机器学习结课报告

图算法在社区发现中的应用

作　　者：	崔博涵
学　　院：	统计与大数据研究院
专　　业：	应用统计班
年　　级：	2021 级
学　　号：	2021103835
指导教师：	严兴、张坤
论文成绩：	
完成日期：	2022.06.21

摘要

本文首先对图算法进行分类概述；社区发现作为图算法的一个重要应用场景，从聚类等一般社区检测技术再到深度神经网络，经历很长一段时间的发展；通用社区发现方法大体可以分为四类：传统聚类，基于边介数的分裂算法、基于模块度优化的聚合算法，基于图表示学习与神经网络的无监督聚类，为探讨上述几种方法在各场景下的适用性，本文按照复杂程度，分别构造和选取了由简到繁的三种数据集，初步得出结论：基于深度学习的无监督聚类算法在各数据集上的表现效果都较好，尤其是在社区数目较少，社区区分度较低的图数据集上；相对而言，louvain 等算法的划分效果不如图表征学习 + 聚类；由此可见，关于图神经网络在社区发现中的应用，还有很长的一段路要走。

关键词：社区发现　模块度　louvain　图神经网络　无监督聚类

目录

1 图算法综述

1.1 图算法分类

近年来，图算法在具有图结构的数据集上如社交分析，交通预测，推荐系统，生物学及计算机视觉等领域取得广泛且成功的应用。按照结构和性能分，图算法可以分为图表示学习，通过随机游走学习节点或者边的嵌入，常用于图数据预处理阶段；图挖掘算法，可以看做是图的传统机器学习算法，应用场景较为广泛，多为无监督学习；图神经网络，本质为节点信息的传递，将节点嵌入与下游任务通过网络联系起来，因其在节点分类等任务上常出现"sota result", 因此成为图算法研究邻域的一大热门方向。

图算法分类	学习类型	常见算法	核心思路	应用场景
图表示算法	无监督学习	Word2vec(词向量嵌入NLP领域) Deepwalk(DFS+word2vec) Node2vec(BFS+Deepwalk)	在图上随机游走学习图的拓扑结构进行图节点表示	图数据预处理
图挖掘算法	多无监督学习	节点度量：h-index, pagerank	由图的拓扑结构计算各节点的影响力	网页排序，关键点检测
		社区发现:louvian, LPA（标签传播）	基于模块度对节点进行社区划分	反欺诈中的团伙挖掘
		推荐召回：Swing	节点表示+相似度度量	推荐系统算法
		链路预测：Commonfriends	节点->边，编码（Graph embedding），解码（节点相似度度量，构造边）	社交网络中潜在关系挖掘（潜在黑产团伙挖掘） 推荐算法
图神经网络	监督,半监督，无监督都有	GCN, Graphsage（多跳邻居采样）,HGAT（异构图） RGCN（异构图）,GAT（图注意力网络）	图表示学习+下游任务	根据下游任务可以应用到很多场景中去

图 1: 图算法分类

GCN 的优点在于可以捕捉 graph 的全局信息，从而很好地表示节点的特征，然而由于是直推式学习 (Transductive learning) 方式，需要把所有节点都参与训练才能得到节点表示，无法快速得到新节点的表示；GraphSAGE 提出一种多跳邻居采样方法使得在该框架下学到的节点嵌入，是根据节点的邻居关系的变化而变化的，也就是说，即使是旧的节点，若建立一些新的链接，那么其对应的嵌入也会变化，而且也很方便地学到，结构见图 2。RGCN 在 GCN 的基础上考虑到节点之间关系的不同，应用于多关系型图场景中，结构见图 3。

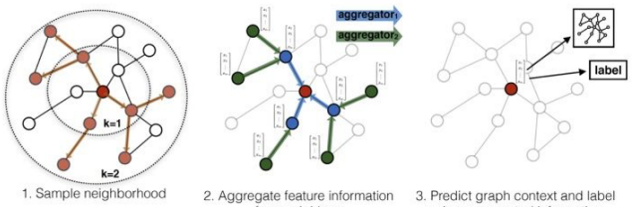


Figure 1: Visual illustration of the GraphSAGE sample and aggregate approach.

图 2: Graphsage 结构示意图

1.2 应用场景简介

通常情况下社交关系可以构成天然图，最常用的社会影响预测侧重于朋友之间行为的影响。例如，如果一些社交网络上的朋友买了一件衣服，他/她会不会也买呢? 如下图 5 所示，以社交图

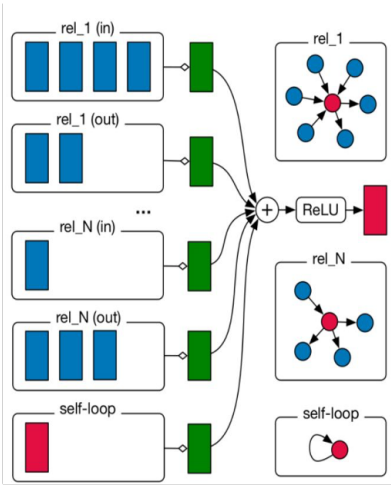


图 3: RGCN 网络结构，in 表示有向图中的入，out 表示有向图中的出

作为输入，DeepInf 为学习用户关系网络嵌入 (一种潜在的社会表征，仅学习图拓扑结构)。与节点其他属性拼接行程 (d) 中的特征，输入 GCN 或 GAT 卷积层，对社会影响进行预测，比如 v 是否也会观看广告片段 (步骤 f)。在训练过程中，将预测结果与真值进行对比，学习这个网络嵌入。

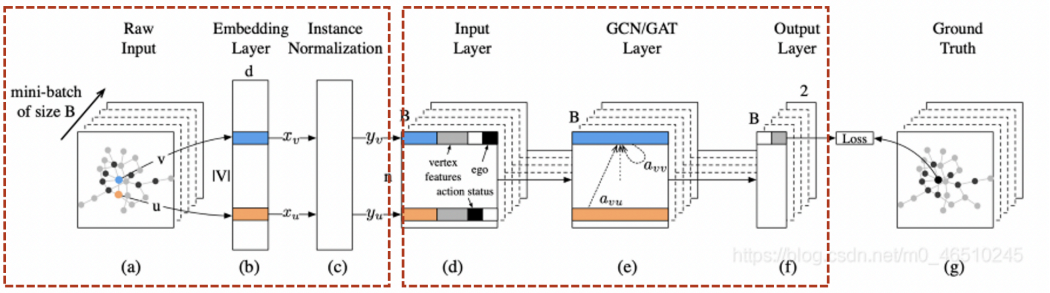


图 4: 图算法在社交网络场景中的应用

基于关系网络，图算法常用到推荐场景中来；一个推荐系统的灵魂在于计算目标变量之间的相似性，比如从历史数据上看，某用户对某些商品感兴趣，既可以寻找与该类商品相似的其他商品对该用户进行推荐，也可以寻找与该用户具有相似购买行为的其他用户推荐该用户曾购买过的商品。图算法在推荐场景中的具体应用总结概括为以下两种框架：

（1）图表示算法计算节点的 embedding 表示，计算节点之间的相似性

（2）二部图表示用户与商品之间的购买、浏览等关系，通过随机游走计算边的得分

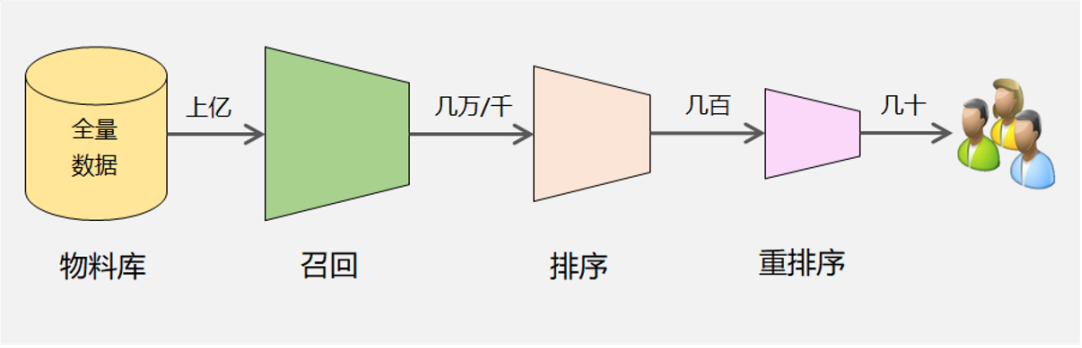


图 5: 推荐系统的简化结构

2 社区发现与图算法

2.1 社区发现的意义和应用

社区发现是图挖掘算法的一个重要应用。社区结构被定义为一组节点，用户相当于每一个点，用户之间通过互相的关注关系构成了整个网络的结构，在这样的网络中，有的用户之间连接较为紧密，有的用户之间连接关系较为稀疏，连接较为紧密的部分可以被看成一个社区，其内部的节点之间有较为紧密的连接，而在两个社区间则相对连接较为稀疏，这便称为社团结构，见图 6。这样的社团结构广泛存在于现实世界中的许多复杂系统中，在这些系统中发现社区（社团）已经成为理解网络结构与系统行为之间关系的主要方法。社区检测作为一种揭示潜在结构的有效技术，已经被应用于许多场景，如在社交媒体中寻找潜在的朋友，为用户推荐产品，分析社会意见。

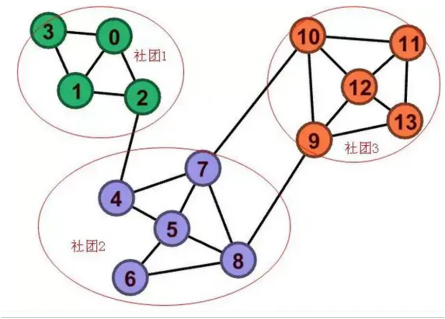


图 6: 社团结构

2.2 社区发现算法概述

常用的社区检测技术包括聚类、基于分裂的 G-N 算法，基于模块度的 louvain 算法。

(1) 层次聚类算法（$hierarchical_clustering$）：分层聚类算法中，认为一个图就是一个大社区，这个大社区包含复杂的层次结构，即社区可能是不同级别的小社区的集合，基于这种思想，通过顶点相似性度量（例如 $jaccarddasguptascore$），不需要预定义的社区规模和数量，使用树状图表示社区结构。

(2) 谱聚类：使用矩阵（常为图的拉普拉斯矩阵）特征向量，根据数据点之间的成对相似性划分数据点集。

(3) 基于分裂的社区检测技术-$Girvan-Newman$ 算法：这种方法基于低相似性删除网络中的簇间的边，从而将社区彼此分离。GN 算法根据边介数（边介数定义为网络中所有最短路径中经过该边的路径的数目占最短路径总数的比例）得分迭代删除边缘，类似于割图法。

(4) 基于模块度优化的社区检测技术-louvain 算法：Louvain 是一种启发式贪婪算法，用于在复杂加权图中发现社群。首先为每个顶点指定不同的社区，根据模块化的增益迭代地合并节点。若没有增益，那么这个节点将保留在它自己的社区中。重复该过程，直到模块度增益不再增加。然后，它以超级节点取代第一阶段确定的社区的方式重建网络继续第二阶段的

迭代计算。模块度公式定义如下：

$$Q = \sum_c \left[\frac{\sum in}{2m} - \left(\frac{\sum tot}{2m} \right)^2 \right]$$
$$= \sum \left[e_c - a_c{}^2 \right]$$

模块度可以理解为：对于一张图中所有已经划分的社区而言，每一个社区的内部的边的权重之和减去所有与社区节点相连的边的权重之和。模块度越大越好。

(5) node2walk-k-means：通过 node2walk 学习图的拓扑结构生成节点表征向量，进一步通过 k-means 对节点完成聚类任务。

(6) GCN-无监督学习-k-means：建立基于 GCN 卷积层的无监督学习任务，目标仍然是生成节点表示，进一步通过 k-means 完成聚类任务。

本文将在上述 6 种社区检测技术上进行实验，比较各方法的适用性。

2.3 实验数据集简介

本次实验基于三种具有社区结构的图数据集：构造图 G, Zachary' s karate club 数据集, American College football；其中，图 G 是人为构造的具有明显 4 个社团结构的图数据。Zachary 网络是通过对一个美国大学空手道俱乐部进行观测而构建出的一个社会网络。网络包含 34 个节点和 78 条边, 其中个体表示俱乐部中的成员, 而边表示成员之间存在的友谊关系。Football 网络是 Newman 根据美国大学生足球联赛而创建的一个复杂的社会网络。该网络包含 115 个节点和 616 条边, 其中网络中的结点代表足球队, 两个结点之间的边表示两只球队之间进行过一场比赛. 参赛的 115 支大学生代表队被分为 12 个联盟。比赛的流程是联盟内部的球队先进行小组赛, 然后再是联盟之间球队的比赛。这表明联盟内部的球队之间进行的比赛次数多于联盟之间的球队之间进行的比赛的次数. 联盟即可表示为该网络的真实社区结构。

表 1: 实验数据集

名称	节点数	边数	社区数
G	13	23	4
Zachary' s karate club	34	78	2
American College football	115	616	12

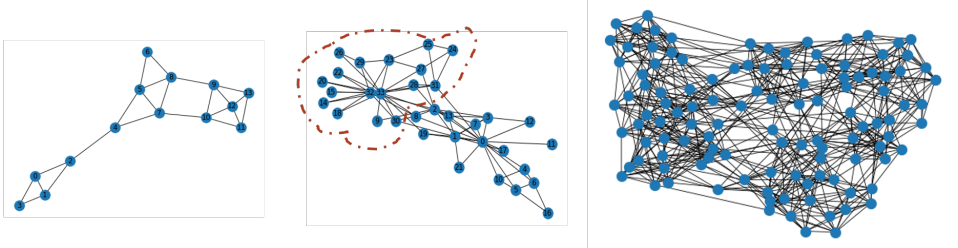


图 7: 从左至右依次是 G、Zachary、football 网络的可视化结构

对于无监督实验，没有明确的评价指标，因此，本文提出以下评价指标，可以较好衡量两社区集合之间的相似度，自定义社区划分结果 A 和社区划分结果 B 的相似度：

$$S(A \mid B)=\frac{\sum_{a_i \in A} \max \{\operatorname{len}(a_i \cap b_i) \mid b_i \varepsilon B\}}{\operatorname{len}(A)}$$

a_i 表示 A 中的社区，b_i 表示 B 中的社区，$\max\{len(a_i \cap b_i) \mid b_i \varepsilon B\}$ 表示 a_i 与 B 中社区交集的最大长度。假设 B 是真实的社区结构，若模型划分结果 A 中的社区个数多于 B 中社区个数，则 $S(A \mid B)$ 值偏高，举一个极端例子，假设 A 中的社区个数就是一个个独立的节点，则 $S(A \mid B)$=1，因此为消除由社区划分数不等带来的影响，进一步定义两社区间的相似度度量公式：

$$\text{similarity} = \frac{S(C_{\text{model}} \mid C_0)+S(C_0 \mid C_{\text{model}})}{2}$$

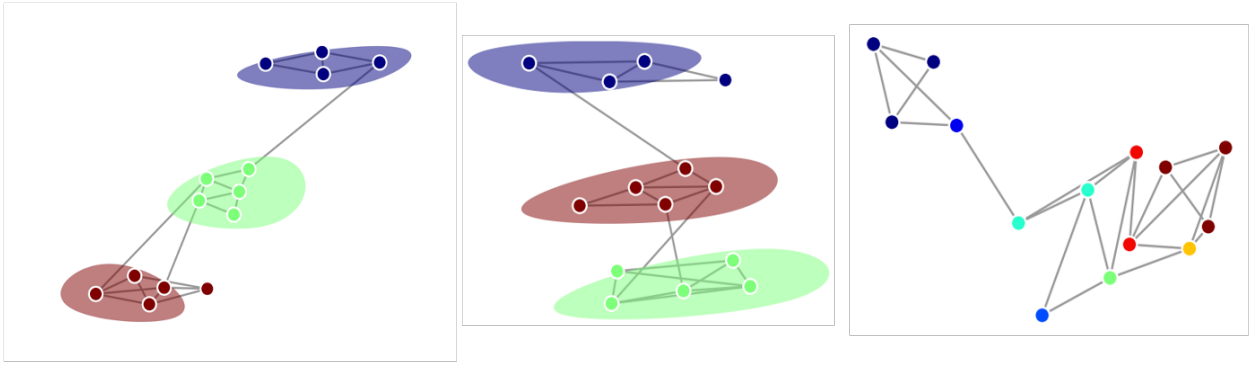
2.4 实验结论

在三种图数据集、6 种社区检测技术上进行实验，实验结果如下表：

表 2: 实验结果

名称	Louvain	g_n	spectral	hierarchical	Node2walk+kmeans	GCN-kmeans
G	1	1	1	1	1	1
Zachary	77.94%	77.94%	94.12%	54.41%	61.76%	97.06%
Football	89.57%	87.83%	80.48%	55.65%	91.74%	91.30%

根据实验结果绘制可视化图像，结合上表下图，可得出以下结论：1. 据可视化结果，本次新定义的相似度指标能较好反映两社区划分之间的差异性 2. 社团结构明显、数据集结构简单的图数据，比如图 G 在各模型上的表现都比较好，然而现实中的图数据却达不到该效果。3. 传统的社区发现聚类算法在复杂图数据集上的效果表现不如在简单图数据集上的表现。随着数据集复杂性的提高，图神经网络的优势逐渐凸显；但复杂数据集在神经网络上或是图表示学习上的效果是否一定强于传统社区发现算法，有待进一步的实验和理论上的探索。



(Louvain& g_n) **(spectral_clustering)** **(hierarchical&b_k)**

图 8: 图 G 在四种算法上的划分结果

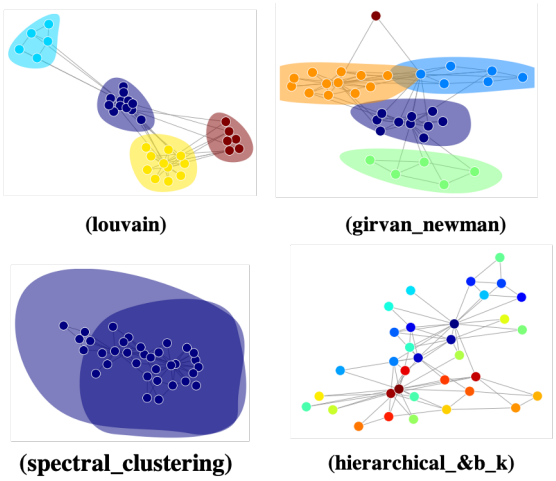


图 9: Zachary 数据集在四种算法上的划分结果

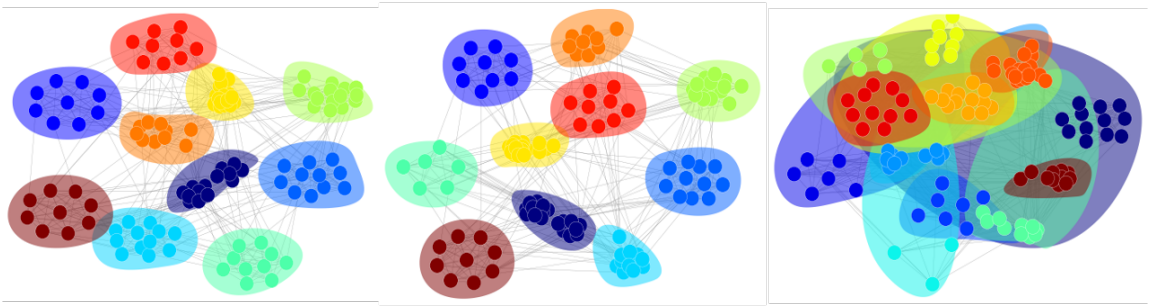


图 10: football 数据集在 louvain/g-n/spectrual 聚类算法上的划分结果

3 社区发现的挑战

纵观已有的社区监测技术，结合现实中的复杂的数据集，社区发现算法仍存在较大的挖掘空间，总结为以下几点：

(1) 社区数未知

长久以来，由于社区数未知而引发的挑战始终没有得到很好的解决。在机器学习领域中，社区发现经常被表示为一种无监督聚类任务。总现实世界的网络中提取出的研究数据大多是没有标签的。因此，我们很难获取有关社区数的先验知识。此外，大多数现有的深度学习社区发现方法（尤其是深度图嵌入），通过评估潜在特征空间中的节点相似度获取分类节点。然而，在后续的聚类算法中，聚类的目标数量仍然需要被事先定义。

(2) 网络异质性

网络的异质性指的是网络中实体类型的显著差异，而各种各样的节点集合和它们之间复杂的联系形成了异质网络。因此，我们应该通过不同于同质网络的方式研究异质网络中的社区发现。在应用和研发深度学习模型和算法时，应该解决异质网络实体上的概率分布的差异。

(3) 社区嵌入

社区嵌入是一个新兴的研究领域，这种方法将对社区而不是每个独立的节点进行嵌入。社区嵌入重点关注对社区进行感知的高阶近似而不是在节点邻居之间的 1 阶或 2 阶近似。未

来，社区嵌入研究面临的挑战有：（1）高昂的计算开销；（2）节点和社区结构之间的关系评估；（3）应用深度学习模型时发生的其它问题，例如社区之间的分部漂移。

(4) 大规模网络

大规模网络指的是拥有数以百万计的节点和边、大规模结构化模式以及高度动态性的大型网络。因此，大规模网络有其固有的规模特性（例如，社交网络中与规模无关的特性，节点度的米率分布特性），这些特性会影响社区发现任务中的聚类系数。此外，通过分解后的有关高维邻接关系的近似度度量，研究人员将分布式计算应用于可扩展的学习，同时他们也面临着鲁棒的学习控制和协作计算的问题。不断变化的网络拓扑进一步增加了近似度估计的难度。总而言之，大规模网络中的社区发现设计上述所有提到的挑战，以及可扩展学习方面的挑战。

(5) 节点特征的使用

已有的社区检测算法倾向于通过网络拓扑结构学习节点的表征，而往往节点的属性对节点的划分也有着重要影响，如何将节点属性特征与表征融洽地结合从而更全面地表示节点也是未来社区发现研究领域中的一个重要课题。

4 图算法的开源项目

4.1 communities

communities 是一个用于检测社区结构的 python 库，包括 Louvain method、Girvan-Newman algorithm、Hierarchical clustering、Spectral clustering、Bron-Kerbosch algorithm 等聚类算法；也可以使用 communities 可视化这些算法的实验结果。Github 链接

4.2 CogDL

CogDL 是将数据处理-图表示-预测自动化起来的强大的 python 库，CogDL 最特别的一点在于以任务 (task) 为导向来集成所有的算法，将每一个算法分配在一个或多个任务下，从而构建了“数据处理-模型搭建-模型训练和验证”一条龙的实现。CogDL 将已有的图表示学习算法主要划分为以下任务：

有监督节点分类任务 (node classification): 包括 GCN，GAT，GraphSAGE，MixHop，GRAND 等。

无监督节点分类任务 (unsupervised node classification)：包括 DGI，GraphSAGE(无监督实现)，以及 Deepwalk，Node2vec，ProNE 等。

有监督图分类任务 (graph classification)：包括 GIN，DiffPool，SortPool 等

无监督图分类任务 (unsupervised graph classification)：包括 InfoGraph, DGK，Graph2Vec 等。

链接预测任务 (link prediction)：包括 RGCN，CompGCN，GATNE 等。

异构节点分类 (multiplex node classification)：包括 GTN，HAN，Metapath2vec 等

Cogdl 跟进 SOTA 跟进最新发布的算法，包含不同任务下 SOTA 的实现，同时建立了不同任务下所有模型的 leaderboard（排行榜），研究人员和开发人员可以通过 leaderboard 比较不同算法的效果。调用方式如下图：

1. 命令行直接运行。通过命令行可以直接指定"task"、"model"、"dataset"以及对应的超参数，并且支持同时指定多个模型和多个数据集，**更方便。**

```
# 监督GraphSAGE
python scripts/train.py --task node_classification --dataset pubmed --model gra
# 无监督GraphSAGE
python scripts/train.py --task unsupervised_node_classification --dataset pubme
# DeepWalk + Node2Vec算法 + BlogCatalog + Wikipedia数据集
python script/train.py --task unsupervised_node_classificatoin --dataset blogca
```

2. 通过API调用。在代码中调用CogDL的数据、模型、任务构建API，方便使用自定义数据集和模型，**更灵活。**

```
# 获取模型/数据/训练的参数
args = get_default_args()
args.task = 'node_classification'
args.dataset = 'cora'
args.model = 'gcn'
# 建立数据集
dataset = build_dataset(args)
args.num_features = dataset.num_features
args.num_classes = dataset.num_classes
args.num_layers = 2
# 建立模型
model = build_model(args)
# 训练+验证
task = build_task(args, dataset=dataset, model=model)
ret = task.train()
```

图 11: 调用方法

参考文献

[1] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J/OL]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008. https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008. DOI: 10.1088/1742-5468/2008/10/p10008.

[2] CHEN M, WEI Z, HUANG Z, et al. Simple and Deep Graph Convolutional Networks[J]. ArXiv e-prints, 2020, arXiv:2007.02133: arXiv:2007.02133. arXiv: 2007.02133 [cs.LG].

[3] GROVER A, LESKOVEC J. Node2vec: Scalable Feature Learning for Networks[J/OL]. CoRR, 2016, abs/1607.00653. arXiv: 1607.00653. http://arxiv.org/abs/1607.00653.

[4] HAMILTON W L, YING R, LESKOVEC J. Inductive Representation Learning on Large Graphs[J/OL]. CoRR, 2017, abs/1706.02216. arXiv: 1706.02216. http://arxiv.org/abs/1706.02216.

[5] SCHLICHTKRULL M, KIPF T, BLOEM P, et al. Modeling Relational Data with Graph Convolutional Networks[J]. ArXiv, 2018, abs/1703.06103.