异质网络社区发现研究进展*

阳 雨, 郭 勇, 李海龙, 邓 波
(北京系统工程研究所, 北京 100101)

摘 要: 异质网络将复杂系统中的信息抽象成不同类型的节点和链接关系,不同于同质网络,基于异质网络的社区发现能够挖掘出更加精确的社区结构。异质网络的社区发现通过对异质网络中的多维结构、多模信息、语义信息、链接关系等信息进行建模表示和提取分析,以发现其中相对紧密稳定的社区结构,对网络信息的获取与挖掘、信息推荐以及网络演化预测具有重要的研究价值。首先对社区发现和异质网络进行了简单阐述;随后结合实例介绍了异质网络社区发现的现有研究方法,包括基于主题模型、基于排序和聚类相结合、基于数据重构和基于降维的方法等,并针对各类方法指出了其特点和局限性;最后讨论了当前该领域在结构复杂性、建模复杂性、数据规模等方面面临的挑战。在将来,基于并行化、可扩展、动态增量的研究更能适应当前的变化环境。

关键词: 异质网络; 社区发现; 网络结构
中图分类号: TP393 **文献标志码**: A **文章编号**: 1001-3695(2018)10-2881-07
doi:10.3969/j.issn.1001-3695.2018.10.001

Survey of community detection in heterogeneous networks

Yang Yu, Guo Yong, Li Hailong, Deng Bo
(*Beijing Institute of System Engineering*, *Beijing* 100101, *China*)

Abstract: Most real systems consist of a large number of interacting, multi-typed components, while most contemporary researches model them as homogeneous networks without distinguishing different types of objects and links in the networks. Compared with homogeneous networks, community detection based on heterogeneous networks could obtain more accurate community structures. By modeling and analyzing various information including multi-dimensional structure, multi-mode information and semantic meaning in heterogeneous networks, community detection is to detect relatively stable community and valuable for network information collection and mining, information recommendations and predicting the evolution of networks. Firstly, this paper introduced the community detection and heterogeneous networks. In community detection of heterogeneous networks, the current mainstream methods included topic model, ranking-based clustering, data reconstruction, dimensionality reduction and so on. This paper summarized the above types of methods and analyzed their performance with practical applications. It also discussed the development trend of the community detection in heterogeneous networks. In the future, researches in the parallel, scalable and incremental dynamic heterogeneous networks will get more attention.

Key words: heterogeneous networks; community detection; network structure

现实世界中的很多复杂系统都可以抽象成网络的形式,对于复杂网络的研究能够加深对于不同系统性质的了解。随着科技、经济、社交等领域的不断快速发展,使得各种网络无处不在,如移动通信网络、蛋白质结构网络、电子商务网络、著作研究合作网络、社交网络、疾病传播网络等。尤其是随着全球互联网、社交媒体的发展和移动设备的普及,使得社交网络呈现爆发式增长。据 2016 年最新报告显示[1],全球网民达到 34.2 亿人,相当于全球人口的 46%,增长 3.32 亿人,年增幅 10%;社交媒体用户达到 23.1 亿人,相当于全球人口的 31%,新增社交媒体用户 2.19 亿人,年增幅 10%;手机用户达到 37.9 亿人,相当于全球人口的 51%,新增手机用户 1.41 亿人,年增幅 4%。社区发现是社会网络分析研究领域的一个研究方向,在网络中挖掘出相似的、同类的社区关系,对网络信息的获取与感知、信息推荐和网络演化具有重要的研究价值。

复杂网络可以分为同质网络和异质网络。目前大部分社区发现方法都是针对同质网络进行的,假定网络中的实体和连接关系都是同种类型的,这为社区发现的研究提供了较大程度的便利。但是实际情况下,网络中存在多种不同类型的节点和边,这就为社区发现提出了新的挑战。相应地,异质网络这一概念应运而生,针对异质网络的社区发现方法也越来越多地受到研究人员的关注。异质网络中的异构性,导致社区发现方法的实施面临诸多的问题与挑战。例如,异质网络中互动噪声很多,导致社区发现算法的性能降低;在不同类型的节点之间的互动或者在同一维度间的互动噪声过大,导致无法划分出有意义的社区;并且,并不是所有实体在所有维度或者对所有类型的节点都是活跃的。另外,在异质网络中存在的高维度和多节点类型的特点,也为异质网络的社区研究提出了新的要求。

1 社区发现概述

社区发现是复杂网络研究中的一项重要研究内容,不同的个体由于某一共同的性质而形成一个社区,如存在相同的兴趣爱好、属于同一蛋白质结构、属于同一个研究领域等。社区发现的研究成果可以广泛应用于不同的领域,例如,购物网站中为具有同一爱好的用户社区个性化推荐相同的产品,为相同领域的研究者推荐同一领域的研究成果,为属于同一疾病传播网络中的个体提供相同的疾病预防治疗措施等。

尽管针对社区发现的研究由来已久,但是目前大部分社区发现算法都存在以下问题:

收稿日期: 2017-07-19; **修回日期**: 2017-09-14 **基金项目**: 国家自然科学基金资助项目(61402486)

作者简介:阳雨(1992-),男,硕士,主要研究方向为数据挖掘(fqyangsjtu@gmail.com);郭勇(1966-),男,研究员,博士,主要研究方向为数据挖掘;李海龙(1976-),男,副研究员,硕士,主要研究方向为数据挖掘、软件测试;邓波(1973-),男,研究员,博士,主要研究方向为数据挖掘、软件测试.

a)算法的时间复杂度较高,无法满足目前大规模数据的要求。随着社交网络数据的迅猛增长,传统算法的时间复杂度呈现出指数级增长的现象。

b)算法主要针对静态网络的社区发现,适用于动态网络的算法较少。然而大部分复杂网络都是基于时间不断变化的,高效的动态增量的社区研究是目前亟待解决的问题。

c)算法主要针对同质网络,缺少针对异质网络的社区发现方法。目前随着大规模网络中异构信息的增加,网络结构中通常包含多种关系和实体,具有多维度、多维复杂关系、多类型节点等特点。因此,迫切需要社区发现算法能够充分利用其中的异构信息,从而提高算法的准确性,然而大部分算法并没有考虑这个问题。

2 网络类型的基本定义

复杂网络是为了便于分析自然界个体间形形色色的连接关系而提出的,即将自然界中的实体抽象成网络中的节点,将实体之间的联系抽象成网络中的边。复杂网络按照不同的标准可以划分为不同的类型,本文按照节点和边的类型将其划分为同质网络和异质网络。首先给出信息网络的定义:

定义 1 信息网络(information networks)[2]。一个信息网络可以定义为图 $G=(V,E)$,其中,实体类型映射函数为 $\varphi: V\rightarrow A$,边类型映射函数为 $\psi: E\rightarrow R$。对于任意一个实体对象 $v\in V$,存在一个特定的对象类型 A 使得 $\varphi(v)\in A$;同样地,对于任意一个边对象 $e\in E$,存在一个特定的边类型 R 使得 $\psi(e)\in R$。若两条边属于同一边类型,则必有这两条边的起点和终点分别属于同一种实体类型。

2.1 同质网络

同质网络(homogeneous networks)是复杂网络中较为简单的一种网络结构,是对复杂系统的一种简单抽象,将其中的实体和连接关系抽象成同一种类型的节点和边。

定义 2 同质信息网络。对于一个信息网络,当网络中的节点类型数量 $|A|=1$ 且边类型数量 $|R|=1$ 时,此时的信息网络称为同质信息网络。

由于同质网络较为简单,比较容易进行建模分析,并且发展的历程较长,所以目前大部分社区发现算法都是针对同质网络来进行研究的。但是大部分真实世界的复杂网络结构都是具有异构(heterogeneous)特征的,网络中的节点和关系不是只有一种类型,而是多种多样的,例如,医疗健康网络中有病人、医生、疾病、药物、治疗方法等不同类型的节点;学术社交网络中有作者、论文题目、发表时间和会议等不同类型的节点,以及合作关系、引用关系等不同类型的连接边类型。因此,简单地抽象成同质网络的社区发现研究方法是不严谨的,会造成大量信息丢失。相反,对于异质网络的社区发现研究更加具有针对性,更能提高精确性,更具有实用价值,但是由于其复杂度较高,也为该领域的研究提出了更大的挑战。

2.2 异质网络

在真实世界的网络结构中,如电子商务、文本挖掘等,复杂网络既包括同种类型节点的链接关系形成的网络,也包括不同种类型节点的链接关系形成的网络,此时抽象而成的复杂网络称之为异质网络(heterogeneous networks)。

定义 3 异质信息网络[2]。对于一个信息网络,当网络中的节点类型数量 $|A|>1$ 或边类型数量 $|R|>1$ 时,此时的信息网络称为异质信息网络。

顾名思义,异质网络即指网络中所含有的信息是不同类型的。显然,含有不同类型节点和连接边的异质网络包含更加丰富的网络结构信息,相对于同质网络来说,针对异质网络的研究更能够得到相对准确、相对丰富的价值信息。

根据异质网络的定义,异质网络可以分为多模网络(multi-mode network)和多维网络(multi-dimensional network)两种类型[3]。其中,多模网络侧重于指网络中存在不同类型的节点,每一模即指一种节点类型。特别地,单模网络即是指前面所说的同质网络,是多模网络最简单最基本的形式。图 1 展示了学术社交网络多模的特点,其中包含作者、论文、期刊和研究领域四种节点类型。

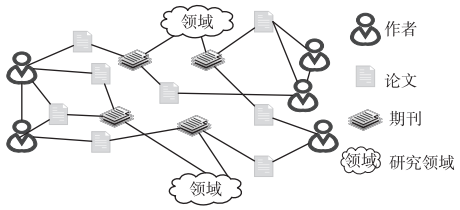

图 1 学术社交网络

多维网络侧重于指网络中存在不同类型的链接关系,即多种关系类型,每一维度表示节点间一种类型的链接关系,即一种交互形式[4]。例如在微博中,用户之间可以私信交流,可以相互关注;一个用户可以点赞、转发或是评论其他用户的微博内容,也可以通过“@”的形式提及另一用户。图 2 是多维网络的示例,展示了针对同一用户群体在不同社交网络平台上的社交关系特点。

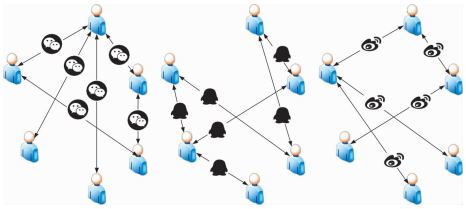

图 2 多维网络示例

3 异质网络社区发现方法

在第 2 章中已经提到,目前大部分社区发现方法主要针对同质网络,然而现实世界中的复杂网络大多带有不同类型节点或不同类型链接关系的信息,这使得抽象成同质网络的社区发现方法并不能准确地反映出其中蕴涵的信息。因此,为了提高研究的精确性和实用性,针对异质网络的社区发现方法逐渐进入研究者的视野。

相对于同质网络,由于异质网络中集成大量多类型的异构信息,以及多维异质网络和多模异质网络中存在的高维度、多节点类型的网络特点,使得传统的同质网络社区发现算法并不适用于异质网络,对异质网络社区发现方法提出了新的挑战。异质网络社区发现是一个逐渐兴起的研究领域,尚处于探索阶段,目前的研究方法还不是很成熟。本章针对现阶段已经提出的大部分异质网络社区发现方法进行一个梳理总结和归类。

3.1 基于主题模型的方法

异质网络中包含大量的文本信息,其中蕴涵了有关社区结构的丰富语义信息,许多研究者采用主题模型(topic model)的方法对其中的文本信息进行集成建模,以提高社区发现结果的准确性。

主题模型也称为话题模型,是在文档集合语义信息中自动发现主题的统计模型。主要包据 LSA(latent semantic analysis)[5]、PLSA(probabilistic latent semantic analysis)[6] 和 LDA(latent Dirichlet allocation)[7] 模型等。其中 LDA 是目前应用最广泛的主题模型,是 PLSA 模型的泛化。

文献[8]提出了 CUT(community-user-topic)模型,将网络

中的语义信息与社区发现相结合。CUT 模型分为 CUT1 和 CUT2 两个模型,其中 CUT1 模型用于主题发现、社区发现和作者主题关注分析,CUT2 模型用于语义社区发现和用户用词习惯分析。这两个模型是两个相互独立的模型,并没有将主题发现、社区发现和社区语义集成到统一的框架中。文献[9]提出了 HLDA(hierarchical-LDA)模型,对抽取出的话题树状关系模型进行分析,从而得到用户的分类,但由于未考虑局部社区,存在社区划分不连通的问题。文献[10,11]采用 ATM(author-topic model)对含有语义信息的学术网络进行论文作者的社区归类。但上述文献仅考虑到了异质网络中的语义信息,忽略了网络结构在社区发现中所含的价值信息。针对上述问题,文献[12~14]分别提出了 NetPLSA(PLSA with network regularization)模型、LapPLSI(Laplacian probabilistic latent semantic indexing)算法和 LTM(locally-consistent topic modeling)模型,将主题建模与网络结构相结合,提高社区发现结果的准确性。

文献[15]指出,尽管上述方法能够处理复杂网络中的文本信息,但并不能定义为异质网络的社区发现,而应该归类为同质网络的研究领域。因此,该文献首先提出 TMBP(topic model with biased propagation)算法将异质网络中的异质信息与主题建模融合到一个统一的框架中;然后在此基础上,根据两个不同的研究观点分别提出了 TMBP-RW 框架(biased random walk framework)和 TMBP-Regu 框架(biased regularization framework)。TMBP-RW 框架中主题建模与随机游走的过程是相互独立的,TMBP-Regu 是在 TMBP-RW 上的改进,将主题建模与网络中的异质信息结合到统一框架中,从而达到协同增强的效果。式(1)以学术社交网络中的作者、文献、会议地点为例展示了 TMBP-Regu 模型。

$$\ell = L(C) - \lambda R(G) =$$
$$\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) -$$
$$\frac{\lambda}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k \mid d_i) - \sum_{u_j \in U_{d_i}} \frac{P(z_k \mid u_j)}{|U_{d_i}|}\right)^2 -$$
$$\frac{\lambda}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k \mid d_i) - \sum_{v_j \in V_{d_i}} \frac{P(z_k \mid v_j)}{|V_{d_i}|}\right)^2 \quad (1)$$

文中分别定义了正则项(regularization term)R_U 和 R_V,R_U 反映了作者($u_l \in U$)与文献($d_i \in D$)的关系,R_V 反映了会议地点($v_j \in V$)与文献($d_i \in D$)的关系;$R(G) = R_U + R_V$ 衡量了主题模型中文本信息与其他类型节点的差异,差异性越大,$R(G)$ 的值越大。从式(1)可以看到,当 $\lambda = 0$ 时,TMBP-Regu 模型可以归结为 PLSA 模型,即未考虑异质网络的结构信息;当 $\lambda > 0$ 时,TMBP-Regu 模型同时考虑了网络中的文本信息和不同类型节点间的异质信息,提高了 PLSA 模型分析的准确性。

部分文献[16~18]认为,大部分主题模型,如 PLSA、LDA、Net-PLSA 等,仅仅是针对异质网络中的文档信息进行分析建模,都是基于不同类型的节点的异质信息是相互独立的假设进行的,并未将文档主题和与其相联系的节点同时进行建模分析,并未考虑相互节点之间的联系。文献[16]提出了一种联合概率主题模型——CTM(collective topic model),同时对异质网络中的不同类型节点的文本内容进行建模。文献[17,18]分别提出了 LSA-PTM 模型和 cluTM 模型,同时对异质网络中的文本信息和节点间的链接信息进行分析建模。

特别地,同样是利用异质网络中的文本信息,但未采用主题模型的方法,文献[19]最近提出了运用通用知识(world knowledge)作为间接监督的方法,然后提出了一种聚类方法用于文档聚类的研究。

3.2　基于排序和聚类相结合的方法

在数据挖掘领域,聚类可以作为独立的研究方向,也可以与其他研究方向相结合进行研究。近年来,在异质网络社区发现的研究中,基于排序的聚类方法逐渐兴起,排序与聚类的研究方法相互促进、相互增强,可以得到较好的研究结果。

在该类方法中,RankClus[20] 和 NetClus[21] 是较早提出的比较经典的方法。RankClus 算法提出了一种异质网络排序聚类方法,该方法基于作者—会议的双类型(bi-typed)异质网络的问题背景,通过对作者和会议类别进行排序,根据目标对象确定聚类对象向量,迭代调整每个对象的类别,最终得到较为准确的作者和会议类别划分。NetClus 算法针对更一般的异质网络结构,即星型网络结构,如图 3(a)表示以论文为中心的学术社交网络的星型结构。与 RankClus 的思想一样,NetClus 也是一个基于排序的迭代方法,即利用排序来提升聚类的效果。但与 RankClus 不同的是,NetClus 能够处理具有星型结构的任意数量的类型对象,而且产生的聚类结果也不是对单个类型对象的集合,而是具有相同拓扑结构的输入网络的子网络集合,如图 3(b)所示。

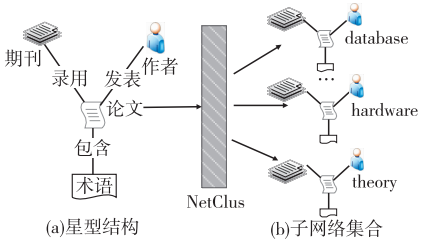

图 3　论文合作网络的聚类过程

文献[22]提出,大部分基于排序的聚类算法仅针对具有异质链接关系的信息网络。而复杂网络中常常表现出混合网络(hybrid network)的特性,即同时具有异质链接关系和同质链接关系的信息,如图 4(a)所示,图 4(b)则反映了混合网络的聚类结果。针对上述问题,该文献提出了 ComClus 算法,应用于带有自循环链接信息的星型异质网络中。

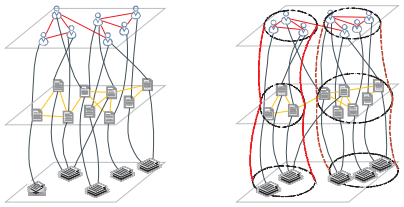

图 4　基于混合网络的聚类

上述算法[20~22]均是针对某一特定的网络结构类型(network schema)提出的,对不同的网络结构类型不具有普适性。针对此研究问题,文献[23,24]提出了对任意网络结构类型均适用的研究算法。文献[23]提出了 HeProjI 算法,将异质网络投射成一系列的子网络,并提出一种信息传导机制保证各个子网络间的一致性。对于每一个子网络,采用基于路径的随机游走策略估计子网络中每一个实体节点的重要性概率,经过反复迭代得到排序和聚类的结果。文献[24]提出了 GPNRankClus 模型,在不需要任何网络结构类型背景的前提下,对不同类型的连接边进行概率建模。但其有一点不足的是,正如文献[22]所指出的,GPNRankClus 仅考虑了异质链接关系的信息,忽略了同种类型节点间的链接信息。

文献[25]提出了 CATHYHIN 框架,利用文本信息和异质链接类型,采用排序与聚类相结合的方法,递归地建立多层级的主题模型结构。该模型也不局限于特定类型的网络结构,并且与 NetClus 算法不同,该方法可以保证最后的结果是收敛的。文献[26]提出了针对有向异质网络的重叠社区发现算法——OcdRank 算法,该算法将重叠社区发现和社区成员排序相结合,具有相对较低的时间复杂度,并且支持增量更新的社

区发现,这在真实场景中具有很大的应用价值。

3.3 基于数据重构的方法

由于异质网络具有多维度、多模的特点,可以将异质网络进行数据重构转换为较简单的网络类型,进而进行社区发现。Liu 等人[27,28]提出的基于链接分析重构的方法,对多维异质网络进行重构,将异质网络中的边或超边作为一类节点,从而将多维异质网络转换为二分图(bipartite graph)。

文献[27]提出,首先将异质网络转换为二分网络(bipartite network),异质网络的顶点转换为二分网络的顶点节点,异质网络的边或者超边转换为二分网络的连接节点。经过此转换过程,当且仅当原异质网络中的顶点和边相连时,二分网络的顶点节点和连接节点相连。如图 5 所示,其中,a、b、c、d 表示顶点节点,边 α 和超边 β,γ 表示连接节点,可以看出同类型节点不存在连接关系,连接边只存在于不同种类型的节点之间,即最终重构成了二分图的结构。

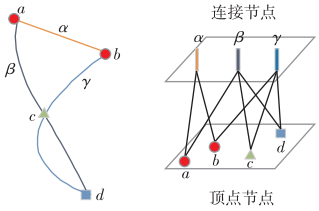

图 5 异质网络转换为二分网络[27]

与传统的二分网络不同,该转换得到的二分网络相对比较稀疏,并且具有不对称性。因此,传统的二分网络中基于模块度优化的社区发现方法并不适用,采用 Murata[29] 提出的模块度优化的方法和 NetFinder 算法[30] 进行该二分网络的社区发现。但是,该方法时间复杂度较高,不适用于大规模网络的社区发现。

文献[28]提出了一种基于模块度优化的异质网络社区发现方法,针对异质网络提出了复合模块度(composite modularity)的概念,在此基础上提出一种快速的启发式算法,即 Louvain-C 算法,进行模块度的优化研究,进而执行社区划分。文中将异质网络 G 划分成一系列的子网络,如式(2)所示,$G^{[s]}$ 表示可以用模块度衡量的子网络,如同质网络、二分网络或 K-模(K-partite)网络等。最终,复合模块度的定义如式(3)所示。其中,$m^{[y]}$ 表示子网络 $G^{[y]}$ 的连接边的数量,m 表示整个网络 G 的连接边的数量,$Q^{[y]}$ 表示子网络 $G^{[y]}$ 的模块度。

$$G = G^{[1]} \cup G^{[2]} \cup G^{[3]} \cup \cdots \cup G^{[s]} \tag{2}$$

$$Q(L) = \sum_{y=1}^{s} \frac{m^{[y]}}{m} Q^{[y]}(L) \tag{3}$$

其中:$m^{[y]} = |E^{[y]}|$;$m = \sum_{y=1}^{s} m^{[y]}$。

文献[28]所提出的方法具有以下优点:a)该方法适用于任意的网络结构类型,也适用于具有超边结构的异质网络;b)事先不需要指定社区个数等参数,不需要任何先验知识用于社区发现;c)相对于文献[27],该方法解决了不可扩展的问题,可以适用于大规模的异质网络。但是,同其他基于模块度优化的社区发现方法一样,该方法同样具有分辨率问题的限制,即不能发现大规模网络结构中比较小规模的社区。

3.4 基于降维的方法

与基于数据重构的方法类似,基于降维的方法同样是将异质网络进行转换,采用降维的方法将异质网络转换成同质网络或是简单的二分网络。常用的降维方法有主成分分析(principal component analysis,PCA)、线性降维分析(linear discriminant analysis,LDA)、非负矩阵分解(non-negative matrix factorization,NMF)等。

PCA[31] 使用线性投影的方法,将高维数据映射到低维空间中,当映射后的数据方差最大时,可以保证尽可能地保留最多的原始信息,该方法能够极大地提升无监督特征的学习速度。LDA[32] 是一种有监督的线性降维方法,该方法使得降维后的数据尽可能地容易被区分。NMF[33] 是一种非线性降维方法,是机器学习中的特征提取和降维技术。相对于其他降维方法,NMF 具有实现上的简便性、可解释性和存储空间少等优点,该方法已经广泛应用于社区发现算法中[34~36]。另外,上述提到的基于主题模型的方法也可以看成是一种基于文本的降维方法,可以将文本中的词向量降维为主题向量。文献[37]提出了 cFTM(contextual focused topic model)算法,对 author 和 conference 进行聚类。该算法的新颖之处在于,在传统的主题模型中考虑了 author 和 conference 的词分布,而且主题模型的个数不需要预先设定。

异质网络经过转换为同质网络和二分网络,前者已经有很多的方法解决,对于后者,目前的社区发现方法主要分为两类。第一类方法是将二分图中的两类节点划分到不同社区,每个社区包含同种类型的节点[38],该方法将二分网络投影为同质网络后再进行同质网络的社区发现,即对两个单模网络进行社区划分,得到社区发现结果。显然,该类方法存在信息丢失的问题。第二类方法将不同类型的节点划分到同一个社区中,一个社区含有不同类型的节点[30],该方法可以尽可能多地保留原有的网络结构信息。

3.5 其他方法研究

除上述提到的几种方法外,研究者还从别的角度将异质网络的属性信息加入到社区发现的算法或模型中。

由于网络中节点和边的分布密度是不同的,若未考虑网络结构密度的差异,会导致较差的社区发现效果,文献[39]提出的方法考虑了不同局部的网络密度属性,可以得到一个相对平衡的异质网络局部社区发现结果。在异质网络中,节点所携带的属性信息可能不完整,或者部分节点根本就没有属性信息,并且连接边所携带的语义信息存在着各种各样的差异,这为在聚类过程中判断连接边重要性的差异带来了挑战。文献[40]针对上述问题提出了 GenClus 算法,解决了异质网络中信息的缺失性和不完整性的问题,并且可根据用户需求自动定义不同类型关系的重要性。同样考虑到异质网络信息的差异性,文献[41]提出了 HRF 模型(heterogeneous random field model)检测并过滤掉其中可能会影响社区发现效果的连接边噪声信息。文献[42]考虑到异质网络中的连接信息和节点的属性信息,提出了第一个考虑异质网络中子空间聚类的分析模型,即 TCSC(typed combined subspace cluster)模型,并且该模型可以检测出不同类型社区之间的交互信息。

部分研究者将基于半监督的聚类方法应用于异质网络的社区发现中[43,44]。文献[43]将元路径(meta-path)与聚类相结合,首先需要预先为每一个聚类提供一部分种子节点,在此基础上,系统学习到元路径的权重值,再根据此权重值产生社区。如图 6 所示,在组织(organization)、作者(author)、会议地点(venues)形成的异质网络中,(a)表示基于组织(O)形成的两个作者聚类{1,2,3,4}和{5,6,7,8},(b)表示基于会议地点(V)形成的两个作者聚类{1,3,5,7}和{2,4,6,8},(c)表示将两元路径结合后形成的四个作者聚类{1,3},{2,4},{5,7}和{6,8}。基于这种思想,文中提出一种有效的迭代算法——PathSelClus 算法用于此模型系统的建立。

与上述提到的众多方法均不同,文献[45]没有采用传统的如主题模型、聚类算法等方法,而是从一个独特的角度,提出了一种新颖的博弈论框架——GHIN 框架,用于定义和发现异质网络中的社区结构,文中将聚类定义为一种纳什均衡的概念。

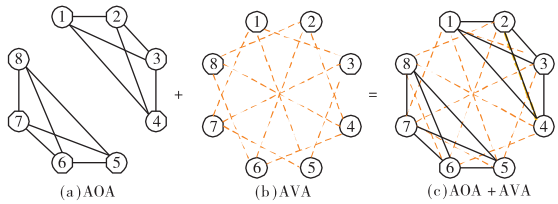

图 6　不同元路径下的聚类过程

4　技术挑战与发展趋势

异质网络的社区发现已经越来越受到研究人员的关注,目前已有比较多的研究方法,但是该领域仍然处于一个探索阶段,具有比较大的发展前景。由于异质网络固有的特性,且目前信息化时代爆炸式的信息增长量,使得该领域的深入研究面临着挑战。本章对该领域的研究热点、技术挑战和发展趋势进行了梳理和总结,并指出未来研究方向。

4.1　网络结构的复杂性

异质网络中,根据网络的结构类型,可以划分为二分图、星型网络、*K*-模网络、多中心网络等,如图 7 所示。若异质网络含有两种类型的节点,且链接关系只存在于不同种类型的节点中,则该网络为二分网络,即二分图。*K*-模网络是二分图的一种扩展,可以包含 k 种类型的节点。二分网络在异质网络的建模中应用较广泛,近年来星型网络结构的应用逐渐增多,如对于 DBLP 数据集[46]和电影网络数据集[47]的研究。为了描述同种类型节点间的链接信息,异质网络也可以用带有自循环的星型网络描述[22],如图 7(c)所示。但是正如上文所提到的,大部分异质网络的社区发现都是针对特定网络结构的[20~22],而相对缺少针对任意结构的方法,文献[23]提出了 HeProjI 方法来解决上述问题。针对任意网络结构的方法,由于其具有的一般性,往往不能很好地对异质网络进行建模,这为此方面的研究提出了较大的挑战。

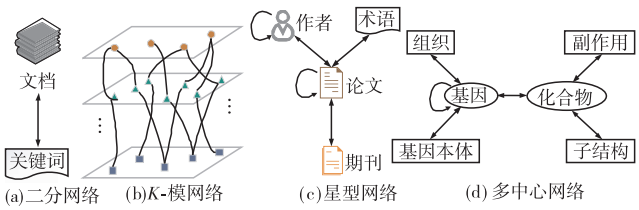

图 7　异质网络结构类型

在实际应用中,网络数据往往更加复杂和不规则。真实网络中的连接关系有时并不能用简单的连接边单独刻画,其中可能含有属性值信息。例如观众与电影的连接关系中,观众可以给电影在 1~5 进行评分,不同的评分表明了观众与电影之间不同的连接信息;在学术社交网络中,作者与文献的关系也有第一作者、第二作者等的区别。在这种情况下,就必须考虑其中的权重信息,从而针对加权异质网络进行研究。此外,在部分实际应用中,节点可能同时存在于多个异质网络中,如对于多维社交网络,用户可能同时存在于 Facebook、微博、微信等社交网络中,为了挖掘出更准确的信息,就需要同时考虑不同社交网络中的连接信息[48]。

目前,出现越来越多的网络可能无法用传统单一的网络结构进行建模分析,如 RDF 数据中包含了大量类型的节点和连接关系,在这种情形下就不能采用常用的网络结构进行分析[19]。随着真实网络结构越来越复杂,需要提出更加有效、更加一般性的网络结构分析方法。

4.2　网络建模的复杂性

针对异质网络的研究中,通常假设该异质网络都是能够明确定义的,其中的链接关系和节点是能清晰建模的。但是实际应用中,该假设往往是不成立的,这就使得对真实数据创建异质网络抽象存在极大的难度。

对于结构化数据,如 DBLP 数据集[46]和电影网络数据集[47]等,可以采用目前常用的网络结构建模方法进行分析。但是,其中的节点和连接边信息可能存在互动噪声,会影响社区发现结果的准确性。其一是,网络中节点之间的连接信息可能并不明确,或者其中的信息不完整,如文献[49]所提及的在 DBLP 数据集中的 advisor-advisee 关系。这种情况下,可以采用链路预测[50]的方法对其中的缺失信息进行填充。其二是,部分节点或链接信息是错误或不可靠的信息,如电子商务网络中不准确的商品信息[51],在多维网络中可能存在具有冲突性的不同信息等。文献[52,53]提出信任模型以从相互冲突的信息源中挖掘出真实可靠的连接信息。

对于非结构化数据,如文本数据、多媒体数据、多语种的数据信息等,这无疑加大了对异质网络表示与建模的难度,更是对社区发现研究的一项挑战。为了构建出准确的异质网络,需要运用信息提取、自然语言处理等关键技术。以文本数据为例,需要采用主题模型等方法将长文本进行压缩,提取出其中重要的短语信息。除此之外,还需要对其中的关系信息进行提取,以构建表示出异质网络。文献[49]提出 TPFG 模型,利用联合似然目标函数对 DBLP 数据集中的 advisor-advisee 关系进行提取。同样地,对于多媒体数据或是多语种数据来说,也需要对其中的实体与连接信息进行提取,以构建出异质网络进行分析。在将来的研究中,怎样高效地提取分析出异质网络中的异质信息会得到越来越多研究者的关注。

4.3　网络的大规模特性

信息化时代的一个显著特点就是,信息呈现出指数级增长的态势,这也使得网络规模增长速度越来越快。针对大规模数据的异质网络社区发现是目前的一个研究热点,同时也是一个研究难点。目前大部分算法都是针对小规模的数据集,并且算法的时间复杂度较高,无法适应目前大规模数据的要求。随着复杂网络数据的迅猛增长,当网络的节点和边的数目增多时,传统算法的时间复杂度呈现出指数级增长的趋势,严重影响在产业界的应用。针对大规模的数据结构,尽管部分文献采用了快速或是并行化的方法对异质网络进行分析[46,54,55],但是针对异质网络社区发现的并行化研究还比较少。此外,针对异质网络的云计算分析也为大规模异质网络的分析提供了一种可行性的研究方法。

近年来,部分研究者提出了并行图挖掘算法[56]和并行图处理平台[57],部分研究者将社区发现的研究聚焦于 Hadoop 或者 Spark 环境的并行化实现[58,59],也取得了一定的成果。特别地,基于 Spark 环境下的 GraphX 分布式图处理框架等开发环境的提出与优化,为大规模异质网络社区发现的并行化研究提供了更加高效的可行性研究平台,极大地促进了该领域并行化研究方法的发展。但是,并行化研究也面临着一系列的难点,例如,在进行异质网络分割时,其中涉及到每个节点的负载均衡问题、不同类型节点的平衡问题,以及节点之间的相互通信问题和其中包含的语义信息等问题。因此,大规模异质网络社区发现的并行化研究必定是以后的研究热点和亟待攻克的难点。

4.4　网络的动态增量特性

目前,大部分异质网络社区发现算法主要针对静态网络,适用于动态网络的算法较少。对于异质网络的研究,通常是将其看成静态图的形式,即网络的数据是整个网络所有数据的集合,或是在某个特定时间点的一个快照数据,该类方法忽略了网络的动态特征。但是现实生活中的网络并不是一成不变的,

随着时间和外界条件的变化,网络也会呈现出相应的变化,例如,新浪微博中增加新的关注或是对某人取消关注,在网络中体现为边的添加和移除;电影网络中,观众对电影的打分情况也是随着时间的推移而变化的。由于静态社区发现未考虑网络的动态增量特性,不能满足对社区发现算法计算精度和效率的要求。所以,需要考虑到时间维度的因素,创建动态异质网络[60],研究适用于动态增量网络的社区发现方法。

对于包含时间序列数据的异质网络,若单纯地将基于静态异质网络的社区发现算法应用于动态异质网络中,会导致在动态网络中频繁地执行静态社区发现算法,造成大量的重复计算,会严重影响到社区发现的时效性。目前对于同质网络中动态网络的研究也相对较少,对于异质网络的动态社区研究则更加具有挑战性,对于动态变化、增量的网络,需要更为高效的算法进行较为精确的社区发现。

5 结束语

本文在对社区发现和异质网络进行简单介绍之后,在较为广泛地分析当前异质网络的社区发现研究的基础上,较为全面地总结了当前异质网络中社区发现的各类方法,分别就其适用条件、算法特点、算法创新以及其不足进行了归纳。本文归纳提炼了基于主题模型、基于排序和聚类相结合、基于数据重构和基于降维四类异质网络社区发现方法,归纳整理了基于属性信息的分析建模、基于半监督的聚类方法等其他比较突出典型的异质网络社区研究方法,并且结合近几年较为出色的相关研究成果进行了较为深入的分析。应该指出,当前异质网络的社区发现研究属于数据挖掘领域的一个研究热点。众多研究方法的出发点是,需要利用异质网络中的什么信息、怎么提取并分析建模存在的信息、如何提高算法效率。

本文对现有研究领域的研究热点、技术挑战和发展趋势进行了梳理分析,指出当前的热点与难点主要集中体现在异质网络中结构的复杂性、网络建模的复杂性、数据的大规模特性和网络的动态增量特性等方面。文章指出:a)目前针对任意网络结构的研究方法较少,且算法效率不高;b)异质网络信息多样,实体和链接信息提取存在一定的难度,并且难以构建合适的异质网络进行分析;c)数据的爆发式增长为异质网络的社区研究在产业界应用提出了新的挑战;d)异质网络其实是基于时间序列的函数,未考虑动态增量特性的研究方法是不准确的。基于上述问题,进一步指出未来该领域的研究方向将会集中在网络结构的普适性算法、高效信息提取、并行化方法研究以及动态增量研究等方面。异质网络的社区研究必定需要考虑学术界和工业界的结合,学术界的研究能为工业界的具体应用提供价值,才能达到研究的目的,考虑到异质网络的特殊性,该领域进一步的研究需要学术界和工业界的共同努力。

参考文献:

[1] We are social: 2016 年全球互联网、社交媒体、移动设备普及情况 [EB/OL]. (2016-12-18). http://www.199it.com/archives/437192.html.

[2] Sun Yizhou, Han Jiawei. Mining heterogeneous information networks: a structural analysis approach [J]. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 20-28.

[3] Tang Lei, Liu Huan. Community detection and mining in social media [M]. San Rafael, CA: Morgan & Claypool Publisher, 2010: 1-137.

[4] Tang Lei, Wang Xufei, Liu Huan. Uncoverning groups via heterogeneous interaction analysis [C]//Proc of the 9th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2009: 503-512.

[5] Papadimitriou C H, Raghavan P, Tamaki H, *et al.* Latent semantic indexing: a probabilistic analysis [J]. Journal of Computer & System Sciences, 2000, 61(2): 217-235.

[6] Hofmann T. Probabilistic latent semantic indexing [C]//Proc of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999: 50-57.

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[8] Zhou Ding, Manavoglu E, Li Jia, *et al.* Probabilistic models for discovering e-communities [C]//Proc of the 15th International Conference on World Wide Web. New York: ACM Press, 2006: 173-182.

[9] Cha Y, Cho J. Social-network analysis using topic models [C]//Proc of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2012: 565-574.

[10] Rosen-Zvi M, Griffiths T, Steyvers M, *et al.* The author-topic model for authors and documents [C]//Proc of the 20th Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia: AUAI Press, 2004: 487-494.

[11] Liu Yan, Niculescu-Mizil A, Gryc W. Topic-link LDA: joint models of topic and author community [C]//Proc of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 665-672.

[12] Mei Qiaozhu, Cai Deng, Zhang Duo, *et al.* Topic modeling with network regularization [C]//Proc of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 101-110.

[13] Cai Deng, Mei Qiaozhu, Han Jiawei, *et al.* Modeling hidden topics on document manifold [C]//Proc of the 17th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2008: 911-920.

[14] Cai Deng, Wang Xuanhui, He Xiaofei. Probabilistic dyadic data analysis with local and global consistency [C]//Proc of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 105-112.

[15] Deng Hongbo, Han Jiawei, Zhao Bo, *et al.* Probabilistic topic models with biased propagation on heterogeneous information networks [C]//Proc of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 1271-1279.

[16] Deng Hongbo, Zhao Bo, Han Jiawei. Collective topic modeling for heterogeneous networks [C]//Proc of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2011: 1109-1110.

[17] Wang Qian, Peng Zhaohui, Jiang Fei, *et al.* LSA-PTM: a propagation-based topic model using latent semantic analysis on heterogeneous information networks [C]//Proc of International Conference on Web-Age Information Management. Berlin: Springer, 2013: 13-24.

[18] Wang Qian, Peng Zhaohui, Wang Senzhang, *et al.* cluTM: content and link integrated topic model on heterogeneous information networks [C]//Proc of International Conference on Web-Age Information Management. Cham: Springer, 2015: 207-218.

[19] Wang Chengguan, Song Yangqiu, El-Kishky A, *et al.* Incorporating world knowledge to document clustering via heterogeneous information networks [C]//Proc of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2015: 1215-1224.

[20] Sun Yizhou, Han Jiawei, Zhao Peixiang, *et al.* RankClus: integrating clustering with ranking for heterogeneous information network analysis [C]//Proc of the 12th International Conference on Extending Database Technology: Advances in Database Technology. New York: ACM Press, 2009: 565-576.

[21] Sun Yizhou, Yu Yintao, Han Jiawei. Ranking-based clustering of heterogeneous information networks with star network schema [C]//Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 797-806.

[22] Wang Ran, Shi Chuan, Yu P S, *et al.* Integrating clustering and ranking on hybrid heterogeneous information network [C]//Proc of

Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2013: 583-594.
[23] Shi Chuan, Wang Ran, Li Yitong, *et al.* Ranking-based clustering on general heterogeneous information networks by network projection [C]//Proc of the 23rd ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2014: 699-708.
[24] Chen Junxia, Dai Wei, Sun Yizhou, *et al.* Clustering and ranking in heterogeneous information networks via gamma-Poisson model [C]// Proc of SIAM International Conference on Data Mining. [S. l.]: SIAM Press, 2015: 424-432.
[25] Wang Chi, Liu Jialu, Desai N, *et al.* Constructing topical hierarchies in heterogeneous information networks [J]. Knowledge and Information Systems, 2015, 44(3): 529-558.
[26] Qiu Changhe, Chen Wei, Wang Tengjiao, *et al.* Overlapping community detection in directed heterogeneous social network [C]//Proc of International Conference on Web-Age Information Management. Cham: Springer, 2015: 490-493.
[27] Liu Weichu, Murata T, Liu Xin. Community detection on heterogeneous networks [C]//Proc of the 27th Annual Conference of Japanese Society for Artificial Intelligence. 2013.
[28] Liu Xin, Liu Weichu, Murata T, *et al.* A framework for community detection in heterogeneous multi-relational networks [J]. Advances in Complex Systems, 2014, 17(6): 1450018.
[29] Murata T. Detecting communities from bipartite networks based on bipartite modularities [C]//Proc of International Conference on Computational Science and Engineering. Picataway, NJ: IEEE Press, 2009: 50-57.
[30] Liu Xin, Murata T. Detecting communities in *K*-partite *K*-uniform (hyper) networks [J]. Journal of Computer Science and Technology, 2011, 26(5): 778-791.
[31] Jolliffe I. Principal component analysis [M]. 2nd ed. New York: Springer-Verlag, 2002.
[32] Mika S, Ratsch G, Weston J, *et al.* Fisher discriminant analysis with kernels [C]//Proc of IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing. Washington DC: IEEE Computer Society, 1999: 41-48.
[33] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401(6755): 788-791.
[34] Wang Wenjun, Jiao Pengfei, He Dongxiao, *et al.* Autonomous overlapping community detection in temporal networks: a dynamic Bayesian nonnegative matrix factorization approach [J]. Knowledge-Based Systems, 2016, 110(10): 121-134.
[35] Psorakis I, Roberts S, Ebden M, *et al.* Overlapping community detection using Bayesian non-negative matrix factorization [J]. Physical Review E, 2011, 83(6): 066114.
[36] Yang J, Leskovec J. Overlapping community detection at scale: a nonnegative matrix factorization approach [C]//Proc of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2013: 587-596.
[37] Chen Xu, Zhou Mingyuan, Carin L. The contextual focused topic model [C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 96-104.
[38] Guimerà R, Sales-Pardo M, Amaral L A N. Module identification in bipartite and directed networks [J]. Physical Review E, 2007, 76(3): 036102.
[39] Aggarwal C C, Xie Yan, Yu P S. Towards community detection in locally heterogeneous networks [C]//Proc of SIAM International Conference on Data Mining. [S. l.]: SIAM Press, 2011: 391-402.
[40] Sun Yizhou, Aggarwal C C, Han Jiawei. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes [J]. Proceedings of the VLDB Endowment, 2012, 5(5): 394-405.
[41] Qi Guojun, Aggarwal C C, Huang T S. On clustering heterogeneous social media objects with outlier links [C]//Proc of the 5th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2012: 553-562.
[42] Boden B, Ester M, Seidl T. Density-based subspace clustering in heterogeneous networks [C]//Proc of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2014: 149-164.
[43] Sun Yizhou, Norick B, Han Jiawei, *et al.* PathSelClus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks [J]. ACM Trans on Knowledge Discovery from Data, 2013, 7(3): article No 11.
[44] Luo Chen, Pang Wei, Wang Zhe. Semi-supervised clustering on heterogeneous information networks [C]//Proc of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2014: 548-559.
[45] Alqadah F, Bhatnagar R. A game theoretic framework for heterogenous information network clustering [C]//Proc of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 795-804.
[46] Sun Yizhou, Han Jiawei, Yan Xifeng, *et al.* PathSim: meta path-based top-k similarity search in heterogeneous information networks [J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.
[47] Shi Chuan, Zhou Chong, Kong Xiangnan, *et al.* HeteRecom: a semantic-based recommendation system in heterogeneous networks [C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 1552-1555.
[48] Zhang Jiawei, Philip S Y. Multiple anonymized social networks alignment [C]//Proc of IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2015: 599-608.
[49] Wang Chi, Han Jiawei, Jia Yuntao, *et al.* Mining advisor-advisee relationships from research publication networks [C]//Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 203-212.
[50] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks [J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.
[51] Wang Guan, Xie Sihong, Liu Bing, *et al.* Identify online store review spammers via social review graph [J]. ACM Trans on Intelligent Systems and Technology, 2012, 3(4): 1-21.
[52] Yin Xiaoyin, Han Jiawei, Philip S Y. Truth discovery with multiple conflicting information providers on the Web [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(6): 796-808.
[53] Zhao Bo, Rubinstein B I P, Gemmell J, *et al.* A Bayesian approach to discovering truth from conflicting sources for data integration [J]. Proceedings of the VLDB Endowment, 2012, 5(6): 550-561.
[54] Shi Chuan, Kong Xiangnan, Huang Yue, *et al.* HeteSim: a general framework for relevance measure in heterogeneous networks [J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(10): 2479-2492.
[55] Meng Xiaofeng, Shi Chuan, Li Yitong, *et al.* Relevance measure in large-scale heterogeneous networks [C]//Proc of Asia-Pacific Web Conference. Cham: Springer, 2014: 636-643.
[56] Cohen J. Graph twiddling in a MapReduce world [J]. Computing in Science & Engineering, 2009, 11(4): 29-41.
[57] Kang U, Tsourakakis C E, Faloutsos C. PEGASUS: a Peta-scale graph mining system implementation and observations [C]//Proc of the 9th IEEE International Conference on Data Mining. Picataway, NJ: IEEE Press, 2009: 229-238.
[58] Buzun N, Korshunov A, Avanesov V, *et al.* EgoLP: fast and distributed community detection in billion-node social networks [C]//Proc of IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2014: 533-540.
[59] Gonzalez J E, Xin R S, Dave A, *et al.* GraphX: graph processing in a distributed dataflow framework [C]//Proc of the 11th USENIX Conference on Operating Systems Design and Implementation. Berkeley, CA: USENIX Association, 2014: 599-613.
[60] Sun Yizhou, Tang Jie, Han Jiawei, *et al.* Community evolution detection in dynamic heterogeneous information networks [C]//Proc of the 8th Workshop on Mining and Learning with Graphs. New York: ACM Press, 2010: 137-146.