在线社会网络的动态社区发现及演化

王 莉[1),2)] 程学旗[2)]

[1)](太原理工大学计算机科学与技术学院 太原 030024)

[2)](中国科学院计算技术研究所 北京 100190)

摘 要 在线社会网络研究中,动态隐含社区或群组结构的发现及演化探测是一个十分关键的核心问题,它对于在中观(Mesoscopic)视图观察在线社会网络隐结构特征、预测演化趋势、掌控网络势态、发现网络异常群体事件等具有重要意义.文中首先分析了动态社区发现和社区演化研究的关系,给出动态社区研究中关键挑战问题;然后根据问题背景的不同,从"同构社会网络的动态社区研究"和"异构社会网络的动态社区研究"两个方面进行国内外相关研究现状的阐述和分析,其中,在"同构社会网络的动态社区研究"中,根据评价方法的差异和关注问题的不同将当前相关研究分为基于时空独立评价、时空集成评价、统一评价和增量式算法 4 大类进行综述,同时对动态社区发现的重要应用——异常群体发现的研究进行介绍;最后对在线社会网络动态社区领域的难点和发展趋势进行分析和展望.

关键词 在线社交网络;动态社区发现;社区演化;统计推断;异常群体发现;社会计算

中图法分类号 TP399 **DOI 号** 10.3724/SP.J.1016.2015.00219

Dynamic Community in Online Social Networks

WANG Li[1),2)] CHENG Xue-Qi[2)]

[1)](*College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan* 030024)

[2)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract It is an important issue in online social networks to detect hidden communities and track their evolution process, which will help understanding the latent topology, predicting its evolution trend, discovering abnormal events and controlling the network. We firstly give the explanation of the relationship between community detection research and community evolution research, and put forward their main challenges. Then we introduce the related research from two different angles, one is dynamic community in homogenous social networks and the other is that in heterogeneous social networks. To clearly state the first area, we introduce the related work by dividing them into 4 classes on the evaluation mechanism: temporal-spatial independent evaluation based, temporal-spatial integrated evaluation based, unified evaluation based and incremental algorithms. An important application is also reviewed that is detection abnormal swarm events. At last some future research topics are given.

Keywords online social networks; dynamic community detection; community evolution; statistical inference; abnormal swarm detection; social computing

---

收稿日期:2013-05-30;最终修改稿收到日期:2014-09-15.本课题得到国家"九七三"重点基础研究发展规划项目基金(2013CB329602)、国家自然科学基金重点项目(61232010)、国家"八六三"高技术研究发展计划项目基金(2014AA015204)、第 53 批中国博士后科学基金资助项目(2013M530738)、山西省自然科学基金项目(2014011022-1)资助.**王 莉**,女,1971 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为人工智能、社会计算、网络通信.E-mail: l_lwang@126.com. **程学旗**,男,1971 年生,博士,研究员,中国计算机学会(CCF)高级会员,主要研究领域为网络科学与社会计算、网络信息安全以及互联网搜索与挖掘.

1 引 言

在线社会网络将用户在网络虚拟空间关联起来,扩展了人类交流、信息共享及社会活动空间,正在成为当前最具影响的一类互联网应用.根据应用类型不同,在线社会网络可以分为社交网络和社会媒体网络两大类.社交网络关系结构一般是现实世界中朋友、同事、亲属等双向亲密关系在网络上的延伸,呈现出强关系结构,具有显式的社区或朋友圈等群组形态,典型应用有 Facebook、人人网、QQ 等即时通信和在线社交类应用服务等.社会媒体网络以关注、引用等单向关系构成关系结构,为弱关系结构,呈现出核心-边缘、聚团等多类型结构的混合形态,典型应用有新浪微博、Twitter、博客、论坛及其他共享空间类等应用[1].

作为人类真实世界在网络虚拟世界的一种延伸,无论社交网络还是社会媒体网络,"物以类聚,人以群分",群组结构是在线社会网络的一个重要结构特征,也是中观(Mesoscopic)尺度观察和理解网络拓扑的一种重要结构.表示在图形中,就是存在这样的子图,子图内个体之间的关系比较紧密,而子图间个体的关系比较稀疏.Newman[2]把满足这一特征的子图结构称之为社区结构.相关术语有社区、聚团或群组结构等[2-5],本文对这些概念不加区分,统一称之为社区.

在线社会网络中,社会个体通过各种连接关系构成"关系结构",基于关系结构大量信息交流、传播形成"交互结构",并影响"关系结构"[1].社区结构在"关系结构"层和"交互结构"层同时存在,但具有不同的形成机制和表现方法,分别称为显式社区和隐式社区.

显式社区发生在"关系结构"层,表现为由各种静态社会关系构成的显式子群网络,例如,熟人、朋友、同学、同事等物理社会关系映射到网络空间的好友群、邮件群组等,因兴趣相似组成的论坛,"关注"关系所构成的微博粉丝群等.对于显式社区的研究主要是基于真实数据进行实证研究,发现线上社区结构随时间的演化情况或社区成员加入、离开社区的情况[3,6],进而进行结构预测或为提高网络管理质量提供辅助决策意见.

隐式社区发生在"交互结构"层,表现为网络上信息交互、通信、传播等各种行为所产生的隐含子群结构特征,例如,QQ 好友中经常聊天的紧密子团,微博转发和评论中的话题团等.隐式社区结构受显式社区结构的限制,一般为包含于显式社区结构中的紧密子集,在更深程度上反映了社会网络中隐含的真实关系,所以,当前社区发现的相关研究大部分是面向隐式社区的发现,如果不加特别说明,本文的社区默认为隐式社区.

社区结构的重要性使得众多研究者们提出各种社区发现方法,他们将网络观察数据进行叠加集成以发现静态社区结构,掩盖了不同时间点节点关系变化情况,无法探测到网络结构的演化过程.所以,动态社区发现及演化的研究成为当前在线社会网络研究的热点.

动态社区是中观视图分析和观察网络演化情况的重要基础结构,在影响力分析、信息传播、网络营销等应用中具有重要价值.另外,网络群体事件的重要影响,也使得发现隐含群体结构、探测隐含群体结构的演化过程,对于网络群体事件监测、舆情预警等具有重要价值.美国亚利桑那大学、斯坦福大学、康奈尔大学、卡耐基梅隆大学、清华大学、中科院等国内外许多大学和研究机构都在此领域展开了深入研究,一些重要的研究成果频频出现在《Science》、《Nature》、PNAS、WWW、KDD、AAAI 等国际重要期刊和会议上.

本文对在线社会网络中动态社区发现及演化的研究进行综述,组织结构为,首先对问题的基本概念和关键挑战性问题进行阐述,然后根据问题背景的不同,从"同构社会网络的动态社区研究"和"异构社会网络的动态社区研究"两个方面进行相关国内外研究现状的阐述和分析,并对动态社区发现的重要应用——异常群体发现的研究进行介绍,最后给出研究难点和发展趋势.

2 基本概念

2.1 动态社区发现和社区演化

根据在线社会网络中观察到的节点属性和交互信息,学者们提出多种方法进行动态社区发现及演化的研究.从研究目标、研究方法上看,动态社区发现和社区演化是有一定区别的.动态社区发现以发现不同时段社区结构为主要目标,主要研究的是用什么方法以揭示社会网络中隐含的不断动态变化的社区结构,尤其是动态网络中核心稳定社区结构[7-19];社区演化以观察隐含社区结构变化过程为目标,主要着眼于评价不同时间片断中社区结构的

变化情况[20-33]. 体现在研究方法上，动态社区发现的研究一般有一个假设前提，即认为社区变化是平缓的，网络变化中存在核心稳定社区结构；而社区演化的研究并没有这个假设，旨在观察社区产生、发展、突变及消失的生存周期.

虽然在研究目标和方法上有所差别，但社区发现和社区演化都是在空间维度发现隐含社区结构的基础上，扩展时间维度，以发现不同时间点或时间窗口的社区信息和社区变化情况，所以，分时间片(snapshot)进行计算是它们共同的技术路线. 另外，社区演化虽然以观察社区变化情况为目标，但它需要首先发现隐含的社区结构，然后才能分析判断社区变化的过程；动态社区发现虽然以揭示不同时间窗口隐含社区结构为目标，但相邻时间窗口的社区结构之间并非毫无关系，是具有一定的时序演化性质的，所以，动态社区发现不能忽略社区演化的基本属性. 因此，当前许多研究工作不再对这两类研究严格加以区分. 本文对这两类问题的研究现状和技术路线也统一进行分析和综述，并统一称之为动态社区研究，其中较为偏重于动态社区发现和演化研究的方法介绍，对于实证研究所发现的动态社区演化的规律不在本文进行具体讲述.

2.2　关键挑战

动态社区研究的主要技术路线为基于时间片划分的算法设计、社区结构评价及分析，主要步骤为：首先对社会网络数据进行时间片划分，按照一定时间窗口集成数据，然后计算各时间片的社区结构，对相邻时间片社区结构的关系进行分析，最后得到社区变化情况、发现结构演化模式和演化异常点. 这种方法简单直观，是社会学、物理学、计算机领域研究者经常采用的方法，但是，在这类研究路线中存在有若干关键挑战：数据质量问题、时间窗口设定问题、评价问题、社区个数设定和演化问题等.

(1) 数据质量问题

在线社会网络中，将动态变化的网络按照时间片划分或时间窗口集成，最直接面对的一个问题就是数据的可用性和够用性问题. 由于分时间窗口收集数据，数据规模的降低使得原来那些因观察或收集数据的技术局限性所带来的噪音数据、缺失数据的影响被放大，在线社会网络结构固有的稀疏性使得分窗口中的数据稀疏性问题更为显著，由此影响了计算结果的有效性.

(2) 时间窗口设定问题

在线社会网络的主体是现实社会中的人，在线网络中人的交互行为受时间、物理环境、网络环境、个体习惯、社会环境等多方面的影响，具有随机性和复杂性. 对动态网络数据进行人为的时间窗口设定，将网络动态演化过程离散为若干窗口的集成数据，时间窗口长度和时间划分点的选择合适与否将会在很大程度上影响社区结构发现和演化分析的质量. 时间窗口设定过大，可能会将社区的重要变化信息淹没在窗口集成数据中；窗口过小，窗口数据可能会非常稀疏，无法发现重要中观结构信息，同时过细划分还会增加计算复杂度. 时间划分点的选取也非常重要，事件的发生会引发一系列的交互行为，形成一定的网络结构，时间划分点选在事件初始、事件中、事件后所得到的网络结构是具有差异性的，所以，抽样时间窗口选择得是否合适，将极大影响算法发现结果的真实性和计算性能.

有一些学者针对特定数据集，采用统计和数据分析的实证方法对时间窗口问题进行了研究，例如 Clauset 等人[21]认为时间窗口受日历等生活周期的影响，提出 4.08 是一个较好的时间窗口，但是由于不同应用背景的数据差异性以及数据背后人们生活习俗等的差异性，该研究结论是否具有普适性，没有定论，也缺乏科学理论依据. 另外，时间窗口设定除了和实际应用情况相关外，与收集的数据集的数据分布也相关，当前研究工作一般综合了应用问题、数据分布状况和人的经验来设定时间窗口.

(3) 评价问题

动态社区研究中，需要对发现的社区质量和相邻时间的社区变化情况进行评价. 不同问题场景中，社区定义不同，社区质量评价方法也会不同，从而设计产生不同的社区发现算法，例如，由模块度评价引发的基于模块度的社区发现算法[34-37]，由流评价引发的基于图切(graph cut)的方法[38]，由谱评价引发的基于谱聚类或基于拉普拉斯矩阵的方法[39]以及基于信息论的方法[40]等. 这样就产生这样一种现象，同一个网络，用不同社区发现算法得到不同的社区结构，不同评价体系中得到不同的最优社区结构，那么哪一种才是真正的隐含结构，成为一个令人困惑的问题. 这折射出两个问题：① 在评价标准中获得高评价的社区结构未必和现实情况拟合；② 评价标准的相互之间不能有效支持.

动态社区演化中另一个重要的评价问题就是社区演化情况的判别问题. 一般方法是建立社区相似度计算方法，并设定一定阈值和规则，然后根据相邻时间点的社区相似度进行评价，判别社区合并、分

裂、缩小、增大、产生和消失等各种情况.当前存在多种社区相似度计算方法,典型的有基于 Jaccard 系数[30]、多结构特征综合评价的方法[31]、归一化互信息 NMI[41](Normal Mutual Information)等,合适评价方法的选择以及评价中阈值的设定等都是当前的难点问题.

无论社区质量评价还是演化评价,这些评价方法的一个共同点是,都事先设定了良性社区结构的特征、相似社区结构特征,忽视了不同应用网络特质的差异性和表现出的结构特点,主观性较大,即使有些评价方法在实际使用中表现出较好的特质,例如,模块度评价方法,但仍存在分辨率问题以及评价结果与真实情况不拟合等问题[42-43].另一方面,将演化评价和社区质量评价割裂开来的动态社区评价方法,无法体现出社区结构变化的本质性和连续性.

(4)社区个数及演化的问题

静态社区发现中,如何通过计算得到真实的社区个数是一个较为难解决的问题.动态网络中,不同时间片段上的社区结构不同,隐含的社区的个数也会有变化,如何不通过人为设定、自动学习出不同时间片段上真实的社区数目更是动态社区发现的一个重要挑战.

(5)异构网络中的动态社区发现

存在于在线社会网络中的实体并非总是单一类型,例如,微博网络中用户和微博信息并存,论文合作网络中作者、文章、会议等多实体并存,形成多模网络;实体间关系往往也是多样的,呈现出关系异构性,如同时存在于用户之间的兴趣关系、好友关系、引用关系等不同关系类型,同时存在于微博网络中的用户引用关系、用户和信息间的发布关系、信息间主题相似关系等,从而构成异构网络.这些不同类型实体、关系间具有丰富的复杂联系,其隐含的社区结构可能是单模、也可能是多模混合的,社区间关系可能是同质,也可能异质,如何根据特定需求,充分利用丰富信息发现隐含真实的有用社区结构,是近年的研究热点和难点.

(6)计算性能问题

社会网络中用户间的交互是不断动态变化的,社会网络数据呈现出海量性、稀疏性、快速变化性等特征,要发现隐含社区结构需要对不断动态变化的网络拓扑进行分析,这对于计算性能提出了较高要求.

3 同构社会网络的动态社区

数据质量、时间窗口设定、评价问题以及社区个数设定等挑战性问题的研究推动了同构社会网络的动态社区研究的进程,许多新型模型和计算方法被提出.其中,社区评价不仅在算法结果质量判定上起重要作用,而且影响着模型建立和算法设计.根据动态社区研究中社区质量和社区演化评价体系的差别,本文首先将相关国内外研究主要分为三类进行阐述:基于时空独立评价的方法、基于时空集成评价的方法和基于统一评价的方法.同时,针对动态、大数据情况下动态社区发现的计算性能问题,许多增量式解决策略被提出,本文将其归为第四类方法——"增量式动态社区发现方法".

3.1 基于时空独立评价的方法

基于时空独立评价的方法中,社区结构评价和演化的评价完全独立无关,这类方法主要应用于社区演化的研究中,能够发现隐含社区的演化规律,并从社区演变情况中发现突发群体事件和异常变化情况等[7,22-33,44-49].

根据问题背景不同,研究者们采用不同的静态社区发现算法辨识单时间快照上的社区结构,例如,在引文网络分析中,Hopcroft 等人[7]计算文章参考文献的余弦距离,利用分层聚类算法得到单时间片的社区信息.在以超链接关系为主的 Web 社区发现中,Toyoda 等人[32]首先选取所关注的 Web 页面为种子节点集,然后基于页面相关性利用 HITS(Hyperlink-Induced Topic Search)算法发现与种子节点关系紧密的社团结构;Falkowski 等人[33]采用了分层的基于模块度优化的边介数聚类算法发现子社区;Palla 等人[44]利用 CPM(Clique Percolation Method)算法进行单时间段上的社区发现,以研究科学家合作网络和移动用户通信网络的社区演化情况[30].相应的对于时间快照上所发现的社区结构,采用静态社区评价的方法对所发现的社区质量进行评价.

在发现不同时间快照社区的基础上,对相邻时间点的社区结构变化情况进行计算和分析,得到社区演化信息.

对社区演化的分析一般会从网络节点、边、子结构、权值、内容等多方面进行.相关研究内容主要包括:社区相似度计算方法、相邻时间的演化社区匹配对的判断、社区演化模式分类以及社区演化评价指标等.

基于 Jaccard 系数的社区相似度计算方法是进行社区关系分析的一种常用方法.例如,Palla 等人提出,式(1)表示相邻时间段两社区结构的匹配度,

式(2)表征各序列状态的平均关联程度.

$$C_A(t)=\frac{|A(t_0)\bigcap A(t_0+t)|}{|A(t_0)\bigcup A(t_0+t)|} \tag{1}$$

$$\zeta\equiv\frac{\sum_{t=t_0}^{t_{\max}-1}C(t,t+1)}{t_{\max}-t_0-1} \tag{2}$$

其中，$A(t)$表示 t 时刻的社区，t_0 为社区初始产生的时间，$t_{\max}$ 为社区消失的时间，$|A|$ 表示社区 A 的节点数目.

基于信息论的归一化互信息 NMI 方法[41]也是当前一种较为广泛使用的社区结构匹配计算方法，其计算方法为式(3).

$$I(A,B)=\frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{C_A}N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{C_B}N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \tag{3}$$

其中，C_A、C_B分别是不同社区集中的社区数目，N_{ij}、$N_{i.}$、$N_{.j}$分别表 i 社区和 j 社区共同出现的节点数目、i 社区节点数、j 社区节点数.

社区演化分析一般是对相邻时间快照上的社区结构变化情况进行计算，演化分析中首先要辨识出不同时间片间的社区匹配对，一般以具有最大共享边或最大共享点的不同时序上的社区对为匹配序偶，认为是同一社区在不同时间点的呈现形态. 关于社区演化模式，主要分为外部演化和内部演化两大类.

(1) 外部演化描述了不同社区结构间在不同时间快照上的变化情况，一般通过直接计算两个社区共享点或边的情况进行度量，描述的是分布在不同时间快照上不同社区结构间的演化关系，可分为 6 种[26,30,32,45]基本模式，图 1 为文献[30]给出的这 6 种演化模式的示例.

① 生长(growth). 表示某时间快照中的社区在相邻下一时间规模增长的情形.

② 萎缩(contraction). 表示某时间快照中的社区在相邻下一时间规模缩小的情形.

③ 合并(merging). 表示某时间快照中的多个社区在相邻下一时间合并成一个社区的情形.

④ 分裂(splitting). 表示某时间快照中的社区在相邻下一时间分裂成多个社区的情形.

⑤ 产生(birth). 表示在某时间快照中产生了上一相邻时间快照中不存在的社区的情形.

⑥ 消失(death). 表示某时间快照中存在的社区在相邻下一时间快照中消失的情形.

(2) 内部演化描述的是同一社区在不同时间快照上的演化情况. 包括[26]：

① 规模演化. 表现为社区内部的权值变化.

② 压缩演化. 表现为社区内部标准偏差的变化.

③ 位置变化. 表现为社区中心度或社区分布变化.

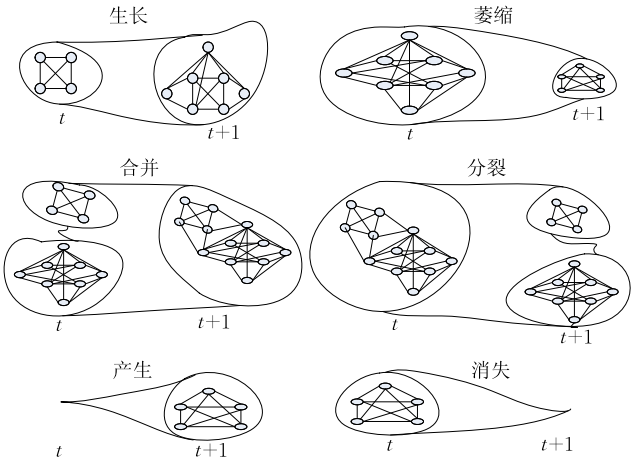

图 1　社区外部演化模式分类[30]

除了外部演化、内部演化这些演化模式分类和度量体系外，根据具体问题需求的不同，研究者们还提出了一些其他度量体系. Asur 等人[27]将中观社区结构演化分析和微观个体交互结构演化分析相结合，定义了一系列事件模式和度量指标来描述、辨识和预测社区演化情况. 他们不仅计算分析了社区合并、分解等模式，还考虑了个体行为对社区的影响，归纳出 4 种个体基本行为模式：个体在网络中的出现、消失、个体加入社区、离开社区. 基于这些行为模式，给出可增量计算的多种度量指标，例如，稳定性——描述了个体间交互行为的不变性(同时，也表征了个体对某社区的归属关系的稳定性)；社会性——描述了个体参与的不同类型交互的类型数；吸引度——描述社区对成员的吸引程度，以某社区中加入个体和离开个体的数目差为度量；影响度——描述个体对别的个体的影响程度，例如影响更多个体加入社区等. 通过计算这些度量指标，辨识网络社区演化情况.

Takaffoli 等人[31]则针对信息网络中的事件发现问题，提出对社区结构的多个方面进行评价的方法，包括社区节点数量(size)、社区边数量(compact)、社区重要节点(leader)的变化以及社区结构稳定性(persistence)等. Toyoda 等人[32]考虑了演化时长对社区演化情况的影响，通过计算不同时间快照上的社区共享边在单位时间内的变化情况，从社区生长率、静止率、新鲜率、消失率、合并率、分裂率等多方面对社区演化情况进行评价. Falkowski 等人[33]提出多方法综合的评价体系进行社区演化分析，该综合评

价体系中包括:静态性(stability),表征在其他时段中保持活跃性的初始成员比例数;密度(density),表征社区内存在的边占完全图边集的比例;粘合度(cohesion),表征了互联微群组的归属度;还有表征两个子图结构等价性的社区重叠度、关联系数以及群组活动等.

基于时空独立评价的方法在社区演化探测和分析中占据着重要地位,通过分析不同领域数据集,一些重要的现象和规律被发现[22-24,28-30]:Leskovec等人[28]以引文网络为对象,研究了图进化过程中平均点度、两点间距离、图导性(conductance)、社区轮廓图等在时间轴上的进化属性,发现引文网络中大团结构的直径随着时间会收缩.Kumar等人[29]则对Flickr和Yahoo!360°两类在线社会网络分析,发现在线社会网络与引文网络类似,大团结构的直径随着时间会收缩.Palla等人[30]以科学家合作网络和移动用户通信网络的真实数据为对象,发现如果经常动态改变大规模群组中成员结构,该群组将会具有较长的生存周期,而小规模群体则相反,其成员结构越稳定,生存周期越长.

3.2 基于时空集成评价的方法

(1)短时平滑性假设

进化聚类是与动态社区研究极为近似的另一研究领域.进化聚类研究中,学者们发现动态网络的聚类结构在短时间内的变化是平缓的[50-51].而在线社会网络中,根据一般观察和经验,网络用户的交互行为演化在很大程度上受隐含关系结构影响,而表现为群体结构的隐含关系结构变化一般较为缓慢,在一定程度上呈现出同样的短时平滑性规律.这一经验发现具有非常重要的意义:

①一方面,短时平滑性意味着短期内的历史交互信息和当前交互信息具有一定相似性,可以将短时历史交互信息和当前交互信息综合作为当前时间片段的网络结构模型,以克服时间窗口划分带来的数据稀疏、数据噪音、信息不全或观察缺失带来的问题.综合历史信息和当前信息的实现方法主要有两种,一种是对有限相邻时段历史信息进行衰减累计,在累计时主要对前一时间段的信息带衰减的累计,或者是建立衰减模型自动学习不同历史信息的衰减因子;一种是认为当前社区间相关度的均值服从给定方差的上一时间点均值的正态分布,对相邻时间点的社区或群组分布建立正态分布的概率模型以反映出平滑性,克服噪音和缺失数据的影响.

②另一方面,短时平滑性可以用来评价短时序范围内所发现的隐式社区结构的合理性和真实性,如果相邻时间段所发现的隐结构差异度极大,则在一定程度表明所发现的隐结构极有可能是错误的,这样,进而,就可以在时间维度上以相邻时间的网络结构差异度最小化为优化目标,帮助设计更合理的动态社区发现算法.

(2)基于时空集成评价的动态社区发现

基于时空集成评价的方法充分利用了短时平滑性特点,将空间社区评价和演化评价集成到一起.其中,空间评价以隐含社区模型上的交互结构和观察到的交互结构差异性最小为目标,演化评价以相邻时间社区结构差异最小化为目标.集成评价的方法认为动态社区的演化和社区的质量是紧密相关的,所以,以一定权值将两个评价相加或一定策略将时空评价综合,建立最优化模型,根据观察数据不断调整和学习,得到最拟合观察数据的隐含结构.

这类方法以实际观察网络为监督进行空间社区评价,避免了不同社区质量评价标准之间不互相支持、评价结果与现实不符等问题;同时将空间评价和时态评价集成在一个统一框架中作为优化目标,反映了隐含社区结构演化的连续属性,适于发现隐含稳定的社区结构及其演化过程.

根据问题模型中优化目标的不同,这类算法可分为4类:

(1)扩展静态社区评价体系的方法

利用空模型进行社区质量评价并进而帮助辨识社区是当前社区发现的重要方法.要辨识动态演化社区,最自然的一个策略就是将空模型延伸到时间轴上,形成时空维度的空模型,从而得到空间质量和时间维度上演化质量综合评价的体系[52-53].这类方法中社区质量的评价仍然是采用了静态社区评价方法,例如,模块度Q等.Mucha等人[52]在2010年的《Science》上发表文章,针对动态社区演化的质量评价问题,模拟静态社区模块度设定机制,结合拉普拉斯动力学(Laplacian dynamics),设计规则生成空模型,提出一种能表征出时间跨度变化、连接类型变化、不同社区尺度质量等的评价方法.该评价体系不仅计算了相邻时间快照上的社区结构关系,而且对任意时间跨度间社区结构的关系进行了考虑,可以发现核心稳定的动态社区结构.但是,模块度评价的局限性也同时引入了该评价体系,在一定程度上可能会影响计算结果质量.图2为Mucha等人的动态社区演化关系示意图.

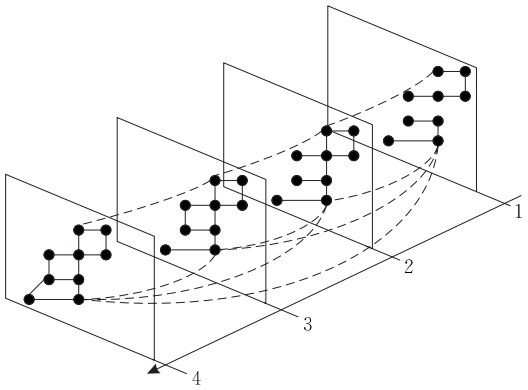

图 2　Mucha 等人[52]的动态社区演化关系

(2) 基于进化聚类的方法

数据挖掘中的进化聚类问题与动态社区发现具有很大的相似性，Chakrabarti 等人[50]在 2006 年第一次提出了进化聚类问题和模型，将进化聚类每一时刻 t 上的任务建模为最大化以下的目标函数(4)，

$$sq(C_t, M_t) - cp.hc(C_{t-1}, C_t) \tag{4}$$

其中，sq 是时间片上聚类质量，C_t、M_t 分别表示 t 时刻的聚类模型和真实网络结构，hc 是相邻时刻聚类模型的平滑性评价，cp 是用户自定义的一个参数. 该模型的等价描述为：以一定权重把时间快照上的社区质量评价和时间轴上演化评价相加作为综合评价指标，对其进行最优化. 优化目标就是，快照质量最大、演化开销最小. 基于此模型，Chakrabarti 等人[50]建立了进化 k-means 算法和分层凝聚聚类的进化算法. Chi 等人[51]则在此基础上提出结构演化的短期平滑性特征，扩展了相似性计算方法，用图切(graph cut)作为测度社区结构和社区进化的指标，第一个建立了带平滑性约束的进化谱聚类算法. Lin 等人[54]借鉴进化聚类的思想，利用信息论相对熵建立了单时间快照社区评价和多快照间演化评价加权和的优化目标(5)：

$$cost = \alpha \cdot D(W_t \parallel \boldsymbol{X}_t \Lambda_t \boldsymbol{X}_t^{\mathrm{T}}) + (1-\alpha) \cdot D(X_{t-1}\Lambda_{t-1} \parallel \boldsymbol{X}_t \Lambda_t) \tag{5}$$

其中，以网络模型和实际网络情况的拟合差异度作为快照上的社区评价(6)：

$$CS = D(W_t \parallel \boldsymbol{X}_t \Lambda_t \boldsymbol{X}_t^{\mathrm{T}}) \tag{6}$$

以相邻时间社区结构差异度为社区演化评价(7)：

$$CT = D(X_{t-1}\Lambda_{t-1} \parallel \boldsymbol{X}_t \Lambda_t) \tag{7}$$

其中，$\boldsymbol{X}_t$ 为 $n \times m$ 矩阵，表示 t 时刻 n 个节点在 m 个社区中发生交互的概率分布，Λ_t 表示 t 时刻社区和社区间的交互. $D(\cdot \parallel \cdot)$ 表示 KL-divergence，在信息论中也被称为相对熵，为式(8)：

$$D(A \parallel B) = \sum_{i,j} \left(a_{i,j} \log \frac{a_{i,j}}{b_{i,j}} - a_{i,j} + b_{i,j} \right) \tag{8}$$

基于进化聚类的算法需要事先设定社区数目，并且一般设定在网络动态变化过程中社区数目不变；对观察网络的起始和结束也有一定限制，为解决此问题，Kim 等人[55]将动态网络建模成由纳米社团构成的系统，社区是该系统中连接稠密的纳米社团群，由于每个纳米社团中含有一定的能表征结构演化趋势的信息，因此，通过对纳米社团的计算可在一定程度上发现社区演变的趋势. 他们使用耗费嵌入技术保证了所发现社区的演化平滑性假设，以模块度优化为目标，建立了基于密度的聚类算法；基于信息论建立了一种映射策略使得可以发现不同时间点上不同个数的社区间的演化关系.

(3) 基于隐空间的方法

基于隐空间的动态社区发现方法[56]的主要理念是将社区结构看做网络结构的隐空间，认为在社区结构的隐空间上，近距离节点较远距离节点间更容易建立连接关系. 同时，基于以下 3 个假设前提：① 相邻时间的隐空间结构变化缓慢(平滑性假设)；② 社区结构演化中 $t+1$ 状态的隐空间仅和 t 时刻隐空间有关，和 t 时刻以前的隐空间无关(与隐马尔可夫链模型假设前提相同)；③ 当前观测值仅和当前隐空间结构相关. 根据这些假设前提，动态社区发现问题可以转化为求最大后验概率的问题，见式(9). 该后验概率模型由 2 部分构成，观察模型和转化模型. 观察模型是同一时间快照上隐含网络结构产生观测网络的概率模型 $P(G_t|X)$，转化模型是相邻时间快照上隐含网络结构间的生成概率模型 $P(X|X_{t-1})$.

$$\begin{aligned} X_t &= \arg\max_X P(X|G_t, X_{t-1}) \\ &= \arg\max_X P(G_t|X) P(X|X_{t-1}) \end{aligned} \tag{9}$$

其中，X_t、X_{t-1} 分别为 t、$t-1$ 时刻的隐含网络结构，G_t 为 t 时刻的观察网络.

在这个框架下，许多类似的算法被提出. Sarkar 等人[56]利用指数分布概率进行单时间快照上的观察模型的建模，利用高斯分布建立时态传递模型，利用经典多维标度(Multidimensional Scaling，MDS)方法对隐空间位置进行初始化，利用核函数进行隐空间近似性度量，建立非线性局部优化的共轭梯度更新规则发现个体的稳定位置，从而监测网络动态变化情况. Lin 等人[54]利用多项式分布建模单时间快照社区质量，用 Dirichlet 分布建模隐含社区演化过程.

基于进化聚类方法的目标是找到满足社区质量

时空评价最大化(或社区时空评价差异度最小化)的隐含社区结构,基于隐空间的方法目标是找到满足后验概率最大化的社区,目标表现形式虽不同,但这两类方法的模型都是由单时间快照中社区结构与观察网络的拟合度评价、相邻时间社区结构的相似性评价两个要素以一定方式组合起来进行建模的.其模型对照情况如图3所示.

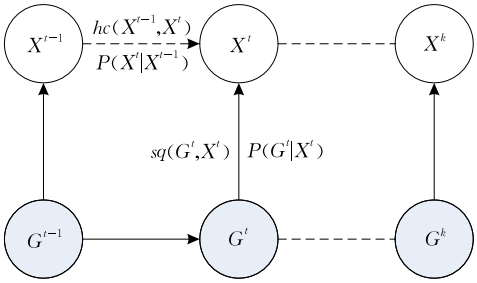

图3 隐空间模型和进化聚类模型的对照
(灰色的圆形 G 代表观察网络,白色圆形 X 代表隐含社区)

Lin 等人[54]证明了这两个模型的等价性,并针对社区重叠性,提出软社区关系,建立了 Facenet 模型,不仅能够发现不同时间段的隐含的重叠和非重叠社区结构,并且能够标示出社区演化的情况.但是,该模型需要事先设定社区数目,并且演化过程中社区数目一般不会发生改变.

(4) 自定义评价指标的方法

另一类建立集成优化目标模型的算法是自定义了一套描述网络结构性质和结构演化的指标,然后根据指标设定所要优化的目标函数.在这些指标中,演化性评价的目标仍然是最小化社区结构变化,即遵循平滑性假设.

Tantipathananandh 等人[57]根据社会经验列举了社区动态变化的各种情况,并构造出相应变化的损耗 *cost* 评价体系,建立了以个体消耗、组消耗、颜色变化消耗3种因素合成的最优化模型.

为了发现网络演化中的异常变化时间点并自动划分时间窗口,Sun 等人[58]基于块模型提出 GraphScope 算法,通过发现网络中社区结构及其变化情况,辨识网络演化关键时间点.他们利用信息论最小描述长度理论,将具有相同性质的节点放在一起,对二部图网络结构行列重排并进行压缩编码,依照"社区变化不大时将相邻时间社区结构压缩表示,变化大时重新进行结构编码"的原则,在时间轴和时间片上分别建立了评价函数,即划分社区的编码耗费和社区表示的编码耗费,求和得到目标函数的表示形式(10),通过找到最小化的编码总耗费得到社区异常变化点,从而也得到合适的时间窗口.

$$C^s=\log^*(t_{s+1}-t_s)+C_p^{(s)}+C_g^{(s)} \tag{10}$$

其中,$t_{s+1}-t_s$ 表示时间分段长度,$C_p^{(s)}$ 表示划分的编码消耗,$C_g^{(s)}$ 表示图的编码消耗.

Chan 等人[59]在 GraphScope 基础上提出 Seqi-Bloc,增加了位置等价性分类和定义,提出两种节点结构演化的等价性定义:块结构保持的等价性和位置保持的等价性,并设计了静态个体编码,基于平滑性变化的个体编码,块保持的序列编码,位置保持的序列编码4种编码耗费的计算方法.通过对各时间段编码耗费和(*cost*)的最小化,达到发现异常点、对社区结构压缩编码的目标.

基于时空集成评价的方法中,评价目标在形式上虽然将时间片上评价和时间轴上演化性评价统一在一个优化框架中,但他们的评价计算仍然是分开的.而且,由于平滑项假设的限制,其优化目标往往是得到核心稳定的社区拓扑,无法及时发现异常演化和由于事件引发的群组突现等.

3.3 基于统一评价的方法

基于时空独立评价和集成评价的方法,都是外部可见数据驱动的、被动的结构分析方法,在实际应用中往往具有普适性差、解决问题有限等问题.最近几年计算机学者们提出利用机器学习中统计推断的技术,从内部模拟网络结构形成和演化的情况,建立生成式概率模型,以实际观察网络为监督,以所建立的隐结构高概率生成网络是否与真实交互网络拟合为评价标准,综合观察网络和一定约束条件建立目标优化模型.这种方法将社区发现问题看作是图重构问题,完全摒弃了社区质量评价和时间轴演化评价割裂的局限性,成为当前最具潜力的热点研究方法.其中,基于贝叶斯推断的概率生成图模型、狄利克雷(Dirichlet)过程模型以及块模型等是这类研究的主要手段.

(1) 基于贝叶斯推断的方法

贝叶斯模型使社区发现的问题转化为隐空间发现问题,提供了将多因素概率关系统一在模型中的可能,是当前社区发现的主要方法之一[60-78].

贝叶斯推断包括两个主要部分:可见信息和统计模型参数.不同模型和算法设计构成了不同的贝叶斯推断方法,其中先验概率的选择和先验概率集成的不同设计构成了不同的优化目标模型.上部分介绍过的时空集成评价中的"基于隐空间的方法"是贝叶斯推断方法的一种,它的模型是社区质量时空评价的乘积.当前,许多研究者提出利用主题模型建

立生成式概率模型进行社区发现的方法[60-66].

McCallum 等人基于隐含狄利克雷分布模型 LDA(Latent Dirichlet Allocation)和作者-主题模型 AT(Author-Topic)[61]，提出在模型中加上交互结构信息，改变 AT 中仅对作者建立话题模型的做法，对信息作者和接收者都建立话题模型，形成一个有向生成图模型 ART[62](Author-Recipient-Topic)，如图 4 所示.

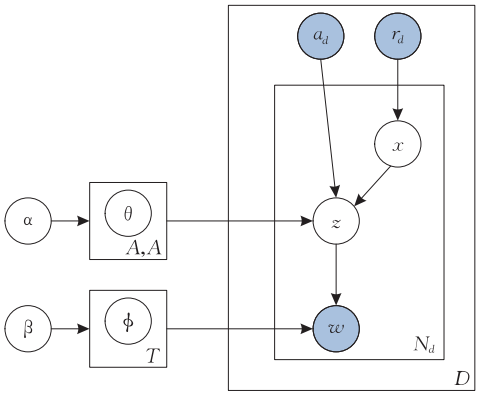

图 4　ART 模型[62]

ART 模型中，α,β 为超级参数，D 表示文档或信息集，每个信息 d 都对应着一个唯一作者和一个接受者集合，分别为 a_d,r_d，为可观察数据；θ 和 ϕ 分别表示混合主题联合分布和词语分布，其中，多项式话题分布 θ 是关于作者和接受者对偶 (a_d,x) 的话题分布，x 为抽样得到的接受者，z 为话题集，w 为单词集. 文档的边际分布表示为式(11)：

$$p(w|\alpha,\beta,a,r)=\iint p(\theta|\alpha)p(\phi|\beta)$$
$$\prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_{x_{d_n}}\sum_{z_{d_n}}p(x_{d_n}|r_d)\cdot p(z_{d_n}|\theta_{a_d,x_{d_n}})p(w_{d_n}|\phi_{z_{d_n}})\mathrm{d}\phi\mathrm{d}\theta \tag{11}$$

利用 ART 模型，不仅可以挖掘隐含话题，还可以根据人的话题分布和交互结构发现社会网络中基于角色分类的群体结构.

Zhang 等人[63]提出 SSN-LDA(Simple Social Network LDA)生成模型进行社区发现. 同时还建立了基于 LDA 的分层贝叶斯模型 GWN-LDA(Generic Weighted Network-LDA)[64]，基于两种不同的网络编码方法发现概率社区. 进一步，针对隐含社区分层问题，他们提出了 HSN-PAM 模型(Hierarchical Social Network-Pachinko Allocation Model)[65]可以发现隐含分层社区，在该模型中，社区被分为超级社区和规则社区两类，模型对这两类社区的生成过程都进行了建模. 除了结构信息，话题内容往往也是促使社区形成的重要因素，将结构和语义综合，Pathak 等人[66]提出一个社区-作者-接受者-话题组成的模型 CART(Community-Author-Receipient-Topic)进行社区抽取. Mei 等人[67]从文本中提取隐含主题并用主题进化图进行主题生命周期分析和时态相关文本挖掘.

另一类基于贝叶斯推断的方法是直接对网络结构生成过程建立先验概率. Hasting[69]将社区发现问题建模成统计推断问题，在给定社区数目和节点连接概率的情况下对社区结构推断. Newman 等人[70]提出一种基于混合概率模型和 EM(Expectation Maximization)的方法发现社区结构，在给定社区数目的情况下，对模型参数和社区结构推断，发现重叠与非重叠社区. 其最大似然概率表示为式(12)

$$\xi=\ln Pr(A,g|\pi,\theta)=\sum_i\left[\ln\pi_{g_i}+\sum_j A_{i,j}\ln\theta_{g_i,j}\right] \tag{12}$$

其中，A 为观察到的网络结构，g 为未知的社区结构，$\{\pi_r,\theta_{r_i}\}$ 为模型参数，π_r 表示随机选择一个点而该点属于社区 r 的概率，θ_{r_i} 表示社区 r 中某点和节点 i 相连的概率. 由于该模型中仅对边存在的情况进行建模，所以计算效率较高.

这些方法都是基于贝叶斯模型的静态社区探测方法，是在给定社区数目情况下进行的.

(2) 基于非参贝叶斯的方法

基于贝叶斯推断的方法有助于发现隐含真实社区结构，但它需要事先设定社区数目. 在不断动态变化的在线社会网络中，隐含社区结构及社区数目也同时在不断动态变化，非参贝叶斯模型有助于解决动态网络中动态社区数目的人为设定问题. 2008 年 Hofman 和 Wiggins[71]针对复杂网络中隐含社区划分和社区数目的发现问题，利用 β 分布和 Dirichlet 分布提出了一种图聚类的贝叶斯模型，设定了同社区节点连边概率和不同社区节点连边概率分布，实现了对社区数目、模型参数(边连接概率等)和社区分配情况的极大似然估计. Xu 等人[72-73]利用 Dirichlet 过程混合模型能够表示无限簇的特点，引入 Dirichlet 过程混合模型对网络演化中的社区数目建立先验概率分布，结合隐马尔可夫模型，建立了两种无参贝叶斯学习模型 DPChain 和 HDP-HTM，发现隐含动态社区并监测社区演化过程. 其中，DPChain 模型是将平滑性假设引入，建立了社区分布的指数衰减因子，和 DPM(Dirichlet Process Mixture)模型结合进行建模. HDP-HTM 模型是将分层 Dirichlet 过程(Hierarchical Dirichlet Process,

HDP)和无限分层马尔可夫状态模型的分层转换矩阵相结合进行建模,运用分层 Dirichlet 过程处理全局和局部社区中的关系问题,运用状态转换矩阵反映不同时间点间社区和社区之间的转换关系.这两个模型不仅能自动发现网络演化过程中的社区数目和社区结构,并且能显式指示出社区演化的因果联系.

(3)基于块模型的方法

块模型[74]能表现出具有相同属性节点间的丰富多样的块结构关系,可用来进行隐含多角色群组发现或重叠社区探测,是社区发现与动态社区演化研究的一类重要方法[74-88].

基本的块模型表现为关系矩阵形式,通过对矩阵行列进行重排,得到连接度紧密以及具有共同性质的群组结构,并能标示出这些群组间的关系.与其他社区发现算法相比,它提供了在统一模型框架中发现紧密团、同质团等多特征群组的功能,成为当前社会网络社区发现、聚类研究中的一个重要方法.

基于节点位置的节点间关系判断是块模型的关键内容,主要有 3 种位置关系等价定义:

①结构等价性(structural equivalence)[75].是最早应用于块模型的关系评价方法.两个节点结构等价的充分必要条件是它们在网络关系矩阵中表现为相同连接关系特征.

②规则等价性(regular equivalence)[76].是 3 种等价关系定义中约束性质最小的关系形式.规则等价的节点集是指该集合中的节点与另外一个规则等价集合中节点具有相同的关系.这种关系度量方法广泛应用于社会关系分析,例如,A 是 B 的女儿,C 是 D 的女儿,则 A 和 C 具有规则等价性,她们的等价性就表现为她们都是"女儿"这个角色,B 和 D 也具有规则等价性,她们的等价性就表现为她们是有女儿的母亲.

③随机等价性(stochastic equivalence)[77-78],主要用于随机块模型,表示两个节点在网络中可以以一定概率分布互换位置.

块模型计算属于排列组合优化问题,数据规模稍有增大,计算性能就无法保证.基于局部优化的生成式随机块模型[79-85]由此产生并成为当前主要模型.

随机块模型的主要基础是无限关系模型 IRM (Infinite Relational Model)[79],即隐含聚类数目不需要事先设定,它会随着观察数据增多而适应性变化.中国餐馆模型 CRP(Chinese Restaurant Process)[83]具有 IRM 性质,经常被用来作为节点的聚类分布模型,即当第 i 个节点被观察到时,它的聚类归属情况由以下概率分布(13)决定,其中,n_a 表示 a 类中成员个数,γ 是一个参数,a 是一个新类.也就是说,来餐馆就餐的人是按一定顺序进来的,每个人在选择餐桌时会根据餐桌上已有人的情况按不同概率进行选择或者添加新的餐桌,这里,一个餐桌的人为一个聚类.

$$P(z_i=a\mid z_1,z_2,\cdots,z_{i-1})=\begin{cases}\dfrac{n_a}{i-1+\gamma}, & n_a>0\\ \dfrac{\gamma}{i-1+\gamma}, & a\text{ 是个新类}\end{cases} \tag{13}$$

随机块模型主要由三部分构成:

节点的聚类分布:

$$z\mid\gamma\sim CRP(\gamma) \tag{14}$$

类间连接分布:

$$\eta_{ab}\mid\alpha,\beta\sim Beta(\alpha,\beta) \tag{15}$$

节点间的连接分布:

$$R_{ij}\mid z,\eta\sim Bernoulli(\eta_{z_iz_j}) \tag{16}$$

优化目标为寻求使 $P(z\mid R)$最大化的 z,即节点的聚类分布结果.

可以看出,基于 CRP 的随机块模型聚类结果是非重叠聚类,为了应用于重叠聚类或社区发现,2008 年 Airoldi 等人[85]提出了一种混合成员的随机块模型方法 MMSB(Mixed Membership Stochastic Blockmodel),它抛弃了 CRP 模型,改变了随机块模型中基于节点的分布建模思想,变为基于关系(边)的分布建模思想,即当一个关系发生时,该连接关系的发起者节点以一定概率在可能属于的多类中选择角色和另一节点相连,而另一节点也以一定概率选择角色来接收发起者的连接请求.2010 年 Xing 等人在 MMSB 上进行扩展以应用到动态网络,他们对所有个体设定了一个有关时间演化的先验单模(单峰 unimodal)假设,即利用一个单模逻辑斯特正态分布(logistic normal distribution)模拟所有个体的角色分布,建立了 dMMSB(dynamic MMSB)模型,发现混合角色的群组结构演化情况.但该模型假设所有个体的多角色分布遵循同样的单峰动态分布,不符合实际网络中个体行为的多样性和个体行为演化的多样性,也无法发现突发行为和异常群体.于是,2011 年,Ho 等人[86]设定了多模混合先验以表征出个体的多角色和演化中的多峰性,建立了状态空间模型集合,结合 MMSB 建立了能发现动态重叠社区的 dM^{3}SB(dynamic Mixture of MMSB)模型,该模

型中一个状态空间轨迹对应着一个群组功能演化的平均轨迹，有多少群组就有多少个先验状态空间模型，表征了群组演化的多样性，能够探测到多样化的社区演化过程，但是需要事先设定社区数目. 同时，dM^{3}SB 模型为了充分利用可得信息，无论节点间是否有联系，对所有边的生成概率都进行了建模，增强了模型表现能力，但对计算性能提出了更高要求.

总之，与其他社区发现方法相比，随机块模型不仅能发现多种群组结构，并且一定情况下其计算结果质量很高[87]，尤其是在用同一种随机块模型生成的网络中，该算法几乎能够完全正确地辨识出隐含块结构. 但是，真实社会网络中节点度差异性很大，度值分布范围广，呈现幂率分布，而随机块模型完全忽略了节点度的差异性，这样，在应用于真实场景中时往往会造成过度学习，丢失重要结构属性，甚至有时会得出错误结论[88].

从模型特征的视角看，基于隐空间、贝叶斯、非参贝叶斯和块模型的动态社区发现方法都属于生成式概率模型推断方法，当前社区动态研究中，基于概率生成模型的方法是最为流行的方法之一，它具有以下优势：

① 适于发现重叠与非重叠社区.

在隐含网络结构未知的情况下，利用图划分的方法进行社区划分，是无法预知应该采用哪种划分方法的. 概率生成图方法以真实网络结构为监督，易于发现隐含的重叠与非重叠社区结构.

② 可以自动生成社区个数，不需要人为指定.

动态社区发现和演化研究中，社区数目的设定和变化是一个重要问题. 由于 Dirichlet 过程支持在混合模型基础上对社区数目添加先验估计，所以，在概率生成模型上引入 Dirichlet 过程，成为当前动态社区发现研究中自动学习社区数目的主要方法.

③ 社区质量的评价更具合理性.

以模型生成的数据与真实交互情况的拟合度为评价指标，免除了经验错误或经验局限性所带来的评价体系偏差以及由这些偏差所引发的社区结构错误.

基于概率生成模型方法面对的主要挑战是过度学习和计算性能问题. 社会网络是一个复杂多尺度网络，忽略网络的宏观属性，仅依靠单纯的机械学习有时会造成过度拟合，从而丢失重要隐含结构. 另一方面，概率生成式模型的结构参数随网络规模增长而增长，大量的模型参数学习对计算性能带来了极大挑战，虽然有研究者们相继提出变分 EM、Gibbs 抽样等不同算法优化学习过程，但是利用生成式模型的算法能够处理的网络规模仍然具有极大局限性.

3.4　增量式动态社区发现方法

在线社会网络数据量大、动态性强，快速准确发现隐含社区结构并得到隐结构演化信息是实际应用的迫切需求. 针对算法效率问题，许多增量式动态社区发现算法被提出，该方法的假设前提仍然是平滑性假设，认为动态社区演化过程中，大部分的拓扑结构保持相对稳定，仅有小部分结构会变化，所以，他们提出，在识别前一时刻社区结构的基础上，仅对改变的网络结构部分进行重计算，而其余结构保持不变，以此提升计算效率[89-97].

当前的增量策略主要分为基于物理学原理的增量策略和基于图特征的增量策略.

(1) 基于物理学原理的增量策略将网络看成复杂物理世界. 受牛顿万有引力定律的启发，Yang 等人[89]将网络节点间关系分为吸引力和排斥力两种，通过迭代增量计算，使得所形成的社区内部吸引力越来越大，社区间的边上的排斥力越来越大，最终连接边断裂，形成分割的社团结构.

(2) 基于图特征的增量式动态社区发现方法一般首先利用静态社区发现算法发现初始时间快照上的社区结构，然后辨识后续网络快照中结构变化的部分，对这小部分改变的结构进行计算，从而避免对全网络重新计算[90-97].

主要的增量策略有基于图分割特征的、基于谱特征的、基于矩阵分解等方法.

单波等人[90]通过设定和识别新增的、改变社区归属条件的顶点，在已知 $t-1$ 时刻社区结构的情况下，仅对增量相关节点进行分析，得到 t 时刻社区结构，避免重新进行社区划分，提高了计算性能. 但该方法没有考虑网络变化过程中节点增加和消失的情况，并且计算过程中社区数目默认为保持不变. 增量式谱聚类方法[91-92]，考虑了节点增加、删除和节点间相似性的变化，通过评估节点和边改变对网络特征值和特征向量的影响，引入转换拉普拉斯矩阵(shifted Laplacian matrix)代替标准拉普拉斯矩阵，用近似特征计算方法(Eigen-approximation method)，利用上一时间点的特征向量结果，近似计算当前时间快照上的最大特征向量，减少了重新聚类，与静态谱聚类算法相比，计算性能提高了 $O(n^{1/2})$；增量式张量分解方法[93]是对 MetaFac(MetaGraph Factorization)[94]方法的扩展，对动态数据分析，发现不同时间上的多划分结构和其低秩核矢量以及相应变化

规则，为了减少计算时间，首先限定了张量的最大规模数，并用时间上的最近邻及时替代过时个体；考虑了历史信息，建立了张量随时间衰减的更新规则，并提出基于 MetaFac 方法的增量式因式分解方法，节省计算时间；增量式 k-clique 聚类[95]通过局部更新最大 k-clique 团集合和 k-clique 团所构成的深度优先搜索森林以提高动态社区发现的性能，与其他增量算法相比，它的最大优势是该算法不需要事先设定社区数目，并且计算准确度高于增量式谱聚类算法.

Dinh 等人[96]针对无尺度网络，根据节点度特征，将节点分为 leader 和 follower 角色，设计增量计算规则，发现动态社区，并给出算法的近似比保障.

总之，增量式动态社区发现方法与基于概率生成模型的方法相比，计算性能要高，但这类算法在计算性能上的提高是以计算结果质量的部分降低为代价的.

其他动态社区发现及演化的研究有，窦炳琳等人[98]通过对 DBLP 和 Facebook 数据进行分析，发现社会网络中存在紧密连接且直径较小的核心结构，规模中等的社区主要呈现星型结构；对动态社区演化分析，发现社区间融合很大程度上取决于社区间直接连接的节点所构成网络的聚类系数，而社区的分裂则与该社区的聚类系数相关，其聚类系数越高，分裂的可能性越小.

Mitra 等人[99]认为，当前按照时间快照分析社区演化的算法生硬地割裂了不同时间快照中节点之间的关联关系，无法发现隐含关联社区结构，提出将不同时间点以及同一时间点节点关系表现在一个统一模型上进行隐含社区发现的方法. 他们以引文网络为主要问题背景，根据论文引用和时间的关系，文章引用发生在被引用文章发表之后，形成一个与时间相关的有向网络. 这种方法记录了交互行为的有序时间，不人为设定累计时域，可以真实记录交互信息. 但是，该方法的扩展性差，不适用于无法直接观察到有向交互或连接关系的网络中，例如，社会媒体中信息的传播，可以观察到信息被报道的情况，但是不能确定该信息是谁传播过来的，这样在问题建模时，无法标示出节点间联结关系. 另外，这种算法对计算性能要求较高.

4 异构社会网络的动态社区

在线社会网络研究中，学者们发现社会网络中往往是多实体并存的，用户间也经常同时并存多种交互活动，而除了交互结构信息，节点属性、节点内容等往往对于社区结构的构成和演化也具有一定的影响，不同实体间的关系演变也在深刻地影响着网络的特性. 所以，异构网络中社区结构发现及演化成为在线社会网络分析中的一项重要内容. 根据网络构成性质的不同，可分为 2 种异构网络模型.

4.1 多模异构网络的动态社区

我们将包含有多种类型个体和多种类型关系的网络称为多模异构网络(multi-mode network)[100-105]. 实际上，包含不同类型个体的网络中，网络的连边关系也是多样的. 由于多模异构网络的复杂性，当前相关研究还较少，已有的相关研究工作也都是对问题进行了简化处理，仅仅考虑节点异构性，不考虑边的异构性，典型研究方法主要有 3 种.

第 1 种是将网络中不同类型的个体同等对待，根据交互关系发现多类型的社区结构，同一社区内个体类型相同. Zhao 等人[100]以学术网为例提出了一个发现和追踪动态社区演化状况的统一框架，首先将一定时间窗口内的对象和关系集成计算，构成异构网络；然后，对时间快照上的各网络抽取快照特征(snapshot-based feature)和时态演化特征(delta-based feature)，利用回归模型得到分层社区结构，然后再利用多分类方法辨识出回归模型无法辨识的多社区结构. 最后，根据具体问题需求和约束条件，对计算结果进行后处理. 该框架简单，缺少对社区结构的评价和演化的评价方法，无法判别算法结果质量. Tang 等人[101]则利用平滑性假设和进化聚类算法框架对多模网络建模，发现动态社区序列. 该方法的不足在于它仅限于发现非重叠动态社区，而且社区数目需要事先人为给定，并在进化过程中默认不变.

第 2 种研究方法是将网络中不同个体类型进行主次之分，选取某一种类型的个体为主要目标对象，其他类型的个体作为该目标对象的属性，目标对象和其他类型个体间关系构成星型结构，形成以目标对象为主体表示的网络结构；基于此结构，Sun 等人[102-104]利用 Dirichlet 过程对隐含社区数目的分布进行先验估计，结合平滑性假设，提出一种过程混合模型的生成模型模拟社区生成，利用该模型可以在每个时间点自动发现能较好解释当前网络和历史网络特征的社区数目和结构，并设计了一个基于 Gibbs 采样的方法进行模型学习和推断. 与文献[100-101]多模网络社区发现中单社区仅包含一类个体不同，Sun 等人的工作所发现的单社区包含有不同类型个体.

第 3 种方法是利用超图模型描述多模异质网络. Lin 等人[105]基于超图理论提出一种新的关系超图——元图(metaGraph)模型，来表示由多维属性的社会节点和节点间不同关系所构成的网络，基于元图模型，他们将社区发现问题转化为一个关于多关系因式分解的优化问题：如何将表示数据关联的张量分解为非负超对角核张量和一个对应于每个因子的非负因子矩阵的乘积，优化目标被表示为关于元图的函数；同时建立了基于增量超图因式分解的在线算法以处理随时间变化的关系网络.

4.2　单模异构网络的动态社区

我们将包含单一类型个体和多种类型关系的网络称为单模异构网络[106-109]. 单模异构网络的动态社区研究中的一个关键挑战是，如何评价和计算这些异构关系在动态网络分析中的作用及其联系.

Tang 等人[106]从社区评价方法入手，将单一关系网络的社区模块度评价扩展成多关系网络的综合模块度评价. 由于对各关系网络模块度求平均的模块度综合评价方法和对各关系网络模块度求和的评价方法对噪音的敏感性，他们提出了一种主模块度优化方法(Principal modularity maximization)：首先对不同关系维度上的模块度进行分析，根据特征值和特征向量抽取出每种关系维度网络中的主结构特征，然后将这些主结构特征关联起来，发现具有鲁棒性的共享社区结构，该共享社区结构应能够使各关系维度上的社区模块度达到最大或近似最大，最后建立了互验证的方法进行结果评价.

张春英等人[107]则是从多属性多关系社会网络中个体间确定性关系和不确定性关系并存的视角入手，采用集对的思想对多关系社会网络进行分析，建立了节点关系同一度、差异度和对立度的定义和基于集对联系关系矩阵的社会网络模型，提出一种点关系满足一定阈值的 α 关系社区概念，然后建立了静态 α 关系社区和动态 α 社区的挖掘算法，以及社会网络集对关系转移矩阵的计算方法，以掌握网络发展态势. 该算法是一种社区时空独立评价的方法，不足之处是缺少社区结构的质量评价和验证.

林旺群等人[108]对多关系社会网络动态社区发现的性能问题进行了研究，他们针对多种关系并存的现象，将社会网络建模为多重图，建立了基于多重图并行分解的静态社区发现算法 P-SNCD(Parallel Social Network Community Discovery)，以快速发现由不同级别社区构成的层次化的社区树；然后，利用复杂动态网络在空间结构上的演化特点：动态社会网络拓扑结构变化相对于社区结构而言具有局部特性，他们对静态社区并行算法 P-SNCD 进行扩展，提出在动态社区计算中仅对算法生成的上一时间快照上的层次化社区树的分枝进行选择性更新，减少无谓的计算，从而提升动态社区发现的时间效率.

单模异构网络中的多关系对于探究隐含网络社区演化具有重要意义，但是这些不同类型关系间是否相互独立、如何关联以及对于隐含拓扑演化的作用的差异性仍然是非常复杂的问题，也是未来多模异构网络演化的一个重要研究内容.

5　异常群体发现

在线社会网络中，复杂的物理世界和网络世界的交融，使得网络突发异常事件、异常群体或异常联系时有发生，如何在检测网络社区或群组演化时识别这些异常情况，是动态社区研究的一个重要应用.

在线社会网络中，除去语义信息，我们能看到的是各种交互信息，事件异常与否的主要表现也体现在交互结构上，例如，突然很快传播并迅速扩展开的信息、非寻常的个体间频繁交互等. 当前在线社会网络异常事件检测的方法可以分为 2 类：

一种研究方法是机器学习的方法，主要对两种异常进行监测. 一种是监测动态网络中异常演化的社区. 首先在网络数据中抽取出正常社区演化的训练集，学习出社区正常演化特征，然后据此检测异常的社区演化[109]. 另一种是探测社区演化过程中的异常个体行为，Gupta 等人[110]将社区发现、社区匹配和异常检测迭代集成，以发现社区演化中异于其他个体的异常个体.

另一种是针对社会网络事件的特点，对信息传播中的单因素累积数量随时间的变化情况进行监测，如通过对微博转发量的监测发现异常热点事件，通过个体间交互频度的突然变化发现异常突现的关系等. 例如，突然快速增加、突然快速减少、突然增加又突然快速减少蕴含了三类异常情况等. 文献[111]则是通过发现隐含社区结构，当某成员的社区行为和社区公共行为不一致时认为该成员是异常者.

在“群体异常”的评估和判断方面，主要是根据相邻时间点社区变化情况进行判断，当社区差异度大到一定程度就认为是“异常”，这类方法中，社区相似性评价方法和阈值的设计会影响异常群体发现的质量.

Kumar 等人[112]以 blogosphere 为研究对象，把 blog 构建成图模型，根据其入度、出度、强连通组件等性质的变化情况研究 blog 社区变化、突变等性质. GraphScope[59]和 SeqiBloc[60]方法利用最小描述长度表示节点关系，建立动态社区发现模型，并发现异常时间点. 由于对社区发现方法和演化评价不同，群体异常的判断标准也有很大差别. 因此，需要科学的测度标准和评价方法.

6 研究难点和发展趋势

隐含动态社区结构发现及演化的研究对于在中观视图观察在线社会网络动态隐结构特征、预测演化趋势、掌控网络态势具有重要意义. 网络信息的丰富性、稀疏性、隐私保护与信息残缺性、动态变化性等使得基于可见信息的隐含动态社区结构发现成为可行但又具有一定挑战的研究内容，虽然已有许多动态社区发现及演化的工作被报道，但是，动态社区研究方法仍然存在许多没完全解决的问题. 总结起来，未来的研究难点和发展趋势主要分为以下 3 点：

(1) 平滑性假设的局限性

对国内外相关研究分析可以看出，在最近较多的动态社区发现及演化的研究中，学者们普遍接受了短时平滑性假设. 这是因为虽然受个体、环境等因素影响，交互结构复杂多变，但交互结构的发生很大部分因素是由其背后所蕴含的关系结构决定的，而关系结构体现在群组层面上，其随时间的变化是具有短时平滑性的. 所以，交互结构的群组演化也具有平滑性特征. 但是，在线社会网络发展过程不是封闭过程，物理世界的环境、事件等不断有新的刺激使得网络个体间的关系发生变化，有可能是现实世界社会关系的改变带来网络关系结构的变化，通过网络事件或行为反映在交互结构中；也可能是消息事件的特殊属性带来交互行为的突然改变，并影响了底层关系结构. 总之，社会网络不是封闭系统，由内因和外因(外界刺激)所引起的个体域及关系的变化总在发生，当突发事件发生或异常发生时，关系网络的演化过程也许并非总是平滑，即使认为无论突变还是正常演变都具有相对短期平滑性，但是这个平滑性假设成立的短期的限定是和领域紧密相关的，多短合适也是当前以平滑性假设为前提的算法在现实应用中不得不面对的一个重要挑战.

(2) 多源异质信息集成的动态社区发现

社会网络的个体在网络中呈现的属性可能是多刻面的、高维的、稀疏的，社会网络关系的变化是和群体交互、事件等紧密相关的，如何将动态变化的语义信息、社会信息、交互信息等多源信息有效综合进行动态社区发现、群体事件发现和预测将成为未来的重要研究领域.

(3) 面向网络大数据的统计推断概率模型设计和算法优化

基于可见事实、后验推断网络隐含社区结构及动态社区结构的统计推断方法，已经成为当前在线社会网络社区结构发现和动态社区发现的主要方法，但是复杂的模型、大量的参数学习使得该类算法的运算效率都较低，不能处理较大规模数据. 在线社会网络应用的极大扩展带来复杂海量的大数据，由此，简单的概率生成模型、高效的面向大数据的机器学习算法设计，尤其是基于统计推断模型的高效算法设计成为当前及未来的一项重要研究内容.

7 结束语

在线社会网络是一个开放复杂的动态世界，人类个体的多样性、复杂性使得在线社会网络行为拓扑关系演化呈现出规则与随机混合的复杂性，面对海量的社会网络数据，发现隐含中观结构，探测中观视图上的结构演化规律，对于观察在线社会网络隐结构特征、揭示在线社会网络演化机理、掌握网络发展态势、识别群体事件、挖掘用户观点、分析信息传播过程、优化网络应用服务具有重要意义.

参 考 文 献

[1] Wang Li, Cheng Su-Qi, Shen Hua-Wei, Cheng Xue-Qi. Structure inference and prediction in the co-evolution of social networks. Journal of Computer Research and Development, 2013, 50(12): 2492-2503(in Chinese)
(王莉，程苏琦，沈华伟，程学旗. 在线社会网络共演化的结构推断与预测. 计算机研究与发展，2013，50(12)：2492-2503)

[2] Newman M. Detecting community structure in networks. European Physical Journal B, 2004, 38(2): 321-330

[3] Kairam S, Wang D, Leskovec J. The life and death of online groups: Predicting group growth and longevity//Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM). Washington, USA, 2012: 673-682

[4] Xu Tianbing, Zhang Zhongfei, Yu P S, Long Bo. Dirichlet process based evolutionary clustering//Proceedings of the IEEE International Conference on Data Mining (ICDM). Pisa, Italy, 2008: 648-657

[5] Wang Li. SoFA: An expert-driven, self-organization peer-to-peer semantic communities for network resource management. Expert Systems with Applications, 2011, 38(1): 94-105

[6] Backstrom L, Huttenlocher D, Kleinberg J, Lan X. Group formation in large social networks: Membership, growth, and evolution//Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD). Philadelphia, USA, 2006: 44-54

[7] Hopcroft J, Khan O, Kulis B, Selman B. Tracking evolving communities in large linked networks. Proceedings of the National Academy of the Sciences of the United States of America (PNAS), 2004, 101(Suppl 1): 5249-5253

[8] Gong Mao-Guo, Zhang Ling-Jun, Ma Jing-Jing, Jiao Li-Cheng. Community detection in dynamic social networks based on multi-objective immune algorithm. Journal of Computer Science and Technology, 2012, 27(3): 455-467

[9] Wang Li, Bi Yuanjun, Wu Weili, Biao Lian, Xu Wen. Neighborhood-based dynamic community detection with graph transform for 0-1 observed networks//Proceedings of the International Computing and Combinatorics Conference (COCOON). Hangzhou, China, 2013: 821-830

[10] Berger-Wolf T Y, Saia J. A framework for analysis of dynamic social networks//Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining (SIGKDD). Philadelphia, USA, 2006: 523-528

[11] Berger-Wolf T Y, Lahiri M, Tantipathananandh C, Kempe D. Finding structure in dynamic networks//Proceedings of the 1st Workshop on Information in Networks (WIN-09). New York, USA, 2009: 1-4

[12] Tantipathananandh C, Berger-Wolf T Y. Finding communities in dynamic social networks//Proceedings of the IEEE 11th International Conference on Data Mining (ICDM). Vancouver, Canada, 2011: 1236-1241

[13] Habiba H, Yu Y, Berger-Wolf T Y, Sala J. Finding spread blockers in dynamic networks//Proceedings of the Advances in Social Network Mining and Analysis. Las Vegas, USA, 2008: 55-76

[14] Giatsoglou M, Vakali A. Capturing social data evolution using graph clustering. IEEE Internet Computing, 2013, 17(1): 74-79

[15] Tantipathananandh C, Berger-Wolf T Y. Constant-factor approximation algorithms for identifying dynamic communities //Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining (SIGKDD). New York, USA, 2009: 827-836

[16] Delvenne J C, Yaliraki S N, Barahona M. Stability of graph communities across time scales. Proceedings of the National Academy of the Sciences of the United States of America (PNAS), 2010, 107(29): 12755-12760

[17] Asur S, Parthasarathy S. A viewpoint-based approach for interaction graph analysis//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2009: 79-88

[18] Mansoureh T, Farzad S, Justin F, Osmar R Z. A framework for analyzing dynamic social networks//Proceedings of the 7th Conference on Applications of Social Network Analysis (ASNA). Zurich, Switzerland, 2010: 1-14

[19] Takaffoli M, Sangi F, Fagnan J, Zäıane O R. MODEC—Modeling and detecting evolutions of communities//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM). Barcelona, Spain, 2011: 626-629

[20] Chen Zhengzhang, Wilson K A, Ye Jin, Hendrix W. Detecting and tracking community dynamics in evolutionary networks//Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW). Sydney, Australia, 2010: 318-327

[21] Clauset A, Eagle N. Persistence and periodicity in a dynamic proximity network//Proceedings of the Workshop on Computational Methods for Dynamic Interaction Networks. New Jersey, USA, 2007: arXiv: 1211.7343

[22] Kossinets G, Watts D J. Empirical analysis of an evolving social network. Science, 2006, 311(5757): 88-90

[23] Newman M, Barabasi A-L, Watts D J. The Structure and Dynamics of Networks. New Jersey: Princeton University Press, 2006

[24] Goldberg M, Magdon-Ismail M, Nambirajan S, Thompson J. Tracking and predicting evolution of social communities//Proceedings of the IEEE 3rd International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE 3rd International Conference on Social Computing (SOCIALCOM). Boston, USA, 2011: 780-783

[25] Takaffoli M. Community evolution in dynamic social networks—Challenges and problems//Proceedings of the IEEE 11th International Conference on Data Mining Workshops (ICDMW). Vancouver, Canada, 2011: 1211-1214

[26] Spiliopoulou M, Ntoutsi I, Theodoridis Y, Schult R. MONIC: Modeling and monitoring cluster transitions//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 706-711

[27] Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Transactions on Knowledge Discovery from Data (TKDD), 2009, 3(4): 16

[28] Leskovec J, Kleinberg J, Faloutos C. Graphs over time: Densification laws, shrinking diameters and possible explanations//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, USA, 2005: 177-187

[29] Kumar R, Novak J, Tomkins A. Structure and evolution of online social networks//Yu P S, Han Jiawei, Faloutsos C eds. Link Mining: Models, Algorithms, and Applications. New York: Springer, 2010: 337-357

[30] Palla G, Barabasi A L, Vicsek T. Quantifying social group evolution. Nature, 2007, 446(7136): 664-667

[31] Takaffoli M, Sangi F, Fagnan J. Tracking changes in dynamic information networks//Proceedings of the International Conference on Computational Aspects of Social Networks (CASON). Guilin, China, 2011: 94-101

[32] Toyoda M, Kitsuregawa M. Extracting evolution of Web communities from a series of Web archives//Proceedings of the 14th ACM Conference on Hypertext and Hypermedia. Pennsylvania, USA, 2003: 28-37

[33] Falkowski T, Bartelheimer J, Spiliopoulou M. Mining and visualizing the evolution of subgroups in social networks//Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Hong Kong, China, 2006: 52-58

[34] Wang Li, Wang Jiang, Shen Hua-Wei, Cheng Xue-Qi. Improvement on fast uncovering community algorithm. Chinese Physics B, 2013, 22(10): 108903

[35] Shen Huawei, Cheng X, Cai K, Hu Mao-Bin. Detect overlapping and hierarchical community structure in networks. Physica A, 2009, 388(8): 1706-1712

[36] Guimera R, Sales-Pardo M, Amaral L. Modularity from fluctuations in random graphs and complex networks. Physical Review E, 2004, 70(2): 025101

[37] Clauset A, Newman M, Moore C. Finding community structure in very large networks. Physical Review E, 2004, 70(6): 066111

[38] Flake G, Lawrence S, Lee Giles C, Coetzee F. Self-organization and identification of Web communities. IEEE Computer, 2002, 35(3): 66-70

[39] White S, Smyth P. A spectral clustering approach to finding communities in graphs//Proceedings of the 5th SIAM International Conference on Data Mining. Philadelphia, USA, 2005: 76-84

[40] Martin R, Carl T. An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of the Sciences of the United States of America (PNAS), 2007, 104(18): 7327-7331

[41] Danon L, Duch L, Guilera A D, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005, 9: P09008

[42] Lancichinetti A, Fortunato S. Limits of modularity maximization in community detection. Physical Review, 2011, 84(6): 066122

[43] Fortunato S, Barthélemy M. Resolution limit in community detection. Proceedings of the National Academy of the Sciences of the United States of America (PNAS), 2007, 104(1): 36-41

[44] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(3): 814-818

[45] Greene D, Doyle D, Cunningham P. Tracking the evolution of communities in dynamic social networks//Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM). Odense, Denmark, 2010: 176-183

[46] Nguyen N P, Dinh T N, Shen Y, Thai M T. Dynamic social community detection and its applications. PLoS ONE, 2014, 9(4): e91431

[47] Cuzzocrea A, Folino F, Pizzuti C. DynamicNet: An effective and efficient algorithm for supporting community evolution detection in time-evolving information networks//Proceedings of the 17th International Database Engineering & Applications Symposium (IDEAS). Barcelona, Spain, 2013: 148-153

[48] Takaffoli M, Sangi F, Fagnan J, Zäıane O R. Community evolution mining in dynamic social networks. Procedia-Social and Behavioral Sciences, 2011, 22: 49-58

[49] Nguyen N P, Dinh T N, Ying Xuan, Thai M T. Adaptive algorithms for detecting community structure in dynamic social networks//Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM). Shanghai, China, 2011: 2282-2290

[50] Chakrabarti D, Kumar R, Tomkins A. Evolutionary clustering//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 554-560

[51] Chi Y, Song X, Zhou D, et al. Evolutionary spectral clustering by incorporating temporal smoothness//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 153-162

[52] Mucha P J, Richardson T, Macon K, et al. Community structure in time-dependent, multiscale, and multiplex networks. Science, 2010, 328(5980): 876-878

[53] Bassett D S, Porter M A, Wymbs N F, et al. Robust detection of dynamic community structure in networks. Chaos, 2013, 23(1): 013142

[54] Lin Yu-Ru, Chi Yun, Zhu Shenghuo, et al. Facetnet: A framework for analyzing communities and their evolutions in dynamic networks//Proceedings of the 17th International Conference on World Wide Web. New York, USA, 2008: 685-694

[55] Kim Min-Soo, Han Jiawei. A particle-and-density based evolutionary clustering method for dynamic networks//Proceedings of the 2009 International Conference on Very Large Databases (VLDB). Lyon, France, 2009: 622-633

[56] Sarkar P, Moore A W. Dynamic social network analysis using latent space models. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 31-40

[57] Tantipathananandh C, Berger-Wolf T, Kempe D. A framework for community identification in dynamic social networks//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 717-726

[58] Sun J, Faloutsos C, Papadimitriou S, Yu P S. GraphScope: Parameter-free mining of large time-evolving graphs//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 687-696

[59] Chan J, Liu Wei, Leckie C, et al. SeqiBloc: Mining multi-time spanning blockmodels in dynamic graphs//Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD). Beijing, China, 2012: 651-659

[60] Yang T, Chi Y, Zhu S, et al. Detecting communities and their evolutions in dynamic social networks — A Bayesian approach. Machine learning, 2011, 82(2): 157-189

[61] Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada, 2004: 487-494

[62] McCallum A, Corrada-Emmanuel A, Wang X. Topic and role discovery in social networks//Proceedings of the 19th International Joint Conference on Artificial Intelligence. Beijing, China, 2005: 786-791

[63] Zhang H, Qiu B, Giles C L, et al. An LDA-based community structure discovery approach for large-scale social networks //Proceedings of the IEEE International Conference on Intelligence and Security Informatics. New Jersey, USA, 2007: 200-207

[64] Zhang H, Giles C L, Foley H C, Yen J. Probabilistic community discovery using hierarchical latent Gaussian mixture model//Proceedings of the National Conference on Artificial Intelligence(AAAI). British Columbia, Vancouver, Canada, 2007: 663-668

[65] Zhang H, Li W, Wang X, et al. HSN-PAM: Finding hierarchical probabilistic groups from large-scale networks// Proceedings of the 7th IEEE International Conference on Data Mining (ICDM Workshops). Los ALamitos, USA, 2007: 27-32

[66] Pathak N, Delong C, Banerjee A, Erickson K. Social topic models for community extraction//Proceedings of the 2nd ACM Workshop on Social Network Mining and Analysis (SNA-KDD). Las Vegas, USA, 2008: 8

[67] Mei Q, Zhai C. Discovering evolutionary theme patterns from text: An exploration of temporal text mining//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2005: 198-207

[68] Negoescu R. Modeling and Understanding Communities in Online Social Media Using Probabilistic Methods [Ph. D. dissertation]. Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland, 2011

[69] Hastings M. Community detection as an inference problem. Physical Review E, 2006, 74(3): 035102

[70] Newman M, Leicht E. Mixture models and exploratory analysis in networks. Proceedings of the National Academy of the Sciences of the United States of America (PNAS), 2007, 104(23): 9564-9569

[71] Hofman J, Wiggins C. Bayesian approach to network modularity. Physical Review Letters, 2008, 100(25): 258701

[72] Xu T, Zhang Z, Yu P, Long B. Evolutionary clustering by hierarchical dirichlet process with hidden Markov state// Proceedings of the International Conference on Data Mining (ICDM). Pisa, Italy, 2008: 658-667

[73] Xu Tianbing, Zhang Zhongfei, Yu P S, Long Bo. Generative models for evolutionary clustering. ACM Transactions on Knowledge Discovery from Data (TKDD), 2012, 6(2): 7

[74] Faust K, Wasserman S. Blockmodels: Interpretation and evaluation. Social Networks, 1992, 14(1): 5-61

[75] Lorrain F, White H C. Structural equivalence of individuals in networks. The Journal of Mathematical Sociology, 1971, 1(1): 49-80

[76] Borgatti S P, Everett M G. Two algorithms for computing regular equivalence. Social Networks, 1993, 15(4): 361-376

[77] Holland P W, Laskey K B, Leinhardt S. Stochastic blockmodels: First steps. Social Networks, 1983, 5(2): 109-137

[78] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. New York: Cambridge University Press, 1994

[79] Kemp C, Tenenbaum J, Griffiths T, et al. Learning systems of concepts with an infinite relational model//Proceedings of the 21st National Conference on Artificial Intelligence (AAAI). Boston, USA, 2006: 381-388

[80] Doreian P. Some open problem sets for generalized blockmodeling//Batagelj V, Bock H, Ferligoj A, Žiberna A eds. Data Science and Classification. Berlin: Springer-Verlag, 2006: 119-130

[81] Doreian P, Batagelj V, Ferligoj A. Generalized Blockmodeling. New York: Cambridge University Press, 2005

[82] Doreian P, Batagelj V, Ferligoj A. Generalized blockmodeling of two-mode network data. Social Networks, 2004, 26(1): 29-53

[83] Pitman J. Combinatorial Stochastic Processes. Berlin: Springer, 2006

[84] Choi D S, Wolfe P J, Airoldi E M. Stochastic blockmodels with growing number of classes. Biometrika, 2012, 99(2): 273-284

[85] Airoldi E M, Blei D M, Fienberg S E, Xing E P. Mixed-membership stochastic blockmodels. Journal of Machine Learning Research (JMLR) 2008, 9(9): 1981-2014

[86] Ho Qirong, Song Le, Xing E P. Evolving cluster mixed-membership blockmodel for time-evolving networks. Journal of Machine Learning Research (JMLR), 2011, 15(8): 342-350

[87] Bickel P J, Chen A. A nonparametric view of network models and Newman-Girvan and other modularities. Proceedings of the National Academy of the Sciences of the United States of America (PNAS), 2009, 106(50): 21068-21073

[88] Karrer B, Newman M. Stochastic blockmodels and community structure in networks. Physical Review E, 2011, 83(1): 016107

[89] Yang B, Liu D. Force-based incremental algorithm for mining community structure in dynamic network. Journal of Computer Science and Technology, 2006, 21(3): 393-440

[90] Shan Bo, Jiang Shou-Xu, Zhang Shuo, et al. IC: Dynamic social network community structures incremental recognition algorithm. Journal of Software, 2009, 20(Supplement): 184-192(in Chinese)
(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. 软件学报, 2009, 20(增刊): 184-192)

[91] Ning Huazhong, Xu Wei, Chi Yun, et al. Incremental spectral clustering by efficiently updating the eigen-system. Pattern Recognition, 2010, 43(1): 113-127

[92] Dhanjal C, Gaudel R, Clémençon S. Incremental spectral clustering with the normalised laplacian//Proceedings of the 3rd NIPS Workshop on Discrete Optimization in Machine Learning (DISCML). Granada, Spain, 2011: 1-6

[93] Bockermann C, Jungermann F. Stream-based community discovery via relational hypergraph factorization on evolving networks//Proceedings of the Workshop on Dynamic Networks and Knowledge Discovery (DyNaK). Barcelona, Spain, 2010: 41-52

[94] Lin Y, Sun J, Castro P, et al. MetaFac: Community discovery via relational hypergraph factorization//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). Paris, France, 2009: 527-536

[95] Duan Dongsheng, Li Yuhua, Li Ruixuan, Lu Zhengding. Incremental k-clique clustering in dynamic social networks. Artificial Intelligence Review, 2011, 38(2): 129-147

[96] Dinh T N, Nguyen N P, Thai M T. An adaptive approximation algorithm for community detection in dynamic scale-free networks//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Turin, Italy, 2013: 55-59

[97] Falkowski T, Barth A, Spiliopoulou M. Studying community dynamics with an incremental graph mining algorithm//Proceedings of the 14th Americas Conference on Information Systems (AMCIS). Toronto, Canada, 2008: 29

[98] Dou Bing-Lin, Li Shu-Song, Zhang Shi-Yong. Social network analysis based on structure. Chinese Journal of Computers, 2012, 35(4): 741-753(in Chinese)
(窦炳琳, 李澍淞, 张世永. 基于结构的社会网络分析. 计算机学报, 2012, 35(4): 741-753)

[99] Mitra B, Tabourier L, Roth C. Intrinsically dynamic network communities. Computer Networks, 2012, 56(3): 1041-1053

[100] Zhao Q, Bhowmick S, Zheng X, Kai Y. Characterizing and predicting community members from evolutionary and heterogeneous networks//Proceedings of the 17th ACM Conference on Information and Knowledge Management. California, USA, 2008: 309-318

[101] Tang Lei, Liu Huan, Zhang Jianping, Nazeri Z. Community evolution in dynamic multi-mode networks//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2008: 677-685

[102] Sun Yizhou, Tang Jie, Han Jiawei, et al. Community evolution detection in dynamic heterogeneous information networks//Proceedings of the 8th Workshop on Mining and Learning with Graphs (MLG). Washington, USA, 2010: 137-146

[103] Sun Y, Yu Y, Han J. Ranking-based clustering of heterogeneous information networks with star network schema//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Paris, France, 2009: 797-806

[104] Sun Yizhou, Tang Jie, Han Jiawei, et al. Co-evolution of multi-typed objects in dynamic star networks. IEEE Transactions on Knowledge and Data Engineering, 2013, (99): 1

[105] Lin Y-R, Sun J, Sundaram H, et al. Community discovery via metagraph factorization. ACM Transactions on Knowledge Discovery from Data (TKDD), 2011, 5(3): 17

[106] Tang Lei, Wang Xufei, Liu Huan. Uncovering groups via heterogeneous interaction analysis//Proceedings of the International Conference on Data Mining (ICDM). Miami, USA, 2009: 503-512

[107] Zhang Chun-Ying, Guo Jing-Feng. The α relationship communities of set pair social networks and its dynamic mining algorithms. Chinese Journal of Computers, 2013, 36(8): 1682-1692(in Chinese)
(张春英, 郭景峰. 集对社会网络 α 关系社区及动态挖掘算法. 计算机学报, 2013, 36(8): 1682-1692)

[108] Lin Wang-Qun, Deng Lei, Ding Zhao-Yun, et al. A new dynamic hierarchical parallel computing community. Chinese Journal of Computers, 2012, 35(8): 1712-1725(in Chinese)
(林旺群, 邓镭, 丁兆云等. 一种新型的层次化动态社区并行计算方法. 计算机学报, 2012, 35(8): 1712-1725)

[109] Gupta M, Gao Jing, Sun Yizhou, Han Jiawei. Community trend outlier detection using soft temporal pattern mining//Proceedings of the European Conference on Macine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). Bristol, UK, 2012: 692-708

[110] Gupta M, Gao Jing, Sun Yizhou, Han Jiawei. Integrating community matching and outlier detection for mining evolutionary community outliers//Proceedings of the International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 859-867

[111] Chen Y, Nyemba S, Malin B. Detecting anomalous insiders in collaborative information systems. IEEE Transactions on Dependable and Secure Computing, 2012, 9(3): 332-344

[112] Kumar R, Novak J, Raghavan P, Tomkins A. On the bursty evolution of blogspace//Proceedings of the 12th WWW Conference. Budapest, Hungary, 2003: 568-576

WANG Li, born in 1971, Ph. D., professor. Her research interests include artificial intelligence, social computing and network communication.

CHENG Xue-Qi, born in 1971, Ph. D., professor. His main research interests include network science and social computing, Internet information security, Web search and mining.

Background

Online social network is a new popular platform. It is an important media for users to maintain social relationship and propagate information. And it takes numerous changes into human life. There are a lot of latent rules that we did not know, especially its dynamic evolution progress.

Community is an important structure in social networks. It influences the network function and its evolution progress. Research on community offers a way for us to understand social network topology and its evolution. So we write this survey paper to show a full view for academicians about work of detecting latent dynamic community and understanding its evolution.

This paper is supported by the project of the National Basic Research Program (973 Program) of China under Grant No. 2013CB329602, the National High Technology Research and Development Program (863 Program) of China under Grant No. 2014AA015204, the National Natural Science Foundation of China under Grant No. 61232010, the Natural Science Foundation of Shanxi Province under Grant No. 2014011022-1 and the National Science Foundation for Post-doctoral Scientists of China under Grant No. 2013M530738. These projects aim to uncover some important principles and offer key technologies to help guiding public sentiment, responding emergency in time and offer more better services than ever. The authors have been working on dynamic network computing and dynamic community evolution. Some methods and software tools have been produced to analyze online social networks. Many papers have been published in international and domestic conferences and journals.