软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2699−2709 [doi: 10.3724/SP.J.1001.2013.04474]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于随机块模型的快速广义社区发现算法*

柴变芳[1,2], 于 剑[1], 贾彩燕[1], 王静红[3]

[1](交通数据分析与挖掘北京市重点实验室(北京交通大学),北京 100044)

[2](石家庄经济学院 信息工程系,河北 石家庄 050031)

[3](河北师范大学 信息技术学院,河北 石家庄 050024)

通讯作者: 于剑, E-mail: jianyu@bjtu.edu.cn

摘 要: 随机块模型可以生成各种不同结构(称作广义社区,包括传统社区、二分结构、层次结构等)的网络,也可以根据概率对等原则发现网络中的广义社区.但简单的随机块模型在网络生成过程建模和模型学习方面存在许多问题,导致不能很好地发现实际网络的结构,其扩展模型 GSB(general stochastic block)基于链接社区思想发现广义社区,但时间复杂度限制其在中大型规模网络中的应用.为了在无任何先验的情形下探索不同规模网络的潜在结构,基于 GSB 模型设计一种快速算法 FGSB,更快地发现网络的广义社区.FGSB 在迭代过程中动态学习网络结构参数,将 GSB 模型的参数重新组织,减少不必要的参数,降低算法的存储空间;对收敛节点和边的参数进行裁剪,减少每次迭代的相关计算,节省算法的运行时间.FGSB 与 GSB 模型求解算法有相同的结构发现能力,但 FGSB 耗费的存储空间和运行时间比 GSB 模型求解算法要低.在不同规模的人工网络和实际网络上验证得出:在近似相同的准确率下,FGSB 比 GSB 模型求解算法快,且可发现大型网络的广义社区.

关键词: 随机块模型;广义社区;时间复杂度;复杂网络

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 柴变芳,于剑,贾彩燕,王静红.一种基于随机块模型的快速广义社区发现算法.软件学报,2013,24(11):2699−2709. http://www.jos.org.cn/1000-9825/4474.htm

英文引用格式: Chai BF, Yu J, Jia CY, Wang JH. Fast algorithm on stochastic block model for exploring general communities. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2699−2709 (in Chinese). http://www.jos.org.cn/1000-9825/4474.htm

Fast Algorithm on Stochastic Block Model for Exploring General Communities

CHAI Bian-Fang[1,2], YU Jian[1], JIA Cai-Yan[1], WANG Jing-Hong[3]

[1](Beijing Key Laboratory of Traffic Data Analysis and Mining (Beijing Jiaotong University), Beijing 100044, China)

[2](Department of Information Engineering, Shijiazhuang University of Economics, Shijiazhuang 050031, China)

[3](College of Information Technology, Hebei Normal University, Shijiazhuang 050024, China)

Corresponding author: YU Jian, E-mail: jianyu@bjtu.edu.cn

Abstract: A stochastic block model can produce a wide variety of networks with different structures (named as general community, including traditional community, bipartite structure, hierarchical structure and etc); it also can detect general community in networks according to the rules of stochastic equivalence. However, the simple stochastic block model has some problems in modeling the generation of the networks and learning the models, showing poor results in fitting the practical networks. The GSB (general stochastic block) model is an extension of the stochastic block model, which is based on the idea of link community and is provided to detect general communities. But its complexity limits its applications in medium and large networks. In order to explore the latent structures of networks

* 基金项目: 国家自然科学基金(61033013, 61370129); 教育部创新团队项目(K10JB00440); 北京市自然科学基金(4112046); 中央高校基本科研业务费专项资金; 河北省科技厅项目(13210702D); 河北省教育厅项目(ZD2010128); 民航局科技基金项目(K12I00051)

收稿时间: 2013-04-22; 修改时间: 2013-07-17; 定稿时间: 2013-08-27

© 中国科学院软件研究所 http://www.jos.org.cn

with different scales without prior knowledge about networks, a fast algorithm on the GSB model (FGSB) is designed to explore general communities in networks faster. FGSB dynamically learns the parameters related to the network structure in the process of iterations. It reduces the storage memory by reorganizing parameters to cut down unnecessary parameters, and saves the running time by pruning the related parameters of converging nodes and edges to decrease the computing time of each iteration. FGSB has the same ability of structure detection as the GSB model, but its complexities of time and storage are lower. Tests on synthetic benchmarks and real-world networks have demonstrated that FGSB not only can run faster than the algorithm of the GSB model in the similar accuracy, but also can detect general communities for large networks.

Key words: stochastic block model; general community; time complexity; complex network

互联网上存在许多大型在线网络,如微博用户网、作家合作网、网页引用网,发现这些网络的结构有助于人们分析网络,进而利用分析结果指导实际应用,如舆情监测、个性化推荐.已有的社区发现方法可用来发现团内紧密、团间稀疏的网络结构(以下称该类结构为传统社区)[1],研究表明,此类方法存在这样的问题:(1) 传统的社区发现方法大多采用启发式方法发现紧密子图[2],缺乏理论依据;(2) 流行的传统社区发现方法——模块化方法,存在分辨率和尺度问题[3],这类方法发现的结构不能反映实际网络的中观结构;(3) 人们关于实际在线网络结构规律的先验很少,不清楚网络中是否存在链接紧密的子图,是否还存在其他类型的结构,以及这些结构间以何种方式重叠和组织,因此传统社区发现方法在实际网络中没有传统社区时就不能发挥作用[4].最新研究结果表明,实际网络中确实存在多种类型的结构[5−7],因此,能够发现更多类型结构的网络结构发现模型更符合实际在线网络结构发现的应用需求.

随机块模型 SBM(stochastic block model)是由社会学领域角色分析模型——块模型(block model)发展而来的,根据概率对等性(stochastic equivalent)将具有相似角色的节点分类.该模型不对网络结构做任何假设,可较好地发现未知先验的网络的结构[8,9],是目前很有前景的网络结构发现模型,在社会学、统计物理学、计算机科学、生物学等领域引起人们的关注.SBM 是一个网络生成模型,分两步生成网络:先为所有节点指派隶属类;然后根据类及类间链接概率决定任意两个节点间是否产生链接.类间链接概率矩阵的引入,使该模型可以灵活地产生各种类型的网络结构,如:对角链接概率矩阵可产生由独立子图组成的网络;对角元素较大、非对角元素较小的链接概率矩阵可产生同配(assortative mixing)网络结构(即传统社区结构);非对角元素较大、对角元素较小的链接概率矩阵可产生异配(disassortative mixing)网络结构(如多分图);变换矩阵形式可产生更丰富的混合网络结构.以下称 SBM 根据概率对等性定义的类为广义社区,它包括传统社区.除了能够产生多种类型结构的网络以外,SBM 还可以完成比传统社区发现更广的网络结构发现任务.其利用模型假设和参数生成的网络我 拟合实际观测网络,并通过 EM 算法、变分 EM 算法、Gibbs 采样等方法求解网络节点的社区指派参数及社区链接概率矩阵参数.由社区指派参数可得网络广义社区划分结果,由链接概率矩阵参数可得社区间复杂交互模式,这使 SBM 成为网络结构发现的一个有效工具.简单 SBM 不能很好地拟合实际观测网络,许多研究对 SBM 的参数学习和类个数选择方法进行改进,但最多处理几百个节点的网络[10−16].此外,这类研究生成网络的过程和简单 SBM 一样,在生成链接的过程中没有考虑一些实际因素,如不同节点在网络生成过程中产生边的差异.近年来,有研究者致力于改进 SBM 的生成过程,典型的研究有:2008 年,Airoldi 等人提出的混合隶属度随机块 MMSB (mixed membership stochastic block)模型[17],2011 年,Karrer 等人提出的度纠正随机块模型(degree-corrected stochastic block model)[2],中科院计算所沈华伟等人提出的扩展随机块 GSB(general stochastic block)模型[5].

GSB 模型基于随机块模型和链接社区思想,对有向网络的生成过程建模,在生成过程中考虑节点的不同角色,可以有效地求解网络广义社区及社区间的交互规律.基于 GSB 模型的广义社区发现算法不仅能够发现传统社区,而且能够发现传统社区之外的更多类型的结构以及这些结构间的交互模式.但 GSB 模型求解算法的复杂度限制其只能发现中小型网络上的广义社区,影响了该模型的广泛应用.为了有效地发现未知先验的网络的广义社区,本文对 GSB 模型参数求解进行变型,从理论上推导变型后求解的参数与 GSB 模型求解参数结果一致;并设计一种快速广义社区发现算法 FGSB,利用参数变换消除不必要的参数,降低了算法的空间复杂度;利用参数裁剪方法裁剪迭代中收敛的节点和边的相关参数计算,节省了算法运行时间.在不同规模的人工网络和实际

© 中国科学院软件研究所 http://www.jos.org.cn

网络上,证明该算法与 GSB 模型求解算法有相近的准确率,却有更快的运行速度,且能处理大规模的网络.

本文第 1 节总结和分析随机块模型的相关工作.第 2 节给出 GSB 模型的生成过程和参数求解方法.第 3 节阐述快速广义社区发现算法 FGSB 的设计动机、原理及算法细节.第 4 节给出实验设置及结果分析.第 5 节总结全文并展望随机块模型未来的研究方向.

1 相关工作

随机块模型是由社会学领域的块模型(block model)发展而来的,块模型的先验模型已知节点类别属性,根据类别将相同属性的节点划分到一个块.该模型的判别形式依据结构对等性(structural equivalent)[18]和规则对等(regular equivalent)[19]将网络节点分组,在社会学的角色分析中有广泛应用.节点具有结构对等性,当且仅当这些节点与网络中剩余的所有节点链接模式相同;节点具有规则对等性,当且仅当它们与规则对等节点有相似的链接.1981 年,Fienberg 和 Wasserman 扩展块模型的结构对等性为概率对等性[8];随后,Wasserman 和 Anderson 提出了随机块模型 SBM[20].SBM 的求解和假设存在一些问题,大致出现了两类相关研究:改进 SBM 模型学习方法和扩展 SBM 网络生成过程.

改进 SBM 模型学习方法的研究主要包括模型参数估计和模型选择.这类研究试图更快、更准确地学习 SBM 的社区划分和社区链接概率矩阵参数.这些方法的主要区别在于模型参数的先验设置、类的个数选择以及算法的加速方式等.1997 年,Snijders 和 Nowicki 研究了无向图的二分类随机块模型的参数求解[10],其后,又将模型扩展为有向网络的多类随机块模型[11].这两个模型假设节点的社区变量服从以各社区节点比例为参数的多项式分布,网络链接服从以节点社区指派和社区链接概率为参数的伯努利分布,并分别对参数设置先验,用吉布斯采样估计参数.这两个模型可发现较复杂的网络结构规律,但只能处理节点数小于 200 的网络.2008 年,Daudin 等人在给定类个数和观测网络的前提下,利用变分 EM 算法求解 SBM 参数以提高算法性能,并利用贝叶斯模型选择方法 ICL(integrated classification likelihood)进行模型选择[12].该方法的缺点是低估了网络的社区个数.Latouche 等人用变分贝叶斯方法求解参数,解决模型过拟合问题,用 IL_{Vb} 进行模型选择,解决 ICL 估计社区个数保守的问题[13].为了发现重叠社区,Latouche 等人扩展随机块模型以处理重叠社区发现问题[14].已有方法假设社区链接矩阵为任意形式,该复杂形式导致参数求解速度慢.Zanghi 等人限制 SBM 链接概率矩阵参数为传统社区结构形式,并通过典型 EM 算法在线学习参数以提高算法收敛速度[15],但该方法丧失了随机块模型链接概率矩阵可表示任意结构网络的优势.McDaid 等人设计的 Collapsed SBM 模型精确推导整个数据的似然,以避免已有方法近似似然导致的偏差,用变维 MCMC 方法求解参数以提高参数的求解速度,同时将社区个数作为模型的变量,用精确推导得出社区个数的估计,最后证明其在解决万数节点时速度优于已有的求解方法[16].

已有的 SBM 学习方法大多有相应的 R 程序包(http://cran.r-project.org),有很好的理论基础,运行速度较快,模型选择方法比较鲁棒.但这些方法假设生成链接只与节点的社区指派有关,导致生成的网络不能很好地拟合实际幂律度分布的网络,一般的划分效果并不理想.主要原因是:随机块模型在生成网络的过程中只考虑边的两个端点的社区指派及社区间链接概率的影响,还有许多其他因素(如节点度)也影响网络链接的生成过程.因此,最近有些研究通过改进随机块模型的网络生成过程来提高网络结构发现的性能.

一些扩展随机块模型在网络生成过程中考虑更多影响因素,以便生成更实际的网络,即度服从幂律分布的网络;进而用生成过程得到的网络拟合观测网络,利用各种参数估计方法(如最大似然估计、贝叶斯估计、collapsed gibbs sampling)求得网络划分结果.这些扩展模型在发现的社区类型、重叠性等方面都有差异.PHITS (probabilistic hypertext-induced topic selection)[21],PCL(probabilistic conditional link)[22],PPL(popularity and productivity link)[23],SPAEM(simple probabilistic algorithm for community detection employing expectation maximization)[24]等扩展模型假设相同社区中的节点链接概率较大,只考虑同一社区的两个节点生成的链接,主要用来识别网络中重叠的传统社区.Karrer 和 Newman 提出的度纠正随机块模型 DCSB(degree-corrected stochastic block model)[2]在网络链接生成过程中考虑节点度的影响,可以更好地识别网络中的非重叠广义社区.但 DCSB 模型求解方法采用启发式 MCMC 方法,速度较慢,且不能发现重叠社区.混合隶属度随机块 MMSB 模

© 中国科学院软件研究所 http://www.jos.org.cn

型(mixed membership stochastic block model)[17]假设每个节点隶属多个社区,生成边的概率与两个节点社区及社区间链接概率相关,可识别网络的重叠广义社区.但MMSB模型只能生成社区重叠部分边密集的网络,而实际存在重叠社区的网络也有不符合这种情况的结构.文献[25,26]假设网络边的两个端点服从对称的分布,利用主题模型对有向稀疏网络建模,并用 Collapsed Gibbs Sampling 方法求解参数.但文献[25]只能识别传统社区.文献[26]提出的模型可识别网络中的同配或异配重叠社区,但其关于有向链接节点角色一致的假设不符合实际,且参数求解算法需要设置多个超参数,影响实际应用.中国科学院计算技术研究所沈华伟等人提出的扩展随机块GSB(general stochastic block)模型[5]利用链接社区思想扩展随机块模型,在网络生成过程中考虑节点的不同角色.与已有的扩展随机块模型相比,GSB 模型可以生成更符合实际的网络,但是该模型的复杂度较高,限制其广泛应用.Brian 等人[27]利用链接社区思想设计一种无向网络的重叠传统社区发现算法,并给出一种快速实现方法,但其只能发现无向网络的传统社区结构.因此,本文拟在已有研究的基础上,对扩展随机块模型 GSB 的参数求解算法进行改进,以扩展其应用场景.

2 扩展随机块模型 GSB

本节首先介绍简单随机块模型 SBM,其次描述扩展随机块模型 GSB 的生成过程及参数求解方法.

简单随机块模型 SBM 假设网络 A 的链接生成过程分两步:

1) 将网络中每个节点 i 以概率 θ_k 指派到第 k 个社区,θ 为 K 维向量,节点 i 的社区变量用 Z_i 表示:

$$Z_i \sim Mult(\theta);$$

2) 每对节点产生链接 A_{ij} 的概率服从伯努利分布,$A_{ij} \sim Bernoulli(\omega_{z_i z_j})$,$\omega$ 为 $K\times K$ 维矩阵,每个元素 ω_{rs} 表示社区 r 和 s 间生成链接的概率.

根据已有参数求解方法可估计 Z 和 ω,从而获得网络节点的社区指派及社区间的交互矩阵.

简单随机块模型假设生成链接的概率只与两个端点的社区相关,社区选择各节点的概率相同.而实际的链接生成过程与很多因素有关,如有些节点是权威的节点,其被社区选择的概率相对要大.另外,简单随机块模型只能用来进行非重叠社区发现.因此,简单随机块模型不符合实际应用需求.扩展随机块模型 GSB(generative stochastic block model)为解决此问题提供了一种有效方法,其采用链接社区的思想(即如果一个节点有多种类型的边,则其隶属于多个社区),对有向网络生成过程建模,该假设保证模型能够生成重叠社区.GSB 模型可发现广义社区,该定义与随机块模型的定义相同,即社区 r 中节点与社区 s 中节点的链接概率相同.其关于社区的定义保证 GSB 模型能够发现多种类型的网络结构.GSB 模型在有向链接的生成过程中,假设节点被选择的概率与节点在本社区中的重要性相关.该假设保证模型能产生更实际的网络.假设有向网络 A 有 K 个社区,社区 r 中节点和社区 s 中节点以相同的概率 ω_{rs} 产生链接,则网络产生每个链接$\langle i,j\rangle$的过程如下:

1) 以概率 ω_{rs} 为链接$\langle i,j\rangle$选择社区对$\langle r,s\rangle$,ω_{rs} 满足限制 $\sum_{rs}\omega_{rs}=1$.

2) 以概率 θ_{ri} 从社区 r 中选择节点 i,θ_{ri} 满足限制 $\sum_{i}\theta_{ri}=1$.

3) 以概率 ϕ_{sj} 从社区 s 中选择节点 j,ϕ_{sj} 满足限制 $\sum_{j}\phi_{sj}=1$.

根据生成过程,可得网络的对数似然函数如下:

$$\ln P(A\mid \omega,\theta,\phi)=\sum_{ij}A_{ij}\ln\left(\sum_{rs}\omega_{rs}\theta_{ri}\phi_{sj}\right) \tag{1}$$

基于对数似然及参数的限制条件,用 EM 算法求解最大对数似然函数的参数,E 步计算每条边的潜在社区对指派概率分布 q,M 步计算参数:社区交互概率矩阵 ω,产生链接的节点中心度 θ 和接收链接的节点中心度 ϕ.潜在变量和参数计算公式如公式(2)~公式(5)所示:

$$q_{ijrs}=\frac{\omega_{rs}\theta_{ri}\phi_{sj}}{\sum_{rs}\omega_{rs}\theta_{ri}\phi_{sj}} \tag{2}$$

© 中国科学院软件研究所 http://www.jos.org.cn

$$\theta_{ri}=\frac{\sum_{js}A_{ij}q_{ijrs}}{\sum_{ijs}A_{ij}q_{ijrs}} \tag{3}$$

$$\phi_{sj}=\frac{\sum_{ir}A_{ij}q_{ijrs}}{\sum_{ijr}A_{ij}q_{ijrs}} \tag{4}$$

$$\omega_{rs}=\frac{\sum_{ij}A_{ij}q_{ijrs}}{\sum_{ijrs}A_{ij}q_{ijrs}} \tag{5}$$

利用ω可得到网络各社区间的交互矩阵,利用θ和ϕ可计算节点的软社区隶属度,进一步可利用这3个参数进行重叠和非重叠社区发现.令 GSB 模型中的θ和ϕ相等,可处理无向网络的广义社区发现问题;令 A_{ij} 为[0,1]区间的任意值,可处理有权网络的广义社区发现问题.

3 快速广义社区发现算法 FGSB

GSB 模型采用 EM 算法进行参数求解,拟对算法空间复杂度和时间复杂度进行分析,以提高算法效率:

1) 存储隐变量 q 的空间复杂度为 $O(m\times K\times K)$,存储模型参数的空间复杂度为 $O(N\times K)$,其中,m 为网络边数,K 为社区个数,N 为网络节点个数.当网络较大时,通常边数 m 远大于节点个数 N,因此,算法存储空间主要由隐变量 q 占用,拟通过减少 q 的空间来减少算法空间复杂度.
2) GSB 参数求解算法每次迭代都需要计算隐变量及模型参数,时间复杂度为 $O(m\times K\times K)$,如果 EM 算法通过 T 次迭代以达到收敛,运行 R 次 EM 算法来逼近全局最优,则 GSB 模型参数学习算法需要耗费 $O(R\times T\times m\times K\times K)$时间才能得到较优的结果.在参数学习算法动态迭代过程中,许多参数没有等到迭代结束已经收敛,或者有些计算涉及的参数值很小,对参数不起作用.因此,拟通过减少每次迭代不必要的参数计算,以减少算法运行时间.

下面从参数存储空间和参数计算时间两个方面对模型参数求解设计改进策略,最后给出基于这些策略设计的快速广义社区发现算法 FGSB(fast general stochastic block model).

3.1 存储空间的改进策略

GSB 模型参数的求解采用 EM 算法,需要计算并存储两类数据:模型参数ω,θ和ϕ及隐变量 q.计算模型参数时依赖于所有边的社区对概率分布,即隐变量 q,因此需要存储 q 的整个分布.但算法最终不需要输出空间复杂度为 $O(m\times K\times K)$的 q.这里重新组织算法参数,使算法不用存储整个隐变量分布 q,仅存储计算当前边涉及的社区对的概率分布,从而减少算法的空间复杂度.为实现此目的,需要引入变量 B,E 和 W,计算公式如公式(6)~公式(8)所示:

$$B_{ir}=\sum_{js}q_{ijrs} \tag{6}$$

$$E_{js}=\sum_{ir}q_{ijrs} \tag{7}$$

$$W_{rs}=\sum_{ij}q_{ijrs} \tag{8}$$

其中,B_{ir} 表示节点 i 从社区 r 发出的平均边数,E_{js} 表示节点 j 在社区 s 接收的平均边数,W_{rs} 表示社区对$\langle r,s\rangle$间从 r 到 s 的平均边数.公式(6)~公式(8)分别为公式(3)~公式(5)的分子.q 与 GSB 中的含义一致,表示每条边$\langle i,j\rangle$在社区对$\langle r,s\rangle$上的概率分布.不同于 GSB 的是:变量 q 为临时存储变量,每条边的 q 分布值根据 B,E 和 W 计算,只需要耗费 $O(K\times K)$的临时空间存储当前边的社区对概率分布.每条边$\langle i,j\rangle$的 q 计算如下:

$$q_{ijrs}=\frac{B_{ir}E_{js}W_{rs}}{DK_rP_s} \tag{9}$$

$$D=\sum_{rs}\frac{B_{ir}E_{js}W_{rs}}{K_rP_s} \tag{10}$$

GSB 中,θ分母的计算相当于公式(11),ϕ分母的计算相当于公式(12):

© 中国科学院软件研究所 http://www.jos.org.cn

$$K_r = \sum_i B_{ir} \tag{11}$$

$$P_s = \sum_j E_{js} \tag{12}$$

根据 B 和 E 可计算 GSB 的参数θ和ϕ,如式(13)和式(14)所示.为计算方便,每次迭代都对 W 进行均一化,均一化的 W 与参数ω相等.

$$\theta_{ri} = \frac{B_{ir}}{K_r} \tag{13}$$

$$\phi_{sj} = \frac{E_{js}}{P_s} \tag{14}$$

给定前一次迭代得到的参数值 B,E 和 W,利用公式(9)计算当前边$\langle i,j\rangle$的临时社区对分布 q,然后将当前边的社区对分布按公式(6)~公式(8)累加到两个端点的当前平均边变量 NB 和 NE 及 NW 上.

利用上述改进策略可使新设计的参数求解算法实现 GSB 参数求解算法的相同功能,存储空间由 $O(m\times K\times K)$变为 $O(N\times K)$,即前一次迭代的平均边变量 B,E,W 和当前迭代的平均边变量 NB,NE,NW 所占的空间.

3.2 运行时间的改进策略

GSB 模型参数的求解采用 EM 算法,在 E 步计算所有边的社区对分布 q,在 M 步利用相关 q 更新平均边参数.改进的参数求解策略遍历网络一次,更新一次节点的平均边参数 B,E 和 W,算法迭代遍历多次网络更新平均边参数,直到算法收敛或达到最大运行次数.改进的参数求解策略需要迭代多次,每次迭代需要遍历所有边,遍历每条边$\langle i,j\rangle$时,根据公式(9)计算当前边在 $K\times K$ 个社区对上的概率分布;然后将 q_{ijrs} 累加到节点 i 和 j 的相关参数 B,E 和 W.迭代涉及的边和边上的计算耗费大量时间,且边和节点的参数在计算过程中会收敛.为了节省计算时间,从以下 3 个方面对算法加以改进:

a) 减少每条边社区对分布 q 的计算.计算每条边$\langle i,j\rangle$的 q_{ijrs} 时,必须保证 B_{ir},E_{js} 和 W_{rs} 都不为 0,因此只计算 B_{ir},E_{js} 和 W_{rs} 不为 0 的相关 q.判断这些参数是否为 0 也需要耗费时间,如果社区个数较大,则节省的计算量很显著;反之,如果社区个数较小,则节省计算量赢得的时间反而有可能抵偿不了判断耗费的时间.因此,如果社区个数很小时,则计算当前边所有 q 有可能更节省时间.

b) 裁剪值较小的参数 B,E 和 W 为计算 q 准备数据.当边$\langle i,j\rangle$的相关参数 B_{ir},E_{js} 和 W_{rs} 小于阈值时,将它们设置为 0.该策略可以节省计算 q_{ijrs} 的时间.

c) 裁剪收敛边,减少每次边集合上的迭代需要计算的参数.当边$\langle i,j\rangle$的两个端点都自收敛到社区 r 和 s 时,节点 i 只在社区 r 中产生边;类似地,节点 j 只在社区 s 中接收边.此时,边$\langle i,j\rangle$的社区对分布也只有 q_{ijrs} 不为 0,在以后的迭代中,其不会改变 B_{ir},E_{js} 和 W_{rs} 的值.因此,这些边对参数的更新不再有贡献,则在以后的迭代中不需要考虑该边上的计算,可以从以后迭代的边集合中删除.

3.3 快速广义社区发现算法FGSB

根据时间和空间的改进策略,设计有向网络上的快速广义社区发现算法 FGSB.算法 1 给出其详细描述.

算法 1. 快速广义社区发现算法 FGSB.

输入:网络边集合 *EdgeSet*、最大迭代次数 *iterMax*、似然收敛阈值 *conThre*、参数裁剪阈值 *thre*.

输出:收敛参数 B,E,W.

```
1.  初始化参数 B,E,W 和 NB,NE,NW;所有节点置未收敛状态;初始边集 newEdgeSet=EdgeSet:
2.  Repeat
3.  {  对 newEdgeSet 中的每条边⟨i,j⟩作如下计算:
4.     {
5.        根据 i,j 收敛情况和运行时间改进策略 c),判断当前边是否收敛,并修改 newEdgeSet;
6.        根据 i,j 收敛情况、改进策略 a)及上次迭代参数 B,E,W 计算相关 q;
7.        利用 q 更新当前参数 NB,NE,NW.
```

© 中国科学院软件研究所 http://www.jos.org.cn

8. }
9. 对未收敛的节点 I 做如下计算:
10. {
11. 判断 i 是否收敛,若收敛改变收敛状态;
12. 根据改进策略 b)修改 NB,NE,NW;利用 NB,NE,NW 更新 B,E,W.
13. }
14. 利用 NB,NE,NW 计算网络的最大似然;
15. } Until (迭代次数>*iterMax* 或当前似然值与上次似然差值<*conThre*)

我们对单次 FGSB 算法的复杂度进行分析:

1) 根据算法获得的参数 B,E,W 可得节点的中心度θ和ϕ,计算公式如公式(13)和公式(14)所示.根据θ和ϕ可以计算节点的社区隶属度,从而对社区进行软划分或硬划分.该算法空间复杂度由 GSB 的 $O(m\times K\times K)$缩减为 $O(N\times K)$;
2) FGSB 时间复杂度 GSB 为 $O(T\times m\times K\times K)$,虽然没有从数量级上改变 GSB 算法复杂度,但在 T 次迭代过程中处理的边逐渐减少,随着迭代的增加,每次迭代只涉及边集的一部分,相应参数计算也逐渐减少.算法中的裁剪阈值可能使算法性能略差于原算法,当阈值为 0 时与原算法结果几乎相同,但运行速度比原算法快.当阈值为大于 0 的很小值时,准确率会与原算法有较小的差别,但却获得了算法运算速度上的较大提高.

FGSB 是基于 EM 框架的,EM 算法会收敛到局部最优.为避免此问题,许多方法通过多次重启 EM 算法来得到更优的结果.这里假设算法重启 R 次可得到较优的结果.多次重启 FGSB,可采用并行算法实现.因此,可在小于 $O(T\times m\times K\times K)$的时间内获得较好的结果.下面通过实验验证算法性能.

4 实验分析

实验环境为 Windows XP 系统,运行环境采用 Visual studio 2010(C#语言),PC 机配置为主频 3.0 GHz,内存 2.0 GB.用不同规模、不同属性的人工网络和实际网络测试算法性能,有 3 类实验:

1) 比较各算法发现不同网络结构的性能,网络数据为有标定结果的中小型实际网络(http://www-personal.umich.edu/~mejn/netdata/)和人工网络[28].
2) 测试算法随节点个数及混合密度变化时性能的变化规律,网络数据为 LFR 人工网络数据[29].
3) 测试算法的运行效率及随社区个数增加的变化规律,网络数据为大型实际有向网络(http://snap.stanford.edu/data).

文献[27]证明其提出的算法为当前较优社区发现的算法,但其只能发现传统社区,因此,这里只将 FGSB 与该文献算法进行比较.下面将该文献中原 EM 算法记作 LCM,改进的 EM 算法裁剪参数阈值为 0 时算法记作 FLCM,裁剪参数阈值为 0.001 时的算法记作 FLCM_r.原 EM 算法迭代结束条件为最大似然差异小于 10^{-7}.快速算法迭代结束条件也为两次迭代似然差值小于某个阈值 10^{-7}.所有算法最大迭代次数设置为 100 000.度量算法准确性采用 NMI[29].NMI 越大,算法社区划分准确率越高.度量算法运行时间(Time)采用算法 10 次运行时间的平均值,时间的单位为秒.Time 的值越小,算法效率越高.文献[5]已经证明,GSB 模型求解算法用来发现重叠社区时可以很好地量化节点的隶属度,第 3 节分析 FGSB 与 GSB 有相同的划分能力,所以这里只通过算法隶属度得到网络非重叠社区的划分结果,通过非重叠社区划分效果度量算法性能.

(1) 不同结构网络上算法性能比较

为了验证 FGSB 算法与 GSB 模型参数求解算法有几乎相同的结构发现能力,测试网络采用两类具有标准划分结果的网络:一类为有传统社区结构的网络,如 Karate,Dolphin,Football,Political blogs;另一类为具有二分图结构的无向网络,如英语词汇链接网络 Adjnoun,文献[28]生成的二分网络 E1,E2,E3.实验测试结果见表 1、表 2.表中的 NMI 和 Time 的值是 10 次运行结果的平均,其中,“–”表示运行时间太长,统计比较已没有意义,0 表示运

© 中国科学院软件研究所 http://www.jos.org.cn

行时间很小可以忽略,m 表示网络边数,n 表示节点个数,K 表示社区个数.

Table 1 Comparisons of algorithms on medium- and small-scale networks with traditional communities

表 1 具有传统社区的中小型网络上算法比较

Algorithms	Karate (m=78, n=34, K=2)		Dolphin (m=159, n=62, K=2)		Football (m=613, n=115, K=12)		Political (m=19025, n=1490, K=2)	
	NMI	Time (s)	NMI	Time (s)	NMI	Time (s)	NMI	Time (s)
LCM	1	**0**	1	1	**0.906 9**	102	–	–
FLCM	1	**0**	1	**0**	0.905 3	0	**0.509 9**	**0**
FLCM_r	1	**0**	1	**0**	0.904 7	0	0.507 7	**0**
GSB	1	7	1	10	0.889 7	127	–	–
FGSB	1	**0**	1	**0**	0.885 0	28	0.509 5	25
FGSB_r	1	**0**	1	**0**	0.873 5	**16**	0.491 9	11

Table 2 Comparisons of algorithms on medium- and small-scale networks with bipartite structures

表 2 具有二分图结构的中小型网络上算法比较

Algorithms	Adjnoun (m=425, n=112, K=2)		E1 (m=372, n=116, K=2)		E2 (m=2231, n=494, K=2)		E3 (m=2144, n=525, K=2)	
	NMI	Time (s)	NMI	Time (s)	NMI	Time (s)	NMI	Time (s)
LCM	0.002 9	32	0.005 5	5	0.000 3	62	0.000 8	53
FLCM	0.002 7	0	0.003 2	0	0.000 3	0	0.000 8	0
FLCM_r	0.002 7	0	0.002 3	0	0.000 2	0	0.000 6	0
GSB	**0.525 8**	99	**0.184 9**	29	**0.199 9**	218	**0.912 5**	139
FGSB	**0.525 8**	**0**	**0.184 9**	**0**	0.197 5	6	0.905 2	2
FGSB_r	0.508 4	**0**	**0.184 9**	**0**	0.191 7	**2**	0.902 0	**0**

分析表 1 和表 2 数据可以得出如下结论:

1) 在具有传统社区结构的网络上,如表 1 所示,GSB 模型与 LCM 模型参数求解算法获取社区结构的准确率近似相同,但是 GSB 模型求解算法耗费时间比 LCM 要多,原因是其试图发现传统社区结构以外的其他隐含结构,GSB 时间复杂度为 $O(m\times K\times K)$,LCM 为 $O(m\times K)$.

2) 在发现传统社区以外的网络结构上,和文献[5]一样,主要测试如表 2 所示的二分结构.GSB,FGSB,FGSB_r 可以比较准确地发现二分图结构,LCM,FLCM,FLCM_r 则不能发现此类网络结构.另外,GSB 模型还可以发现其他结构.目前,除了社区结构的标准测试网络以外,还没有更多其他结构的测试网络.

3) FGSB 获取社区结构的准确率与 GSB 相近,但 FGSB 的运行速度低于 GSB,尤其是在网络边数较多的情况下,如网络 Political 上,FGSB 运行速度提升效果较明显,用 LCM 和 GSB 模型参数求解算法不能在 1 小时内获得准确结果.

4) 设定不同的裁剪参数阈值,会不同程度地提高算法的速度,表 1 和表 2 数据显示,算法 FGSB_r(裁剪阈值为 0.001 时)运行时间小于 FGSB(阈值为 0)的运行时间,表明适当的阈值可以裁减某些参数的相关计算,从而减少计算量.FGSB_r 的运算准确率可能要比 FGSB 差些,FGSB 与 GSB 有几乎相同的准确率;FLCM_r,FLCM 和 LCM 的运算准确率和时间有类似的关系.FGSB 的运行时间要比 FLCM 长,因为 FGSB 算法需要更广义的社区结构,复杂度高于 FLCM.

(2) 不同清晰度传统社区网络上算法性能比较

为了比较具有不同清晰度的传统社区网络上的算法性能,通过不同规模的人工网络,测试算法随网络节点个数及混合参数变化时的变化规律,采用 LFR 标准生成两组人工网络.节点个数分别为 n=100(边数为 4 000 左右),n=500(边数为 20 000 左右).其他参数设定为:平均度 k=38,最大度 max$k$=50,度指数 $t1$=−2,社区指数 $t2$=−1.人工网络的混合参数 mu 反映网络社区结构的清晰程度,该值越小,社区越清晰.分别在各组网络上变化混合参数 mu 的值,由于 mu 为 0.8 和 0.9 时社区结构已经不明显了,只生成 mu 为 0.1~0.7 的测试网络.通过实验测试随着网络规模(节点个数和边数)的增加,各规模网络随 mu 增长,GSB,FGSB,FGSB_r 的 NMI 及运行时间的变化规律.表 3 和表 4 给出各组网络上算法 10 次运行结果的 NMI 和运行时间 Time 的平均值,其中,“–”表示算法运行

© 中国科学院软件研究所 http://www.jos.org.cn

时间超过 12 小时.

Table 3 Comparisons of algorithms on networks with different mixed parameters (n=100)

表 3 不同混合参数的网络上算法比较(n=100)

mu	*K*	GSB		FGSB		FGSB_r	
		NMI	Time (s)	NMI	Time (s)	NMI	Time (s)
0.1	2	1	44.93	1	19.2	1	**3.8**
0.2	2	1	52.3	1	24	1	**10.8**
0.3	3	1	143.78	1	**25.7**	1	30.1
0.4	3	1	511.49	1	**29.4**	1	37.8
0.5	4	1	887.57	1	152.5	1	**59.7**
0.6	5	0.910 1	959.2	0.902 9	350.5	0.893 7	**111.3**
0.7	7	0.614 4	2 886.6	0.614 4	1 150.8	0.625 7	**211.6**

Table 4 Comparisons of algorithms on networks with different mixed parameters (n=500)

表 4 不同混合参数的网络上算法比较(n=500)

mu	*K*	GSB		FGSB		FGSB_r	
		NMI	Time (s)	NMI	Time (s)	NMI	Time (s)
0.1	12	0.957 6	8 834.4	0.957 6	4 450.4	0.957 6	**801.4**
0.2	14	0.908 7	43 250.5	0.907 7	8 588.6	0.905 7	**2172.7**
0.3	13	–	–	0.907 2	36 025.3	0.907 2	**6914.3**
0.4	16	–	–	–	–	0.903 2	**21 112.2**
0.5	15	–	–	–	–	0.858 3	**26 784.1**
0.6	14	–	–	–	–	0.853 2	**17 379.5**
0.7	13	–	–	–	–	0.682 9	**11 986**

比较表 3 和表 4 数据可得出如下结论:

1) 在相同规模的网络上,随着混合度 *mu* 值的增大,算法 GSB,FGSB,FGSB_r 的社区识别能力(通过 NMI 判断)逐渐降低;随着 *mu* 和社区个数增加,运行时间逐渐增加,这与传统的社区发现算法一致.表 4 中,FGSB_r 运行时间 *mu*=0.7 比 *mu*=0.6 小,*mu*=0.6 的比 *mu*=0.5 的小,原因是社区个数对运行时间的影响,在相同的 *mu* 值下,社区个数大的网络比社区小的耗时多.

2) 在相同规模的网络上,随着混合度 *mu* 值的增大,算法 GSB,FGSB_r 运行时间逐渐增加,FGSB_r 增加的幅度小于 GSB 增加的幅度.随着网络的节点数、边数、社区个数及网络规模的增长,FGSB_r 算法运行时间明显增大,但增长幅度小于 GSB 的增长幅度.GSB 在边数接近 2 万、社区个数较大的网络上显得尤其慢,如表 4 所示,在 *n*=500,*mu*>0.1,*K*>12 的网络上已经不能在 12 小时内给出运行结果.在节点个数为 500 的网络上,当社区个数较大时,FGSB_r 的运行时间在 7 小时左右.

3) 在社区个数较小时,FGSB 在某些网络上运行的时间小于 FGSB_r,如节点个数为 100、混合度为 0.3,0.4、社区个数为 3 的网络.FGSB_r 在网络社区个数大时运行效率普遍优于 FGSB,主要原因是社区个数小时,为减少参数所进行的比较,耗费的时间有可能大于直接计算参数的时间.

(3) 不同社区个数的大规模网络上算法性能比较

为了验证 FGSB 算法在实际大型有向网络上的运行效率,从斯坦福 SNAP 图库选用不同数量级的网络进行测试.由于大型网络没有实际的划分结果,且算法设计时已证明其与 GSB 有近似的准确度,这里仅对其运行时间进行度量.这些网络没有固定的社区个数 *K*,分别测试 *K*=2,3,4、FGSB 阈值为 0.001 时,在各个网络上算法 5 次运行的平均时间,见表 5.由表中数据可发现:虽然在大型网络上运行时间较长,但 GSB 关于万数以上边的网络不能在有意义时间得到结果,相比之下,该算法至少能在有限的时间内发现网络结构.随着网络社区个数的增长,相同网络上的运行时间也随之增长.并且算法耗费的内存也随之增长,当 *K*=3 时,socLiveJournal 网络耗费的内存过大,本实验采用的 PC 机已经不能满足其内存需求.与表 4 的 500 个节点、20 000 条左右边的网络的运行时间相比,大型稀疏网络在社区个数较小时运行时间优于小型多社区稠密网络的运行时间,如在 *K*=4 时,9 分钟左右就可处理边为 103 689 的网络;而在社区个数为 15,边数不到 20 000 的混合参数较小的网络上运行时间要 7 个多小时,与 amazon0601 数据上 *K*=4 时的时间相当.一般对在线网络进行划分时,人们希望将网络组织为层次结果,因此对于社区个数更大的网络,可分层将网络划分为更细粒度的社区,也可避免社区个数过大造成时间上

© 中国科学院软件研究所 http://www.jos.org.cn

的负荷.因此这里仅测试大网络社区个数较小时社区发现的时间.

Table 5 Comparisons of running time of FGSB on big directed networks

表 5 大型有向网络上 FGSB 运行时间比较

	wikiVote	socSlashdot	amazon0601	socLiveJournal
n	7 115	82 168	403 394	4 847 571
m	103 689	948 464	3 387 388	68 993 773
K=2	47.7	2 054.8	3 143.7	9 073.2
K=3	165.3	4 362.3	10 451.9	–
K=4	553.2	12 005.2	27 802.6	–

5 结 论

GSB 模型能够发现网络潜在结构规律,但其计算复杂度影响其在实际应用中的推广.本文设计了一种基于 GSB 的快速网络广义社区发现算法 FGSB,可用来快速发现不同规模网络的各种潜在结构及社区间的交互规律,其与 GSB 有相同的结构识别能力,但速度有一定的提升.FGSB 与传统社区发现算法一样可发现传统社区结构,还可快速发现 GSB 难以处理的大型稀疏网络的结构,但这类算法仍需继续改进.将来的工作主要从以下几个方面进行:(1) 改进 GSB 模型的网络生成过程,如考虑节点入度和出度分布对边生成的影响,使模型生成更接近实际的网络;(2) 考虑并行处理模型参数,以便使用并行技术;(3) 为模型的参数选择适当的先验以避免过拟合问题,并尝试采用 collapsed Gibbs 方法或变分方法求解参数;(4) 利用模型选择方法解决广义社区发现的社区个数选择问题;(5) 融合网络链接和节点内容属性发现网络中潜在的语义结构.

致谢 中国科学院计算技术研究所沈华伟老师为本文工作提供了宝贵意见,在此表示感谢.

References:

[1] Fortunato S. Community detection in graphs. Physics Reports, 2010,486(3):75−174. [doi: 10.1016/j.physrep.2009.11.002]

[2] Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Physical Review E, 2011,83(11):016107. [doi: 10.1103/PhysRevE.83.016107]

[3] Cheng XQ, Shen HW. Community structure of complex networks. Complex Systems and Complex Science, 2011,8(1):57−70 (in Chinese with English abstract).

[4] Newman MEJ, Leicht EA. Mixture models and exploratory analysis in networks. Proc. of the National Academy of Sciences, 2007, 104(23):9564−9569. [doi: 10.1073/pnas.0610537104]

[5] Shen HW, Cheng XQ, Guo J. Exploring the structural regularities in networks. Physical Review E, 2011,84(5):056111. [doi: 10.1103/PhysRevE.84.056111]

[6] Yang B, Liu J, Liu DY. Characterizing and extracting multiplex patterns in complex networks. IEEE Trans. on Systems, Man, and Cybernetics, 2012,42(2):469−481. [doi: 10.1109/TSMCB.2011.2167751]

[7] Chai BF, Yu J, Jia CY, Yang TB, Jiang YW. Combining a popularity-productivity stochastic block model with a discriminative-content model for general structure detection. Physical Review E, 2013,88(3):012807.

[8] Fienberg SE, Wasserman S. Categorical data analysis of single sociometric relations. Sociological Methodology, 1981,12:156−192. [doi: 10.2307/270741]

[9] Holland PW, Laskey KB, Leinhardt S. Stochastic blockmodels: First steps. Social Networks, 1983,5(2):109−137. [doi: 10.1016/0378-8733(83)90021-7]

[10] Snijders T, Nowicki K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of Classification, 1997,14(1):75−100. [doi: 10.1007/s003579900004]

[11] Nowicki K, Snijders T. Estimation and prediction for stochastic block structures. Journal of American Statistical Association, 2001, 96(455):1077−1087. [doi: 10.1198/016214501753208735]

[12] Daudin J, Picard F, Robin S. A mixture model for random graphs. Statistical Computing, 2008,18(2):173−183. [doi: 10.1007/s11222-007-9046-7]

[13] Latouche P, Birmele E, Ambroise C. Variational Bayesian inference and complexity control for stochastic block models. Statistical Modeling, 2012,12(1):93−115. [doi: 10.1177/1471082X1001200105]

© 中国科学院软件研究所 http://www.jos.org.cn

[14] Latouche P, Birmele E, Ambroise C. Overlapping stochastic block models with application to the French political blogosphere. The Annals of Applied Statistics, 2011,5(1):309−336. [doi: 10.1214/10-AOAS382]

[15] Zanghi H, Ambroise C, Miele V. Fast online graph clustering via Erdös-Rényi mixture. Pattern Recognition, 2008,41(12): 3592−3599. [doi: 10.1016/j.patcog.2008.06.019]

[16] McDaid AF, Murphy B, Friel N, Hurley NJ. Improved Bayesian inference for the stochastic block model with application to large networks. Computational Statistics and Data Analysis, 2013,60:12−31. [doi: 10.1016/j.csda.2012.10.021]

[17] Airodi EM, Blei DM, Fienberg SE, Stephen E, Eric XE. Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 2008,9(1):1981−2014.

[18] Lorrain F, White HC. Structural equivalence of individuals in social networks. Journal of the American Statistical Association, 1971,1(1):49−80.

[19] Everett MG, Borgatti SP. The centrality of groups and classes. Journal of Mathematical Sociology, 1999,23(3):181−201. [doi: 10.1080/0022250X.1999.9990219]

[20] Wasserman S, Anderson C. Stochastic a posteriori blockmodels: Construction and assessment. Social Networks, 1987,9:1−36. [doi: 10.1016/0378-8733(87)90015-3]

[21] Cohn D, Chang H. Learning to probabilistically identify authoritative documents. In: Pat L, ed. Proc. of the 17th Int'l Conf. on Machine Learning. Morgan Kaufmann Publishers, 2000. 167−174.

[22] Yang TB, Jin R, Chi Y, Zhu SH. Combining link and content for community detection: A discriminative approach. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2009. 927−936.

[23] Yang TB, Chi Y, Zhu SH, Gong YH, Jin R. Directed network community detection: A popularity and productivity link model. In: Bing L, Srinivasan P, Chandrika K, eds. Proc. of the SIAM Conf. on Data Mining. SIAM, 2010. 742−753.

[24] Ren W, Yan GY, Liao XP, Xiao L. Simple probabilistic algorithm for detecting community structure. Physical Review E, 2009, 79(3):036111. [doi: 10.1103/PhysRevE.79.036111]

[25] Sinkkonen J, Aukia J, Kaski S. Inferring vertex properties from topology in large networks. In: Paolo F, Kristian K, Koji T, eds. Working Notes of the 5th Int'l Workshop on Mining and Learning with Graphs. 2007.

[26] Gyenge A, Sinkkonen J, Benczur A. An efficient block model for clustering sparse graphs. In: Ulf B, Lise G, Sofus AM, eds. Proc. of the 8th Workshop on Mining and Learning with Graphs. New York: ACM, 2010. 62−69. [doi: 10.1145/1830252.1830261]

[27] Ball B, Karrer B, Newman MEJ. An efficient and principled method for detecting communities in networks. Physical Review E, 2011,84(3):036103. [doi: 10.1103/PhysRevE.84.036103]

[28] Hintze A, Adami C. Modularity and anti modularity in networks with arbitrary degree distribution. Biology Direct, 2010,5(32): 1−25. [doi: 10.1186/1745-6150-5-32]

[29] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008,78(4):046110. [doi: 10.1103/PhysRevE.78.046110]

附中文参考文献:

[3] 程学旗,沈华伟.复杂网络的社区结构.复杂系统与复杂科学,2011,8(1):57−70.

柴变芳(1979－),女,山西运城人,博士生,讲师,CCF 学生会员,主要研究领域为复杂网络分析,文本挖掘.
E-mail: chaibianfang@163.com

贾彩燕(1976－),女,博士,副教授,主要研究领域为复杂网络分析,生物信息学.
E-mail: cyjia@bjtu.edu.cn

于剑(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,图像处理.
E-mail: jianyu@bjtu.edu.cn

王静红(1967－),女,博士,教授,主要研究领域为智能信息处理.
E-mail: wangjinghong@126.com

© 中国科学院软件研究所 http://www.jos.org.cn