文章编号:1671-4598(2018)07-0242-04 DOI:10.16526/j.cnki.11—4762/tp.2018.07.054 中图分类号:TP274 文献标识码:A

基于层次覆盖的多层网络社团发现算法

王 林,李 阳,周媛媛,于文涛
(西安理工大学 自动化与信息工程学院,西安 710048)

摘要:如何检测多层网络的局部社团是近年来的热门问题之一;现有算法多针对于单层网络衡量指标的设计与改进,但节点往往处于多种复杂关系之中;为了精确的划分多层网络社团结构,一种基于层次覆盖的多层网络社团发现算法被提出;该算法结合 RA 相似度提取每层的内外连接的拓扑信息,并通过比较每层的拓扑信息关系来提取社团结构;实验结果表明,与 CLECC 和 CLEDCC 两种算法相比,提出的算法不仅降低了时间复杂度,而且在划分社团的准确度方面也有所提高,同时可以确定多层网络中无直接相连节点的划分关系。

关键词:多层网络;拓扑信息;覆盖;节点相似度

A Multilayer Network Community Detection Algorithm Based on Hierarchical Cover

Wang Lin, Li Yang, Zhou Yuanyuan, Yu Wentao
(College of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: How to detect the local community of multilayer network is one of the hot issues in recent years. Existing algorithms focus on the design and improvement of monolayer network, However, a node often exists in various relations. Therefore, how to find the community structure of multilayer network is a more meaningful research issue. A multilayer network community detection algorithm based on hierarchical cover is proposed . The algorithm combines the RA similarity to extract the topological information of the inner and outer connections of each layer, and extracts the community structure by comparing the topological information of each layer. The experimental part tests the effectiveness of the algorithm, The test results show that compared with CLECC and CLEDCC algorithms, the proposed algorithm not only reduces the time complexity, but also improves the accuracy of detecting the community. Meanwhile, we can determine the relationship between the nodes without direct connections in the multilayer network.

Keywords: multilayer network; topological information; cover; node similarity

0 引言

现实世界中的许多复杂系统可以用复杂网络来表示,社团结构是复杂网络的一个重要拓扑特性[1],它具有同一类节点之间联系紧密,不同类节点之间联系稀疏的特性。社团发现是根据复杂网络里隐含的拓扑信息来找出其中的社团结构,它可应用于信息标签化、预防病毒,预测行为等。研究复杂网络社团结构的性质不仅有助于分析复杂网络的功能,对研究生物学、医学、工程学、计算机科学等也具有十分重要的意义。因此,对社团发现算法的研究受到了国内外许多学者的广泛关注[2]。

目前,已经存在的社团发现算法多针对于单层网络,例如:以图分割[4]、GN[5]算法和标签传播算法(LPA)[6]为代表的从网络的整体到局部的社团发现算法,当然也有以 Newman 快速算法[7]、谱聚类[8]算法和 CNM[9]算法为代表的从网络的局部到整体的社团发现算法。对多层网络的研究最初起源于社会科学领域,后来发展到医学、计算机科学等。近年来,多层网络的研究逐步发展起来,继而出现了许多多层社团发现算法:CLEDCC 算法[10]、多层 α—核散列聚类的异常数据社团发现算法[11]、基于多层粒子群的社团发现算法[12]、CLECC 算法[13]、多层网络局部社团发现算法[14]以及通过比较节点度之间的关系来发现多层网络中的局部社团结构[15]等。

为了提高目前多层网络社团发现算法社团划分的准确度,以及对于一些没有直接相连的节点作出准确的划分。本文提出一种新的算法,通过结合 RA 相似度提取拓扑信息,从而间接的提取社团结构。这种基于层次覆盖的多层网络社团发现算法在时间复杂度方面得到了改善。实验结果表明,该算法能较准确地划分出多层网络中的社团结构,避免了对多层网络划分的局部性。

收稿日期:2017-12-07; **修回日期**:2018-10-08。
基金项目:陕西省科技计划重点项目资助(2017ZDCXL—GY—05—03)。
作者简介:王 林(1962-),男,江苏东台人,博士,教授,主要从事无线传感器网络、复杂网络社团发现、大数据、数据挖掘方向的研究。

1　算法

1.1　多层网络

定义单层网络 $G=<V,E>$，其中节点集为 $V=(v_1,v_2,\ldots,v_n)$，n 是节点数目，边集为 E，(v_i,v_j) 表示节点 i 与 j 之间的边。图 1 是一个简单的无权无向图，假设它是某实验室的成员图，每个节点代表一个成员，每个边代表成员之间的关系。

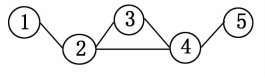

图 1　单层网络模型

一个特定复杂系统构成一个单层网络，人们在生活中会碰到各种各样的单层网络，比如在日常通信中会用到的微信这一通讯工具，微信里的朋友圈就构成了一个单层的复杂系统。然而，随着社会的快速发展，人们在平时的社交过程中不仅仅局限于微信这一种通信工具，还会涉及到微博、E－mail、Twitter、Facebook 等，每一种通讯工具都会构成一个单层的复杂系统，因此，日常生活中人们处在多个单层的复杂系统之间，这就构成了一个多层次复杂系统即多层网络。

多层网络的每一层都可以用一个图来表示，由于图与图的某些节点是相互对应的关系，且多层网络是由多个相互对应的关系组成的网络，因此了解每一个多层网络里不同图节点之间的对应关系很重要。下面给出多层网络的定义：多层网络是一个单层网络的集合，多层网络 $G=<V_\ell,E_\ell,V,E,\ell_i>$，每一层网络的层次序号 $\ell=<e_1,e_2,\ldots,e_l>$，$\ell_i=<V_i,E_i>$，$i\in 1,\ldots,l$，$V_i$ 和 E_i 分别是第 i 层网络的节点集和边集. 图 2 是一个柱形多层网络的模型，第一层与第二层之间的节点都是一一对应的。假设第一层是微信网络，第二层是微博网络，如果微信里的用户在微博里都有注册，那么在微博网络里这些用户之间的关系可以代表他们在微信网络里的关系，在这里需要注意每个用户的账户只能在他所在的层次登录，不能在其他层次登录。图 3 是一个较为复杂的普通多层网络，假设第一层是 Twitter 网络，第二层是 Facebook 网络，第一层的节点 2 是 Twitter 网络里的某用户，那么第一层的节点 2 对应的第二层的两个节点 2 是他在 Facebook 网络里注册的不同的账户，因此，不同节点不一定代表不同的用户，而是不同的账户。

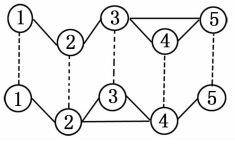

图 2　柱形多层网络

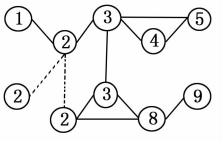

图 3　普通多层网络

1.2　RA 相似度

由于多层网络的多重性，且为了准确的找到不同层次任意两个节点之间的连接密切关系，本文采用 RA 相似度[16]公式：

$$Sim(a,b)=\sum_{i\in\varphi(a)\cap\varphi(b)}\frac{1}{k(i)} \tag{1}$$

公式（1）中的 $\varphi(i)\cap\varphi(j)$ 表示节点 i 和节点 j 的共同邻居节点集合，$k(i)$ 表示两个节点共同邻居的节点的度. 该相似度公式与以往的相似度公式不同之处在于：在比较两个节点之间的相似度时，已经存在的相似度公式仅考虑的是共同的邻居节点数目，如 Jaccard 相似度[17]，而 RA 相似度基于网络中资源配置的原理，通过比较两个节点之间的共同的邻居节点的特征来反映这两个节点之间的相似性。RA 相似度通过公式（1）计算两个节点的共同邻居节点集合里每个节点的度，以得到两个节点之间的相似性。RA 相似度的方法避免了局部相似度存在的一些问题，能更准确的比较两个节点之间的相似性，下面我们通过一个例子来介绍一下 RA 相似度公式：

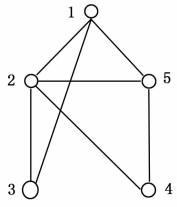

图 4　节点图

图 4 所示 5 个节点之间的连接关系，由图可知，节点 3 与节点 4 和节点 5 都没有直接相连，节点 3 与节点 4 的共同邻居节点是节点 2，节点 3 与节点 5 的共同邻居节点是节点 1 和节点 2，我们用公式（1）计算节点 3 与节点 4 的相似度可得 $Sim(3,4)=\frac{1}{4}$，且节点 3 与节点 5 的相似度为 $Sim(3,5)=\frac{1}{3}+\frac{1}{4}=\frac{7}{12}$，比较两个相似度值能得到节点 3 与节点 5 的相似度大于节点 3 与节点 4 的相似度，与图中相符，也就说明相比节点 3 与节点 4 之间的关系，节点 3 与节点 5 之间的关系更加密切。

1.3　算法核心

给定一个多层网络 $G=<V_\ell,E_\ell,V,E,\ell_i>$，定义 C 为某一层的内部初始核心节点社团，定义 $D=\{v\in C\mid\exists(u,v)\in E_\ell,u\in S\}$ 为社团 C 的边界子集合，定义 $S=(v\in V\setminus C\mid(u,v)\in E_\ell,u\in C)$ 为外层节点集合，E^D 代表集合 D 与集合 S 中节点之间的边，$E_i^D\in E^D$ 是第 ℓ_i 层的边集，定义 L 为局部社团相似性测度，表示为社团内部节点连接拓扑信息关系与社团间节点连接拓扑信息关系之间的比值。首先随机选取一个节点 v 作为初始社团 C，通过比较给出两个公式来表示社团内连接关系和社团间连接关系：

$$L^{\text{int}}(C)=\frac{1}{C}\sum_{v\in C}\sum_{L_i\in\ell}\sum_{(u,v)\in E^i\wedge u\in C}sim(u,v) \tag{2}$$

公式（2）表示社团内部节点连接拓扑信息关系，其中 $\frac{1}{C}$ 中的 C 表示社团 C 里含有的节点数量. $sim(u,v)$ 是在公式（1）给出的计算节点 u 和节点 v 之间的相似度公式。

$$L^{ext}(C)=\frac{1}{D}\sum_{v\in D}\sum_{L_i\in\ell}\sum_{(u,v)\in E''\wedge u\in S}sim(u,v) \tag{3}$$

公式（3）表示社团间节点连接拓扑信息关系。

1.4 判断条件

为了判断一个外层节点加入内部社团 C 后是否能加强社团的紧密性，定义局部社团相似性测度 L 来评价外层节点加入内层节点之后的效果，公式如下所示：

$$L=\frac{L^{\text{int}}(C)}{L^{ext}(C)} \tag{4}$$

若外层节点 u 加入 C 后满足以下条件：

$$L(C\bigcup\{u\})>L(C) \tag{5}$$

$$L^{\text{int}}(C\bigcup\{u\})>L^{\text{int}}(C) \tag{6}$$

公式（5）、（6）表示若节点 u 加入社团 C 后，局部社团相似性测度 L 值变大，L 值越大表明外层节点加入内层社团 C 后，C 更紧密，且内层社团与外层社团连接更加稀疏，社团结构更加明显。且社团内部节点连接更紧密，此时将节点 u 加入到 v 所在的社团 C 中。在判断的过程中，每次将 L 取最大值时的外层节点加入到社团 C 中，迭代地判断每个节点，直到 L 不再增大。

用 L 作为划分节点的评判标准，将每次使得 L 值最大的外层节点划分到内层社团 C 中，直到不存在符合条件的节点出现为止，但这种方法在面对一些异常节点时通常不能有很好的划分效果，对于这种情况，若外层节点符合以下两个条件，就将它们划分到社团 C 中：

$$L_1^{int}(C\bigcup\{u\})>L^{int}(C)\text{ 且 }L_1^{ext}(C\bigcup\{u\})<L^{ext}(C) \tag{7}$$

$$L_1^{int}(C\bigcup\{u\})>L^{int}(C)\text{ 且 }L_1^{ext}(C\bigcup\{u\})>L^{ext}(C) \tag{8}$$

公式（7）表示加入节点 u 后，社团内连接系数较没加入 u 之前增大，社团间连接系数变小。明显地看出，公式（7）所述情况 L 值会增大，且符合（5）、（6）两个条件，此时，将节点 u 划分到社团 C 中。如果加入节点 u 后遇到公式（8）这种情况，即加入节点 u 后，社团内连接系数和社团间连接系数较之前都有增大，社团内部连接以及社团之间的连接都更加紧密，此时节点 u 有两种可能：

（1）节点 u 符合以上条件，且 u 不是核心节点，可以与社团 C 内的节点进行合并；

（2）节点 u 可能是一个核心节点，它与社团内和社团外的节点都有大量的连接。

对于（1）中的节点 u，将它划分到社团 C 中。对于（2）中的节点 u，暂时将它加入到 C 中，直到所有节点被划分到相应的社团后，再将这些疑似的核心节点从 C 中移除，此时，再返回到条件（5）、（6）对这些节点进行判断。

1.5 算法流程

该算法首先随机选取一个节点 v 作为覆盖第一层的中心节点，在初始阶段，集合 D 和集合 C 里只有节点 v，集合 S 是外层节点集合，不断地从 S 集合里随机选出节点 u，如果 u 加入到 C 中使得 L 值较大且能满足公式（4）以及公式（5）中的两个条件，则将 u 加入到 v 所在的集合里，迭代上述过程，在每一步迭代中，都要更新集合 D、集合 C 和集合 S，且直到 L 值不再增大，算法结束。具体算法流程如下所示。

算法：基于层次覆盖的多层网络社团发现算法：

输入：多层网络 G；

输出：网络 G 的社团划分结果。

（1）初始化：随机选取一节点 v 加入集合 C 与集合 D 中，S 为外层节点集合。

（2）从集合 S 中随机选取一节点 u 加入集合 C，分别计算 $L^{\text{int}}(C)=\frac{1}{C}\sum_{v\in C}\sum_{L_i\in\ell}\sum_{(u,v)\in E''\wedge u\in C}sim(u,v)$，$u$ 加入 C 后的社团内连接系数和社团间连接系数 $L^{ext}(C)=\frac{1}{D}\sum_{v\in D}\sum_{L_i\in\ell}\sum_{(u,v)\in E''\wedge u\in S}sim(u,v)$。

（3）计算局部社团相似性测度 $L=\frac{L^{\text{int}}(C)}{L^{ext}(C)}$，判断节点 u 加入集合 C 后是否同时满足以下两个条件：

1）$L(C\bigcup\{u\})>L(C)$ 加入节点 u 后，相似性测度 L 变大；

2）$L^{\text{int}}(C\bigcup\{u\})>L^{\text{int}}(C)$ 加入节点 u 后，社团内部连接系数变大。

（4）若节点 u 同时满足条件 1）和 2），将 u 加入到社团 C 中，若节点 u 为疑似核心节点，也将其放入 C 中，返回第（2）步，直到遍历所有节点，L 不再变大，再将疑似的核心节点从 C 中移除，返回（3）对其进行重新判断。

（5）合并集合 C 与集合 D，并计算不同层次的模块度 Q。

2 实验分析

2.1 多层网络数据

为了测试算法的性能，会用到的不同的多层网络数据集，下面先来对这些数据集做一个简单的介绍。

MIT Reality Mining[18]网络数据集是通过给麻省理工学院 87 个移动用户安装一个软件，记录用户之间的数据交互信息，网络的每一层分别代表从现实中采集到的用户的物理位置、蓝牙交互和通话记录等用户之间的互动行为。

E－mail[19]网络数据集是一个记录 Enron 公司员工之间电子邮件往来的数据集，该网络有 150 个节点，每个节点代表 1 个用户，每条边代表 2 个用户之间发的一条电子邮件，该网络包括用户之间发送邮件的时间、发送主题、发送者账户以及接收者账户等。网络中的每一层分别代表员工之间的关系和邮件信息内容的相似性。

IMDB[20]网络数据集是一个互联网电影数据集，该数据集包含 300 个节点，每个节点代表一个一位演员，每条边代表这两个演员一起演了一部戏，该网络数据集的每一层分别代表第一年演员之间的合作、最后一年演员之间的合作、演员的平均收入和门票卖出的平均数量。

在实验仿真部分，用以上三个多层网络数据集对算法进行测试，测试了算法在三种网络上的运行时间，并用模块度 Q[21]来评价社团划分结果，之后又将该算法与

CLECC、CLEDCC 两种算法进行了对比，结果表明本文算法的准确度更高，运行时间更少。

2.2　评价标准

为了评价社团划分的结果，采用 Newman 和 Girvan 提出的模块度 Q。其定义式如下：

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \| e^2 \| \qquad (9)$$

其中：$\| e \|$ 表示矩阵 e 中所有元素之和，e_{ij} 代表连接社团 i 和社团 j 之间边的总数，$Tre = \sum_i e_{ii}$ 是矩阵 e 对角线元素的和，表示相同社团内节点之间边的集合，$a_i = \sum_j e_{ij}$ 表示连接社团 i 的边的总数。Q 值取值范围在 0～1 之间，而在现实网络中，Q 值的范围在 0.3～0.7 之间。算法在划分过程中会出现不同的 Q 值，且 Q 值越大代表社团划分结果越好，内部节点之间的连接更紧密，内部节点与外部节点之间的连接越稀疏，社团结构比较明显，因此，在划分过程中选择使 Q 值最大的划分为最终划分结果。

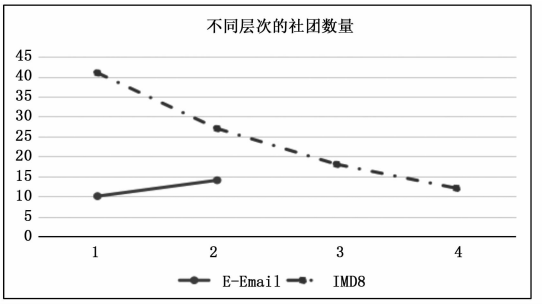

图 5　随着层数的变化，社团数量的变化

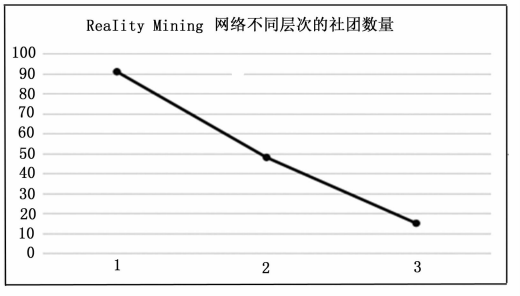

图 6　随着层数的变化，社团数量的变化

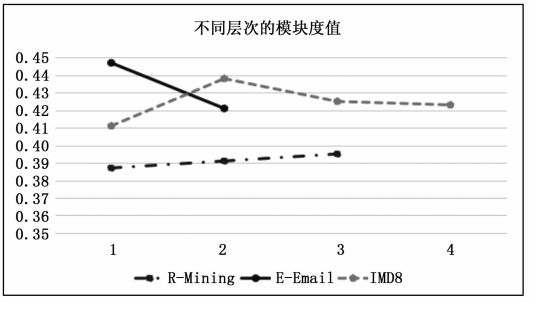

图 7　不同层次的模块度 Q 值的变化

图 5、6 代表算法对三种网络的划分随着层数的变化，社团数量的变化，其中 x 轴代表某一层次，y 轴代表社团数量。图 7 代表不同层次的模块度 Q 值的变化，其中 x 轴代表社团数量，y 轴代表模块度 Q 值。

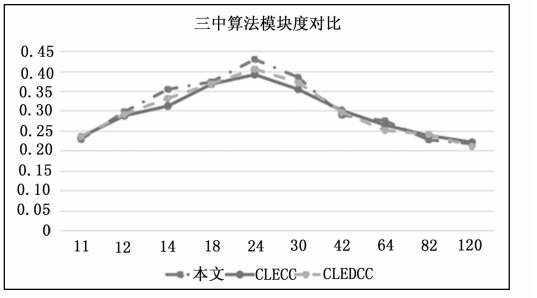

图 8　E－mail 网络模块度 Q 值对比

图 8 和图 9 分别是本文算法、CLECC 算法以及 CLEDCC 算法在 E－mail 网络上以及 IMDB 网络上划分的模块度 Q 值的对比，其中 x 轴代表社团数量，y 轴代表模块度 Q 值。从图 8 可以看出，将 E－mail 网络划分在 24 个社团左右时，有最大的模块度 Q 值，此时的划分效果较好。从图 9 可以看出，将 IMDB 网络划分在 95 个社团左右时，会得到较为明显划分结果。

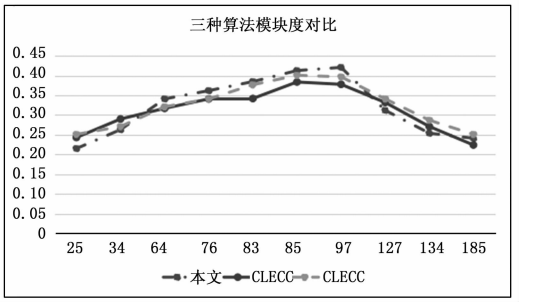

图 9　IMDB 网络模块度 Q 值对比

由图 8 和图 9 可以看出，不同算法在 E－mail 网络和 IMDB 网络上的模块度 Q 值不尽相同，本文算法相比 CLECC 算法和 CLEDCC 算法模块度更高，划分也更准确。

表 1　三种算法运行时间比较(ms)

	CLECC	CLEDCC	本文
R－Mining	68	54	37
E－Email	117	96	74
IMDB	142	114	107

表 1 是本文算法和 CLECC、CLEDCC 三种算法对三个网络划分时间的对比，可以看出本文算法需要的运行时间更少，效率更高。

3　结论与展望

文中采用 RA 相似度和一种多层网络社团结构检测的模型，提出了一种基于层次覆盖的多层网络社团发现的新算法，并将算法在几个经典的多层网络进行了性能测试，均取得了不错的划分结果。实验的后一部分将本文算法与

(下转第 250 页)