基于节点影响力和局部中心度的社区发现方法

朱 叶[1] 章 敏[2] 吴 璇[1]
ZHU Ye ZHANG Min WU Xuan

摘 要 一方面，社区信息沿着最短路径传播且随着传播逐渐衰减，信息传到较远位置可能性很小。另一方面，在信息量一定的情况下，在不同路径长度下，每条边累积信息量不同。由此两方面的考虑，引入节点影响力和局部中心度，结合 GN 算法删除最大边介数的核心思想，得到一种新的社区发现算法 WLCD（weighted local community detection，WLCD）。实验证明，在三种真实网络数据集中，WLCD 算法对比其他几种经典社区检测算法更好，在模块度、调整兰德系数、标准互信息以及准确率等评价指标方面都有比较好的结果。

关键词 节点影响力；局部中心度；社区；社区发现；复杂网络

doi：10.3969/j.issn.1672-9528.2024.02.010

0 引言

复杂网络与现实生活密切相关。如果把互联网看成是一个复杂网络，那么在互联网中，每个节点都代表一个实体，可以是个人、网站、服务器等，而节点之间的连线则代表这些实体之间的联系和交互。此外，复杂网络还可以用来描述和分析现实生活中的各种系统。社区这一概念可以在很多不同的环境下进行特殊场景的研究，如电力网络、社交网络、遗传网络、运输网络、书籍引用网络、足球队网络等 [1-3]，而社区发现的目的是揭示社区这一结构在复杂网络中的存在。通过对比研究，可以察觉到复杂网络中节点和边的丰富和不一致性。与此同时，不同节点和边在复杂网络中的位置不同，其节点属性和边权也不尽相同。因此，应该从复杂网络的本质出发，灵活地根据社区发现应用场景的不同进行算法改进。

社区发现算法具有广泛的应用前景和优势，可以帮助人们更好地理解和分析各种网络结构和行为，提高网络的性能和安全性，促进社会的稳定和发展。在个性化服务方面，可以用于构建高效的推荐系统，根据用户的兴趣、行为和位置等信息，为其推荐个性化的服务和产品。在网络安全方面，可以帮助人们更好地识别和防范网络攻击，提高网络的安全性和稳定性。在城市规划方面，可以帮助人们更好地了解城市的空间结构和人口分布，从而为城市规划提供科学依据。在公共卫生方面，可以帮助人们更好地了解和预测疾病的传播趋势和影响，从而制定更加有效的防控措施。在社交分析方面，可以帮助人们更好地理解社交网络的结构和行为，从而更好地分析和预测社交趋势和事件。

总之，社区发现在不同研究领域都发挥着重要的作用，不仅有助于理解和分析各种网络结构和功能，还可以帮助预测和控制各种网络行为和过程，具有研究意义。

1 复杂网络概述

1.1 基础理论

新时代是复杂性和网络性并存的时代，即复杂网络时代。在现实的背景之中，个体与环境都是息息相关、相互作用的，把个体当成是节点，节点间的关联看成是联系，即组成网络结构。而网络通常是错综复杂的，这也是网络最基本的性质。

复杂网络领域是当前研究热点之一，其应用与生活息息相关。本文重点研究在无向不加权静态网络中的改进边介数中心度计算，结合社区发现 GN 算法进行社区检测。下面分别从网络机制模型、介数中心度及社区发现算法三个方面进行介绍。

目前，规则网络、随机网络、小世界网络和无标度网络等是较为完善的几种模型。中心度反映复杂网络中各个节点相对重要性，在复杂网络分析中，对某个节点中心度的表征主要包括度中心度、介数中心度、接近度中心度和特征向量中心度。目前，划分社区发现算法的依据有很多，如根据是否随时间推移产生变化分为静态和动态算法，根据是否能检测重叠社区分为重叠和非重叠算法，本文涉及的主要是静态非重叠社区检测算法。

1. 江西科技学院信息工程学院 江西南昌 330098
2. 九江职业技术学院信息工程学院 江西九江 332099
[基金项目] 江西科技学院自然科学技术项目（23ZRYB05）；江西省教育厅科学技术研究项目（GJJ2202622）

1.2 评价指标

定义 1：模块度 Q（modularity）[4]。

$$Q=\sum_i(e_{ii}-a_i^2) \tag{1}$$

式中：$\sum_i e_{ii}$ 和 $\sum_j e_{ij}$ 分别指的是相同社区的节点边数与总边数的比值以及社区 i 中节点相连的边数与总边数的比值。模块度越大说明结果越好，一般取值在 0.3 ～ 0.7 之间。

定义 2：标准化互信息（normalized mutual information，NMI）[5]。

$$NMI(X;Y)=2\frac{I(X;Y)}{H(X)+H(Y)} \tag{2}$$

$$I(X;Y)=\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

$$H(X)=\sum_i h(P(X=X_i)) \tag{4}$$

$$h(x)=-x\cdot\log(x) \tag{5}$$

式中：两个随机变量 $(X;Y)$ 的联合分布为 $p(x,y)$，边缘分布分别为 $p(x)$、$p(y)$。NMI 值越大表示社区划分效果越好。

定义 3：调整兰德系数（adjusted rand index，ARI）[6]。

$$ARI=\frac{RI-E(RI)}{\max(RI)-E(RI)} \tag{6}$$

$$RI=\frac{a+b}{C_m^2} \tag{7}$$

式中：a 表示节点真实情况和预测情况同为错的个数，b 表示节点真实情况和预测情况同为对的个数，m 代表节点个数。ARI 值越大表示社区划分结果与真实情况越吻合。

2 基于节点影响力和局部中心度的社区发现算法 WLCD（weighted local community detection）

2.1 WLCD 算法基本思想

GN 算法初始时默认所有节点同属于一个社区，计算并找出中心度最大的边，从中移除且记录此时连通分量，若有所增加，则计算并记录此时模块度的值，直到所有边被移除才能结束算法，否则将重复进行上述步骤。算法结束之后，可以得到模块度最大时的结果。

GN 算法核心是中心度计算，需要对整个网络中所有点对之间最短路径进行广度优先遍历计算，然后沿着最短路径累积计算每条边的中心度。一方面，信息会沿着最短路径传播，且随着传播会逐渐衰减，则信息传到较远位置可能性很小，因此在计算中心度时，将距离较远的点对也纳入计算可能会引入误差。另一方面，在信息量一定的情况下，不同路径长度下每条边累积信息量不同。由此两方面的考虑，引入长路径对中心度重要性度量的干扰以及不同路径长度下对中心度重要性影响。因为这两方面的原因，本文引入 K 步范围内的局部中心度和节点影响力，提出一种基于节点影响力和局部中心度的社区发现算法 WLCD。首先，将所有节点放在同一社区，对所有边的加权局部中心度进行计算。然后，删除加权局部中心度最大的那条边，直到所有边被移除才能结束算法，否则将重复进行上述步骤。最后，算法结束即可得到模块度最大时的结果作为算法划分结果。

2.2 基本概念

定义 4：最短路径（short path）。在复杂网络图 G 中，任意两个节点 v_0 和 v_k 点对之间有最短路径个数为：

$$\sigma_{v_0v_k}=\prod_{0\le i\le k}\sigma_{v_iv_{i+1}} \tag{8}$$

定义 5：局部中心度（local centrality）。在图 G 中，K 是常量，表示公式为：

$$C_B^K(e)=\sum_{s,t\in V,e\in E,d_G(s,t)\le k}\frac{\sigma_{st}(e)}{\sigma_{st}} \tag{9}$$

定义 6：网络直径（network diameter）。网络中所有点对间的最长距离则为网络直径，可表示为：

$$D=\max_{i,j}d_{ij} \tag{10}$$

定义 7：平均路径长度（average path length）。任意两点之间距离的平均值是平均路径长度，可表示为：

$$L=\frac{1}{\frac{1}{2}n(n+1)}\sum_{i\ge j}d_{ij} \tag{11}$$

3 多种算法的对比分析

本文算法所采用的是空手道俱乐部网络、书籍销售网络、足球队网络等真实网络。实验涉及网络的基本信息如表 1 所示。

表 1 实验涉及网络的基本信息表

Data set	Nodes	Edges	Network diameter	Average path length	Communitys
Karate	34	78	5	2.4	2
Polbooks	105	441	7	3.079	3
Football	115	613	4	2.508	12

三种真实网络关于路径距离的 CDF 函数情况如图 1 所示。

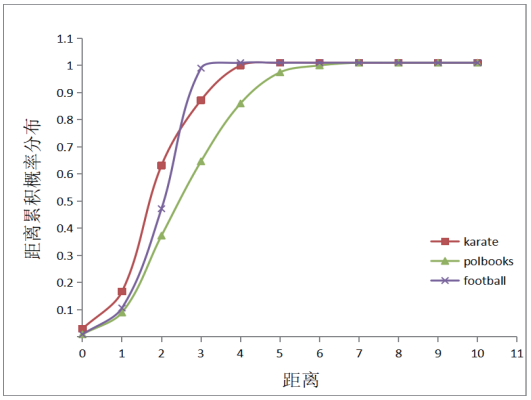

图 1 距离—CDF 函数情况图

图 1 表示的是距离—CDF 函数情况，易知，空手道俱乐部网络、书籍销售网络和足球队网络分别在 3、5 和 3 步内可到达网络中 90% 节点，但是足球队网络由于当步长为 3 时计

算的是近似全局中心度，这与实验目的不一致，因此取 2 更为合理。

图 2 表明现实网络在不同 K 值下的模块度大小的变化趋势，当 K 值超出点对之间最长距离时，计算的是全局边介数，因此 K 取值小于网络直径即可。考虑到效率因素，K 越小所需计算量越少。Karate 网络在 K 为 3 之后时模块度最大且不变，取其中最小 K 值，故 K=3 最优。Polbooks 网络在 K 为 4 时模块度最大，故 K=4 最优。Football 网络在 K 为 2 之后模块度最大且不变，故 K=2 最优。

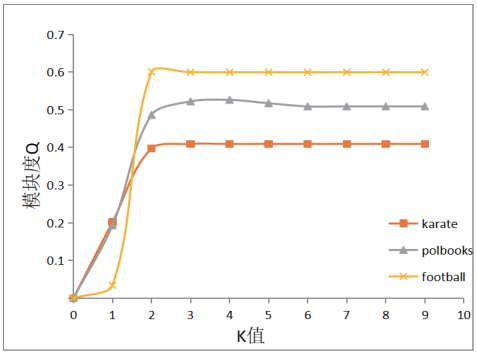

图 2 真实网络 K 值与模块度 Q 变化趋势图

空手道俱乐部网络这一数据集在本文的 WLCD 算法中被划分成 5 个社区，分别是社区 1 包括节点 0、1、2 、3、7、12、13、17、19、21，社区 2 包括节点 4、5、6、10、16，社区 3 包括节点 8、14、15、18、20、22、23、24、25、26、27、28、29、30、31、32、33，剩下的节点 9 和节点 11 都分别单独为一个社区，算法结果如图 3 所示。

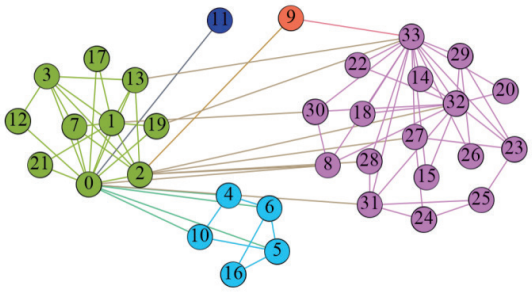

图 3 空手道俱乐部网络算法结果

各算法在空手道俱乐部网络上结果进行对比，结果如图 4 所示。

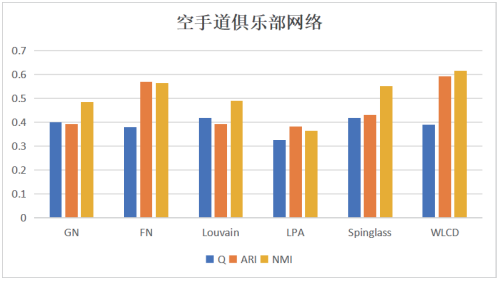

图 4 各算法在空手道俱乐部网络上的结果对比图

本文采用 Q、ARI 和 NMI 等标准进行算法结果的对比，空手道俱乐部网络在真实情况下是存在着两个社区，而 FN 和 LPA 算法的划分社区结果数目是 3，与真实情况最为接近。但是这两种算法在任何一个标准方面上的结果都不如本文算法的结果好。就模块度而言，Louvain 算法的结果最佳，本文算法模块度结果 0.391 1 与其结果相差甚小，而且 Louvain 算法在另外两个标准上的结果明显不如本文算法的结果。就 ARI 和 NMI 两个标准而言，本文算法结果是最优。因此，综合而言，相比于其他算法，本文算法整体上效果更好。

书籍销售网络这一数据集在本文的 WLCD 算法中被划分成 4 个社区，分别是社区 1 包括节点 0、1、2、4、5、6、7、29，社区 2 包括节点 3、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、53、54、55、56、57，社区 3 包括节点 51、52、58、64、65、67、68、69、103、104，剩下的节点为一个社区，算法结果如图 5 所示。

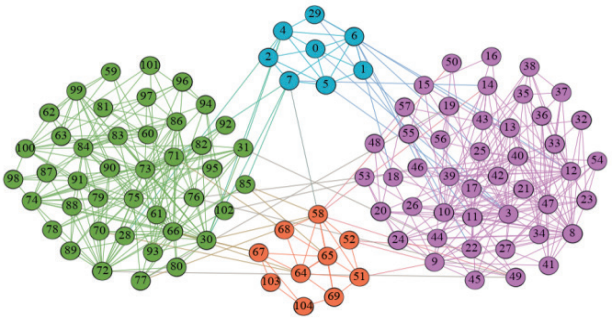

图 5 书籍销售网络算法结果

各算法在书籍销售网络上的结果进行对比，结果如图 6 所示。

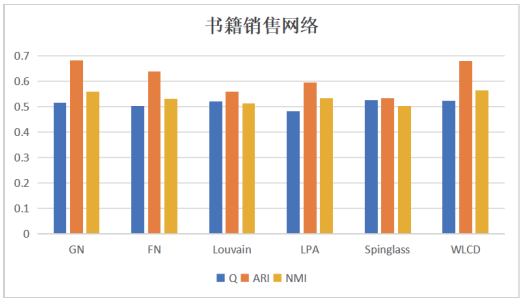

图 6 各算法在书籍销售网络上的结果对比图

本文采用 Q、ARI 和 NMI 等标准进行算法结果的对比，书籍销售网络在真实情况下是存在着三个社区，而 FN、LPA、Louvain 以及本文算法的划分社区，本文算法的划分社区结果数目是 4，与真实情况最为接近。但是这三种算法在任何一个标准方面上的结果都不如本文算法的结果好。从模块度这一个角度看，Spinglass 算法的结果最佳，但是本文算法的结果是 0.522 2，与其结果相差甚小，而且在另外两个标

准上的结果明显不如本文算法的结果。就 ARI 而言，GN 算法结果是最优，本文算法结果 0.680 2，与其结果相差甚小。就 NMI 而言，本文算法结果最优。因此，综合而言，相比于其他算法，本文算法整体上效果更好。

足球队网络这一数据集在本文的 WLCD 算法中被划分成 12 个社区，分别是社区 1 包括节点 0、4、9、16、23、41、93、104，社区 2 包括节点 1、25、33、37、45、89、103、105、109，社区 3 包括节点 2、6、13、15、32、39、47、60、64、100、106，社区 4 包括节点 3、5、10、40、52、72、74、81、84、98、102、107，社区 5 包括节点 7、8、21、22、51、68、77、78、108、111，社区 6 包括节点 11、24、28、50、69、90，社区 7 包括节点 12、14、18、26、31、34、38、42、43、54、61、71、85、99，社区 8 包括节点 17、20、27、56、62、65、70、76、87、95、96、113，社区 9 包括节点 19、29、30、35、55、79、80、82、94、101，社区 10 包括节点 19、29、30、35、55、79、80、82、94、101，社区 11 包括节点 44、48、57、66、75、86、91、92、112，剩下的节点为一个社区，算法结果如图 7 所示。

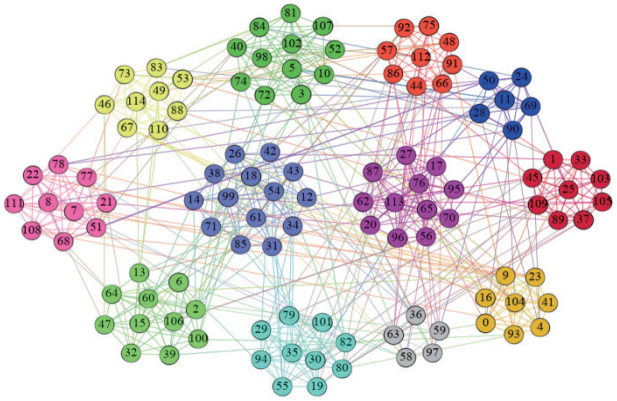

图 7 足球队网络算法结果

本文 WLCD 算法与其他算法在书籍销售网络上的结果进行对比，对比结果如图 8 所示。

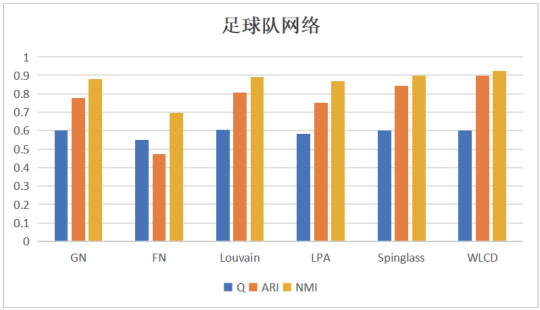

图 8 各算法在足球队网络上的结果对比图

本文采用 Q、ARI 和 NMI 等标准进行算法结果的对比，书籍销售网络在真实情况下是存在着十二个社区，与本文算法结果一致。就模块度而言，Louvain 算法的结果最佳，但是本文算法模块度结果 0.600 5 与其结果相差甚小，而且在另外两个标准上的结果明显不如本文算法的结果。就 ARI 和 NMI 两个标准而言，本文算法结果是最优。因此，综合而言，相比于其他算法，本文算法整体上效果更好。

4 结语

一方面，现实网络一般具备小世界这一性质，社区信息会沿着最短路径传播，且随着传播会逐渐衰减，则信息传到较远位置可能性很小。另一方面，在信息量一定的情况下，在不同路径长度下，每条边累积信息量不同。由此两方面的考虑，引入节点影响力和局部中心度，结合 GN 算法删除最大边介数的核心思想，得到一种新的社区发现算法 WLCD（weighted local community detection，WLCD）。实验证明，在三种真实网络数据集中，WLCD 算法对比其他几种算法更好，在 Q、ARI、NMI 等方面的综合结果更好，进一步分析验证了本文算法。

参考文献：

[1] LESKOVEC J, LANG K J, DASGUPTA A, et al. Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters[J]. Internet mathematics, 2008, 6(1):29-123.

[2] THOMAS R, MURAT T, BENOIT C, et al. Proteome-scale map of the human interactome network[J].Cell, 2014, 159: 1212-1226.

[3] BO Y, LIU J, FENG J. On the spectral characterization and scalable mining of network communities[J]. IEEE transactions on knowledge & data engineering, 2012, 24(2):326-337.

[4] NEWMAN M E . Fast algorithm for detecting community structure in networks[J]. Phys rev e stat nonlin soft matter phys, 2004, 69(2):66-103.

[5] VINH N X, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance[J]. Journal of machine learning research, 2010,11(10):2837-2854.

[6] DANON L, DUCH J, DIAZ-GUILERA A , et al. Comparing community structure identification[J]. Journal of statal mechanics, 2005, 2005(9):9008.

【作者简介】

朱叶（1996—），女，江西南昌人，硕士研究生，研究方向：复杂网络、社区发现等。

章敏（1996—），江西南昌人，硕士研究生，研究方向：机器学习、自然语言处理等。

吴璇（1996—），女，江西鹰潭人，硕士研究生，研究方向：算法优化、传感器网络等。

（收稿日期：2023-10-22）