

复杂动态网络演化社团结构探测分析的研究进展^{*}

杨波^{1a}, 游新冬^{1b}, 段文奇²

(1. 杭州电子科技大学 a. 管理学院; b. 计算机学院, 杭州 310018; 2. 浙江师范大学 经济与管理学院, 浙江金华 321004)

摘要: 针对复杂动态网络中演化社团结构的探测, 综述了该新领域的研究进展。首先对演化社团结构探测进行了问题描述, 总结了四种研究思路。重点介绍了其中一些有代表性的分析方法及其特点, 以及衡量探测方法好坏的基准图。最后指出该领域值得进一步探索的研究方向。

关键词: 复杂网络; 动态网络; 社团结构; 网络演化

中图分类号: TP393; N94 **文献标志码:** A **文章编号:** 1001-3695(2013)05-1292-05

doi: 10.3969/j.issn.1001-3695.2013.05.003

Progress on analysis for detecting evolutionary community structure in complex dynamical networks

YANG Bo^{1a}, YOU Xin-dong^{1b}, DUAN Wen-qi²

(1. a. School of Management, b. School of Computer Science & Technology, Hangzhou Dianzi University, Hangzhou 310018, China; 2. School of Economics & Management, Zhejiang Normal University, Jinhua Zhejiang 321004, China)

Abstract: This paper reviewed the new progress of detecting evolutionary community structure in complex dynamical networks. Firstly, it described the problems of the detection of evolutionary community structures and summarized four types of research thought. Then it primarily introduced focusing on some representative methods and their features, and also recommended the benchmark in detecting evolutionary communities. Finally, it pointed out some future directions of this active area.

Key words: complex networks; dynamical networks; community structure; network evolution

近年来在复杂网络研究领域, 网络社团结构受到越来越多的关注。相比于网络中其他部分, 社团结构由于其通常具有的直观的现实涵义以及在影响诸如扩散、抗毁、合作、同步等网络功能方面的显著作用, 而在网络中扮演着重要角色。目前有关网络社团结构的研究主要涉及对社团结构的定义、探测、结构特征及功能的考察^[1], 其中社团结构探测作为此类研究的基础问题而尤其受到关注^[2]。当前研究集中针对于静态网络中的探测算法设计, 相关研究进展已有多篇综述性论文进行了详细的介绍^[3~6]。随着研究的深入, 对复杂动态网络中演化社团结构的探测分析开始受到研究者的关注。事实上, 从静态网络分析转向对动态网络的演化研究是近年来复杂网络研究的一个新趋势。研究者指出, 随时间演化的动态性是实际复杂系统内部关联的重要特点^[7,8]。现有的复杂网络研究仍大多针对某时刻截取的网络快照或多个时间段数据构成的网络合成快照, 其研究视角是静态的, 没有考虑随时间演化网络中的动态信息。上述问题在网络社团研究中同样存在, 社团结构探测也正经历从静态研究向动态演化分析的转变。然而, 受研究者对静态网络社团探测算法的持续关注以及随时间演化的网络数据获取困难的影响, 目前有关动态网络演化社团结构探测的研究仍处于起步阶段^[1]。近期随着一些动态网络数据集的公开化, 研究者开始能够监测实际系统的动态演化过程, 这也使得对网络中社团结构的形成、演化和解体等现象的研究成为可

能。本文关注到复杂网络社团结构探测从静态转向动态研究的趋势, 对目前动态网络演化社团结构的探测中一些有代表性的研究进行了综述, 以期让读者能够了解该新领域的进展并开展相关研究。

1 演化社团结构探测的问题描述及方法思路

对复杂动态网络而言, 当网络动态变化时, 网络中的社团可能随时间推进而发生生成、解体、合并、分裂等事件, 这些事件将导致社团的结构形态发生动态演化。如果把 t 时刻网络 G_t 的社团划分记为 $\mathcal{C}_t$, 其为该时刻网络中所有的社团 C_i 构成的集合, 社团结构探测的任务就是要确定整体的社团划分或者集合中的某个社团在下一时刻的可能结构形态 $\mathcal{C}_{t+1}$ 或 C_{t+1} 。从初始网络社团结构识别开始, 通过探测方法最终识别出动态网络 $\{G_t\}$ 中全部的演化社团序列。

目前文献中演化社团结构探测的研究思路基本可以划分为四种: a) 各自单独识别不同时刻网络社团划分, 并从 $t+1$ 时刻社团出发寻找与之满足一定相似条件的 t 时刻社团; b) 先整合相邻时刻网络再分离 t 和 $t+1$ 时刻的社团; c) 从节点行为分析出发, 根据节点行为对网络的影响来确定社团结构的可能演化; d) 独立地对一定时刻网络快照进行社团划分, 并按发生的时间先后顺序自然排列成演化社团序列。第三种研究思路由

收稿日期: 2012-08-29; **修回日期:** 2012-10-10 **基金项目:** 国家教育部人文社会科学研究资助项目(09YJCZH029); 国家自然科学基金资助项目(71271193)

作者简介: 杨波(1979-), 女, 四川江油人, 副教授, 博士, 主要研究方向为复杂网络、创新及信息扩散(yangbo@hdu.edu.cn); 游新冬(1979-), 女, 讲师, 博士, 主要研究方向为网络计算、复杂网络; 段文奇(1976-), 男, 副教授, 博士, 主要研究方向为复杂网络理论及电子商务领域应用。

于体现预测的思想,较难实现。目前的分析方法基本基于前两种及第四种研究思路。

2 演化社团结构的探测方法

2.1 基于相邻时刻相似度直接比较的演化社团探测

基于相邻时刻相似度直接比较的演化社团探测方法采用第一种分析思路,这也是确定动态网络演化社团序列 $\{C_t\}$ 最直接的方法。其一般做法为:首先对相邻时刻网络 G_t 和 G_{t+1} 分别利用静态网络社团探测算法,确定两个网络中各自的社团划分;然后对所探测到的相邻时刻网络中的社团进行比较,确定在 t 时刻网络中与 $t+1$ 时刻社团 C_{t+1} 满足一定相似条件的社团 C_t ;在此基础上,将 C_{t+1} 加入 C_t 所在的演化社团序列当中。文献[9]最早对动态网络演化社团展开研究,采用的正是上述做法。文中分析对象为动态引文网络,数据取自 NEC CiteSeer 数据库 1990—2001 年信息。首先利用层次聚类算法对动态网络序列中的网络快照进行社团探测。该算法的核心思想是,由相似度最高的社团开始合并,直到所有元素都归入一个社团为止。文中利用信息科学领域常用的相似度测量指标——夹角余弦分别定义了节点间相似性及社团间距离:

$$\text{similarity}(i, j) = \cos(r_i, r_j) = \frac{r_i \cdot r_j}{\|r_i\| \|r_j\|} \quad (1)$$

$$\text{dis}(C, C') = \sqrt{\frac{n_C n_{C'}}{n_C + n_{C'}} (1 - \cos(r_C, r_{C'}))} \quad (2)$$

其中: n_C 为社团 C 的节点规模, r_i 为网络节点(即论文) i 的所有参考文献构成的属性向量, r_C 为社团内所有节点属性向量的归一化和。此外文中把在凝聚过程得到的树状图中,当网络受到小部分节点和边被删除的扰动时,仍仅受轻微影响的社团定义为自然社团。针对利用层次聚类算法所确定的每个网络快照中的自然社团,文中定义了社团间匹配度,并据此寻找相邻时刻最佳匹配自然社团,从而实现对演化社团的追踪。社团间匹配度利用式(3)计算:

$$\text{match}(C_t, C_{t+1}) = \min\left(\frac{|C_t \cap C_{t+1}|}{|C_t|}, \frac{|C_t \cap C_{t+1}|}{|C_{t+1}|}\right) \quad (3)$$

上述演化分析方法的不足主要来自对层次聚类算法的应用。由于该算法聚类结果不稳定,少数次的聚类可能不能从层次树中分类挑选出有价值的社团,故需经多次聚类再分析。此外,文中方法主要针对引文网络,在面向其他类型动态网络的应用中相似性及距离等定义不一定适用,可能需要修正。

文献[10]给出了一个更具一般性的方法描述。定义 $\mathcal{C}_t$ 为 t 时刻网络 G_t 全部社团的集合, $\mathcal{D}$ 为动态网络序列 $\{G_t\}$ 中全部演化社团序列的集合。分析过程如下:

a) 运用静态网络社团结构探测算法从网络 G_1 中确定 $\mathcal{C}_1$ 。对于每个 $C_{1i} \in \mathcal{C}_1$,建立一个以 C_{1i} 作为起始状态社团的新演化社团序列 $D_i = \{C_{1i}\}$ 。初始化 $\mathcal{D} = \{D_1, \dots, D_i, \dots\}$ 。

b) 对后续时刻 $t(t > 1)$ 的网络 G_t ,运用静态网络社团结构探测算法从网络中确定 $\mathcal{C}_t$ 。

c) 对每个 $C_{ta} \in \mathcal{C}_t$ 按下述过程进行处理:首先将 C_{ta} 与每个演化社团序列 D_i 中距 t 最近时刻的社团(记为 F_i)进行相似性匹配比较,寻找满足条件 $\text{sim}(C_{ta}, F_i) > \theta$ 的 D_i ,其中 θ 为匹配阈值, $\text{sim}(C_{ta}, F_i)$ 为两个社团的 Jaccard 相似性系数:

$$\text{sim}(C_{ta}, F_i) = \frac{|C_{ta} \cap F_i|}{|C_{ta} \cup F_i|} \quad (4)$$

如果经比较后均不匹配,则生成一个新的演化社团序列,其以

C_{ta} 为起始状态社团;否则,将 C_{ta} 加入至寻找到的每个与之匹配的社团序列 D_i 当中(按时间顺序 C_{ta} 应排至 D_i 中最后位置)。若出现某演化社团序列 D_k 与两个以上社团匹配,则从 D_k 中分裂生成一个新的演化社团序列 $D_{k'}$,其与 D_k 在 t 时刻之前社团序列的成员完全相同,而从 t 时刻开始二者出现不同。

d) 重复上述 b) 之后的步骤,直到所有的 G_t 都得到了处理。

文献[11]给出了一个与上述分析过程类似的演化分析方法,对其中各种类型社团演化事件给出了更为明确的逻辑定义。

总体而言,上述研究都是在各自单独识别不同时刻网络社团划分的基础上,利用对相邻时刻社团相似度的直接比较来确定演化社团序列。此类方法直观性强,容易理解和实施;但方法的准确度存在一定问题,易产生误判情况^[12]。例如,给定网络 G_t 中两个不存在节点重合的社团 A_t 和 B_t ,当 A_t 生长扩张至足够大时,由其扩张生成的 A_{t+1} 可能会与 B_{t+1} 发生节点重合(但在静态探测算法下二者被判属于不同社团)。这种社团间重合的存在有可能使得 A_{t+1} 与 B_t 的相似度要高于 A_{t+1} 与 A_t 的相似度,因此按照相似度最优原则, A_{t+1} 应被归入 B_t 所在演化社团序列,从而导致了演化社团序列成员的误判。大量经验研究显示,实际复杂网络,特别是社会网络存在明显的社团间重合现象^[1,6],因此利用此类方法来确定这种类型网络中的演化社团其准确性存在问题。

2.2 基于演化聚类分析的演化社团探测

基于演化聚类分析的演化社团探测方法也属于前述第一种分析思路,但与 2.1 节方法存在显著不同。2.1 节方法实际采用的是两阶段的方式,即先确定好不同时刻网络的社团划分,再通过分析相邻时刻社团之间的相似度来确定演化社团的序列。这种两阶段方式下最终所确定的演化社团序列中,相邻时刻社团划分之间可能呈现显著的结构变化。由于这一结构变化可能并非完全来自于网络动态演化所带来的对社团的影响,据此分析社团结构的演化特征会带来偏差。文献[13]在静态网络社团结构探测的聚类技术基础上提出了一个演化聚类分析框架,可用于对规模不变的动态网络序列的演化社团检测。其中演化聚类的基本原则为,根据当前时刻网络结构以及前一时刻网络的社团划分来确定对当前时刻网络的社团划分,其结果既是对当前网络的较好划分,又与先前时刻网络的社团划分的差异较小。令 $\mathcal{C}_t$ 为 t 时刻网络某一社团划分的聚类结果,分别定义聚类 $\mathcal{C}_t$ 的快照质量(sq)和历史成本(hc),前者测度此聚类结果对于 t 时刻网络社团划分的准确程度,后者测度 $\mathcal{C}_t$ 相对于前一时刻所采用聚类结果 $\mathcal{C}_{t-1}$ 的距离/相异性。动态序列 $\{\mathcal{C}_t\}$ 的总质量由各时刻快照质量和历史成本的组合构成。理论上,一个好的划分应具有高快照质量(即它能够对 t 时刻数据进行很好的聚类)以及低历史成本(即它应当与前一时刻的划分相似)。为此,文中提出 t 时刻网络最优社团划分的确定准则,即满足使

$$\text{sq}(\mathcal{C}_t, \mathbf{M}_t) - cp \times \text{hc}(\mathcal{C}_{t-1}, \mathcal{C}_t) \quad (5)$$

最优的聚类 $\mathcal{C}_t$ 即为该时刻网络的社团划分。其中可调参数 cp 为相对权重, $\mathbf{M}_t$ 为 t 时刻网络中节点的距离/相似度矩阵。利用上述准则,可从 $\mathcal{C}_{t-1}$ 和 t 时刻关系数据中确定 $\mathcal{C}_t$,从而得到演化社团序列 $\{\mathcal{C}_t\}$ 。上述演化聚类过程的输入参数为 $\{\mathbf{M}_t\}$ 。在静态网络聚类技术中,距离/相似度有多种定义方法^[3],如最短距离法、最长距离法、平均距离法等,演化聚类分析过程中视聚类技术的要求选择采用。实际应用当中,还可采用此距

离/相似度矩阵的改进版本,对跨特定时间窗口的数据取(加权)平均,以增强关系数据对于噪声的鲁棒性,也使得聚类过程的结果更可靠。此外,最优准则中的快照质量和历史成本也可视实际应用的需要而选择采用不同方式测量。

在文献[13]提出的演化聚类的分析框架下,那些已有的在静态网络上适用的聚类技术均能够进行重新表述。文献[13]提供了层次聚类技术和K-均值聚类技术的演化分析版本。其他的研究如文献[14,15],则在演化聚类分析框架下对谱分析聚类技术进行了重新表述。此外,文献[16]引入了一个可允许节点同时属于多个社团的分析框架 FaceNet。该文中快照质量采用 M_i 与 C_i 间的 Kullback-Leibler (KL) 散度来测度,历史成本则采用 C_{i-1} 和 C_i 间的 KL 散度来测度。与文献[13]仅针对规模不变的动态网络序列的演化研究相比,FaceNet 的应用范围可扩展至处理网络规模和社团数量可变的情况。但它不能够解释社团的生成和解体;由于为达到收敛而需重复进行的矩阵计算次数较多,该方法也不适用于大规模系统的分析。

总体而言,上述基于演化聚类技术的演化社团分析把对网络最优社团划分的探测和演化社团序列的成员确定结合起来同时进行。所得演化社团序列结果既保证了序列中相邻时刻社团之间的紧密联系,又确保了其对网络划分的准确度。但该方法给出的是网络社团整体结构的演化序列,而非针对社团划分中某个社团的具体演化分析。

2.3 基于派系过滤算法的演化社团探测

演化社团探测的第二种分析思路来自于文献[12]基于派系过滤(CPM)算法提出的一个新的演化社团探测方法。该方法的设计初衷之一是针对前述直接比较方法当中由于社团间重合导致的演化社团序列成员的误判问题。该方法虽然也是考虑社团相似度最大,但不是直接比较相邻时刻网络社团,而是通过先将相邻时刻网络进行组合,再实施分解。

文献[12]的方法中使用到派系过滤算法,该算法是静态网络社团检测的重要方法,可用于分析具有重合性的社团结构。CPM 算法认为社团是由一系列相互可达的 k -派系(规模为 k 的完全子图)组成的。算法通过合并相邻的 k -派系来实现社团发现,而那些处于多个社团中的节点即是社团间重合的部分。基于 CPM 算法,文献[12]的方法首先将网络动态序列 $\{G_t\}$ 中任相邻两个网络 G_t 和 G_{t+1} 合并为新网络 $G_{t,t+1}$ 。对网络 $G_{t,t+1}$ 运用 CPM 算法进行社团探测,对识别出的每个 CPM 社团 M_k ,从 M_k 中根据节点和边的归属分离出分别属于网络 G_t 和 G_{t+1} 的 CPM 社团 C_k^t 和 C_k^{t+1} 。 $\{C_k^{t+1}\}$ 中任一社团在时刻 t 的形态就是 $\{C_k^t\}$ 中与它有最大相似度(文中称其为相对重合量)的社团。相似度由式(6)计算:

$$S(C, C') = \frac{|C \cap C'|}{|C \cup C'|} \quad (6)$$

利用所提出的演化社团探测方法,文中对电话用户呼叫网络和论文合著网络的动态数据进行了演化研究。

由于所采用的 CPM 算法可识别存在的重合性社团,上述演化社团探测方法避免了此类情景下误判断情况的发生,对演化社团序列成员的识别更为准确。该方法的不足之处主要在于,方法实施过程中网络中的社团只能通过 CPM 算法进行探测,这使得需通过运用其他的社团划分技术来对演化社团序列结果进行交叉验证,以提高结果的可信度。

2.4 基于节点行为趋势分析的演化社团探测

基于节点行为趋势分析的演化社团探测方法属于前述演化社团探测的第三种分析思路,即将社团的动态演化归因为节点的行为作用,从节点行为分析出发,根据节点行为对网络的影响来确定社团结构的可能演化。文献[12]较早针对动态网络研究社团内节点弃团及居留社团行为。文中将社团内节点的关联边划分为两种类型:a)该节点与社团外节点的关联边(社团外边);b)其与社团内节点的关联边(社团内边)。文章认为社团内节点弃团概率及节点花在社团的时间均与节点对社团的承诺水平有关。为考察这些变量间关系,采用指标 $w_{out}/(w_{out} + w_{in})$ 来刻画节点对社团的承诺水平,其中 w_{out} 为该节点社团外边的权重和, w_{in} 为其社团内边的权重和。该指标值越大,节点对社团的承诺水平越低。研究发现,此时节点弃社团概率越大,同时花在社团的时间越短。在此基础上,进一步考察了社团解体概率及生命周期与指标 $W_{out}/(W_{out} + W_{in})$ 之间的关系,分别发现存在正相关和负相关性,其中, W_{out} 为所有社团成员节点的全部社团外边的权重之和; W_{in} 为所有社团成员节点的全部社团内边的权重之和。文献[17]分析了免费在线社区 LiveJournal 的社团演化动态,重点考察了个体加入社团的行为,研究发现行为概率与已在社团内的朋友/合著者的个数及他们的互连程度成正比。进一步分析得到社区中社团的生长概率与包括社团规模、社团成员的社团外朋友数以及二者的比率在内的因素组合正相关。

从节点演化行为分析出发,文献[18]提出了一个更为完整的分析框架——基于事件的网络演化分析框架。文章将网络中演化事件区分为涉及社团的演化事件和涉及节点的演化事件。除合并、分裂、生成、解体这些已被定义的社团演化事件外,文中另引入节点演化事件,并区分为出现型(节点首次加入网络中)、消失型(t 时刻网络中的节点 $t+1$ 时刻不再在网络内)、进入型($t+1$ 时刻集合组中节点 t 时刻不在该集合组内)、退出型(与加入型相反)。基于这些事件类型的定义,提出了四个测度指标,试图捕捉促成网络演化的节点行为趋势,包括:a)稳定性指数,测度节点随时间演化保持与相同节点关联的趋势;b)社交能力指数,测度节点参与不同关联的数目,可用该节点所涉进入型和退出型事件的次数来刻画;c)人气指数,该指标针对社团设计,用给定时段内被该社团吸引的节点数来刻画;d)影响力指数,测度节点对其他节点的影响力,用受该节点影响而离开或加入社团的节点数来刻画。通过对论文合著网络和临床试验网络的动态数据分析显示,上述行为测度指标可对网络随时间演化的情况作出可靠预测,包括诸如对缺失边的推断。

总体而言,目前上述研究还并非主要针对演化社团预测,而更多集中于对节点演化行为特征的分析。但从根本上来说,网络及社团的动态是由节点的演化行为促成的,如由于节点从网络(社团)的退出而导致网络(社团)在下一时刻结构发生改变。因此如果能够清楚节点这些与社团相关的行为特征,则可以对下一时刻社团结构的可能变化作出预测。可以预期,这种通过对节点行为趋势的分析来确定网络中社团的可能演化结构的思路,将促成新的演化社团检测方法的提出。

2.5 基于信息压缩方法的演化社团探测

基于信息压缩方法的演化社团探测方法采用了演化社团检测的第四种研究思路,即独立地对一定时刻网络快照进行社

团划分,并按发生时间先后顺序自然排列成演化社团序列。该方法利用到了信息理论。信息论在静态网络社团结构检测研究中已有应用^[6],它将网络的模块化描述看做对网络拓扑结构的一种有损压缩,从而将社团划分问题转换为信息论中的一个基础问题:对拓扑结构有效压缩方式的寻找。上述信息压缩的方法也可应用于演化社团检测。

文献[19]运用最小码原理(MDL)来探寻描述网络时间序列及社团划分的最小编码成本,在此基础上确定动态网络序列中变异点位置,并识别由变异点分割开的网络快照中的最优社团划分。文章分析的网络对象为随时间演化的二分图,其实际例子如消费者—产品网络、出发地—目的地交通网络、发送者—接收者通信网络等。文中所谓变异点是指在网络动态演化过程中发生大的结构改变的短时周期,这种结构改变通常对应于一些极端事件的发生。根据变异点所处位置,网络时间序列可分割为若干片段,每个片段包含系统中一定数目的连续快照。同一片段中的网络被认为具有相同的模块结构,因为它们代表了系统发展历程中的相同时期。文中定义网络片段的编码成本为两部分之和,包括片段中网络的划分编码成本及片段经划分后形成的子网络片段的压缩熵。而网络时间序列的总编码成本则为各片段编码成本之和。通过最小化总编码成本,不仅可发现每个片段的最佳模块划分(高模块度对应于划分的低编码成本),也可发现将网络时间序列分为多个片段的最为紧致的分割方式,使得同一片段中的各网络相互间强关联。而后一特点可允许在系统的发展历程中识别出变异点的位置。

鉴于总编码成本的最小化是个NP难题,文献[19]进一步提出了一个近似方法,称为GraphScope。该方法包括两步,首先寻找每个片段内网络的最佳划分,然后寻找这些片段的最佳分割,两种情形下的最佳结果均对应于编码成本最小化。其中,为避免全局搜索方式下搜索参数空间过大而无法穷尽分析的不足,文中采用局部搜索方式来寻找编码成本最小化的最佳社团划分。GraphScope最大的好处在于,不需要诸如社团规模和个数的任何输入参数;并且其也适合在随系统演化新的网络配置会实时加入的流式传输环境中应用,因为该方法下处理一个网络快照的计算复杂度平均而言随时间是稳定的。通过对包括电子邮件联系网络、手机通信网络等实际演化数据集的分析显示,GraphScope能够有效发现有意义的社团以及时间序列中的变异点。

2.6 基于自助重取样技术和冲击流图的演化社团探测

文献[8]引入自助重取样技术(bootstrap resampling)和冲击流图(alluvial diagram),结合静态网络的传统聚类技术,提出了一个针对大规模动态网络结构变化的分析方法。文章首先肯定了目前采用的网络聚类技术对于识别和刻画网络中社团的重要作用。同时指出,网络方法所研究的很多实际系统,本质上而言是独一无二的,不能复制观察,如某年度的全球航空运输网络有且仅有一个。因此,无法通过分析多个样本来构建统计显著性。此外也不能依赖于时间稳定性,虽然那些随时间保持不变的结构或许在统计上显著,但研究者通过寻找稳定特征无法发现那些重要的变化。

鉴于上述问题,文献[8]提出引入自助重取样技术和冲击流图。自助重取样技术的主要特点是通过从所观察对象的经验分布中重新取样来评估估计的准确性。应用在该文中,是通过在现有网络样本基础上重新取样生成新的网络,从而扩大样本容量以评估各时刻网络社团结构检测结果的准确性,获得统

计上显著的社团划分。此外,冲击流图是该文为简化并突出对动态网络中演化社团结构变化的刻画而采用的图示方式,图中横向每一栏对应于不同时刻网络具有的显著性的社团划分,纵向上不同颜色块对应于该划分下的社团,块高度代表社团规模。通过绘制冲击流图,可发现社团合并、分裂等演化趋势。

整体上而言,文献[8]的分析方法采用的也是演化社团检测的第四种研究思路,其完整的演化分析过程共有四个步骤:a)对每个时刻所观察到的原网络进行聚类,得到原网络的社团划分;b)对每个时刻的网络运用自助重取样技术生成多个复制网络,对这些网络进行聚类得到社团划分;c)将这些网络的社团划分与原网络社团划分进行比较,确定原网络社团划分的显著程度;d)针对具显著性的原网络社团划分,生成刻画不同时刻间结构变化的冲击流图。利用上述演化分析方法,作者考察了约7 000种科学期刊之间十年的引用模式,数据来自1997—2007年Thomson-Reuter期刊引文报告。

3 动态评判基准图

在动态网络演化社团的研究中,基准图的构建是一项非常重要的工作。利用一些已知社团结构特性的基准图可以实现对各种算法性能的检验和评估^[3,4]。目前文献中的动态基准图多通过理论模型的构建来生成。较早的动态基准图构建主要采用经过成员资格的简单转换而合成的人工数据。文献[15]提出了一个能够生成小规模的多模式动态网络合成数据的方法。该方法首先产生一个由若干社团的集合构成的网络,并在每一时刻随机改变一定比例社团内成员的资格。与上述做法相似,文献[20]生成了一个随机加权有向网络的合成数据序列,该序列中每个网络都内嵌有社团结构,每隔四个时间片段社团结构发生一次改变。文献[16]也在两个已有网络基础上,通过社团内成员的资格转换以及相应改变节点边,生成具有不同模块度和社团个数的演化网络,以及稳定性不同的动态网络序列。

上述动态基准图的构建过程易于理解,但方式过于简单,仅考虑了节点层面上的进入或退出演化事件。这种方式下,虽然可以跟踪节点行为带来的社团结构变化的演化路径,但不能明确地呈现出诸如社团的生成、解体、合并、分裂等更高层次的社团演化现象。而要检验方法效果,需要考察其对各种社团演化现象下演化社团序列的识别程度。因此,有效的动态基准图应能够清晰呈现多样化的演化现象。

文献[10]提出了一个在合成网络数据基础上内嵌多种社团演化事件而生成动态基准图的方法。文中用来生成动态基准图的基础合成网络采用了文献[21]的静态网络基准图生成模型(简称LF模型)。该模型的突出特点是允许节点度和社团规模的异质性,允许社团间节点重合,并具备良好的扩展性,能生成无权无向网络、有向网络、加权网络等多种形式的基准图。利用LF模型,文献[10]通过参数设置生成了三种类型基准图作为基础网络数据,包括具有非重合社团的基准图、具有少量社团重合同时社团间连接水平较高的基准图以及具有高度社团重合的基准图。在此基础上,文中针对四种类别社团演化事件分别构建动态基准图序列:a)针对间歇社团序列(演化路径上可能缺失部分时刻的社团)的刻画,规定每一时刻从初始社团划分 $\mathcal{C}_1$ 的集合中随机隐去一定比例(10%)的社团;b)针对社团扩张和缩减现象,规定每一时刻从基础网络中随机选择一定数目(40个)社团,将社团规模扩张或缩减25%,其中

的扩张过程中新的社团成员从其他社团内随机选择;c)针对社团生成和解体现象,规定每一时刻通过从其他既有社团中移除节点的方式额外构建一定数量(40个)新的社团,同时随机移除同样数量的既有社团;d)针对社团合并和分裂现象,在初始网络社团结构的基础上,规定后续每一时刻随机指定一定数量(40个)既有社团进行分裂,同时指定同样数量的成对社团合并。除上述设计外,文中还规定四种情况中均通过在每一时刻随机变更一定比例节点的社团成员资格来仿真那些随时间演化个体在社团间的自然运动现象。

4 结束语

现有大量研究已证实,社团结构对于理解网络整体结构和网络功能具有重要作用。鉴于网络动态性的客观存在,有关动态网络社团结构的演化分析日益受到研究者的关注。追踪社团结构的时间演化对于揭示社团的产生及随时间相互作用的内在机理非常重要,而这首先就要解决对演化社团的探测问题。本文总结了目前演化社团探测的四种研究思路,并概述了不同思路下一些具体的探测分析方法的做法和特点。总体而言,虽然对动态网络演化社团的探测分析已引起研究者的关注,但这方面的研究还刚起步,尚存在很多需深入的问题。

a)需明确在确定演化社团序列的过程中,更应关注的对象。是更侧重于相邻时刻社团间联系的探测,还是更关注于特定时刻网络的最优社团划分,而仅按时序将其自然排列为演化社团序列,忽视对各种形态的具体社团演化现象的探析?解答上述问题,是设计有效的演化社团探测方法需明确的前提之一。笔者认为,有效的演化社团探测方法需要既能准确刻画当前时刻网络的社团结构,又确保其与先前时刻的社团联系紧密。目前的研究中,基于演化聚类技术的方法有此效果,虽然其还限于分析社团结构整体的演化,不能识别多样化的单个演化社团序列,但其可以作为后续探测更为有效的演化社团探测方法的有利出发点。

b)在静态网络的社团探测算法研究中,研究者已明确指出,针对不同类型网络应选择恰当的算法,以降低算法复杂度并提高结果有效性^[1,6]。上述观点对于演化社团的探测而言,同样是成立的。一些因素诸如网络规模、动态性的强度、社团重合度等都可能影响演化社团探测方法的选择。因此,有必要研究针对特定类型网络,如何选择更为有效的探测方法。由于目前动态网络演化社团研究仍处于起步阶段,研究者还侧重于探测方法的提出,对于方法的复杂度、有效性的分析及不同方法的对比及选择等方面的关注还有待加强。

c)对于动态网络演化社团结构的探测而言,基准图的构建也非常重要。与静态网络中可用基准图的丰富类型相比,动态网络演化社团的基准图还很缺乏,不多的几项研究也仅限在个别文献中应用,并以理论模型为主。如何完善现有理论模型,同时探寻从现实世界中构建有效的动态基准图的可能性,均值得进一步研究。

d)在动态网络演化社团的研究中,探测方法是重要的基础问题但并非全部。在识别得到的动态网络演化社团序列基础上,对网络中一些重要群体的演化行为进行跟踪分析,探索其形成和演化的内在机理十分必要,也是演化研究未来重要的探索方向。

参考文献:

- [1] FORTUNATO S. Community detection in graphs[J]. *Physics Reports*, 2010, 486(3-5): 75-174.
- [2] AHN Y Y, BAGROW J P, LEHMANN S. Link communities reveal multiscale complexity in networks[J]. *Nature*, 2010, 466(7307): 761-764.
- [3] 李晓佳, 张鹏, 狄增如, 等. 复杂网络中的社团结构[J]. *复杂系统与复杂性科学*, 2008, 5(3): 19-42.
- [4] 汪小帆, 刘亚冰. 复杂网络中的社团结构算法综述[J]. *电子科技大学学报*, 2009, 38(5): 537-543.
- [5] LANCICHINETTI A, FORTUNATO S. Community detection algorithms: a comparative analysis[J]. *Physical Review E*, 2009, 80(5): 056117.
- [6] 骆志刚, 丁凡, 蒋晓舟, 等. 复杂网络社团发现算法研究新进展[J]. *国防科技大学学报*, 2011, 33(1): 47-52.
- [7] Da COSTA L F, Jr OLIVEIRA O N, TRAVIESO G, et al. Analyzing and modeling real-world phenomena with complex networks: a survey of applications[J]. *Advances in Physics*, 2011, 60(3): 329-412.
- [8] ROSVALL M, BERGSTROM C T. Mapping change in large networks[J]. *PLoS ONE*, 2010, 5(1): 86-94.
- [9] HOPCROFT J, KHAN O, KULIS B, et al. Tracking evolving communities in large linked networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101(1): 5249-5253.
- [10] GREENE D, DOYLE D, CUNNINGHAM P. Tracking the evolution of communities in dynamic social networks[C]//Proc of the 2nd International Conference on Advances in Social Networks Analysis and Mining. Washington DC: IEEE Computer Society, 2010: 176-183.
- [11] TAKAFFOLI M, SANGI F, FAGNAN J, et al. Community evolution mining in dynamic social networks[J]. *Procedia Social and Behavioral Sciences*, 2011, 22(10): 49-58.
- [12] PALLA G, BARABASI A L, VICSEK T. Quantifying social group evolution[J]. *Nature*, 2007, 446(7136): 664-667.
- [13] CHAKRABARTI D, KUMAR R, TOMKINS A. Evolutionary clustering[C]//Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 554-560.
- [14] CHI Yun, SONG Xiao-dan, ZHOU Deng-yong, et al. Evolutionary spectral clustering by incorporating temporal smoothness[C]//Proc of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 153-162.
- [15] TANG Lei, LIU Huan, ZHANG Jian-ping. Identifying involving groups in dynamic multimode networks[J]. *IEEE Trans on Knowledge and Data Engineering*, 2012, 24(1): 72-85.
- [16] LIN Yu-ru, CHI Yun, ZHU Sheng-huo, et al. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks[C]//Proc of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 685-694.
- [17] BACKSTROM L, HUTTENLOCHER D, KLEINBERG J, et al. Group formation in large social networks: membership, growth, and evolution[C]//Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 44-54.
- [18] ASUR S, PARTHASARATHY S, UCAR D. An event-based framework for characterizing the evolutionary behavior of interaction graphs[C]//Proc of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 913-921.
- [19] SUN Ji-meng, FALOUTSOS C, PAPADIMITRIOU S, et al. GraphScope: parameter-free mining of large time-evolving graphs[C]//Proc of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 687-696.
- [20] DUAN Dong-sheng, LI Yu-hua, JIN Ya-nan, et al. Community mining on dynamic weighted directed graphs[C]//Proc of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management. New York: ACM Press, 2009: 11-18.
- [21] LANCICHINETTI A, FORTUNATO S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. *Physical Review E*, 2009, 80(1): 016118.