

一种新的重叠社区发现算法^{*}

马兴福, 王 红

(山东师范大学 信息科学与工程学院, 济南 250014)

摘 要: 基于节点的中心度和节点对社区的适应度, 提出了一种新的重叠社区发现算法。该算法以中心度很大的节点作为初始社区, 然后访问社区的邻居节点, 把对社区适应度最大的节点加入到社区。如果节点对多个社区都具有很大的适应度, 则这些节点归属于多个社区。考虑到社区之间的重叠性, 将社区相似度很大的社区合并为一个社区。将该算法应用到 Zachary 空手道俱乐部网络和海豚社会网络中, 实验表明该算法能够很好地划分出网络中的重叠社区。

关键词: 复杂网络; 重叠社区; 社区发现; 中心度; 适应度

中图分类号: TP393; TP301.6

文献标志码: A

文章编号: 1001-3695(2012)03-0844-03

doi:10.3969/j.issn.1001-3695.2012.03.011

New algorithm for detecting overlapping communities

MA Xing-fu, WANG Hong

(School of Information Science & Engineering, Shandong Normal University, Jinan 250014, China)

Abstract: Based on the node of centrality degree and the node's fitness to community, this paper proposed a new algorithm to detect overlapping communities. This algorithm took the nodes with a great centrality degree as the initial communities, then visited community's neighbor nodes, and joined the node with a biggest fitness to the community. If the node had the very big fitness to many communities, then it belonged to several communities. Considering the overlapping property of community, it combined communities with great similarities into one community. It applied this algorithm into the Zachary's Karate Club Network and the Dolphins Social Network, and the experiment indicates this algorithm can divide overlapping communities efficiently in the network.

Key words: complex network; overlapping communities; community discovery; centrality degree; fitness

0 引言

复杂网络是复杂系统的抽象, 现实中许多复杂系统都可以用复杂网络的相关特性进行描述和分析。其中网络中的节点表示复杂系统中的个体, 而边则表示个体之间的关系^[1]。如社会关系网络、新陈代谢网络、科研合作网络、演员合作网络、万维网、城市交通网络、电力网等都可以作为复杂网络来研究。近几年来, 对复杂网络的研究一直是许多领域的研究热点, 也取得了许多重要进展。其中社区结构是复杂网络中的一个普遍特征, 整个网络是由许多个社区组成的, 同一社区内的节点与节点之间的连接很紧密, 而社区与社区之间的连接比较稀疏^[1]。如图 1 所示, 图中的小型网络中包含三个社区, 对应图中的三个椭圆, 在这些社区内部, 节点与节点之间的联系非常紧密, 而社区之间的连接则要稀疏很多。

在复杂网络的研究领域中, 社区发现是一个复杂而有意义的过程, 它对研究复杂网络的特性具有重要作用。近几年, 发现及分析复杂网络中的社区结构得到了许多学者的关注, 也提出了很多社区发现算法。其中两个较早的经典社区发现算法是基于 Laplace 矩阵的谱平分法和基于贪婪算法原理的 Kernighan-Lin 算法^[2]。近几年来研究人员提出了许多基于模块

度的社区发现算法, 其中 Newman 等人^[3]提出的 GN 算法是一种基于边介数的分裂算法, 该算法不断地删除网络中边介数最大的边, 直到每个节点成为一个社区为止, 选取对应模块度值最大的结果为最终社区划分结果。该算法复杂度较高, 为 $O(n^3)$ 。随后 Newman 等人提出了 Newman 快速算法, 该算法在初始化时, 定义每个节点为一个独立的社区, 然后每次循环都选择使模块度值增量最大的两个社区进行合并, 直到整个网络成为一个社区, 选择其中对应着模块度值最大的社区划分为最佳的网络社区结构^[4]。

上述算法都将每个节点严格地划分到某个社区, 每个节点属于且仅属于一个社区。但在现实网络中, 许多节点通常同时属于多个社区, 因此便形成了相互重叠的社区^[5], 如图 2 所示, 其中节点 6 和 7 归属于两个不同的社区。重叠社区结构日益受到学者的重视, 成为近年来研究的热点, 如何发现重叠社区受到更多学者的关注。其中首个重叠社区发现算法是 Palla 等人提出的 K-clique 派系过滤算法^[6], 它将相邻的 K-clique 进行合并, 而那些属于多个 K-clique 的节点便构成了网络中社区的重叠部分。

本文提出了一种新的基于节点中心度和节点对社区适应度的重叠社区发现算法。实验表明, 该算法可以有效地发现网

收稿日期: 2011-09-04; **修回日期:** 2011-10-28 **基金项目:** 国家自然科学基金资助项目(60970004, 60975081); 山东省研究生教育创新计划资助项目(SDY10059)

作者简介: 马兴福(1988-), 男, 山东临沂人, 硕士研究生, 主要研究方向为复杂网络(maxingfu0531@163.com); 王红(1966-), 女, 天津人, 教授, 博导, 博士(后), 主要研究方向为移动社会性软件、复杂网络、工作流。

络中的重叠社区。

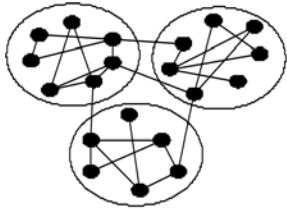


图1 具有社区结构的小型网络示意图

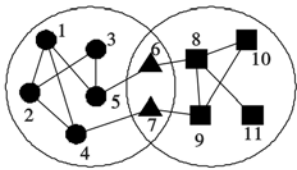


图2 具有重叠社区结构的网络示意图

1 重叠社区发现算法

1.1 相关概念

1) 节点中心度 $C(i)$

$$C(i) = \frac{1}{\sum_{j=1}^n d(v_i, v_j)}, j \neq i$$

其中: n 为网络中节点的数目; $d(v_i, v_j)$ 为节点 v_i 到 v_j 的最短路径长度, 如果节点中心度越大, 则节点越处于网络的中心位置。

2) 社区适应度 $f^{[7]}$

$$f = \frac{k_{in}}{k_{in} + k_{out}}$$

其中: k_{in} 为社区内部度, 它等于社区内部所有连边数的两倍; k_{out} 为社区外部度, 它等于社区内节点与社区外节点的连边数。

3) 最佳适应度 f_{max}

这里引入最佳适应度来判断当前所划分的社区结构是否为最佳状态。

4) 节点 j 对社区 c 的适应度 f_c^j

f_c^j 为节点 j 加入到社区 c 后, 社区 c 适应度的变化量。

5) 社区相似度 e

这里定义社区 c_i 和 c_j 的相似度 e 为社区 c_i 和 c_j 拥有共同节点的数量占两社区总节点数的比例^[8]。

1.2 基本思想

在实际网络中存在一些中心度很大的节点 i , 这些节点处于网络中的重要位置, 可以称这些节点为种子节点。以这些节点为初始社区 c_i , 然后访问社区 c_i 的邻居节点, 将那些对社区 c_i 适应度最大的节点加入到社区 c_i , 当最佳适应度 f_{max} 达到最大值时, 此时便形成社区 c_i 。如果节点对于多个社区都有很大的适应度, 则这些节点被划分到多个社区^[9], 因此这些社区拥有共同的节点, 形成重叠社区。最后合并那些相似度很高的社区, 得到最终的社区。

1.3 算法过程

1.3.1 发现初始社区

a) 计算网络中各个节点的中心度, 将中心度很大的节点 i 作为种子节点, 将节点 i 并入集合 V' 中。以这些种子节点为初始社区 c_i , 并初始化社区 c_i 的最佳适应度 $f_{max}^i = 0$ 。

b) 找出社区 c_i 的邻居节点, 依次访问 c_i 的每个邻居节点 j , 求出每个邻居节点 j 对社区 c_i 的适应度 $f_{c_i}^j$, 将适应度最大的节点并入社区 c_i 中, 同时将该节点并入集合 V' 中, 并令社区 c_i 的最佳适应度 $f_{max}^i = \max(f_{max}^i, f_{c_i}^j)$ 。

c) 若 f_{max}^i 连续三次未发生变化, 则转向步骤 d), 否则继续执行步骤 b)。

d) 执行回退操作, 将最后加入到社区 c_i 的节点从社区 c_i

中移出, 并从集合 V' 中删除, 最终得到社区 c_i 。

e) 检查集合 V' , 如果 V' 中包含了网络中的所有节点, 则算法结束; 否则说明网络中还有未加入到任何社区的节点, 依次访问这些剩余节点。如果节点只与存在的一个社区相连, 则直接将该节点并入该社区; 如果节点与存在的多个社区相连, 则计算节点对每个相连社区的适应度。这里引入阈值 β , 如果适应度大于或等于 β , 则将节点并入相应的社区; 否则选择节点的最大适应度, 将该节点并入相应的社区。

1.3.2 合并相似社区

从已发现的社区中合并那些相似度很高的社区。这里定义阈值 θ , 当社区 c_i 和 c_j 的相似度 $e \geq \theta$ 时, 合并社区 c_i 和 c_j 为一个社区。当任意两个社区之间的相似度小于 θ 时, 停止合并。经过合并, 相似社区最终得到整个网络的社区结构。

2 实验分析

为了验证算法的有效性, 将该算法应用到 Zachary's Karate Club Network 和 Dolphins Social Network 两个真实的网络中。实验结果表明, 由本文算法所划分的社区结构具有较好的效果, 可以有效地发现网络中的重叠社区。

2.1 Zachary's Karate Club Network

在复杂网络社区结构的研究中, Zachary's Karate Club Network 是经常被使用的经典数据集, 它反映的是美国一所大学空手道俱乐部成员之间的社会关系^[10]。该数据集包含 34 个节点、78 条边, 其中节点表示俱乐部的成员, 边表示成员之间的社会关系, 网络结构如图 3 所示。方形和圆形节点代表了该俱乐部分裂后小俱乐部中的各个成员。

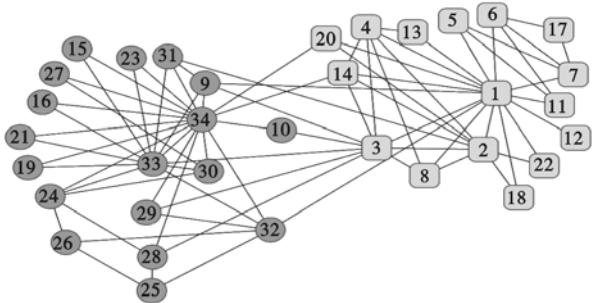


图3 Zachary's Karate Club Network

将本文算法应用到 Zachary's Karate Club Network 中, 这里设定阈值 $\beta = 0.5, \theta = 0.8$, 得到的社区结构如图 4 所示。其中三角形节点 9 和 10 被划分到两个社区, 是两个社区的共同节点。

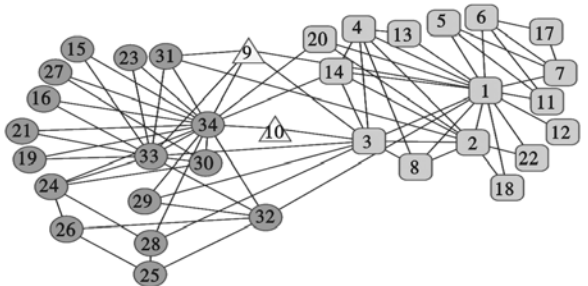


图4 本文算法所划分的社区结构

2.2 Dolphins Social Network

Dolphins Social Network 反映的是频繁交往的海豚关系网络。该数据集包含 62 个节点、159 条边。具体网络结构如图 5 所示。

将本文算法应用到 Dolphins Social Network 中,同样设定阈值 $\beta=0.5, \theta=0.8$,得到的社区结构如图 6 所示。其中形状为三角形的节点 PL、TSN83 和 Beak 为重叠社区的公共节点。

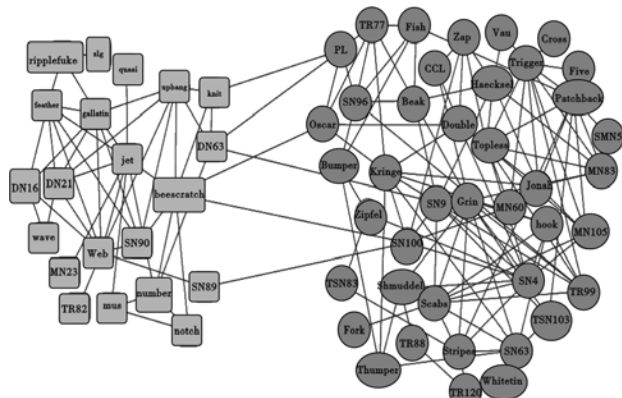


图5 Dolphins Social Network

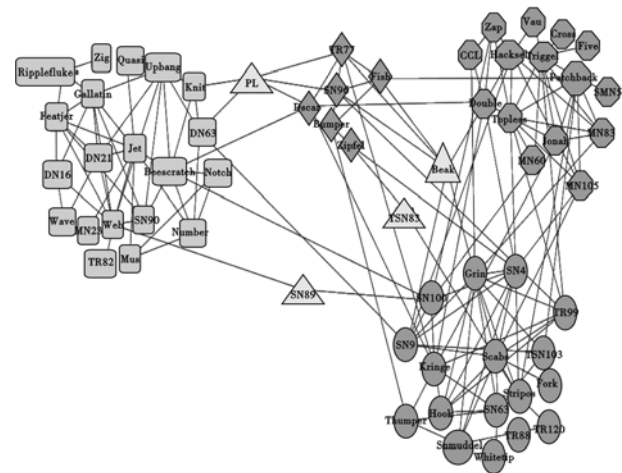


图6 本文算法所划分的四个社区

3 结束语

本文定义了节点的中心度和节点对社区的适应度,提出了

一种新的重叠社区发现算法,并将该算法应用到数据集 Zachary's Karate Club Network 和 Dolphins Social Network 中。实验表明,该算法可以有效地发现网络中的重叠社区。未来的工作将致力于提高算法的效率和精确度。

参考文献:

[1] 汪小帆,李翔,陈关荣.复杂网络理论及应用[M].北京:清华大学出版社,2006:162-188.

[2] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences,2002,99(12):7821-7826.

[3] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2):026113.

[4] CLAUSET A. Finding local community structure in networks[J]. Physical Review E, 2005,72(2):026132.

[5] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks[J]. Physical Review E,2004,70(6):066111.

[6] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E,2004,69(6):066133.

[7] LANCICHINETTI A, FORTUNATO S, KERTESZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics,2009,11:033015.

[8] CHEN Duan-bing, SHANG Ming-sheng, LV Ze-hua, et al. Detecting overlapping communities of weighted networks via a local algorithm[J]. Physica A,2010,389(19):4177-4187.

[9] PALLA G, DERENYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005,435(7043):814-818.

[10] SHANG Ming-sheng, CHEN Duan-bing, ZHOU Tao. Detecting overlapping communities based on community cores in complex networks[J]. Chinese Physics Letters, 2010,27(5):058901.

(上接第 828 页)

[3] GRAN B A, FREDRIKSEN R, THUNEM A P J. Addressing dependability by applying an approach for model-based risk assessment[J]. Reliability Engineering & System Safety, 2007, 92(11):1492-1502.

[4] LAKHINA A, CROVELLA M, DIOT C. Characterization of network wide anomalies in traffic flows, BUCS220042020 [R]. Boston: Boston University, 2004.

[5] STIJN V C. Threat modeling for Web application using the STRIDE model [D]. London:University of London, 2004.

[6] SHEYNAR O, HAINES J, JHA S, et al. Automated generation and analysis of attack graphs [C]//Proc of IEEE Symposium on Security and Privacy. Washington DC: IEEE Computer Society, 2002: 273-284.

[7] WANG Lin-zhang, WONG W, XU Dian-xiang. A threat model driven approach for security testing [C]//Proc of the 3rd International Workshop on Software Engineering for Secure Systems. Washington DC: IEEE Computer Society, 2007:204-213.

[8] DEWRI R, POOLSAPPASIT N, RAY I, et al. Optimal security hardening using multi-objective optimization on attack tree models of

networks [C]//Proc of the 14th ACM Conference on Computer and Communications Security. 2007:204-213.

[9] WYSOPAL C, NELSON L, ZOVİ D D, et al. The art of software security testing: identifying software security flaws [M]. Boston: Addison-Wesley Professional, 2006.

[10] 王红兵. Web 应用威胁建模与定量评估[J]. 清华大学学报, 2009, 49(2):2108-2112.

[11] 张红斌,裴庆祺,马建峰. 内部威胁云模型感知算法[J]. 计算机学报, 2009, 32(4):784-793.

[12] 陈小峰. 可信平台模块的形式化分析与测试[J]. 计算机学报, 2009, 32(4):646-653.

[13] 李梁,黄新芳,赵霖. 威胁树模型在风险评估中的应用[J]. 计算机应用, 2003, 23(2):188-191.

[14] 何可,李晓红,冯志勇. 面向对象的威胁建模方法[J]. 计算机工程, 2011, 37(4):21-26.

[15] 王伟,李春平,李建彬. 信息系统风险评估方法的研究[J]. 计算机工程与设计, 2007, 28(14):3473-3475.

[16] 郭春霞,袁雪红. 嵌入式系统安全的研究与设计[J]. 电子科技, 2005, 191(8):49-53.

[17] 陈火旺,王戟,董威. 高可信软件工程技术[J]. 电子学报, 2003, 31(12A):1933-1938.