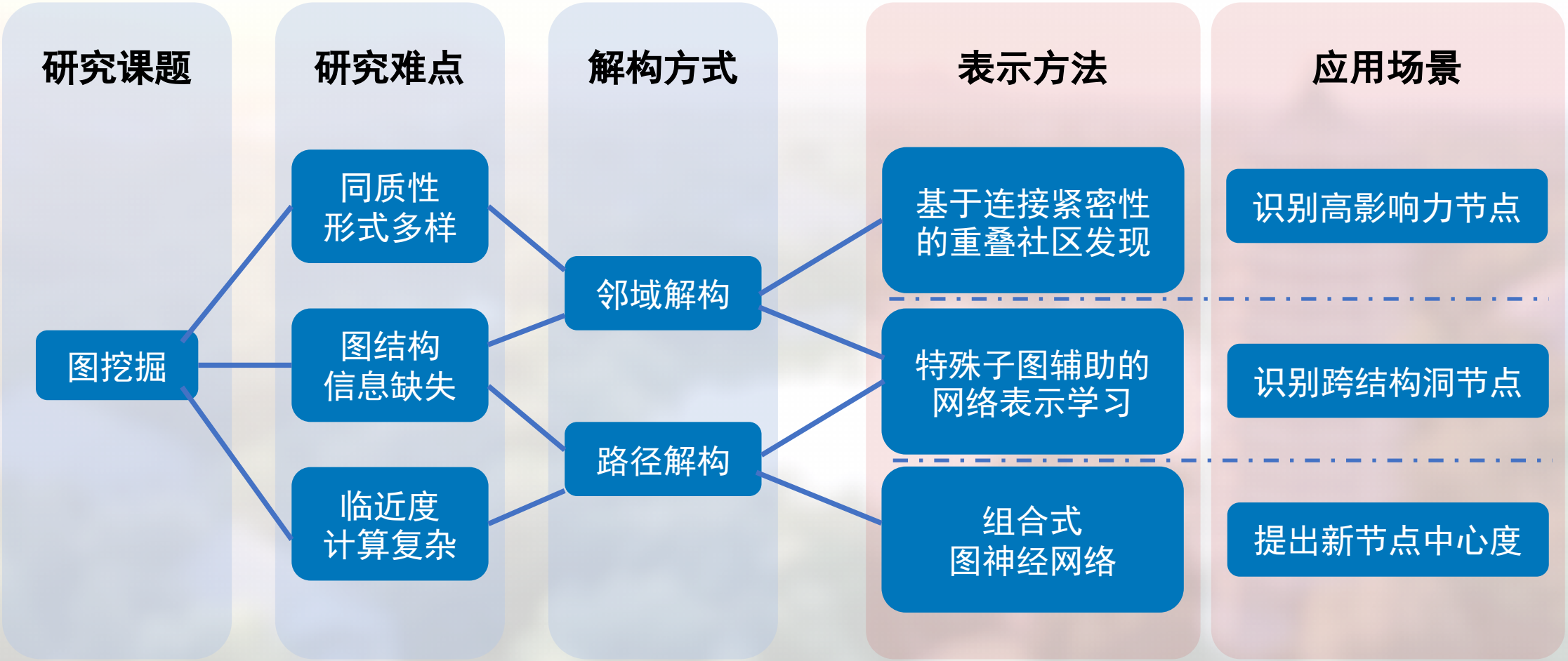
  总结已有的图挖掘经典算法中与**分治思想**相似的设计方法，并将其称为图的解构策略。

- **解构方式**
  - **邻域解构**

    子问题围绕着原图中包含目标节点的**导出子图**进行计算
  - **路径解构**

    子问题围绕着的原图中包含一对目标节点的**游走路径**进行计算

引言 — 主要工作


研究课题
研究难点
解构方式
表示方法
应用场景
图挖掘
同质性
形式多样
图结构
信息缺失
临近度
计算复杂
邻域解构
路径解构
基于连接紧密性
的重叠社区发现
特殊子图辅助的
网络表示学习
组合式
图神经网络
识别高影响力节点
识别跨结构洞节点
提出新节点中心度

相关研究 — 图的表示方法

- 用向量表示图/边/**节点**，方便机器学习模型处理的数据形式
- 图的**离散**表示：社区发现
- 图的**连续**表示：网络表示学习、图神经网络

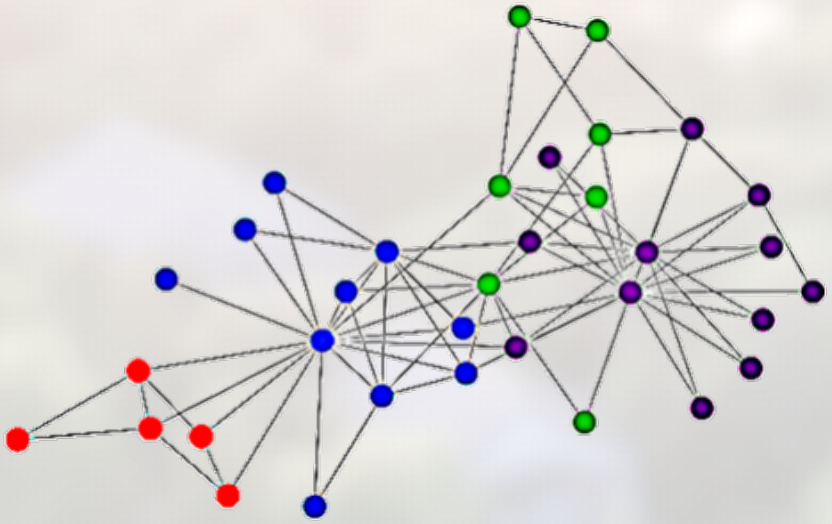


Fig. 图数据

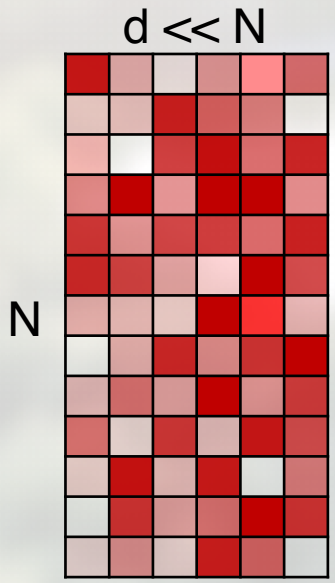


Fig. 图的向量表示

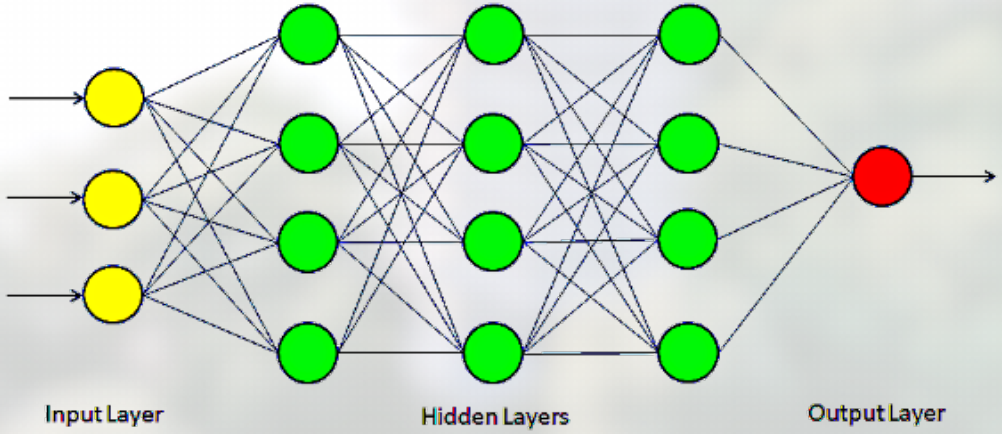


Fig. 神经网络

相关研究 — 图的表示方法 — 社区发现

完全子图 [1]	⚙ 简单易于理解 ⚙ 仅适用于读书分布平均的稠密图
非负矩阵分解 [2]	⚙ 可解释性强 ⚙ 社区的重叠部分连接更加紧密
局部扩散及最优化 [3]	⚙ 目标函数决定复杂度和质量 ⚙ 社区发现结果是冗余的
随机块模型 [4]	⚙ 可解释性不强 ⚙ 算法假设比较少

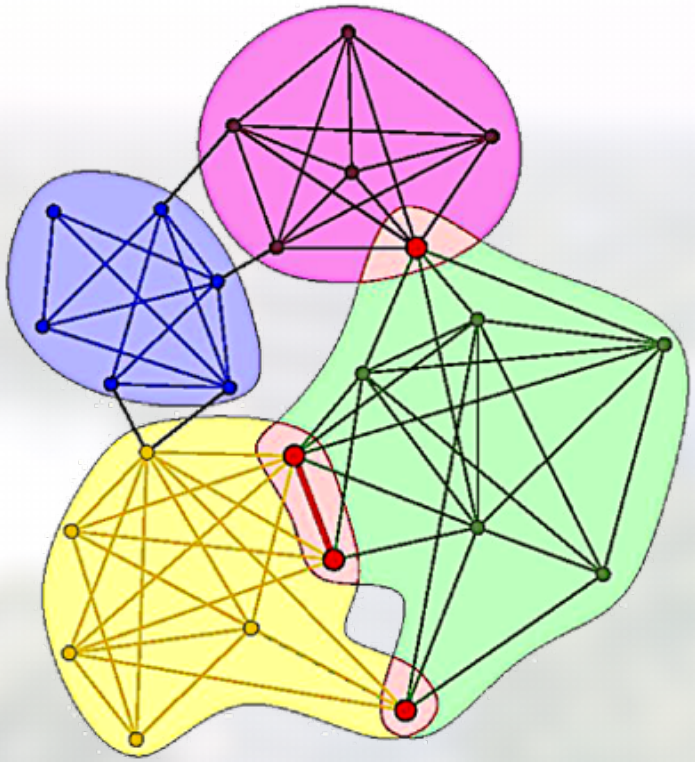


Fig. 社区发现

[1] B. Adamcsek, G. Palla, I. J. Farkas, I. Derenyi, and T. Vicsek, “Cfinder: locating cliques and overlapping modules in biological networks,” Bioinformatics, vol. 22, no. 8, pp. 1021–1023, 2006.
[2] J. Yang and J. Leskovec, “Overlapping community detection at scale: A nonnegative matrix factorization approach,” in WSDM’13, pp. 587–596
[3] J. J. Whang, D. F. Gleich, and I. S. Dhillon, “Overlapping community detection using neighborhood-inflated seed expansion,” TKDE, vol. 28, no. 5, pp. 1272–1284, 2016.
[4] P. K. Gopalan and D. M. Blei, “Efficient discovery of overlapping communities in massive networks,” PNAS, vol. 110, no. 36, pp. 14 534– 14 539, 2013

相关研究 — 图的表示方法 — 图表示学习

算法	捕捉的结构性质	对应的子图
DeepWalk [1]	连接的紧密性	随机游走路径
LINE [2]	一度、二度邻接关系	长度为1或2的短路径
GraRep [3]	高度的邻接关系	直径为定长的子图
SNS [4]	局部环境的相似性	节点数为3-5的小子图

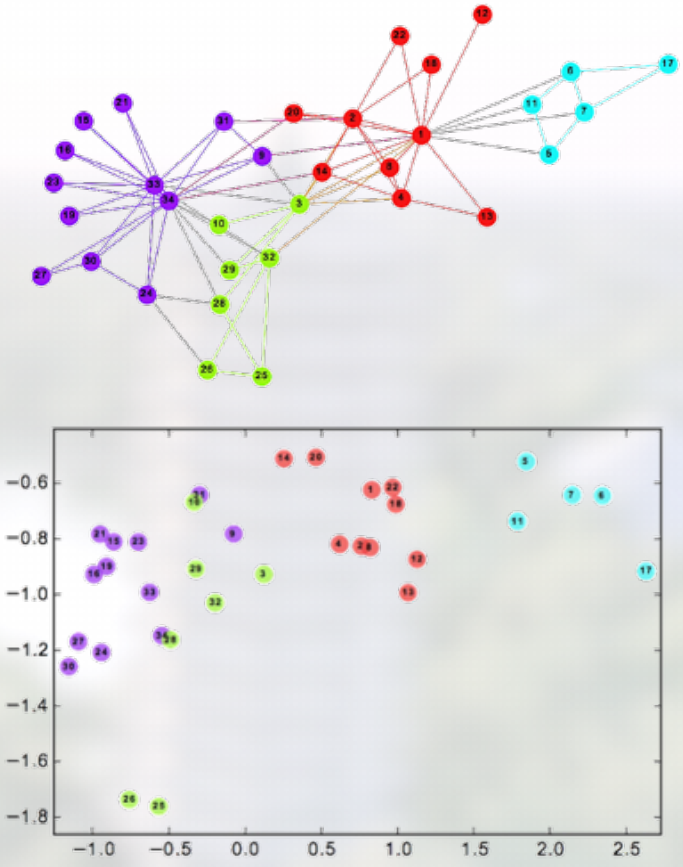


Fig. 图表示学习

[1] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 701-710.
[2] Tang J, Qu M, Wang M, et al. Line: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015: 1067-1077.
[3] Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 891-900.
[4] Lyu T, Zhang Y, Zhang Y. Enhancing the network embedding quality with structural similarity[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017: 147-156.

相关研究 — 图的表示方法 — 图卷积网络

$$\mathbf{Z} = f\left(\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}\right)$$

计算步骤	含义	相关工作
$\mathbf{X}\mathbf{W}$	节点特征向量的线性变换	GraphSAGE [1]
$\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}$	节点邻域间传递信息	GAT [2]
$\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}$	归一化每个节点接收的信息	DiffPool [3]
$f\left(\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}\right)$	用非线性的激活函数处理信息	DGCNN [4]

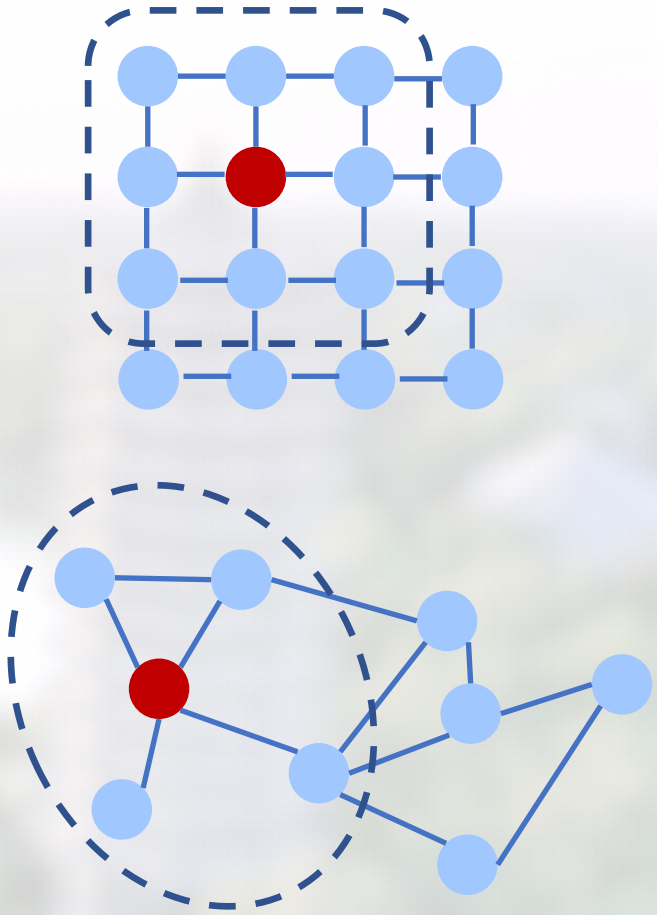


Fig. 2D卷积与图卷积

[1] Inductive Representation Learning on Large Graphs. W.L. Hamilton, R. Ying, and J. Leskovec arXiv:1706.02216 [cs.SI], 2017.
[2] P. Veličković, G. Cucurull, A. Casanova et al. “Graph attention networks”. arXiv preprint arXiv:1710.10903, 2017.
[3] Ying, Zhitao, et al. "Hierarchical graph representation learning with differentiable pooling." *Advances in neural information processing systems*. 2018.
[4] Zhang, Muhan, et al. "An end-to-end deep learning architecture for graph classification." *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

相关研究 — 解构策略

定义 1 (分治) 设 P 是待求解的问题。将 P 归约为 k 个彼此独立的子问题 $P_1, P_2, \cdots, P_k$。然后依此递归地求解这些子问题，得到解 $y_1, y_2, \cdots, y_k$。最后将这 k 个解归并得到原问题的解。

解构策略的优势 [2,3]

- 提升分类效果
- 对大数据的高扩展性
- 提升可解释性
- 实现模块化
- 适于并行计算
- 提升模型选择的灵活度

异构网络的元路径　集成学习　MoE框架　…

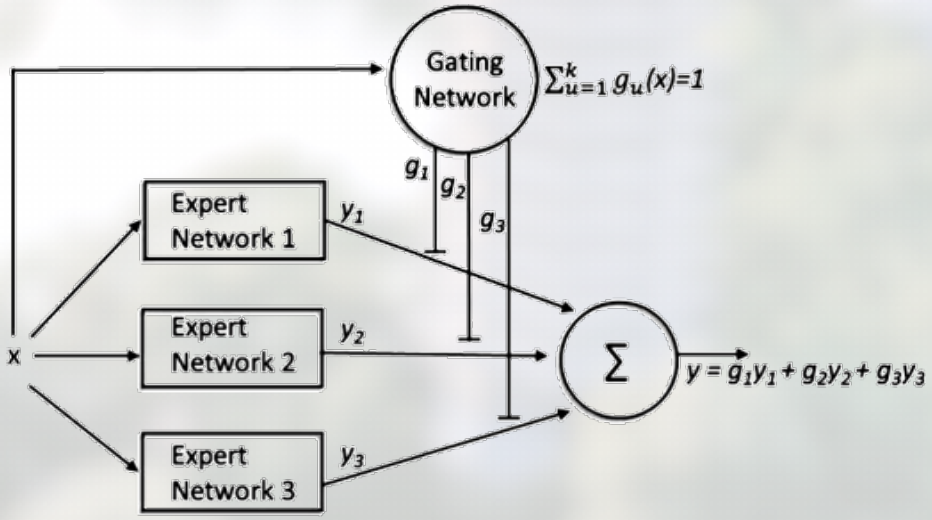


Fig.神经网络中的解构策略举例:Mixture-of-Experts 框架 [1]

[1] D. J. Miller and H. S. Uyar. “A mixture of experts classifier with learning based on both labelled and unlabelled data”. In: Advances in neural information processing systems. 1997: 571–577.
[2] A. J. C. SHARKEY. “On combining artificial neural nets”. Connection Science, 1996, 8(3-4): 299–314.
[3] L. Y. Pratt, J. Mostow, C. A. Kamm et al. “Direct Transfer of Learned Information Among Neural Networks.” In: AAAI. 1991: 584–589.

研究内容 — 图的解构策略 — 邻域解构

定义 3 (邻域解构) 设 P 是待求解的图挖掘问题。将 P 划分为 k 个彼此独立的子问题 $P_1, P_2, \cdots, P_k$，每个子问题围绕着原图中包含目标节点的**导出子图**进行计算。整理子问题的解 $y_1, y_2, \cdots, y_k$ 可以得到原问题的解。

邻域解构：适用于节点属性的建模

- “物以类聚，人以群分”
- “近朱者赤，近墨者黑”

- 邻域解构应对**同质性形式多样**
  - 切断与非邻域节点的联系
  - 刻画节点地位
- 邻域解构缓解**图结构采集缺失**
  - 淡化邻域内部结构的影响
  - 扩大邻域

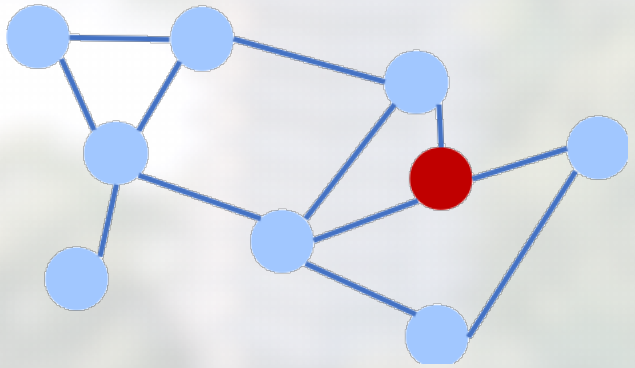


Fig.邻域解构围绕目标节点的k-邻域进行计算

研究内容 — 图的解构策略 — 路径解构

定义 2 (路径解构) 设 P 是待求解的图挖掘问题。将 P 划分为 k 个彼此独立的子问题 $P_1, P_2, \cdots, P_k$，每个子问题围绕着的原图中的一条**游走路径**进行计算。整理子问题的解 $y_1, y_2, \cdots, y_k$ 可以得到原问题的解。

路径解构：适用于节点对之间的关系建模

- 基于**最优路径**刻画节点对关系的劣势
  - 有悖实际情形
  - 计算复杂
- 基于**可行路径数目**刻画节点对关系的劣势
  - 忽略了中继节点的重要性
- 路径解构缓解**图结构采集缺失**
- 路径解构解决**临近度计算复杂**

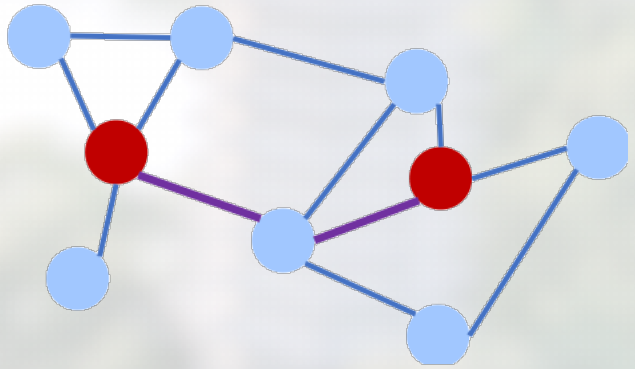
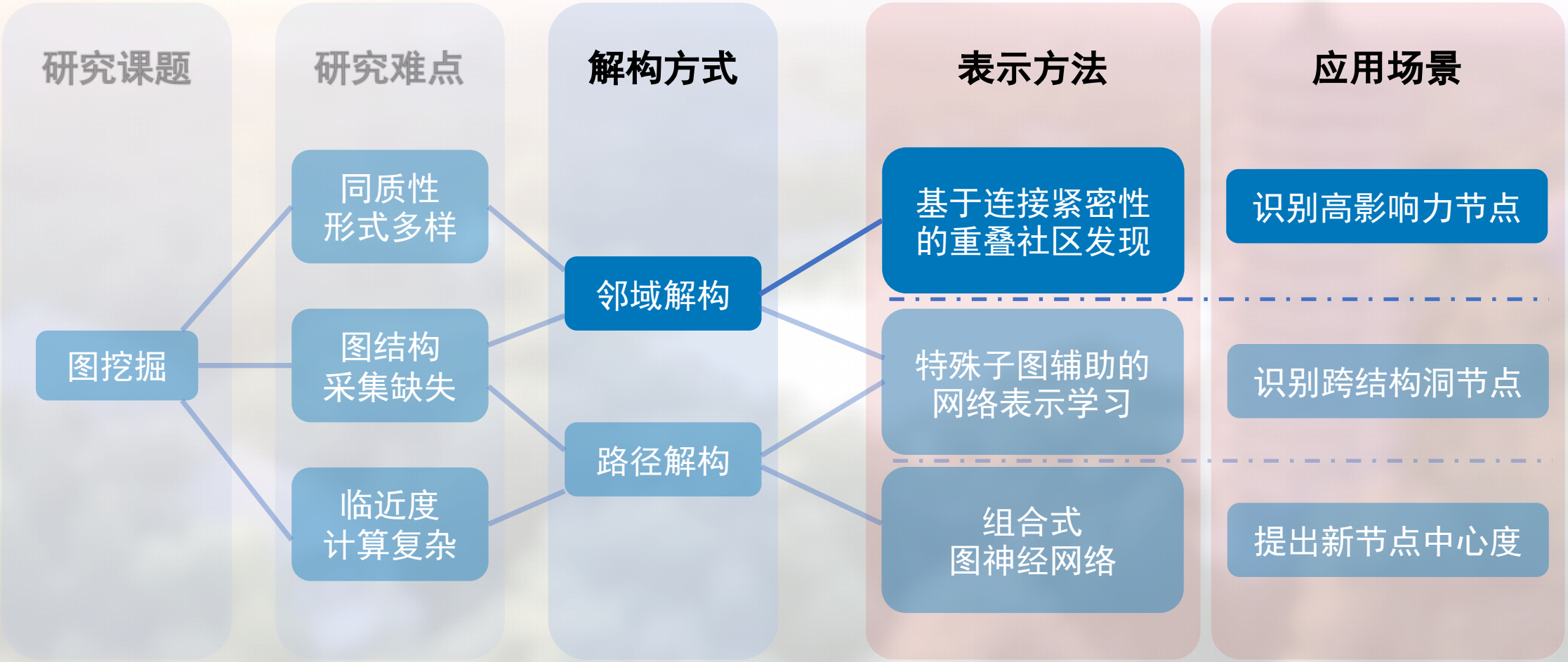


Fig.路径解构围绕多条随机游走短路径进行计算

图挖掘中
邻域解构策略的运用



研究内容 — 邻域解构运用 — Fox

大规模网络上的重叠社区发现算法

- 基于博弈论的社区发现算法 — Fox
  - **挑战：** 已有算法框架复杂度高
  - **策略：** 关注个体与邻居的关系，快速计算
  - **特点：** 符合潜博弈，保证算法收敛
- Fox与邻域解构
  - 社区发现任务 → 判断一对邻接节点是否同属于一个社区
  - 原始图 → 目标节点的自我网络

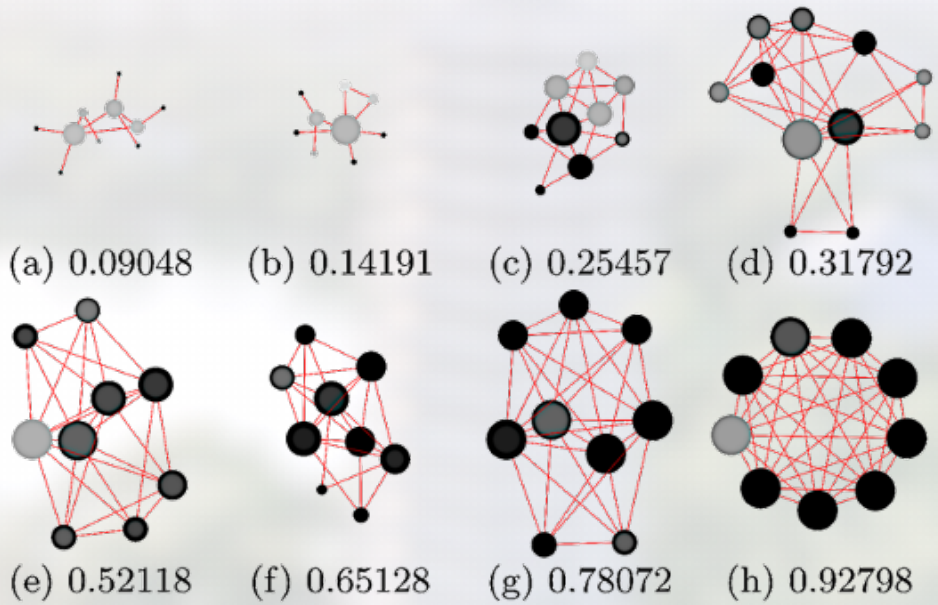


Fig. 用WCC指数刻画连接紧密程度

Tianshu Lyu, Lidong Bing, Zhao Zhang, and Yan Zhang. Efficient and scalable detection of overlapping communities in big networks. In the IEEE International Conference on Data Mining 2016 (ICDM 2016), Barcelona, Spain. December 12-15, 2016

研究内容 — 邻域解构运用 — Fox

大规模网络上的重叠社区发现算法

- 相较已有算法
  - 更优的社区划分
  - 更快的检测速度
  - 更大的网络规模

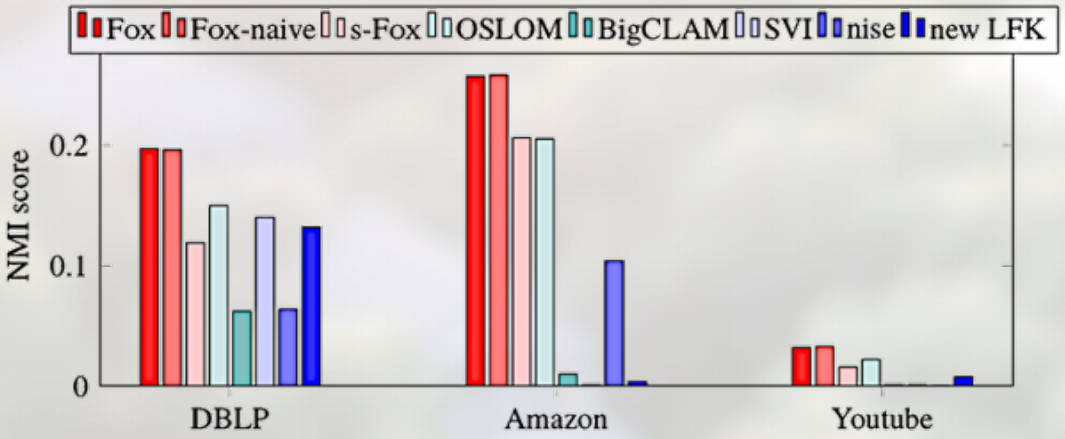


Fig. 各个社区发现算法在不同数据集上的准确度

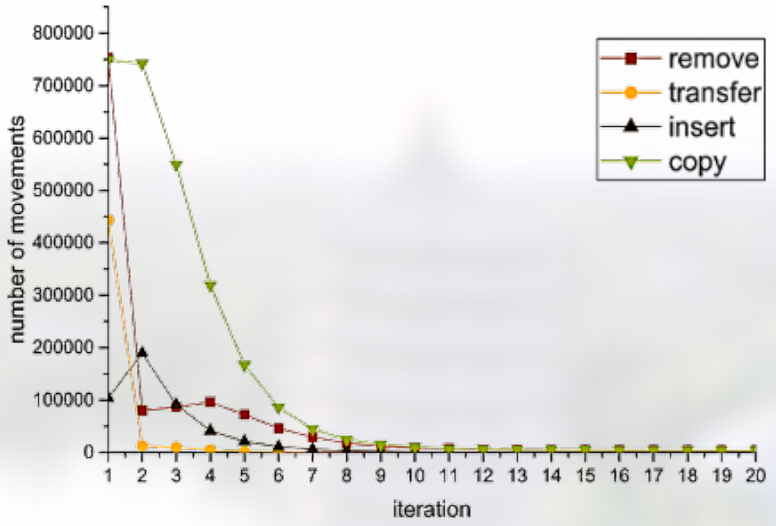


Fig. 迭代增加，移动的节点数目急速下降

	3.9M节点，20.5M边				22.5M节点，127.3M边		
Dataset	phone-call				Google+		
Algorithm	time cost	*Density*	w_c/w_i	Q_{ov}	time cost	*Density*	Q_{ov}
BigCLAM	38 hr.	0.028	0.604	**1.401**	-	-	-
OSLOM	194 min	0.442	1.845	0.621	-	-	-
FOX	**14 min**	**0.607**	2.920	0.758	238 min	0.529	1.044
s-FOX	20 min	0.325	**19.434**	0.936	260 min	0.328	1.334

Table 各个社区发现算法在大规模网络上的表现

研究内容 — 邻域解构运用 — DIM

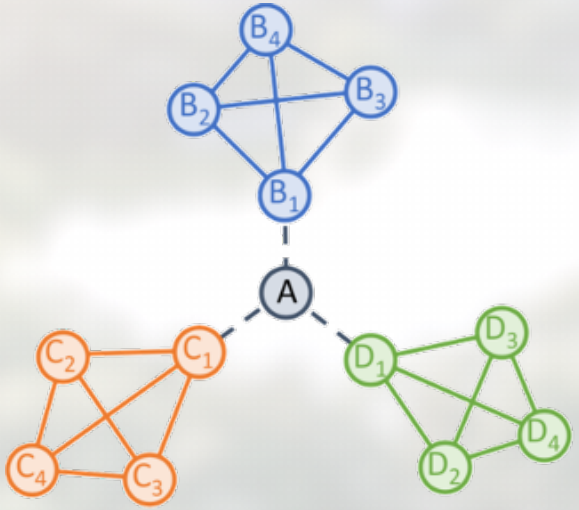
识别高影响力节点

- 差异性影响力最大化问题
  - **挑战：** 影响力高的节点组合在一起会相互影响
  - **对策：** 利用社区划分体现节点影响力的差异性
  - **特点：** 不是仅关注个体
- DIM与邻域解构
  - 图的离散表示的下游任务

Diversified Influence Maximization

Authority the expected cascade size

Diversity community similarity



Single Node Influence

Node	Influence
A	7
B_1	6.5
C_1	6.5
D_1	6.5
...	...

Node Set (k = 3) Influence

Node	Influence
A B_1 C_1	11
A B_1 B_2	9
...	...
B_1 C_1 D_1	12.875
...	...

Fig. 影响力最大化问题中需要考虑节点的diversity

Yu Zhang, Tianshu Lyu, Jiawei Han. Balancing Authority and Diversity in Influential Node Mining. Under review

识别高影响力节点

- 评判传播的多样性
  - 演员共同出演的网络，包含演员国籍、影片等信息

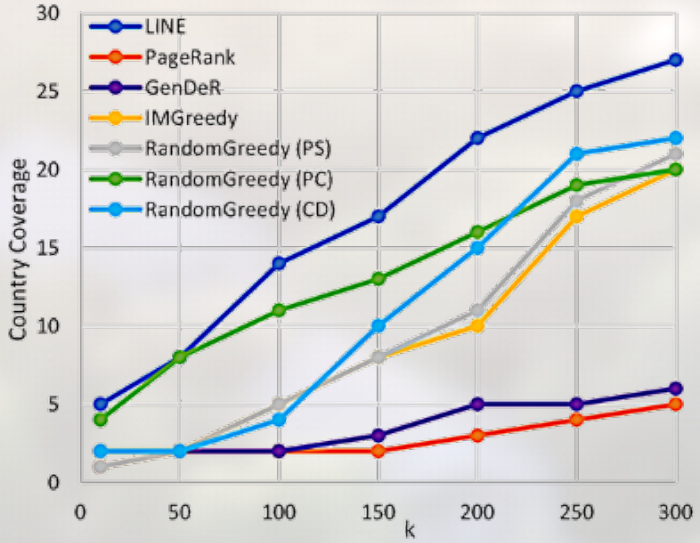


Fig. 国籍覆盖度（数值越高越好）

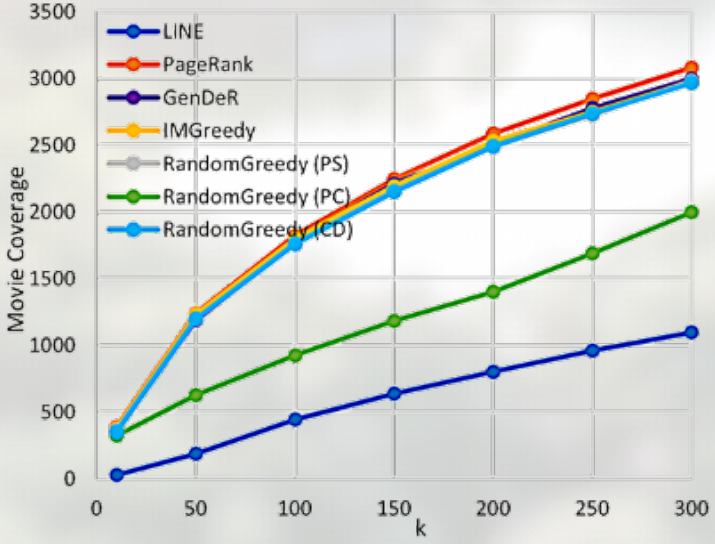


Fig. 影片覆盖度（数值越高越好）

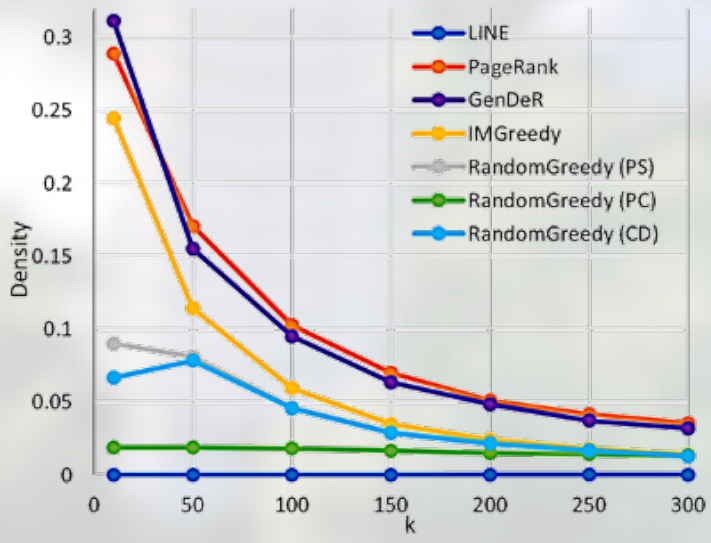
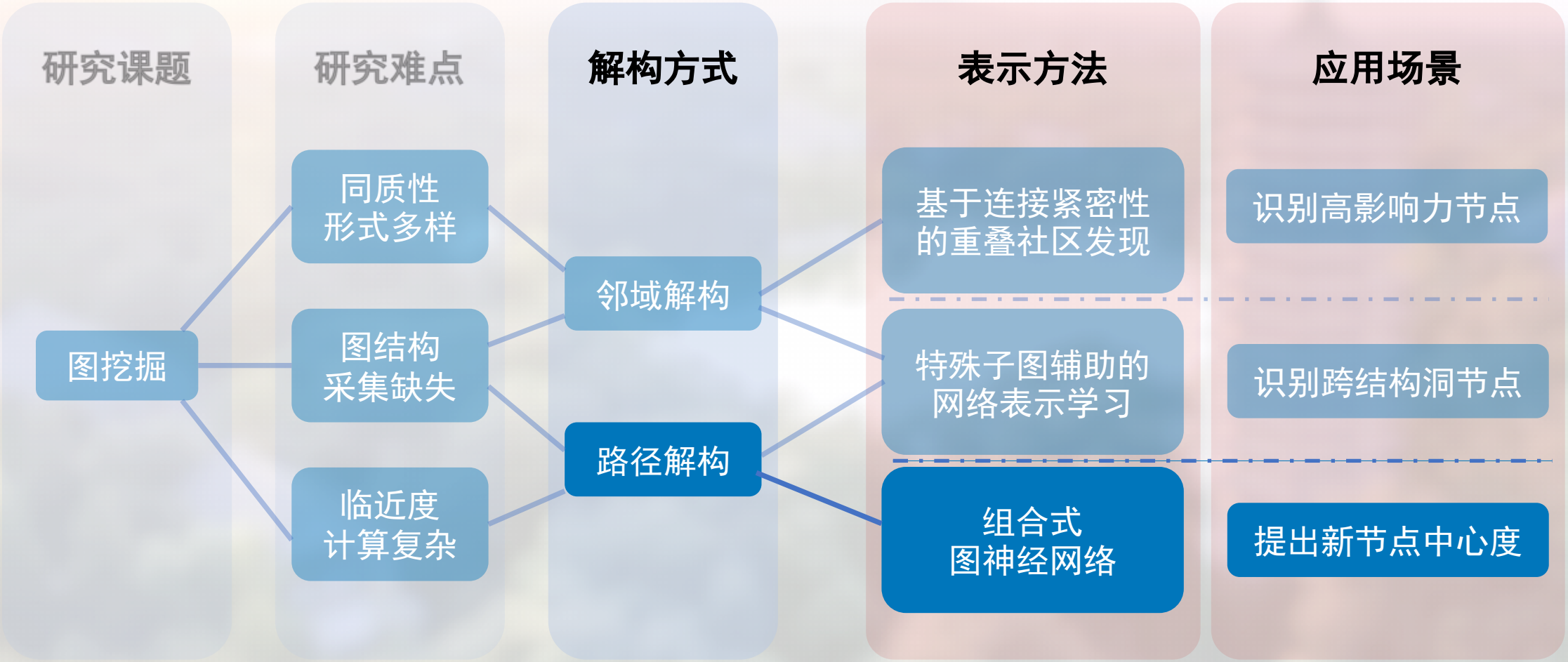


Fig. 密度（数值越低越好）

图挖掘中
路径解构策略的运用



研究内容 — 路径解构运用 — CNE

组合式图神经网络

- 基于组合泛化性的神经网络 — CNE
  - **挑战：** 冷启动问题、异构图问题的根源
  - **对策：** 抛弃ID，利用节点属性
  - **特点：** 结构+属性 联合训练

- CNE与路径解构
  - 刻画一对节点的关系 → 随机游走中，两个节点同时出现在窗口内的概率
  - 原始图 → 经过两个节点的一条随机游走路径

Fig. 图神经网络的两种框架：
传统框架（上）与组合式框架（下）

Tianshu Lyu, Fei Sun, Peng Jiang, Wenwu Ou, and Yan Zhang. Compositional Network Embedding for Link Prediction. In The 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark. September 16th-20th, 2019

研究内容 — 路径解构运用 — CNE

组合式图神经网络

- 相较于已有算法
  - 一般图、冷启动节点、异构图上更优
  - 特征向量反映结构信息

71K节点，1.1M边

算法	20%			40%			60%			80%		
	P@10	P@50	P@100	P@10	P@50	P@100	P@10	P@50	P@100	P@10	P@50	P@100
SGNS	0.033	0.011	0.007	0.034	0.011	0.007	0.035	0.012	0.007	0.036	0.012	0.007
DeepWalk	0.093	0.032	0.019	0.099	0.038	0.023	0.095	0.037	0.023	0.090	0.036	0.022
CANE	0.080	0.029	0.017	0.092	0.033	0.019	0.091	0.033	0.019	0.091	0.033	0.019
TriDNR	0.065	0.022	0.013	0.068	0.024	0.014	0.078	0.028	0.016	0.078	0.029	0.017
GraphSAGE	0.056	0.020	0.012	0.063	0.024	0.014	0.067	0.026	0.016	0.068	0.027	0.016
CNE	0.081	0.029	0.019	0.085	0.034	0.022	0.083	0.033	0.024	0.082	0.036	0.030
CNE$_{MUL}$	**0.120**	**0.040**	**0.022**	**0.128**	**0.041**	**0.022**	**0.134**	**0.047**	**0.027**	**0.136**	**0.054**	**0.033**

Table 边异构网络中的边预测任务准确率

	Rank	Product Title
Click Record	1	Spring green loose mid-sleeve casual T-shirt.
	2	Pierced lace off shoulder 3/4 sleeve loose blouse.
	3	Plus size floral printed slimming princess dresses.
	4	Fake-two-piece pierced lace flowy tank blouse.
DeepWalk	1	Cotton plain loose white t-shirt.
	2	Spring and summer outlet high-waist shorts.
	3	Ethnic style Thailand Napal summer holiday long dress.
CANE	1	Original design fashion loose hip pants.
	2	Ethnic style Thailand Napal summer holiday long dress.
	3	Summer sleeveless wrinkled dress.
TriDNR	1	Puff sleeve elegant floral printed blouse.
	2	Extra size slimming pierced long scarf wrap shawl.
	3	Spring and summer sleeveless casual jumpsuits.
GraphSAGE	1	Korean summer beautiful dress.
	2	Hong-kong embroidery dress.
	3	Korean summer fashion v-neck hoodie.
CNE	1	Summer flower figure-flattering princess dress.
	2	Slimming cold shoulder empire waist fairy dress.
	3	Pink colorful dotted silk long-sleeve blouse.

[1] 相似同义的词语背景色是一致的。{ large size , casual , feminine , hot weather }

Fig. 根据用户的点击历史预测接下来的点击

研究内容 — 路径解构运用 — Node Conductance

提出新节点中心度

- 基于随机游走的节点中心度 — NC
  - **挑战：** 理想路径复杂度高
  - **对策：** 随机游走重返一个节点的概率
  - **特点：** 可利用图表示学习的结果
- NC与路径解构
  - 利用DeepWalk的结果近似计算

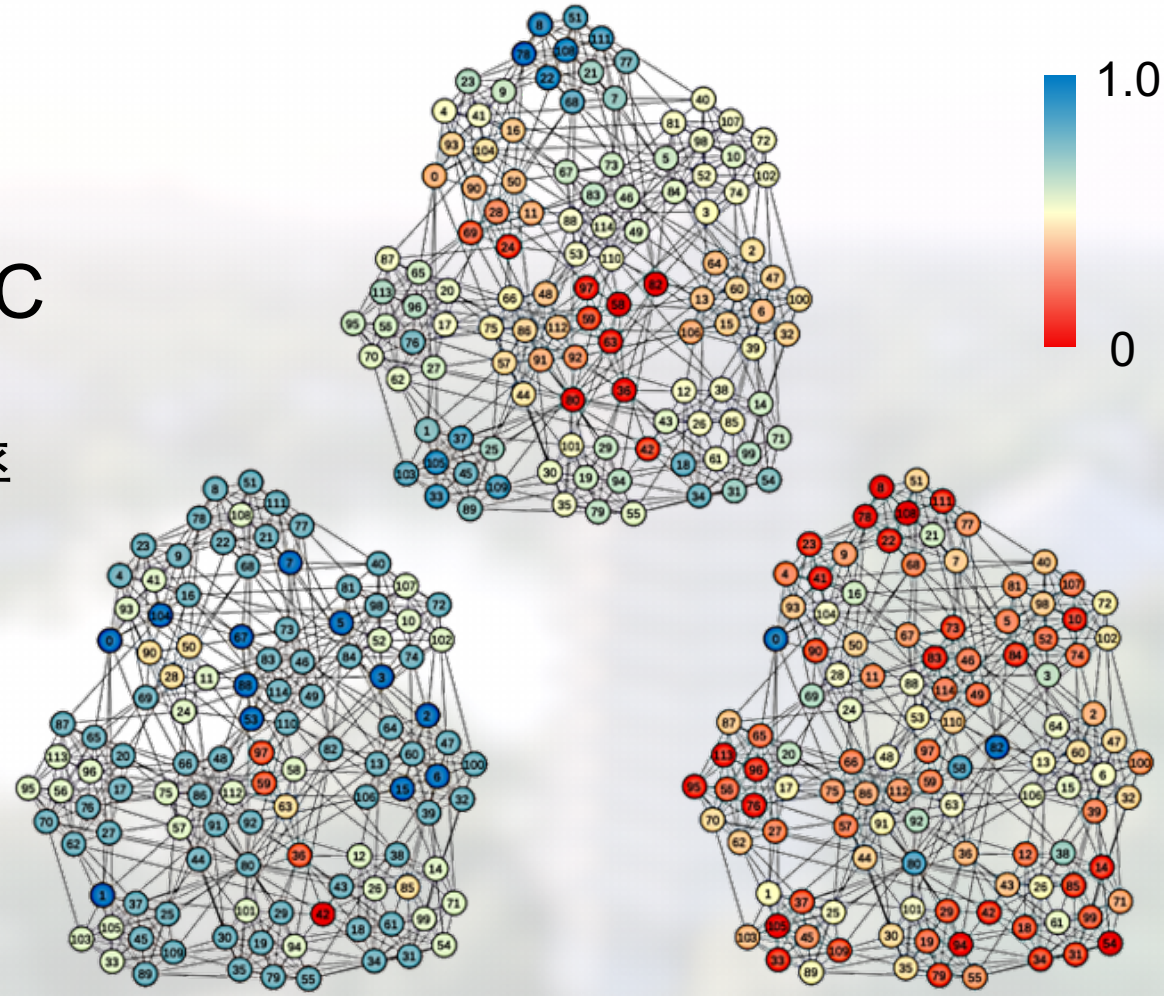


Fig. NC(上图)低的点：度数(左图)低，中介数(右图)高

Tianshu Lyu, Fei Sun, Peng Jiang, Wenwu Ou, and Yan Zhang. Node Conductance: A Scalable Node Centrality Measure on Big Networks. In The Pacific-Asia Conference on Knowledge Discovery and Data Mining 2020 (PAKDD 2020), Singapore. May 16th-20th, 2020

研究内容 — 路径解构运用 — Node Conductance

提出新节点中心度

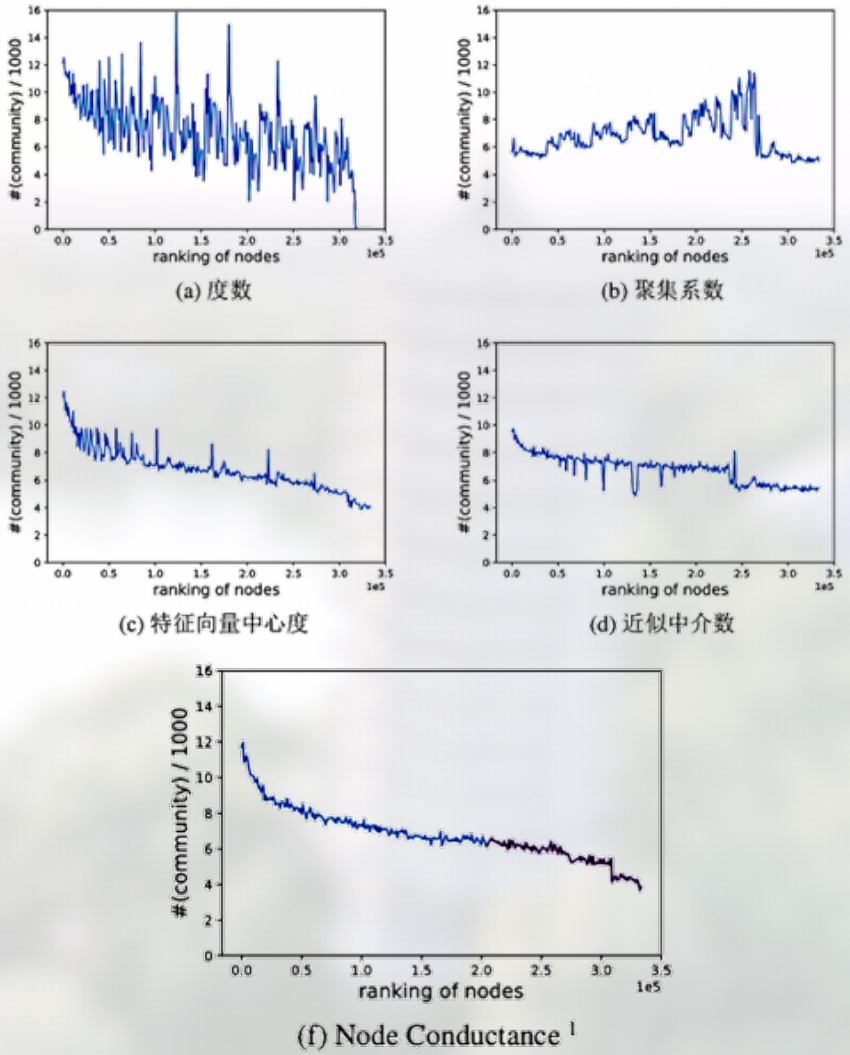
- 相较于已有中心度
  - 有一定差异性
  - 适用于大规模网络
  - 与节点的活跃度、新边生成更为相关

D: 335K节点，926K边 | A: 1M节点，3M边 | Y: 317K节点，1M边

数据集	AP[1]	NC[2]	AB[3]	AE[4]	SC[5]	FB[6]
DBLP	914	985	14268	–	–	–
Amazon	941	988	9504	–	–	–
Youtube	2883	3464	168737	–	–	–

[1] approximate PageRank. [2] Node Conductance.
[3] approximate Betweenness. [4] approximate Eigenvector Centrality.
[5] Subgraph Centrality. [6] Network Flow Betweenness.

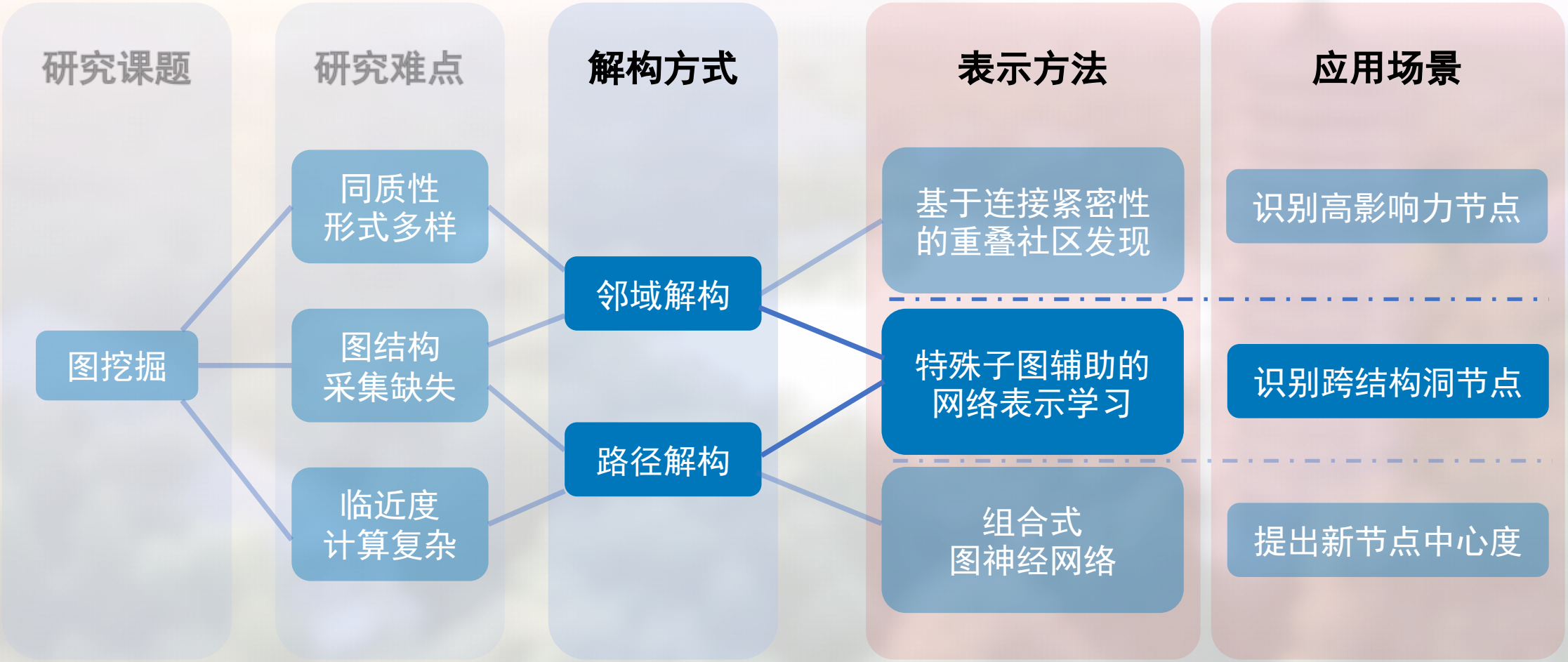
Table 全局节点中心度的运行时间



(a) 度数 (b) 聚集系数 (c) 特征向量中心度 (d) 近似中介数 (f) Node Conductance [1]

Fig. 节点所属的社区数目与节点的不同中心度

图挖掘中邻域解构与路径解构策略的运用



研究内容 — 两种解构综合运用 — SNS

捕捉节点地位信息的图表示学习

- 基于特殊小子图的图表示学习 — SNS
  - **挑战：** 鲜有方法关注节点地位的刻画
  - **对策：** 对节点的邻域进行建模
  - **特点：** 邻域拆解为特殊小子图
- SNS与两种解构
  - 路径解构：节点对连接紧密程度的建模
  - 邻域解构：节点地位的建模

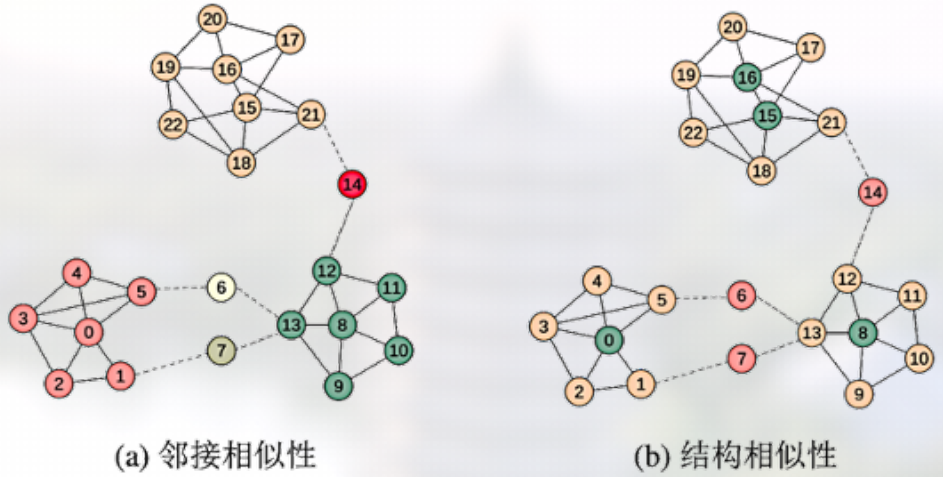


Fig. 节点的两种相似性

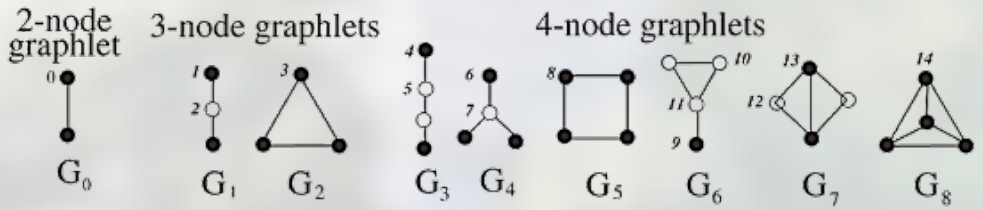


Fig. 用于刻画节点地位的特殊小子图

Tianshu Lyu, Yuan Zhang and Yan Zhang. Enhancing the Network Embedding Quality with Structural Similarity. In the 26th ACM Conference on Information and Knowledge Management (CIKM 2017), Singapore. November 6-10, 2017

研究内容 — 两种解构综合运用 — SNS

捕捉节点地位信息的图表示学习

- 相较于已有算法
  - 表示向量可捕捉地位相似性
  - 更高的节点分类准确率

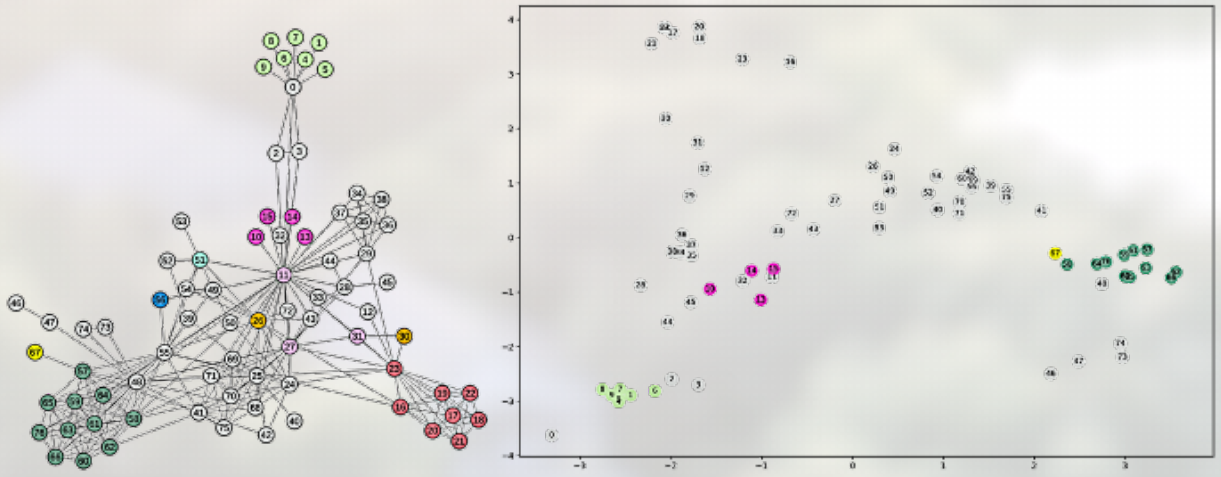


Fig. SNS的表示向量可以捕捉节点的地位相似性

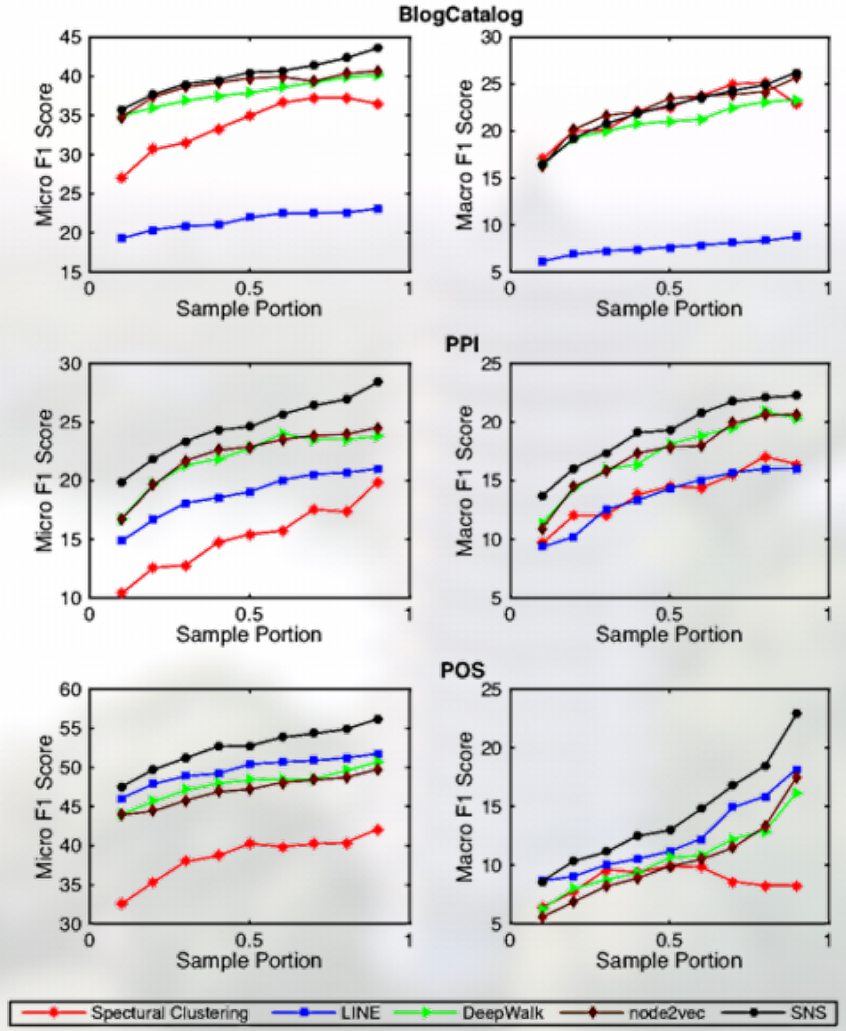


Fig. 节点分类任务的Macro-F1和Micro-F1

研究内容 — 两种解构综合运用 — RWSig

识别跨结构洞节点

- 基于随机游走的结构表示 — RWSig
  - **挑战：** 统计子图刻画结构复杂度高
  - **对策：** 将节点邻域拆解为游走路径
  - **特点：** 步骤简单高效，有图谱理论支撑
- RWSig与两种解构
  - 邻域解构：节点地位的建模
  - 路径解构：邻域节点关系的建模

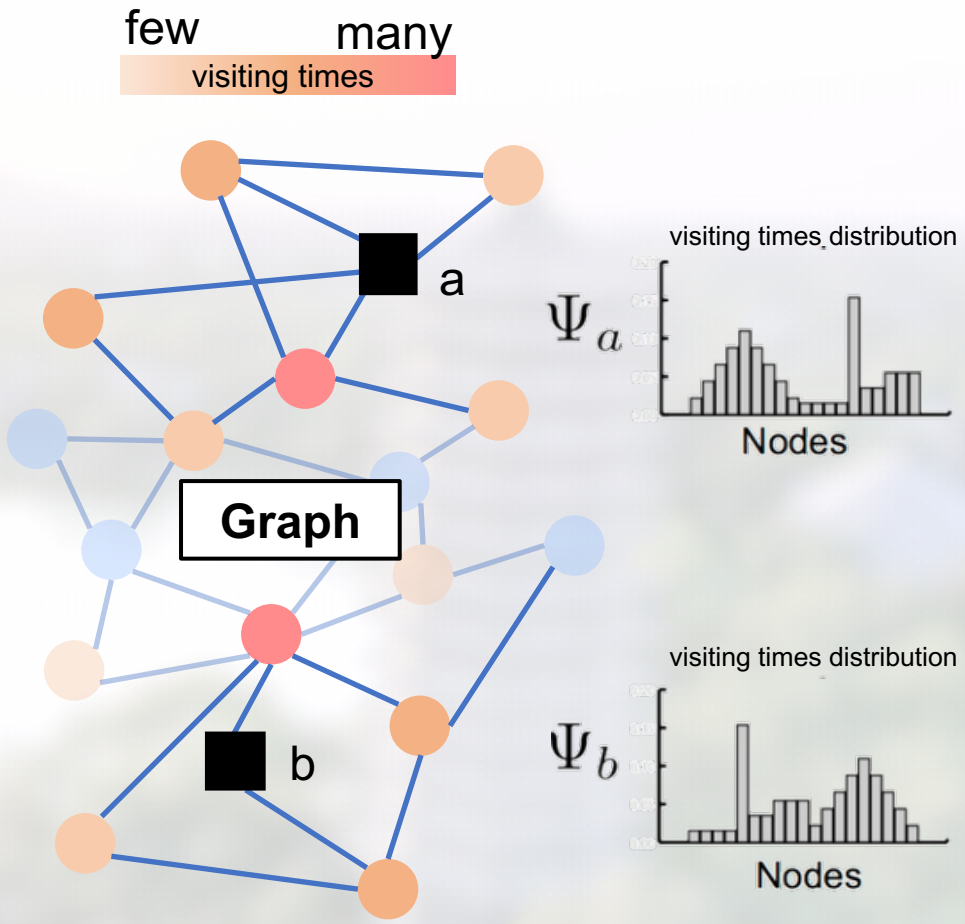


Fig. 利用随机游走访问次数构造表示向量

Tianshu Lyu, Yan Zhang. RWSig: A Network Structural Signature Based on Random Walks. Under Review.

研究内容 — 两种解构综合运用 — RWSig

识别跨结构洞节点

- 相较于已有算法
  - 提升跨结构洞节点的判断准确率
  - 对不活跃节点的判断更准确

335K节点，926K边

方法	指标	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
dw+sig	macro	**0.44**	**0.45**	**0.45**	**0.45**	**0.46**	**0.46**	**0.46**	**0.46**	**0.46**
	micro	**0.45**	**0.45**	**0.46**	**0.46**	**0.46**	**0.46**	**0.46**	**0.46**	**0.46**
sig	maco	0.3	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31
	micro	0.39	0.4	0.4	0.39	0.39	0.4	0.4	0.39	0.39
dw	macro	0.36	0.37	0.37	0.37	0.37	0.37	0.38	0.38	0.38
	micro	0.37	0.37	0.38	0.38	0.38	0.38	0.38	0.38	0.38

Table DeepWalk结合RWSig，识别结构洞节点的能力更强

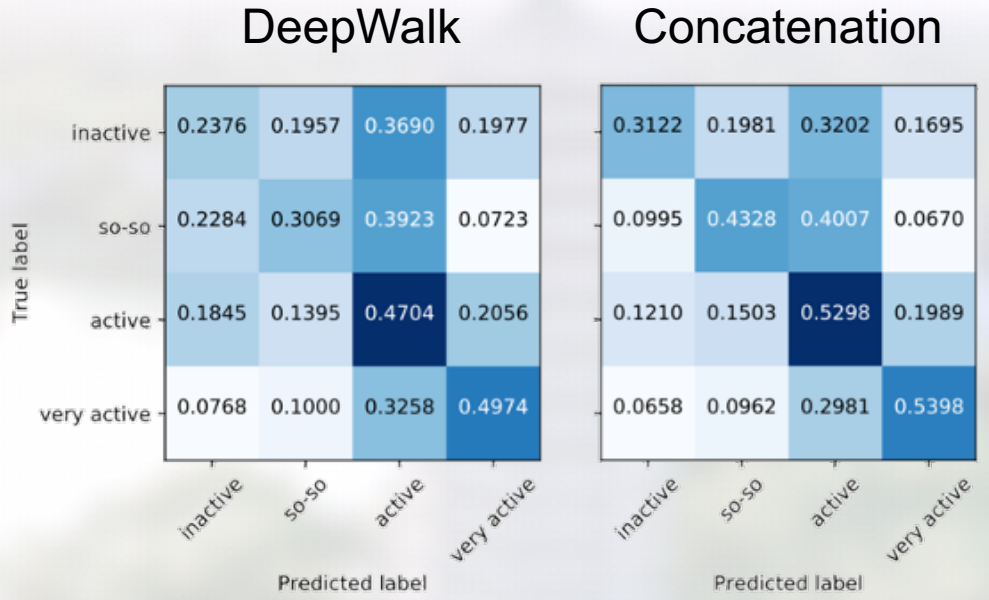


Fig. 混淆矩阵，判断不活跃节点更依赖RWSig

结语

- 邻域解构的运用
- 路径解构的运用
- 邻域解构和路径解构的综合运用

图挖掘任务中的挑战	解构策略赋予算法的能力
同质性形式多样	深入挖掘邻接关系的能力
图结构信息缺失	组合的泛化能力
临近度计算复杂	化繁为简的能力

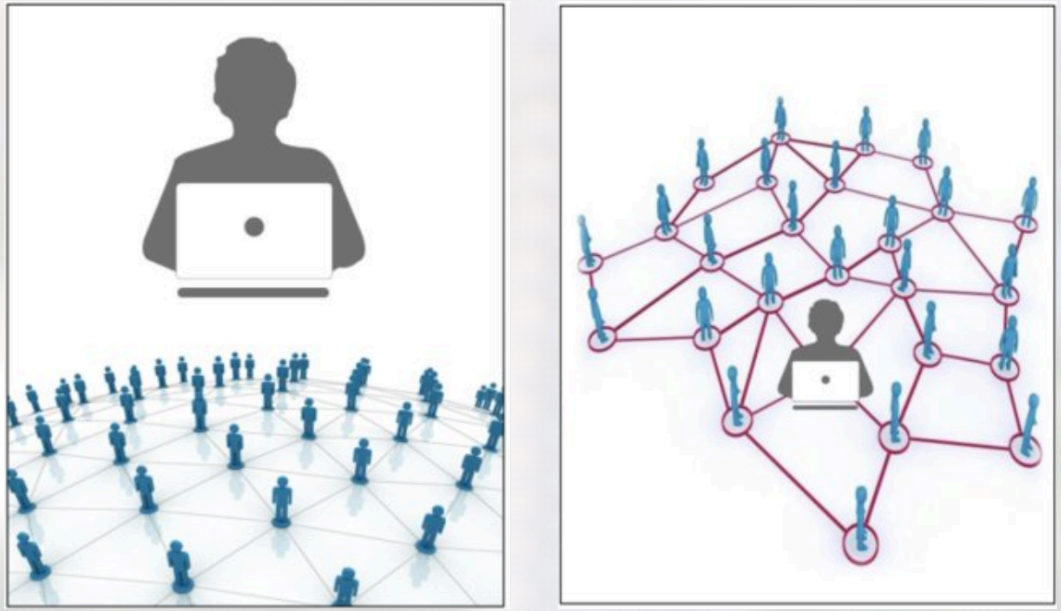


Fig. 图挖掘中的宏观(macro)方法和微观(micro)方法

在学期间研究成果

1. [**TSC** 已接收] **Tianshu Lyu**, Lidong Bing, Zhao Zhang, Yan Zhang . FOX: Fast Overlapping Community Detection Algorithm in Big Weighted Networks
2. [**PAKDD**'20] **Tianshu Lyu**, Fei Sun, Yan Zhang. Node Conductance: A Scalable Node Centrality Measure Based on Deepwalk.
3. [**RecSys**'19] **Tianshu Lyu**, Fei Sun, Peng Jiang, Wenwu Ou, Yan Zhang. Compositional Network Embedding for Link Prediction.
4. [**CIKM**'17] **Tianshu Lyu**, Yuan Zhang, Yan Zhang. Enhancing the Network Embedding Quality with Structural Similarity.
5. [**ICDM**'16] **Tianshu Lyu**, Lidong Bing, Zhao Zhang, Yan Zhang. Efficient and Scalable Detection of Overlapping Communities in Big Networks.
6. [**WISE**'18] Xiaoxuan Ren, **Tianshu Lyu**, Zhao Zhang, Yan Zhang. PUB: Product Recommendation with Users' Buying Intents on Microblogs.
7. [**AAAI**'18] Yuan Zhang, **Tianshu Lyu**, Yan Zhang. COSINE: Community-Preserving Social Network Embedding from Information Diffusion Cascades.
8. [**SIGIR**'17] Yuan Zhang, **Tianshu Lyu**, Yan Zhang. Hierarchical Community-Level Information Diffusion Modeling in Social Networks.
9. 《百面深度学习》，第三章 图神经网络，人民邮电出版社

2020年智能系博士生答辩

感谢！
请各位老师批评指正

答辩人： 吕天舒
导师： 张岩 教授

2020.4.15