基于局部信息检测的多粒度社团挖掘方法

朱 涛 常国岑 郭戎潇 李项军
(空军工程大学电讯工程学院 西安 710077)

摘 要 从复杂性和动态性特征出发,给出了复杂网络局部模块度的定义,并提出了基于局部信息检测的社团发现算法,认为局部模块度值最大的节点集合就是最理想的社团结构。在此基础上提出了多粒度社团挖掘方法,为多视图观察复杂网络结构特征提供了新的研究思路。最后的实验分析表明了方法的有效性和可行性。

关键词 复杂网络,社团挖掘,局部信息检测,多粒度

中图法分类号 TP393 **文献标识码** A

Method of Multi-granularity Community Structure Mining Based on Local Information Detection

ZHU Tao CHANG Guo-cen GUO Rong-xiao LI Xiang-jun
(Telecommunication Engineering Institute, Air Force Engineering University, Xi'an 710077, China)

Abstract Taking consideration of complexity and dynamic of complex networks, a definition of local modularity was proposed, and an algorithm for communication structure mining based on local information detection was given, with the criterion, i.e. the best community is the node group whose local modularity is the largest. Then a method of multi-granularity community structure mining was proposed, which provides new ideas to observe structure characters of complex networks from various angles. Final experiments verify its efficiency and feasibility.

Keywords Complex networks, Community structure mining, Local information detection, Multi-granularity

复杂网络是复杂系统的结构抽象。网络中的节点是系统中的个体,网络中的连边是个体之间的关系。现实世界中包含着各种各样的复杂网络,如国际互联网、科学家合作网、蛋白质折叠网等[1],随着对其网络性质的物理意义和数学特性的深入研究,人们发现这些网络大多具有一个共同的特征"社团性"[2],即整个网络是由若干个节点的集合"社团"组成,其中社团内部的节点之间连接相对紧密,社团之间的连接却比较稀疏。研究网络社团结构的数据挖掘,对于理解网络的结构和功能具有深刻的理论价值和现实意义,为化简复杂网络的拓扑结构、把握复杂关系的本质内涵提供了方法依据。

复杂网络社团结构挖掘的主要思想是利用网络思维,从计算机科学和人工智能的角度,将原始数据之间的相互关系抽象成网络拓扑的形式,结合复杂网络的普遍性质,在多领域、多视图和多粒度下检测网络拓扑的主体倾向,发现组织结构中的社团关系。因此,本文从复杂性和局域性特征出发,提出了一种基于局部信息检测的多粒度社团挖掘方法,以化繁为简的研究思路对网络系统进行多视图下的数据挖掘,进而检测网络拓扑的主体倾向,发现骨干社团的内在联系,以期为不同需求下的观察者提供直观的知识参考。

1 基于局部信息检测的社团发现

目前,社团挖掘方法可以划分为计算机科学中的图形分割法和社会学中的分级聚类法两大类。前者主要采用基于2选1的策略,即在整个网络中寻找最优的两个子网络,然后对两个子网络进行同样处理,反复进行,直到得到足够数目的子图,代表性的有Kernighan-Liu算法[3]和谱分解法[4];后者主要基于同类相近原则,即根据节点之间的相似程度把网络自然划分成组,代表性的有GN算法[2]和Newman快速算法[5]。这些算法的共同点都是从全局性出发挖掘社团结构,因此一个重要的前提条件就是需要知道整个网络的拓扑结构。但这个限制条件对于很多超大型且不断动态变化的网络来说是相当困难的。因此,如何找到一个不依赖于全局信息的方法来划分网络的社团结构具有十分重要的现实意义。

首先,参考Newman提出的模块性[6]概念和Clauset提出的局域社团探测[7]思想,引入一个独立于网络全局信息的"局部模块度"指标来指导网络社团发现,其定义如下:

$$Q=\frac{L_{in}}{L_{in}+L_{out}} \tag{1}$$

其中,L_{in}表示所求社团内的边数,L_{out}表示所求社团与其他社团之间的边数。一般来说,社团内的边数要比社团间的边数多,即$L_{in}>L_{out}$,并且Q的值越大,所求的社团性越好。从定义可知,Q与网络的全局信息无关,仅需了解局部连接关系,因此可以避免全局模块度局限性带来的划分弊端。

在定义了独立于网络全局信息的评价指标之后,社团发

到稿日期:2008-10-14 返修日期:2008-12-18 本文受军队科研基金资助项目(编号:KJ06104)资助。

朱 涛(1982—),男,博士生,主要研究方向为指挥信息系统建模仿真,E-mail:peter_ww99@yahoo.com.cn;常国岑(1945—),男,博士生导师,主要研究方向为指挥信息系统理论与技术;郭戎潇(1981—),女,博士生,主要研究方向为网络信息安全与管理;李项军(1982—),男,博士生,主要研究方向为信息优势相关理论。

现方法也必须是独立于全局网络拓扑信息的，因此需要事先进行一系列的假设：对于网络 G 中已知的节点及其连接关系组成的子图，定义为 C；与之相对的是 G 中的未知区域，仅知道其与 C 的部分连接，定义为 U。如果想要得到更多关于 G 的网络信息，只能通过访问在 U 中的节点来实现，即通过类似“网络蜘蛛”的搜索遍历方式。为此，本文采用一种局部最优化的贪婪凝聚算法，即认为局部模块度值最大的节点集合就是最理想的社团结构。整个社团发现算法可以分解为两个子算法：算法 1 是发现目标节点所属的社团结构，算法 2 是分析网络的全局社团结构。

算法 1 流程如下。

Step1　初始化。将目标节点放到集合 C 中，C 表示目标社团。其余节点全部放到集合 U 中，U 表示网络未知区域。另定义集合 V，表示与 C 中节点相邻的网络区域。初始时刻设 V 为空，C 的局部模块度值 $Q=0$。

Step2　更新 V。在 U 中查找所有与 C 中节点相邻的节点，并放入 V 中。若此时网络中所有节点均被遍历添加至 V，则将 V 中所有节点都加入到 C 中，并转到 Step5。

Step3　更新 C。对于 V 中的每一个节点 V_i，首先假设将其加入 C，并计算此时的局部模块度值 Q_i，然后从计算得到的 Q 值集合中找到使 C 的局部模块度值最大的 Q_{max}，如果 $Q_{max}>Q$，更新 $Q=Q_{max}$，并将此节点加入到 C 中，否则转到 Step5。

Step4　循环查找。重复 Step2 和 Step3 的工作；

Step5　退出，已发现目标社团。

此时，就可以得到两个子网络 C 和 U：C 是为目标节点查找的社团，U 是网络剩余未知区域。

算法 2 是在算法 1 的基础上，在 U 中继续指定目标节点，重复算法 1 的工作，直到所有节点都已划分到目标社团中。这样就可以得到整个网络的社团划分，整个过程可以用图 1 所示的一个二叉树来表示。

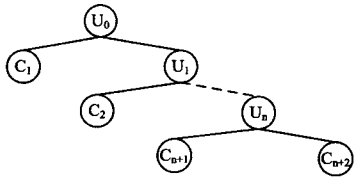

图 1　算法 2 模拟二叉树

由以上分析可以看出，算法 1 可以为任意指定的节点查找其所属的网络社团。算法 2 在此基础之上可以分析整个网络的社团结构，且自动收敛于一个比较好的划分。最重要的是，整个算法利用局部信息执行社团发现，因而避免了基于全局信息算法的先天缺陷。

最后，分析算法的时间复杂度。算法 1 花费的时间代价主要是在获取社团 C 的邻居集合 V 和计算更新社团 C 上。前者的时间复杂度为 $O(N_cK_c/2)$，后者的时间复杂度为 $O(N_c^2K_c/4)$，其中 N_c 表示目标社团的节点个数，K_c 表示目标社团的节点平均度，因此算法 1 的时间复杂度为 $O(N_c^2K_c)$。由此可以很容易地得到算法 2 的时间复杂度 $O(N^2K)$，其中 N 表示网络的节点个数，K 是网络的节点平均度。整个算法在时间复杂度上与极值优化等优良算法[5]相当，适用于大型网络的社团发现，且对于稀疏网络挖掘效率更高，因为此时算法时间复杂度为 $O(N^2)$。

2　多粒度社团挖掘方法

借助有效的社团发现算法就可以得到整个网络的社团结构信息。但当网络节点数量很大时（通常超过 1000），其挖掘得到的社团数目也很大，呈现出的网络视图依然复杂，不易于直观认识和分析。因此提出多粒度社团挖掘方法，即通过将已知社团收缩成点组成新的网络拓扑，继续重复挖掘工作，寻找社团之上的社团，直到符合分析要求为止，其原理流程如图 2 所示。

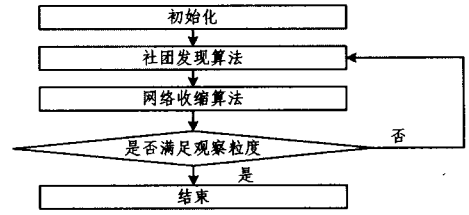

图 2　多粒度社团挖掘方法

显而易见，每次社团收缩后形成的新的网络拓扑就可以看成是该粒度下的网络直观图。其中还需要讨论社团的收缩规则。在本文中，为简化问题讨论，规定每个社团收缩成一个节点；社团间如果存在连边，则认为收缩后节点间有边相连，否则无边。

具体执行流程如下。

Step1　执行社团发现算法，获得网络社团结构信息；

Step2　将网络中的社团收缩成节点，社团之间的联系抽象成边，形成新的网络拓扑图；

Step3　判断网络拓扑图是否满足粒度要求（如以网络节点数量为依据），如满足，转到 Step4，否则转到 Step1；

Step4　退出，已找到符合要求的网络视图。

3　实验分析

3.1　Zachary 网络

Zachary 网络是社会网络分析的一个经典问题。20 世纪 70 年代初，Wayne Zachary 用了两年的时间观察美国一所大学空手道俱乐部成员间的相互社会关系。基于这些成员在俱乐部内部及外部的社会关系，Wayne Zachary 构造了它们之间的关系网，如图 3 所示。事有凑巧，在他调查的过程中，该俱乐部的主管与校长之间因是否抬高俱乐部收费的问题产生了争执。结果，这个俱乐部分裂成了两个分别以校长和主管为核心的小俱乐部。Zachary 网络在复杂网络的社团结构分析中已经成为一个经典的问题，成为了衡量网络社团结构划分算法准确性的标准。

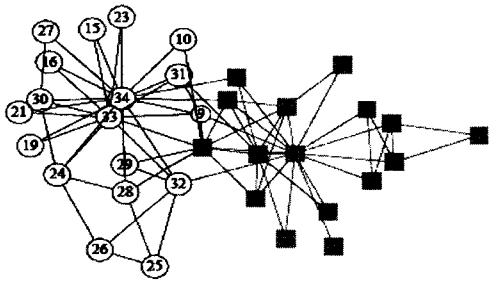

图 3　Zachary 网络拓扑图

首先给出由 Jordi Duch 和 Alex Arenas 提出的 EO 算

法[8]对 Zachary 网络划分的结果，如图 4 所示。这一结果的全局模块度值是 0.4188，大于 Newman 提出的算法给出的0.381、Reichardt 等提出的算法给出的 0.406 和 Donntti 等提出的算法给出的 0.412。之后，从 EO 算法所得到的 4 个社团中各取一个节点，然后执行算法 1，用得到的 4 个社团，再与原来的社团相比较，以检验算法的有效性。

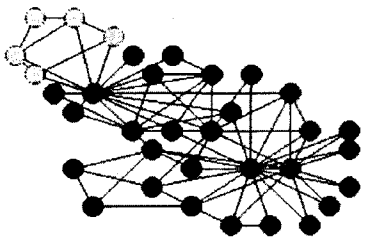

图 4　EO 算法社团划分结构

分别选取节点 1、节点 7、节点 27 和节点 28 作为初始节点执行算法 1，得到以下统计结果：从表 1 中可以看出，节点 7 和节点 28 发现的社团与 EO 算法的结果一致。而节点 1 和节点 27 发现的社团与 EO 算法的结果不一致，对照实际网络图可以看出，社团 1 除了节点 10 和节点 29 以外，社团 3 除了节点 3 以外，正是 Zachary 网络二划分的理想结果。究其原因，都是节点 3 在现实生活中的歧义性所致。但从网络拓扑角度看，对社团 1 和社团 3 的划分都是合理的。而且算法 1 在针对特定点进行社团发现时仅需考虑与节点相关的局部信息，尽可能为其寻找到一个好的社团，这样才可能得到局部最优值，因此表中的局部模块值结果优于 EO 算法，这也说明了该算法的有效性。

表 1　算法 1 与 EO 算法局部社团发现结果比较

网络社团	初始节点	目标社团节点	社团数目	局部模块值	EO 社团数目	EO 模块值
1	1	1-12-13-4-8-14-2-3-18-22-20-17-6-11-5-7-**10-29**	18	0.7727	12	0.6512
2	7	7-5-11-6-17	5	0.6	5	0.6
3	27	27-30-24-26-25-28-32-34-29-10-15-16-33-19-23-31-9-3	18	0.7872	14	0.6667
4	28	28-25-26-24-32-29	6	0.4375	6	0.4118

同时，也发现初始点的不同对划分结果有影响，即算法的稳定性不好。但经过统计发现，虽然初始点的选取会影响划分结果，但是所得结果仍然在可以接受的范围内。图 5 给出了选取不同初始节点对最终划分的影响程度。每个节点做了 50 次测试，取其平均值，所得到的全局模块度值在 0.3702～0.4621。

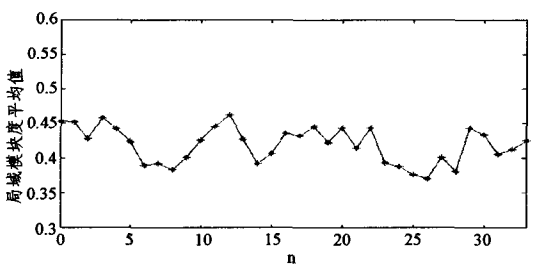

图 5　不同初始节点对社团划分的影响

3.2　计算机模拟网络

计算机模拟生成一系列有已知社团结构的网络，每个网络有 128 个节点，平均分布于 4 个已知的社团中。节点之间的边是随机添加的，同一个社团的两个节点之间有点相连的概率是 p_{in}，不同社团的两个节点之间有边相连的概率是 p_{out}。p_{in} 的选择保证节点度的期望值等于 16，p_{out} 的选择保证节点和其它社团节点之间边的总数为某个指定的值 z_{out}。随着 z_{out} 由 0 逐渐增大，社团结构变得越来越模糊，社团发现变得越来越困难。

由于网络中的社团结构是已知的，给出一个社团划分后，就可以度量有多少个节点是正确划分的。比较发现的社团和已知的社团，对于发现的社团 C_F，把它和各个已知社区 C_R 进行比较，找出重叠度最大的那个 C_R^*，C_F 和 C_R^* 重叠的节点认为是正确划分的节点。对发现的各个社团都做类似的处理，从而得到正确划分的节点总数占所有节点数的比率 FVIC (Fraction of Vertices Identified Correctly)。对于每一个 z_{out}，可以得到一个 FVIC，从而得到一个 FVIC 关于 z_{out} 的函数。

实验时，z_{out} 的取值由 0 到 8。对于 z_{out} 的任意一个取值，为了降低偶然性带来的影响，生成 100 个实验网络。对每个网络执行社团发现算法，发现其中的社区结构，计算 FVIC 值。然后求各自的平均值，得到对应这个 z_{out} 的平均 FVIC 值，以此检验为所有节点划分社团的平均正确性。统计结果如图 6 所示。

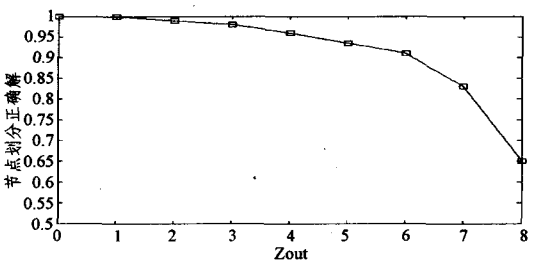

图 6　计算机模拟网络节点社团划分正确率

从图 6 中可以看出，当 $z_{out} \leqslant 6$ 时，能够正确划分 90%以上的节点；当 $z_{out}=7$ 时，也能保证有 80%以上的节点被正确划分；当 $z_{out}=8$ 时，每个节点与社团内部连边数和等于它和社团外部节点的连边数，此时网络的社团结构变得模糊和难以识别，社团划分的正确率明显下降。这说明，对于有已知社区结构的网络，基于局部信息检测的社团发现算法能很好地发现网络的社团结构。

3.3　Schematic 网络

Schematic 网络是一个典型的具有社团结构的网络，它包含 3 个社团，社团内的边很稠密，而社团间的边很稀疏。利用基于局部信息检测的多粒度社团挖掘方法可以得到清晰直观的网络视图，如图 7 所示。

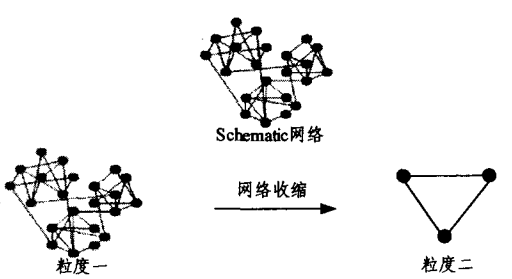

图 7　Schematic 网络多粒度社团挖掘效果

结束语　本文针对全局社团发现算法存在的局限性，给出了局部模块度定义及其局部信息检测算法，并在此基础上

提出了多粒度社团挖掘方法，力图透过错综复杂的交互关系认识系统直观明了的本质联系。通过在 Zachary 网络、计算机模拟网络和 Schematic 网络上的算法分析和方法实践，检验了方法的有效性和可行性。当然，本文只是对社团结构挖掘方法进行了初步的探索，还有很多具体的问题值得关注和研究：(1)如何充分利用全局和局部特征来提高社团挖掘的合理性；(2)如何定量评价社团特征对病毒传播、网络抗毁的影响程度；(3)如何在网络化数据的基础上进行知识发现和数据挖掘。

参 考 文 献

[1] Newman M E J. The Structure and Function of Complex Networks [J]. SIAM Review(S0036-1445),2003,45(2):167-256

[2] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proc. Natl. Acad. Sci. USA, 2002, 99: 7821-7826

[3] Kernighan W, Lin S. An efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49: 291-307

[4] Fiedler M. Algebraic connectivity of graphs [J]. Czech Math J, 1973,23:298-305

[5] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Phys. Rev. E,2004,69:066133

[6] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Phys. Rev. E,2004,69:026113

[7] Clauset A. Finding local community structure in networks [J]. Phys. Rev. E,2005,72:026132

[8] Duch J, Arenas A. Community detection in complex networks using extremal optimization [J]. Phys. Rev. E,2005,72:027104

(上接第 216 页)

跟踪是准确的，图 3 中网球纵坐标的变化情况正说明了网球两次起落的情况。

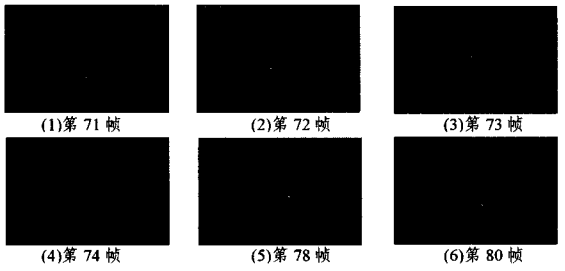

图 1 采用前帧加权采样的粒子滤波算法的网球跟踪结果

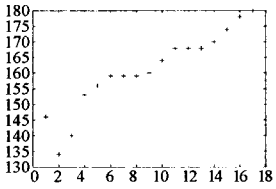

图 2 跟踪目标横坐标的变化情况

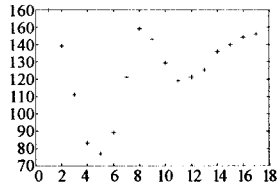

图 3 跟踪目标纵坐标的变化情况

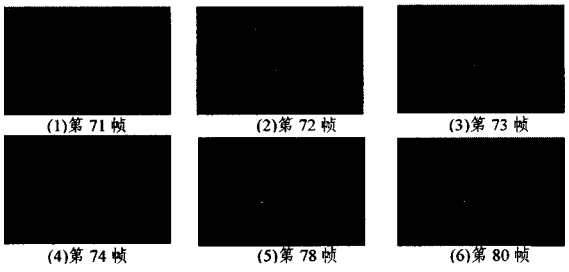

图 4 采用传统采样重要性重采样算法的网球跟踪结果

用传统采样重要性重采样算法的网球跟踪结果显示于图 4，在第 73 帧时似乎还可以勉强跟踪上，但第 74 帧的情况已经说明目标跟丢了，第 80 帧的时候可以明显看到粒子发散，跟踪过程失败。相应网球纵坐标的变化情况也说明跟踪是失败的，预测点不能反映实际的运动情况。

结束语 本文提出一种改进的 SIR 算法，根据相邻帧间信息的强关联性，引入前帧加权采样手段校正预测点的信息。解决了传统 SIR 算法由于引进提议分布而需要严重依赖目标的系统状态模型的问题，可以理想跟踪运动状态不规则的目标，而不用依靠预先建模。本文针对下落并弹跳着的绿色网球视频进行跟踪，在比较了传统 SIR 算法和本文提出的算法的试验基础上，可以明显看出本文提出的算法的跟踪性能优于传统算法，在实际应用中可以用于机器人视觉导航对不规则目标的跟踪。并且由于本文算法的普适性和可操作性，可以推广到粒子滤波在其它方面的类似应用，如机器人三维空间定位和故障诊断等方面。

参 考 文 献

[1] 王亮，胡卫明，谭铁牛. 人运动的视觉分析综述[J]. 计算机学报，2002,25(3):1-16

[2] Horn B. Optical flow[J]. Artificial Intelligence, 1981, 17: 185-203

[3] Nummiaro K, Koller-Meier E, Gool L J V. An adaptive color-based particle filter[J]. Image Vision Comput, 2003, 21(1): 99-110

[4] Vander Merwe R, Doucet A, de Freitas N, et al. The Unscented Particle Filte[R]. CUED/F-INPENG/TR 380. London: Cambridge University Engineering Department, 2000

[5] Gordan N, Salmond D, Smith A. A novel approach to nonlinear/nonGaussian Bayesian state estimation[J]. IEEE Proceedings on Radar and Signal Processing, 2005, 118: 187-113

[6] Doucet A , de Freitas N , Gordon N . Sequential Monte Carlo Methods in Practice[M]. New York: Springer-Verlag, 2001

[7] 汤思维，陈卫东，曹其新. 移动机器人多目标彩色视觉跟踪系统[J]. 机器人，2003,25(1)