第37卷 第18期
Vol.37 No.18

计 算 机 工 程
Computer Engineering

2011年9月
September 2011

·软件技术与数据库·

文章编号：1000—3428(2011)18—0041—03 文献标识码：A 中图分类号：TP311.52

基于完全子图的社区发现算法

骆 挺，钟才明，陈 辉

(宁波大学科学技术学院，浙江 宁波 315000)

摘 要：根据复杂网络中同一社区内节点连接比较紧密，社区之间节点连接比较稀疏的特点，提出一种基于完全子图的社区发现算法，通过判别2个节点是否能在网络中与任意一个节点构成3个节点的完全子图来确认该2点是否属于同一社区。对于有些节点并不满足完全子图，或在不同社区同时满足完全子图的情况，采用节点社区归属度解决该节点的归属问题。该算法不需要任何参数设置，在计算机生成网络和真实网络上进行测试，结果验证了该算法的可行性和准确性。

关键词：复杂网络；社区发现；聚类；完全子图；邻接矩阵

Community Detection Algorithm Based on Complete Subgraph

LUO Ting, ZHONG Cai-ming, CHEN Hui

(College of Science and Technology, Ningbo University, Ningbo 315000, China)

【Abstract】 Nodes in the same community are connected densely, and nodes from different communities are connected sparsely. According to this character, if two nodes and any other one node can constitute three-node complete subgraph, the two nodes are considered in the same community. Some nodes are not satisfied with complete subgraph, or are satisfied with complete subgraph for different communities at same time. A node rate of community ownership is proposed for solving those problems. The algorithm is not requested to set any parameters, and it is tested on the computer-generated and real network. Experimental results show the effectiveness and correctness of the algorithm.

【Key words】 complex network; community detection; clustering; complete subgraph; adjacent matrix

DOI: 10.3969/j.issn.1000-3428.2011.18.014

1 概述

在现实世界中，许多复杂系统都以网络结构形式呈现，如社会系统中的人际关系网、生物系统中的新陈代谢、蛋白质相互作用网、文献引用网络等。这些网络被称为复杂网络(Complex Network)。复杂网络都是由节点集合和边集合构成，其中，节点代表实体，边代表实体间的关联情况。社区结构是复杂网络的一个特性，已经被广泛关注。整个网络是由若干个社区构成的。社区间的连接相对比较稀疏，而社区内部的节点连接相对紧密[1]。社区发现对于深入了解网络结构与分析网络特征有重要的意义。聚类的簇类似于社区，人们也习惯称社区结构发现为网络聚类。

在过去几年中，有很多社区发现的方法已经被大量的研究。Kernighan-Lin算法极小化簇间连接数目与簇内连接数目之差[2]，GN算法的规则则是簇间连接的边介数(Edge Betweenness)大于簇内连接的边介数[3]。以这上些方法都需要参数的设定，如K(社区个数)等。人们提出大量网络聚类算法，还对网络结构评价进行研究，文献[4]提出模块性优化函数，文献[5]提出一种评价新的社区结构模块度。

本文提出一种基于完全子图的社区结构发现的聚类算法。该算法不需要任何的参数设定。

2 基于完全子图的算法

复杂网络可以看作一个无向无权的简单图 $G=(V, E)$，V 是其顶点的集合，E 是边的集合；A 是其对应的邻接矩阵，其中，$a_{ij}=1$ 表示节点 v_i 和 v_j 连接存在；否则 $a_{ij}=0$ 表示节点 v_i 和 v_j 的连接不存在。

2.1 完全子图

完全子图是每个顶点之间都恰有一条边的简单图。在复杂网络中，一般认为能形成一个完全子图的节点属于同一个社区。对于同一个社区，一般每个节点都可以找到同一个社区中的其他点形成一个完全子图。

在图1中，v_1,v_4,v_5 和 v_6 是完全子图，显然它们属于同一个社区。然而 v_1,v_3,v_4 和 v_5 不是一个完全子图，但是它们也属于同一个社区，由2个3节点完全子图组成：v_1,v_3,v_4 一组和 v_1,v_4,v_5 一组。v_3,v_5,v_{11} 和 v_{12} 不是一个完全子图，并且也不能找出3个节点的完全子图，所以它们并不都属于同一个社区。

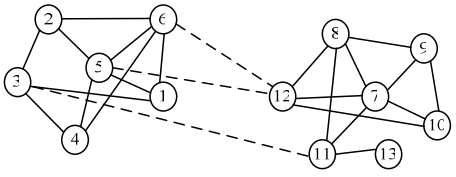

图1 原始数据集

根据以上分析，在同一个社区中，3个节点的完全子图是其基本元素，判断两点是否在同一社区的依据为两点是否能跟网络中的其他任何节点组成3节点的完全子图。

基金项目：浙江省自然科学基金资助项目(Y1090851)；浙江省教育厅科研基金资助项目(Y201016652)；宁波大学校科研基金资助项目(XYL11001)

作者简介：骆 挺(1980－)，男，讲师、硕士，主研方向：数据挖掘，复杂网络，多媒体技术；钟才明，副教授、博士；陈 辉，本科生

收稿日期：2011-04-21 **E-mail：**luoting@nbu.edu.cn

判断两点是否能跟其他一个点组成 3 节点完全子图的依据是：在网络中这两点之间的第二最短路径值为 2。如在图 1 中，节点 v_1 与 v_5 的显而易见的最短距离是 1，计算 v_1 至 v_5 的第二最短距离是 2，则 v_1,v_5 和另外一个节点可以组成一个 3 节点完全子图：v_1,v_5 和 v_6，或者 v_1,v_4 和 v_5。但在图 1 中，v_{11} 和 v_3 的第二最短路径距离大于 2，显而易见 v_{11},v_3 和其他节点不能组成一个完全子图，它们不属于同一个社区。

2.2 节点的社区归属度

在网络中，不是所有节点都完全能由完全子图来决定，节点可以分 2 种情况类似于图 1 的节点 v_{13} 和 v_{12}。由于 v_{13} 不能和其他点组成任意一个完全子图，因此不能由完全子图的方式来决定它的社区归属。v_{12} 既可以和 v_5,v_6 组成一个 3 节点的完全子图，又可以和 v_7,v_8 组成另一个 3 节点的完全子图。然而 v_5,v_6 和 v_7,v_8 分别属于不同的社区，此时 v_{12} 也不能通过完全子图的方式被确定属于哪个社区。

除了完全子图方式，一个节点 v_x 的归属可以由该社区的连接 v_x 的节点在该社区的权重值总和与该社区节点总数比率决定。一个节点在该社区权重值由该节点在本社区中连接其他节点的度来决定。度越大，权重值越大。根据以上思路，本文提出节点的社区归属度来决定以上 2 种情况节点的归属。对于节点 v_x，该节点对应社区 C_y 的归属度的具体公式如下：

$$DC_{x,y}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\{a_{ij}\mid a_{xi}=1,v_i\in C_y\text{且}v_j\in C_y\}}{|C_y|}$$

其中，a_{xi}=1 表示 v_x 和 v_i 相连；$|C_y|$表示社区 C_y 中的节点总数；n 表示网络中节点总数。$DC_{x,y}$ 值越大，则 v_x 属于社区 C_y 的几率越高。

2.3 新算法描述

一个社区有一个或者多个中心节点，中心节点跟社区中大部分节点相连。一个节点 v_x 与其他相连的节点的个数和为该节点的度(D_x)，节点的度可以很好衡量该节点在社区中的权重。D_x 越高，则 v_x 在社区中的权重越高，越容易成为中心节点。

为避免过多运算，在寻找中心节点时给出一个限定条件：其度大于网络中每个节点的平均度 MD。并且为使中心节点尽量能够分布在每个社区，初始选择中心节点时，选取不相连的中心节点。

使用图 1~图 6 对本文算法进行举例演示。图 1 为第一步找出度最大的节点 v_5，计算 v_5 相连的节点与 v_5 的第二最短距离，找出值为 2 的节点，v_5 和它们则为同一社区。图 2 中显示 v_5，v_1，v_2，v_4，v_6，v_{12} 为同一社区 C_1。继续查找剩余度数最大的节点，此时 v_6 的度数跟 v_7 的度数值是一样的，因为 v_6 已经属于社区 C_1，则先操作 v_7。同样找出跟 v_7 相连的，并且与 v_7 的第二最短路径的值为 2 的节点，此时 v_7，v_8，v_9，v_{10}，v_{11} 为同一社区 C_2，如图 3 所示。此时节点 v_{12} 跟 v_7 相连，并且两节点第二最短路径的值也为 2，但是 v_{12} 不能直接并入 C_2，因为 v_{12} 此时已经属于 C_1。此时通过对 v_{12} 2 个社区归属度的计算来决定，$DC_{12,1}$=(4+4)/5；$DC_{12,2}$=(4+3+2)/5。显然 $DC_{12,2}$ 大于 $DC_{12,1}$，则把 v_{12} 归入社区 C_2，如图 4 所示。按照以上思路反复循环操作，当找到最大度的节点为 v_4 时，v_3 与 v_4 的第二最短路径为 2，此时 v_4 已经在社区 C_1，则把 v_3 也归入社区 C_1，如图 5 所示。该循环直到剩余的节点度最大值小于 MD，则跳出循环。此时 v_{13} 还没有归属于任何一个社区。最后计算 v_{13} 社区归属度，显然 v_{13} 属于社区 C_2。最终把该数据集分成 2 个社区，如图 6 所示。

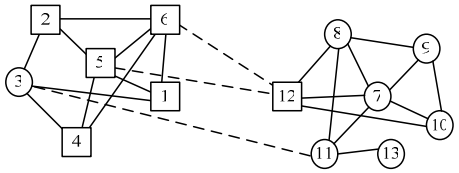

图 2 社区 C_1 节点(方形)

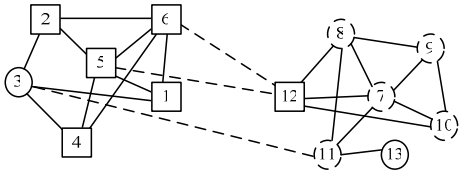

图 3 社区 C_2 节点(圆形虚线)

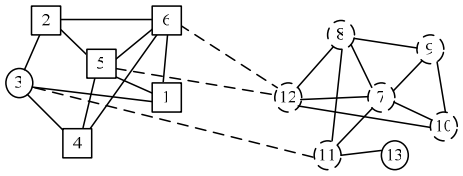

图 4 节点 12 重新划分

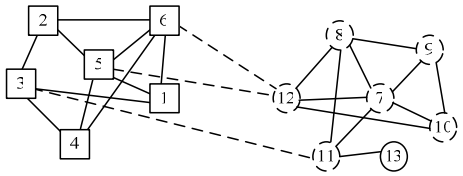

图 5 节点 3 划分

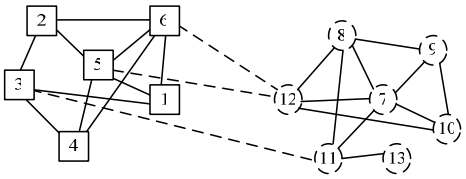

图 6 社区划分最终结果

算法主要步骤如下：

(1)计算网络中的每个节点的度，存储于向量 D；计算度的平均值 MD=mean(D)。

(2)从 D 中选取最大度 $D_{\max,}\geqslant MD$，并且其对应节点 $v_{\max}$ 不归属于任何一个社区。假如不存在该 $v_{\max}$，则选取只满足条件最大度 $D_{\max}>MD$ 对应的 $v_{\max}$。

(3)找出与 $v_{\max}$ 相连的点，计算这些相连的点与 $v_{\max}$ 第二最短路径距离。当第二最短路径距离值为 2 时，则认为该节点与 $v_{\max}$ 属于同一个社区 C_y。当碰到一个节点已经属于其他社区时，则使用节点社区归属度来判断该节点属于哪个社区。

(4)在 D 中删除 $D_{\max}$，回到第(2)步重复运算，直到 $D_{\max}<MD$，则执行下一步。

(5)查找还没有归属社区的节点，计算该节点每个社区的归属度，把该节点归属为社区归属度值最大的社区。

3 实验结果与分析

为测试本算法的可行性，除了对 *DS1* 进行测试，得到的结果跟本文分析相符外，还针对 2 个典型真实模型 Zarchary Karate Club，Dolphin 网络进行计算。结果表明，算法有效，并具有较高的准确度。

Zarchary Karate 网络是美国一所大学中空手道俱乐部成员间的相互社会关系网络。在调查过程中，该俱乐部的主管和校长产生了矛盾，结果该俱乐部分裂成 2 个分别以主管和校长为核心的小俱乐部。本文算法执行该网络能自动发现　2 个社区，并且结果跟实际一致，准确率达到 100%，如图 7 所示。

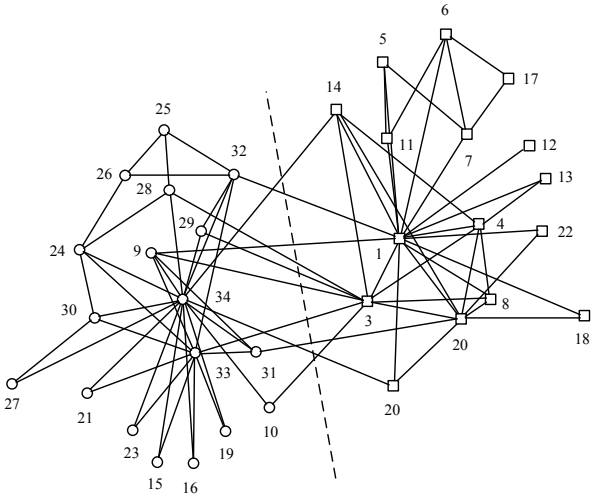

图 7　本文算法对 Zachary 网络的聚类结果

Kernighan-Lin 算法得到的结果也跟实际结果一样，但是该算法必须提前知道 2 个社区的大小分别是 16 和 18。该算法很难应用于实际网络。GN 算法，谱二分方法的结果为节点 3 被误分。此实验证明本算法相比其他算法可以更有效发现网络社区内在结构，并且不用设定任何的参数。

海豚关系网也是社会网分析中常用的一个真实网络，每个节点代表一个海豚，边表示 2 个海豚之间接触频繁。该网络共有 62 个节点，159 条边。实际的网络有 41 个较大的海豚家族，21 个较小的海豚家族。

本文算法把该网络分成 4 个社区，其中一个社区(Diamond 节点)跟实际 21 个海豚的社区完全一样。而另外 3 个社区预示着将来 41 个较大海豚家族有可能分裂成 3 个家族。对于 Kernighan-Lin 和 GN 算法，总有一个节点被误分。实验再次证明本算法的正确性，并且该算法还具有一定的预测性，如图 8 所示。

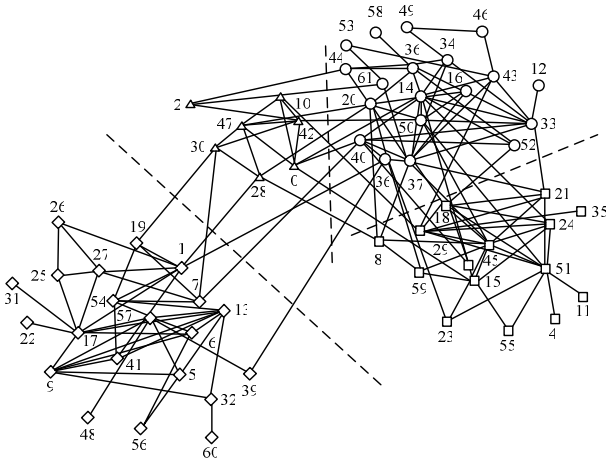

图 8　本文算法对 Dolphin 网络的聚类结果

4　结束语

本文提出一种基于完全子图的社区发现算法，利用节点社区归属度来决定个别节点的归属。该算法不需要任何参数的设置，可以自动识别社区数目。实验结果证明，该算法能够准确识别网络中的社区，具有一定的使用价值。今后的工作将研究社区与社区边界的问题，因为很多实际网络中的某些节点不仅属于一个社区。

参考文献

[1] Santo F. Community Detection in Graphs[EB/OL]. (2010-01-25). http://arxiv.org/abs/0906.0612.

[2] Kernighan B W. An Efficient Heuristic Procedure for Partitioning Graphs[J]. Bell System Technical Journal, 1970, 49(1): 291-308.

[3] Girvan M, Newman M E J. Community Structure in Social and Biological Networks[EB/OL]. (2002-04-06). http://www.pnas.org/content/99/12/7821.full.

[4] Newman M E J, Girvan M. Finding and Evaluating Community Structure in Networks[EB/OL]. (2004-08-11). http://arxiv.org/abs/cond-mat/0308217.

[5] 王　林，戴冠中. 一种新的评价社区结构的模块度研究[J]. 计算机工程, 2010, 36(16): 227-229.

编辑　陈　文

(上接第 40 页)

4　结束语

现有空间挖掘算法不能够有效地提取包含空间布局约束的拓扑关联规则，如果用传统约束关联规则算法进行挖掘，会出现重复候选项和冗余计算的问题，因此，本文提出一种基于空间布局约束的拓扑关联规则挖掘算法，实验结果验证了该算法的有效性。

参考文献

[1] 刘雨露. 基于序号索引的空间关联规则挖掘算法[J]. 计算机工程, 2010, 36(16): 54-56.

[2] 熊　江，方　刚，刘雨露，等. 空间拓扑关联的双向挖掘研究[J]. 计算机工程与应用, 2009, 45(22): 126-128.

[3] 汤小斌，方　刚. 一种用于空间横向挖掘的拓扑关联规则算法[J]. 计算机工程与应用, 2010, 46(1): 109-111.

[4] 罗爱萍. 空间跨层关联规则挖掘算法研究[J]. 西南师范大学学报: 自然科学版, 2009, 34(4): 1-5.

[5] 方　刚，魏祖宽，刘雨露，等. 一种挖掘空间拓扑关联的有效算法[J]. 计算机工程与设计, 2010, 31(6): 1267-1270.

[6] 邵峰晶，于忠清，王金龙，等. 数据挖掘原理与算法[M]. 北京: 科学出版社, 2009.

[7] 方　刚. 一种快速挖掘约束性关联规则的算法[J]. 计算机应用与软件, 2009, 26(8): 268-270.

编辑　陈　文