ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 6,Nov. 2015,pp. 1215－1224
DOI:10.16337/j.1004-9037.2015.06.009

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种新的社区/动态社区优化方法

李亚芳[1,2]　贾彩燕[1,2]　于　剑[1,2]　刘光明[1,2]

(1.北京交通大学计算机与信息技术学院,北京,100044; 2.交通数据分析与挖掘北京市重点实验室,北京,100044)

摘　要:社区结构作为复杂网络的重要拓扑特性之一,成为当前的研究热点。本文提出了一种基于边排序和模块度优化的社区发现方法。该方法首先对初始的静态网络进行稀疏化,然后在稀疏化后的网络上依据边的重要程度对边进行排序,给出了一种模块度最大化、快速边合并的社区发现方法(Fast rank-based community detection, FRCD)。在初始网络社区划分结果的基础上,将该方法推广到动态、实时社区划分上,给出了一种快速、鲁棒的动态社区划分方法(Incremental dynamic community detection, IDCD)。理论分析表明FRCD相对于边具有线性时间复杂度。在实际和人工网络上的实验结果均表明,本文提出的方法无论在静态网络社区划分还是在动态网络社区追踪上都优于已有方法。
关键词:社区发现;模块度;边排序;动态性
中图分类号:TP181　　**文献标志码**:A

Novel Community/Dynamic Community Optimization Algorithm

Li Yafang[1,2], Jia Caiyan[1,2], Yu Jian[1,2], Liu Guangming[1,2]

(1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China; 2. Beijing Key Lab of Traffic Data Analysis and Mining, Beijing, 100044, China)

Abstract: Community structure is one of the most important topological characteristics in the complex network, being a hot research area in different fields. A novel community detection algorithm is proposed based on edges rank and modularity optimization. Local graph is sparsificated and edges are ranked according to the similarity. Therefore, a method called the fast rank-based community detection (FRCD) by maximizing modularity and fast mergement of edges is achieved. Meanwhile the method is also extended to dynamic and real-time community detection on the basis of initial community structure, and a fast and robust dynamic community detection algorithm called the incremental dynamic community detection (IDCD) is presented. Theoretical analysis exhibit that FRCD has linear complexity for network edges. Experimental results in real-world and artificial networks demonstrate the high accuracy and good performance of the algorithm on static community detection and tracking dynamic structure of networks.
Key words: community detection; modularity; rank; dynamic characteristic

引　　言

现实世界中,很多系统都以网络形式存在,如社会系统中的人际关系网、科学家合作网和流行病传

基金项目:国家自然科学基金(61473030,61370129)资助项目;北京市自然科学基金(4112046)资助项目;北京市科委(Z131110002813118)资助项目;中央高校基本科研业务费专项基金(K15JB00070,2014JBM031)资助项目;北大方正集团有限公司数字出版技术国家重点实验室开放课题资助项目。
收稿日期:2014-03-25;**修订日期**:2014-10-23

播网,生态系统中的神经元网、基因调控网和蛋白质相互作用网,科技系统中的电话网、因特网和万维网等。近年来,复杂网络研究已成为最重要的多学科交叉研究领域,吸引了越来越多研究者的目光。研究者发现很多实际的网络具有"模块性"[1],即网络组织结构表现出明显的社区结构,表现为社区内部连接稠密,社区间连接稀疏的中观尺度结构。找到网络中连接紧密的簇结构有助于更深入地了解网络的本质,认识网络结构与其功能之间的关系,发现复杂网络中的规则并对其行为进行预测。社区结构可以揭示出社会系统的组织结构及随时间演化的关系,蛋白质功能和蛋白质物理相互作用间的内在联系以及网页主题和超连接间的内在关系等。

随着对复杂网络中社区结构分析的需求不断增强,目前关于社区发现研究的方法和成果很多。如经典的基于节点分裂的 GN 算法[1]、基于模块度优化的 BGLL[2] 和 CNM 算法[3]、基于标签传播 (Label propagation algorithm, LPA)算法[4]以及基于随机游走和压缩编码的 Infomap 算法[5]等。目前,网络类型多样化,针对网络的不同类型提出的针对多模式和多内容网络的社区检测方法[6-7]。在以上社区发现算法中,由于模块度能够作为一种衡量网络中社区结构的内部指标,最大化模块度以探测网络社区结构的算法(如 BGLL 和 CNM)在很多领域得到了较成功的应用[2-3]。然而,BGLL 和 CNM 等方法只能对一段时间累积下的静态网络进行划分,无法实时追踪网络社区结构的变化,需要对相邻时间快照得到的网络重新进行社区划分,并保存每一时间快照下的社区结构。由于实际的动态社会网络具有时间局部性,相邻时间间隔的网络结构变化相对缓慢[8-9]。因此,对于非常相似的网络反复进行社区划分,会产生了很多重复的操作,从而降低算法的效率。

Shang 等[10]对 BGLL 算法进行了扩展,该方法首先利用 BGLL 算法在初始子网络上得到一个初始的社区划分结果,然后通过随机加边的方式模拟网络的增长,将新增节点不断实时分配到已有社区之中。但是该方法存在以下两个问题:(1)该方法属于两阶段方法,依赖于 BGLL 算法得到的初始社区划分,当初始子网络过小或选择时机不当而不能保证较好的初始划分时,难以得到理想的动态社区划分结果。对不断增长的动态网络,如何选择一个合适的初始子网络是该方法存在的一个问题;(2)由于 Shang 等的算法对下一个时刻增加的边采用随机处理的方法,导致不同的边处理顺序会得到不同的社区划分结果。因此,算法的性能不够稳定。针对以上问题,本文提出一种基于边排序的、模块度优化社区划分算法(Fast rank-based community detection, FRCD)。该算法是一种从头算法方法,对于静态网络能够得到较准确的社区划分结果。并且,其动态网络上的扩展算法(Incremental dynamic community detection, IDCD)对初始子网络依赖性小,在 FRCD 算法在初始子网络已有社区划分的基础上,能实时地处理以不同时间粒度为间隔的动态增加的网络结构,并且具有性能稳定的优点。

1 研究基础

模块度 Q 函数用网络中连接社区内部顶点的边所占的比例与随机网络中连接社区结构内部顶点的边所占比例的期望的差值来衡量网络中社区划分的质量[1,11],其定义如下

$$Q=\frac{1}{2m}\sum_{i,j\in V}(A_{ij}-\frac{k_ik_j}{2m})\delta(c_i,c_j) \tag{1}$$

式中:c_i 为顶点 i 所属的社区;$\boldsymbol{A}$ 为网络的临近矩阵,每个元素 A_{ij} 表示顶点 i 和 j 所在边上的权值,若 i 和 j 没有边,则 A_{ij} 为 0;$\delta(c_i,c_j)$ 为 Kronecker 函数,如果 $c_i=c_j$,则值为 1,否则为 0;m 为网络中边的数目;k_i 为顶点 i 的度;$\frac{k_ik_j}{2m}$ 表示在随机网络中顶点 i 和顶点 j 之间存在连边的概率。

显然,模块度越大、网络的社区结构越明显。因此,对于一个给定的网络,可以通过最大化 Q 函数得到了一个较好的社区划分结果(如 BGLL 算法[2],CNM 算法[3])。

通过对模块度 Q 函数进行变形,可将其转换为以下形式[10]

$$Q=\frac{1}{2m}\sum_{c\in C}(\sum_{\text{in}}^{c}-\frac{(\sum_{\text{tot}}^{c})^2}{2m}) \tag{2}$$

式中:m 为网络中边的数目;$\sum_{\text{in}}^{c}=\sum_{i,j\in V}A_{ij}\delta(c_i,c)\delta(c_j,c)$ 为社区 c 内边的权重总和;in 为社区 c"内部"边权重总和;$\sum_{\text{tot}}^{c}=\sum_{i\in V}k_i\delta(c_i,c)$ 表示所有与社区 c 内节点相连接的边的权重总和,tot 为"所有"与社区 c 内节点相连边的权重总和;$\delta(c_i,c)$ 为 kronecker 函数,如果 $c_i=c$,则值为 1,否则为 0。

Shang 等[10]为了最大化式(2)模块度 Q,将增加的边分为 4 种类型,并对不同类型的边施加以不同的操作来实现对新增节点的社区指派。分别为:(1)全新的边,即新增边的两个节点在社区中没有出现过,对应的操作为将新增的两个节点指派到一个新的社区中;(2)半新的边,即新增边的一个节点在原社区中出现过,则将新增的节点指派到老节点所属的社区中;(3)内部的边,即新增边的两个节点所属同一个社区,那么节点的社区指派不发生变化;(4)交叉的边,即新增边的两个节点分属不同的社区,如果此时满足以下不等式:$(W+w_{ij})(2m+2w_{ij})>(\sum_{\text{tot}}^{c_i}+w_{ij})(\sum_{\text{tot}}^{cj}+w_{ij})$(其中,$i,j$ 是新增边的两个节点,c_i 和 c_j 是其所属的社区,w_{ij} 是两个社区间新增连边的权重和,W 为已存在的社区间的连边权重总和),则将这两个社区合并,否则节点指派不发生变化。

2 FRCD 算法

基于网络局部稀疏和排序的静态社区模块度优化算法 FRCD 可分为 3 个步骤:计算边的相似度、对网络进行局部稀疏化及边排序、节点社区指派及社区合并。本算法假设相似度越高的节点对,其关系越容易建立,在社区中所处的位置越靠近中心。

2.1 边相似度计算

计算边的强度的方法有很多,如边介数[1]、桥系数[12]以及边的聚集系数[12]等,这些方法都可以用于衡量两节点关系的紧密程度。但是这些方法有的基于网络全局结构,有的计算复杂度较高,对于大规模网络的处理代价较大。因此,本文采用一种基于局部的、计算简单的 Jaccard 相似度方法来计算网络中边(i,j)的强度。其中,i 和 j 为与边(i,j)关联的两个节点。具体定义如下

$$\text{Jac}(i,j)=\frac{|N(i)\cap N(j)|}{|N(i)\cup N(j)|} \tag{3}$$

式中:$N(i)$表示节点 i 的邻居节点的集合,$N(i)\cap N(j)$为节点 i 和节点 j 共有邻居个数,$N(i)\cup N(j)$为节点 i 和节点 j 总的邻居个数。

通过 Jaccard 相似度的定义可知:如果两个节点的共有的邻居越多,节点间的相似程度越高,边上的强度越大。根据本文的假设,如果两个节点的相似性越高,这两个节点间的关系更容易建立而且更加稳固,并且这两个节点在社区中所处的位置越靠近社区的中心。因此,首先按照边的排序对最容易建立关系的节点进行社区指派。然后,通过最大化模块度 Q 依次对其他节点进行社区指派。

2.2 网络局部稀疏化

由于在原始网络上进行边排序,其复杂度为 $O(m\log m)$。因此,首先在保证网络社区结构的基础上,对网络进行稀疏化[13]。主要策略为:根据网络中节点相关联边的 Jaccard 相似度,通过选取每个节点的前 d_i^e 条边可以得到一个局部稀疏的网络(至少选择一个与节点相连接的边),其中 d_i 表示节点 i 的度,e 为网络的稀疏化指数。稀疏化后网络的度仍然服从幂率分布,其幂率分布指数为$(\alpha+e-1)/e$,其中 α 为原网络的度的幂率分布指数。

由于不同的网络稀疏指数,得到的结果不同。如果稀疏值设的太小,会造成网络过于稀疏,难以找到原始网络中蕴含的社区结构;设得太大,不能有效加快算法运行速度。因此,通过实验选择 $e=0.5$ 来对网络进行稀疏化,此时在各种网络规模中均能得到较稳定的社区划分结果。显然,通过网络局部稀疏化,能够有效减小运算复杂度,从而提高算法的运行效率。

2.3 社区指派和合并

依据假设,对稀疏化的网络按照边的相似度进行降序排序后,首先将排序在第一位的边上两个节点

指派到同一个社区内,作为初始的社区。然后,按照排序顺序依次对后续的边中的节点进行社区指派,即把相对位于社区边缘的节点进行社区指派。其指派策略采用 Shang 等[10]动态社区指派策略。即根据边的 4 种类型,依次对节点进行社区指派。通过排序,能够保证得到比较稳定的社区结构。社区指派后,需要通过社区的合并对指派的结果进行进一步优化,从而得到更符合实际的划分结果。需要进行社区合并的小社区需要满足以下两个条件。

(1)不满足弱社区的定义[14]。

如果社区 A 不满足

$$\sum_{i\in A} k_i^{\text{in}}(A) > \sum_{i\in A} k_i^{\text{out}}(A) \tag{4}$$

则选取与社区 A 之间连边最多的社区 B 进行合并其中。$k_i^{\text{in}}(A)=\sum_{j\in A} A_{ij}$ 表示社区 A 内部与区 A 中节点 i 与社区外部关联的边数。

(2)社区合并需要保证网络的模块度 Q 值不会减少。

本文提出了一种基于边排序的模块化最大化社区发现算法 FRCD,其整体流程如下。

输入:网络 $G=(V,E)$,稀疏化指数 e。

输出:社区结构,节点列表。

(1)对网络中每个节点关联的边计算相似度并排序;(2)选取每个节点关联的前 d_i^e 条边为稀疏化后的网络;(3)对稀疏后的所有边依据相似度进行全局排序;(4)按边的相似性由高到低地对节点进行社区指派及合并调整,得到最终的社区划分结果,即节点社区归属列表。

2.4 算法的时间复杂度分析

该算法中,对于有 n 个节点,m 条边的网络 G,对每个节点关联的边进行相似度计算并排序,如果采用 Minwise 散列方法处理,时间复杂度为 $O(m)$[13]。对网络稀疏化后,由于只剩余 $\sqrt{m}$ 条边,全局排序的时间复杂度为 $O(\sqrt{m}log\ \sqrt{m})$。社区指派步骤与 Shang 等[10]的方法相似,算法首先需要判断边的类型,对不同类型的边施加不同的操作以得到最终的社区划分结果。前 3 类边的操作复杂度为 $O(1)$,第 4 类边的操作复杂度为 $O(S)$,其中 S 为合并后的社区规模。由于第 4 类边出现的比例较低,因此算法的复杂性接近于 $O(1)$。通过以上分析可知,该算法的总复杂度约为 $O(\sqrt{m}\log\sqrt{m})+O(m)\approx O(m)$。

3 IDCD 算法

任给一个网络,应用 FRCD 算法可以对该网络进行社区划分。相应于 Shang 等的方法,FRCD 算法可以自然地扩展到对动态网络社区划分上,以实时的追踪网络结构的变化。由于现实中的社会网络结构变化都是很缓慢的,合适的相邻采样时刻之间网络变化很小。因此,通过调整时间间隔,能够对不同时间粒度内增加的节点进行社区指派。

如果选取的时间粒度以一条边的增加为粒度,称其为基本元粒度。此时,只需要判断该边所属的类型,根据边的类型采取相应的操作,实现边关联的节点的社区指派。在一定的时间间隔内,当增加的边由多个基本元粒度边构成时,称之为一定时间粒度。此时 IDCD 算法对新增的边根据关联节点是否出现在节点社区指派列表中分为 3 类:全新列表(都没有进行社区指派),半新列表(至少一个节点已经进行了社区指派),已有列表(都进行社区指派)。类似地,根据 Jaccard 相似度和节点的度可以对已有列表和半新的列表中的边进行降序排列,进而由边的重要程度逐个对边关联的节点进行社区指派,并更新节点社区归属列表,算法流程如下。

输入:$t-1$ 时刻网络的社区结构,原网络以及新增的网络,节点列表。

输出:t 时刻的社区结构,更新的节点列表。

(1)对新增的节点和边,根据节点列表分成 3 类;(2)根据原网络对新增边计算相似度,按照相似度和度进行降序排序;(3)在 $t-1$ 时刻得到的社区结构基础上,对新组合的新增列表中的节点依次进行社

区指派;(4)更新节点社区归属列表。

4 实验结果与分析

为了定量地测试本文提出算法的有效性,分别在真实网络数据和人工生成数据上测试了新给出的FRCD算法与基于模块度优化的BGLL算法以及CNM算法的社区划分效率。同时,给出了FRCD算法在不进行网络稀疏化时的划分结果(记为FRCD*)。在人工网络上模拟了当网络动态变化时,新给出的动态社区划分方法IDCD和Shang等给出方法的社区划分的差异。

由于这两类数据集都已知社区结构,因此选取了标准化互信息(Normalized mutual information,NMI)[14]以及准确率[15]作为社区划分结果优劣的评价指标。其中,NMI的定义为

$$\mathrm{NMI}(C_A,C_B)=\frac{\sum_{i=1}^{k_A}\sum_{j=1}^{k_B}n_{ij}\log\frac{n_{ij}n}{n_i^A n_j^B}}{\sqrt{\sum_{i=1}^{k_A}(n_i^A\log\frac{n_i^A}{n})}\sqrt{\sum_{j=1}^{k_B}(n_j^B\log\frac{n_j^B}{n})}} \tag{5}$$

式中:C_A 为真实的社区结构,C_B 为通过算法得到的社区结构,k 为社区个数,n 为网络中节点的个数,n_{ij} 为根据真实的社区 i 在算法得到的社区 j 中正确指派的节点个数,n_i^A(n_j^B)为社区 i(或 j)中节点的个数。通常NMI值越大,表明算法得到的结果越准确。Accuracy为正确划分的节点在所有节点中所占的比例。显然,Accuracy越大,社区划分也越准确。

4.1 静态社区结构的实验比较

4.1.1 真实实验数据

在本节实验中用到的真实数据包括经典的美国大学足球赛网络Football[16],Zachary空手道俱乐部网络Karate[17]]以及海豚关系网络Dolphins[18]。除了这3个经典实际网络,为了测试该算法在不同规模真实网络中的社区划分效果,本文还选取其他3个实际网络,包括美国政治书籍网络Political Books[19],博客网络Blogs[20]以及蛋白质交互网络PPI[21],各网络相关属性如表1所示。

表1 真实网络数据集

Table 1 Real-world data sets

编号	网络	节点	边
1	Karate	34	78
2	Dolphins	62	159
3	Political Books	105	441
4	Football	115	613
5	Blogs	1 490	17 138
6	PPI	1 628	11 249

由于本文提出算法是基于模块度优化的社区发现方法,首先将FRCD与BGLL和CNM算法进行实验比较。Shang等人提出的方法缺乏稳定性,导致程序每次运行的结果不同,因此选择10次运行结果中得到 Q 值最大的结果进行对比,得到表2~4所示的实验结果。

表2 真实网络中的NMI对比结果

Table 2 NMI comparison in real-world data sets

算法	1	2	3	4	5	6
FRCD*	0.60	0.58	0.57	0.91	0.68	0.94
FRCD	0.71	0.60	0.48	0.91	0.55	0.95
BGLL	0.58	0.46	0.51	0.93	0.33	0.95
CNM	0.69	0.56	0.53	0.70	0.45	0.94

表 3 真实网络中的 Accuracy 对比结果

Table 3 Accuracy comparison in real-world datasets

算法	1	2	3	4	5	6
FRCD*	0.71	0.60	0.85	0.86	0.94	0.72
FRCD	0.74	0.56	0.59	0.90	0.88	0.79
BGLL	0.64	0.50	0.72	0.90	0.79	0.88
CNM	0.74	0.69	0.81	0.57	0.80	0.82

表 4 真实网络中的 Q 对比结果

Table 4 Q comparison in real-world data sets

算法	1	2	3	4	5	6
FRCD*	0.42	0.52	0.52	0.60	0.42	0.88
FRCD	0.42	0.50	0.50	0.60	0.42	0.85
BGLL	0.42	0.52	0.53	0.60	0.43	0.88
CNM	0.38	0.50	0.50	0.55	0.43	0.88

通过实验结果比较可以发现：(1)FRCD 算法在各网络中得到的最优模块度值与 BGLL 大致相同，都具有较高的模块度；(2)在 Karate，Dolphins，Political Books 以及 blogs 网络上得到的 NMI 以及准确率要优于 BGLL；(3)网络的稀疏化能够减少社团间的部分连边，同时会对社团内部节点间的连边进行稀疏，此时得到的结果与未稀疏化网络相比，会优于或者略差于未稀疏化的网络。在 Karate 和 Dolphins 网络上得到的 NMI 以及在 Football 网络上得到的准确率明显优于未稀疏化后的网络，但是在 Political Books 网络上效果略差，因此稀疏化的程度会对算法的准确性产生一定的影响。FRCD 算法能够较准确地发现网络中蕴含的社区结构。

4.1.2 人工网络数据

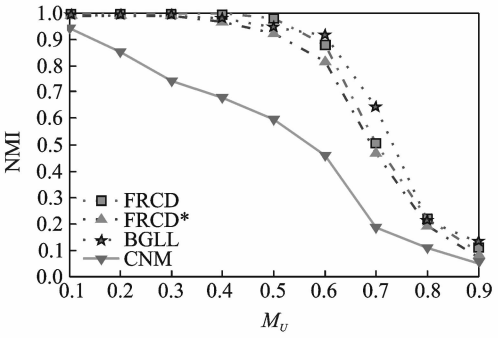
为进一步实验验证本文算法的效果，本文选用 Lancichinetti 等提出的 LFR 人工网络[22]进行比对，表 5 给出了生成的 4 个 LFR 人工网络的参数设置。

表 5 中，N 表示生成的网络中节点的个数；M_U 是网络的混合参数，表示与社区内节点关联的边中处于社区之间边的比例，该值越小表明生成的网络的社区结构越清晰，取值为 0.1 到 0.9；k 表示网络的平均度；$k_{\max}$ 表示网络中节点的最大度；min$c$ 表示网络中最小社区中节点个数；max$c$ 表示网络中最大社区中节点的个数；t_1 为度分布的负指数值；t_2 为网络中社区规模分布的负指数值。通过与 BGLL 和 CNM 进行比较，得到图 1～4 的实验结果。

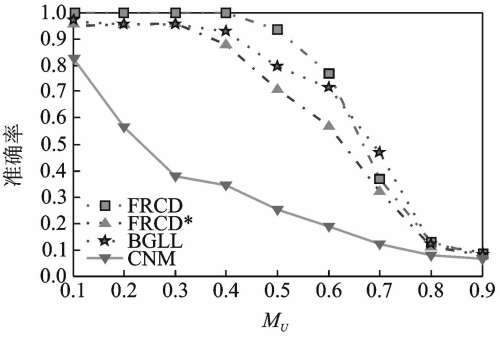
表 5 LFR 人工网络中的参数

Table 5 Parameters in artificial networks

参数	LFR1	LFR2	LFR3	LFR4
N	1 000	1 000	5 000	5 000
M_U	0.1-0.9	0.1-0.9	0.1-0.9	0.1-0.9
K	20	20	20	20
$k_{\max}$	50	50	50	50
min$c$	10	20	10	20
max$c$	50	100	50	100
t_1	2	2	2	2
t_2	1	1	1	1



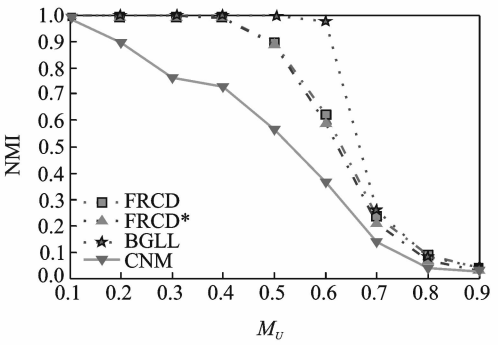
(a) LFR1上的标准互信息比较
(a) NMI results on LFR1



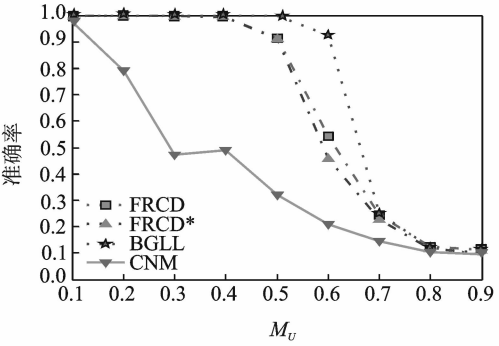
(b) LFR1上的准确率比较
(b) Accuracy results on LFR1

图 1 LFR1 上的实验比较

Fig. 1 Experimental results on LFR1



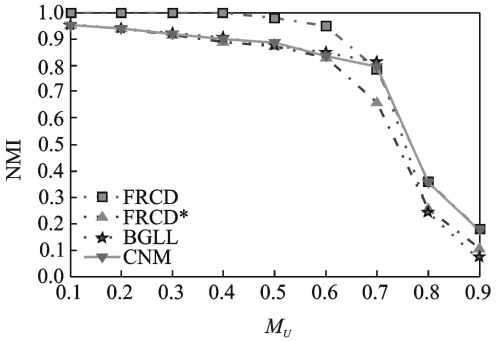
(a) LFR2上的标准互信息比较
(a) NMI results on LFR2



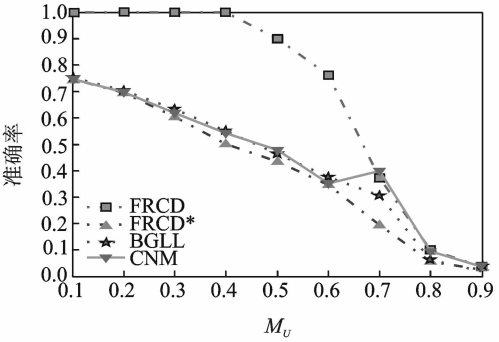
(b) LFR2上的准确率比较
(b) Accuracy results on LFR2

图 2 LFR2 上的实验比较

Fig. 2 Experimental results on LFR2



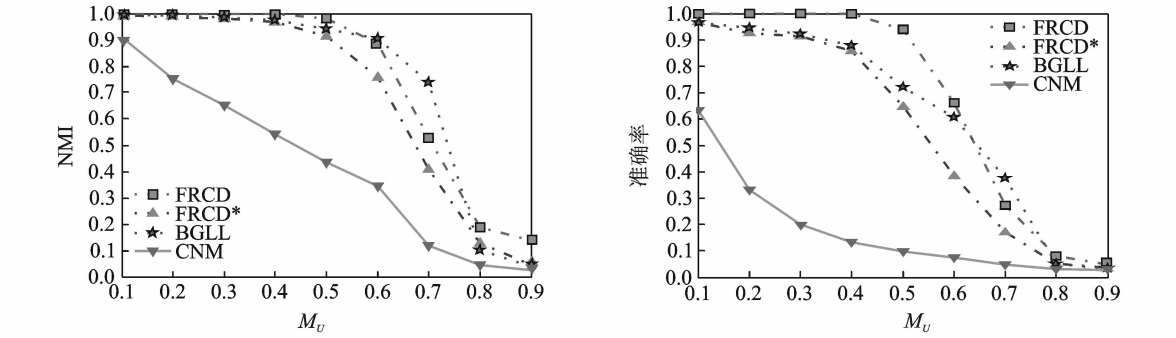
(a) LFR3上的标准互信息比较
(a) NMI results on LFR3



(b) LFR3上的准确率比较
(b) Accuracy results on LFR3

图 3 LFR3 上的实验比较

Fig. 3 Experimental results on LFR3



(a) LFR4上的标准互信息比较 (a) NMI results on LFR4

(b) LFR4上的准确率比较 (b) Accuracy results on LFR4

图 4 LFR4 上的实验比较

Fig. 4 Experimental results on LFR4

通过在人工生成网络上的实验比较可知，当网络混合参数取不同值时，FRCD 算法大部分情况优于 BGLL 和 CNM，尤其是 $M_U=0.1\sim0.4$ 时能得到完全准确划分。但是当 $M_U=0.7$ 时，此时网络社区结构不够清晰，效果不如 BGLL。整体的实验结果可知 FRCD 在静态社区结构发现中，能得到较准确的社区结构。

4.2 动态网络的实验比较

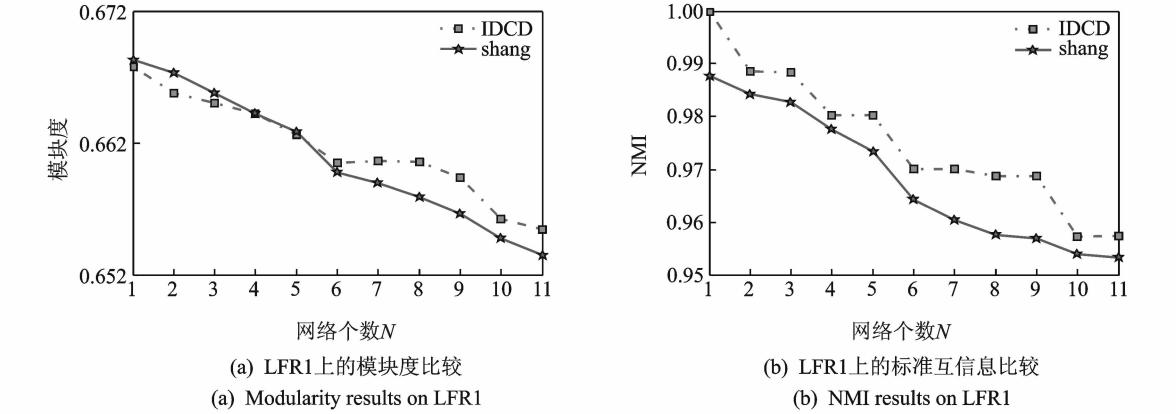
由于现实中的社会网络结构变化都很缓慢，相邻采样时刻之间网络变化很小，因此通过对人工网络随机抽取的方法，模拟增加的网络。具体生成方法如下。

输入：含有 n 个节点的网络 G_0，增量网络个数 N。

输出：初始静态网络，以及后续的增量网络 $G_1,\cdots,G_N$，其中 G_r 为第 r 个增量网络，初始 $r=1$。

(1)随机选取 $\beta_2\times n$ 个节点，随机选择每个节点的 $\sqrt{d_i}$ 条关联的边(d_i 为节点 i 的度)；(2)随机选择 $\beta_2\times n$ 个节点，与其关联的所有边；(3)将步骤 1 和 2 得到的边构成网络 G；(4)将 G_1 从 G_0 中删除，更新 G_0；(5)重复步骤 1～4，直到 $r=N$。

在实验中，设 $\beta_1=0.01$，$\beta_2=0.005$，$N=10$。分别在 LFR1 和 LFR3 中，当 mu=0.3 时，和 Shang 等人提出的方法进行实验比较，由于 Shang 方法具有随机性，因此采用 10 次实验结果的平均值和提出的 IDCD 算法进行比较，横坐标为增加各子集个数，实验结果如图 5，6 所示。

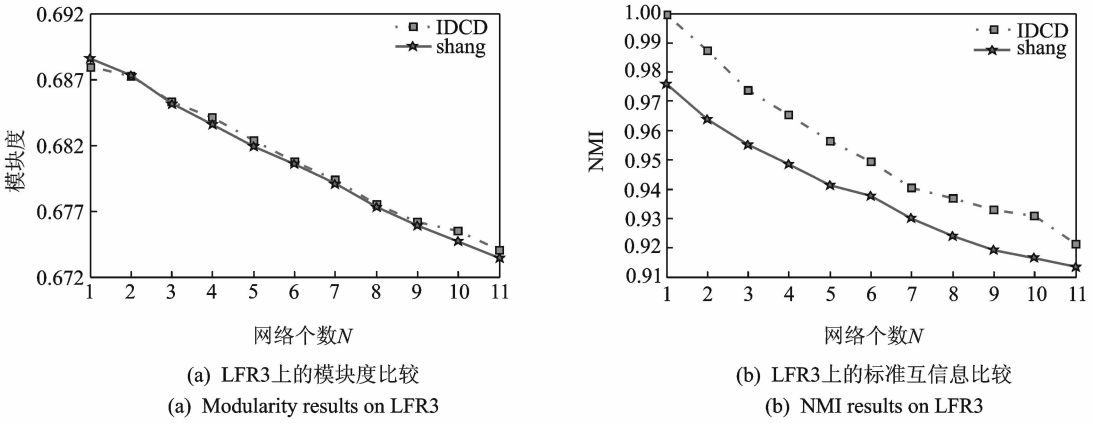


(a) LFR1上的模块度比较 (a) Modularity results on LFR1

(b) LFR1上的标准互信息比较 (b) NMI results on LFR1

图 5 LFR1 上动态社区结果

Fig. 5 Dynamic community detection results on LFR1



(a) LFR3上的模块度比较
(a) Modularity results on LFR3

(b) LFR3上的标准互信息比较
(b) NMI results on LFR3

图 6 LFR3 上动态社区结果

Fig. 6 Dynamic community detection results on LFR3

通过与 Shang 等提出的方法进行比较,可以看到:随着时间的推移,增量式的动态社区发现方法由于存在错误率累积的问题,得到的结果整体呈现变差趋势;本文提出的 IDCD 方法在起初阶段效果略差于 BGLL,但是随着网络规模的增大,得到的效果逐渐优于 Shang 等人的方法,从而说明本文算法在处理动态网络时的有效性。

5 结束语

本文提出了一种新的基于模块度优化的社区发现方法,该方法不仅能够发现静态网络中的社区结构,而且可以对动态增加的网络进行实时的社区结构追踪。通过在实际网络和人工生成网络中与基于模块度最大化的 BGLL 和 CNM 算法进行比较,表明本文提出的静态社区发现方法 FRCD 具有较高的划分准确率。进一步通过在动态网络中与 Shang 等人的方法进行实验对比,说明 IDCD 算法能够有效追踪网络社区结构的变化,对 Shang 方法存在的问题进行了有效改进。另外,本文所给出的算法是一种基于模块度优化的社区发现方法,可以用于优化其他的目标函数,如模块度密度[23]、Surprise[24]等。但本文中的算法在进行动态网络追踪时,只考虑了网络规模不断增大的情形。虽然,这在一定程度上符合实际网络规模的变化,但是在现实世界中也存在部分节点和边在下一个时刻消失的情况。目前,这在动态网络研究中仍是一个研究难点,也是未来要探索解决的问题之一。

参考文献:

[1] Girvan M, Newman M E J. Community structure in social and biological networks[J]. PNAS, 2002,9(12):7821-7826.

[2] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008.

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

[4] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007,76(3):036106.

[5] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences, 2008,105(4):1118-1123.

[6] Tang L, Liu H, Zhang J, et al. Community evolution in dynamic multi-mode networks[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008:677-685.

[7] Tang L, Liu H, Zhang J. Identifying evolving groups in dynamic multimode networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2012,24(1):72-85.

[8] 单波,姜守旭,张硕,等. IC:动态社会关系网络社区结构的增量识别算法[J]. 软件学报,2009,20:184-192.
Shan Bo, Jiang Shouxu, Zhang Shuo, et al. IC: Incremental algorithm for community identification in dynamic social networks[J]. Journal of Software, 2009,20:184-192.

[9] 周耀明,李弼程. 一种自适应网络舆情演化建模方法[J]. 数据采集与处理,2013,28(1):69-76.
Zhou Yaoming, Li Bicheng. Adaptive evolution modeling method of internet public opinions[J]. Journal of Data Acquisition and Processing, 2013,28(1):69-76.

[10] Shang J, Liu L, Xie F, et al. A real-time detecting algorithm for tracking community structure of dynamic networks[C]// The 6th SNA-KDD Workshop. New York, USA: ACM, 2012.

[11] Park J, Newman M E J. The origin of degree correlations in the Internet and other network[J]. Phys Rev E, 2003,68(2): 026112.

[12] Cheng X Q, Ren F X, Shen H W, et al. Bridgeness: A local index on edge significance in maintaining global connectivity[J]. Journal of Statistical Mechanics: Theory and Experiment,2010(10):P10011.

[13] Satuluri V, Parthasarathy S, Ruan Y. Local graph sparsification for scalable clustering[C]// Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. NewYork, NY, USA: ACM, 2011:721-732.

[14] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004,101(9):2658-2663.

[15] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification[J]. Journal of Statistical Mechanics: Theory and Experiment, 2005(9):P09008.

[16] 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.

[17] Zachary W. An information flow model for conflict and fission in small groups1[J]. Journal of Anthropological Research, 1977,33(4):452-473.

[18] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations[J]. Behavioral Ecology and Sociobiology, 2003,54(4):396-405.

[19] Newman M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences, 2006,103(23):8577-8582.

[20] Adamic L A, Glance N. The political blogosphere and the 2004 US election: Divided they blog[C]// Proceedings of the 3rd International Workshop on Link Discovery. NewYork, NY, USA: ACM, 2005:36-43.

[21] Vlasblom J, Wodak S J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs[J]. BMC Bioinformatics, 2009,10(1):99.

[22] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E, 2008,78(4):046110.

[23] Li Z, Zhang S, Wang R S, et al. Quantitative function for community detection[J]. Physical Review E, 2008,77(3): 036109.

[24] Aldecoa R, Marín I. Deciphering network community structure by surprise[J]. PLoS One, 2011,6(9):e24195.

作者简介:

李亚芳(1988-),女,博士研究生,研究方向:数据挖掘,复杂网络分析,E-mail:cyjia@bjtu.edu.cn。

贾彩燕(1976-),女,副教授,研究方向:数据挖掘、生物信息学以及复杂网络分析等。

于剑(1969-),男,教授,博士生导师,研究方向:机器学习、数据挖掘以及图像分割等。

刘光明(1987-),男,博士研究生,研究方向:数据挖掘。