计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20200158
57(8):1650-1662,2020

一种度修正的属性网络随机块模型

郑忆美[1,2] 贾彩燕[1,2] 常振海[3] 李轩涯[4]
[1](北京交通大学计算机与信息技术学院 北京 100044)
[2](交通数据分析与挖掘北京市重点实验室(北京交通大学) 北京 100044)
[3](天水师范学院数学与统计学院 甘肃天水 741000)
[4](百度在线网络技术(北京)有限公司 北京 100085)
(ymzheng@bjtu.edu.cn)

A Degree Corrected Stochastic Block Model for Attributed Networks

Zheng Yimei[1,2], Jia Caiyan[1,2], Chang Zhenhai[3], and Li Xuanya[4]
[1](*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)
[2](*Beijing Key Laboratory of Traffic Data Analysis and Mining (Beijing Jiaotong University), Beijing* 100044)
[3](*School of Mathematics and Statistics, Tianshui Normal University, Tianshui, Gansu* 741000)
[4](*Baidu Online Network Technology (Beijing) Co., Ltd, Beijing* 100085)

Abstract Community detection is an important task in complex network analysis. The existing community detection methods mostly focus on utilizing the simple network structure, while the methods of integrating network topology and node attributes are also mainly aimed at the traditional community structure, which fails to detect the bipartite structure, mixed structure, etc. However, the degree of each node in the network will affect the composition of the links in the network, as well as the distribution of the community structure. This paper proposes a method called DPSB_PG for attributed networks community detection based on the stochastic block model. Unlike other generative models for attributed networks, in this method, the generation of node links and node attributes both followes the Poisson distribution, and considers the probability between communities based on the stochastic block model. Moreover, the idea of degree corrected is integrated in the process of generating node links. Finally, in order to obtain the community membership of nodes, the expectation-maximization algorithm is used to infer the parameters of the model. The experimental results on the real networks show that the DPSB_PG inherits the advantages of the stochastic block model and can detect the general community structure in networks. Since the introduction of the idea of degree corrected, this model has a good data fitting ability. Overall, the performance of this model is superior to other existing state-of-the-art community detection algorithms for both attributed networks and non-attributed networks.

Key words degree corrected; Poisson distribution; stochastic block model; general structure; attributed networks

收稿日期:2020-03-13;**修回日期**:2020-05-14

基金项目:国家自然科学基金项目(61876016,61632004);中央高校基本科研业务费专项资金项目(2019JBZ110);百度松果计划开放研究基金项目

This work was supported by the National Natural Science Foundation of China (61876016, 61632004), the Fundamental Research Funds for the Central Universities (2019JBZ110), and the Baidu Pinecone Program.

通信作者:李轩涯(lixuanya@baidu.com)

摘　要　社区检测是复杂网络分析中的重要任务,现有的社区检测方法多侧重于利用单纯的网络结构,而融合节点属性的方法也主要针对传统的社区结构,不能检测网络中的二部图结构、混合结构等情况.此外,网络中每个节点的度会影响网络中链接的构成,同样会影响社区结构的分布.因此,提出一种基于随机块模型的属性网络社区检测方法 DPSB_PG.不同于其他属性网络中的生成式模型,该方法中节点链接和节点属性的产生均服从泊松分布,并基于随机块模型考虑社区间相连接的概率,重点在节点链接的生成过程中融合度修正的思想,最后利用期望最大化 EM 算法推断模型中的参数,得到网络中节点的社区隶属度.真实网络上的实验结果显示:模型继承了随机块模型的优点,能够检测网络中的广义社区结构,且由于度修正的引入,具有很好的数据拟合能力,因此在属性网络与非属性网络社区检测性能上优于其他现有相关算法.

关键词　度修正;泊松分布;随机块模型;广义结构;属性网络

中图法分类号　TP301

现实世界中各种各样的人、物及二者之间的联系构成了诸多复杂系统,这些复杂系统常常以复杂网络的形式进行刻画.其中,网络中的节点表示数据对象,节点之间的边(节点链接)表示数据对象之间的关系,节点度的大小表示对象与网络中其他对象的交互紧密程度.社交关系网络、蛋白质交互网络、文献引用网络等都是复杂网络的典型代表.

人以类聚、物以群分,复杂网络中同样存在着聚集现象,常被描述为社区结构.依据网络的拓扑关系,网络中社区结构表现为:社区内部节点连接紧密、而社区之间节点连接相对稀疏[1].这种结构也常被称为同配结构,它具有度大的节点倾向于与网络中其他度大的节点相连的性质.

在现实世界中,网络中不仅存在上述传统的社区结构,也存在其他类型的结构,如与社区结构相对的子图内连接稀疏而子图间连接相对紧密的二部图结构.以英文文章中名词形容词搭配网络为例,名词和形容词分别构成二部图中的一个子图(社区),由于名词、形容词之间经常搭配出现,而很少自身相邻,因此 2 个子图间往往存在更多的边.与同配结构相对,我们称二部图结构为异配结构(可简单理解为违背同配结构原理的其他类型网络结构).除此之外,星形结构、核心-外围结构、混合结构等都是异配结构的典型代表[2].本文,我们将同配结构及异配结构统称为广义结构.

现有许多传统社区检测方法仅针对典型社区结构进行,而对真实世界中的网络,我们往往事先未知网络中潜在的结构.因此,不限定网络结构的类型,检测不同类型网络中的广义结构具有重要的理论和现实意义.如社交关系网络中基于用户兴趣群体检测进行产品推荐、互联网中网页主题预测、蛋白质交互网络中蛋白质功能分析等.现有的社区检测方法依据目标函数的不同可以分为:启发式方法和基于概率模型的生成式方法.

典型的启发式方法主要包括:基于模块度优化的 FN 算法[3]、Louvain 算法[4];基于图分割的 KL 算法[5]、归一化割算法[6];基于非负矩阵分解的 BIGCLAM 算法[7];最小描述长度优化方法 Infomap[8]以及最近的多目标演化聚类方法[9]等.这些方法的共同特点是:利用对社区的直观经验假设和机器学习中的聚类模型设计度量函数,进而找到近似最优解,一般适用于发现网络中的经典社区结构.

基于随机块模型的概率生成式方法,包括 SBM[10],NMM[11]和 DCSBM[12]等.这类方法利用网络拓扑关系将网络中的社区分布建模为隐变量,将实际网络看成以社区分布为隐变量的概率生成过程,从而通过最大化似然函数求解网络结构.与文本分析中的主题模型(如 PLSA[13],LDA[14]等)类似,该类方法具有较好的可解释性,数学形式优美.同时由于随机块模型的性质,该类方法可以用于检测相对于经典社区结构更一般的广义结构,受到了人们的关注.

以上模型都只利用了网络中链接反映的拓扑结构,然而,真实网络中的节点间除了存在链接关系,节点本身还具有描述自身特性的属性信息.如社交网络中,对用户的性别、年龄、学历等的描述,本文将节点含有属性信息的网络简称为属性网络.在属性网络中,网络节点间的关系可能就是因为节点属性信息的相似性诱导的.如具有相同兴趣且物理地址比较接近的人与人之间更容易形成朋友关系;具有相同研究兴趣的学者更有可能在作者合作网络中形成边.因此,网络的链接关系和节点的属性信息之间

可能具有一定的一致性和互补性,单纯利用一方面的信息会忽略另一方面的影响.已有研究证明:节点属性信息不但有助于提高网络社区检测的准确度,还可以帮助我们更好地理解网络中存在的结构,对检测到的网络结构提供更好的可解释性[15].

目前越来越多的生成式方法也逐渐面向属性网络展开.主要有:PCL_DC[16],CESNA[17],PPSB_DC[18]及 NEMBP[19]等.其中,PCL_DC 和 CESNA 是面向经典社区结构而设计的,对于网络中的其他异配结构检测适应性较差.同时,PCL_DC 和 PPSB_DC 采用了两阶段投影求解方法,我们发现这种求解策略不能保证算法的收敛性[20].而 NEMBP 学习节点社区与属性隐含主题之间的关系,需要同时给出网络中隐含的社区个数和属性主题个数,且相关研究在实验中假设二者一致,缺乏广泛性.

针对以上问题,我们给出了一种融合节点属性与网络拓扑信息的生成模型 PSB_PG[20].该模型可以检测网络中的广义结构,且在模型参数求解过程中使用了期望最大化(expectation-maximization,EM)算法[21],具有较好的收敛性.同时该模型只关心社区与节点属性的关系,允许一个社区中含有多个属性主题,具有较好的灵活性.但 PSB_PG 模型在生成过程中没有考虑真实网络中节点度的幂律分布规律,不能准确地反映观测网络的结构.因此,本文在其基础上提出了一种度修正的属性网络随机块模型 DPSB_PG,结合节点度信息更好地拟合真实网络.

DPSB_PG 根据随机块模型中的块结构假设和节点属性的稀疏性假设,令节点之间的链接关系和节点属性的产生都服从泊松分布,但彼此之间独立(即独立同分布).同时隐含假设节点的"社区结构"归属度和节点属性的"主题"归属度共享隐变量,进而将链接与属性隐性相关联.并且,我们在该模型中引入了节点的度,通常具有较大度的节点更倾向于与其他节点相连,进而以一定概率影响网络链接的生成.该模型可以反映现实世界中节点度的幂律分布,具有更好的数据拟合能力、可解释强.最后,我们采用 EM 算法[21]推断模型参数,保证了模型求解过程的收敛性.与已有同类方法及 PSB_PG 模型在多个真实网络上的实验对比显示:所给方法 DPSB_PG 不但保持了 PSB_PG 的良好性能:适用于广义结构检测、收敛性好、可解释性强等,且相对于 PSB_PG 等其他方法在性能上有一定程度的提升,不仅能应用于属性网络社区检测,针对非属性网络也具有良好的适用性.

1 相关工作

1.1 基于生成式模型的网络结构检测

真实网络中节点之间的链接关系是对网络拓扑结构最直观的反映.近年来,有大量的算法仅通过利用网络中拓扑关系发现其中的"社区".

2007 年 Newman 等人首先提出了混合模型 NMM(Newman mixture model)[11],在给定社区数目的情况下,对观测网络中链接的生成构建概率模型,通过最大化似然函数,得到节点所在的社区.该模型定义了节点对社区的隶属度,其思想被广泛应用于之后的生成式方法中.同时模拟从某个"社区"中的节点向其他节点(可以来自当前"社区",也可以来自其他"社区")发边的概率构建网络,具有识别广义结构的能力.

随机块模型(stochastic block model, SBM)[10]是近年来生成式方法中的主流模型.该模型定义了节点所在社区间连边的概率,其中边的数量的产生服从泊松分布,进而生成网络.由于块结构假设:2 个结构间有边的概率不受节点间差异的影响,只与结构块有关.因此该类方法可以根据随机块矩阵中对角线与非对角线元素的大小识别网络中多种不同的结构.但由于其中忽略了节点的异质度信息,可能导致与真实网络不符.2011 年 Karrer 和 Newman 在标准 SBM 的基础上引入了节点期望度对网络生成过程进行了控制,改善了 SBM 对真实网络的拟合效果.并根据节点之间连边服从泊松分布,设计得到度修正随机块模型(degree corrected stochastic block model, DCSBM)[12],利用基于贪心策略的社区标签交换思想最大化目标函数,对网络中社区进行刻画.

除对随机块模型中加入节点度信息进行修正外,许多随机块模型的扩展方法相继被提出.如将节点的混合隶属度与随机块模型结合,设计得到 MMSB(mixed membership stochastic blockmodel)模型[22],以检测网络中的重叠社区.2011 年 Shen 等人提出广义随机块模型(general stochastic blockmodel, GSB)[23].该模型假设网络中产生一条边的概率与边的 2 个端点所在的社区及社区间的连接概率有关.同时,模型结合最小描述长度原则(minimal description length principle, MDL)[24]解决了网络中社区数量选择的问题,其提出使得对网络没有过多先验知识情况下也能发现多种网络结构.

1.2 基于生成式模型的属性网络社区检测

属性网络中,除了网络中链接关系能反映社区结构或其他结构外,节点属性也体现出一定的簇或主题结构.这 2 种结构具有一定的一致性和互补性,同时利用这 2 种信息,能够设计更有效的属性网络社区检测算法.最近提出的 Metacode 算法[25],虽然没有完全利用节点属性信息提高社区检测的性能,但其也进一步证明了社区结构与属性的相关性,推动了属性网络社区检测的进展.

2009 年 Yang 等人认为节点之间的连边不仅与尾节点所在社区有关,同时也受节点流行度的影响,即尾节点流行度越大,越容易被节点连接,进而对网络中边的产生过程进行建模得到 PCL(popularity-based conditional link)模型.进一步,在其基础上引入内容判别模型(discriminative content model, DC)对节点内容进行逻辑回归分析,通过共享社区和内容主题隶属度将链接的生成过程与节点内容主题的回归分析过程线性加和,构建了生成模型 PCL_DC[16].最后,利用两阶段投影算法对模型求解.但该模型只考虑了尾节点的流行度对网络生成过程的影响,且较适用于发现传统社区结构,对网络中含有二部图、混合结构等其他异配结构的社区检测问题适应性较差.

为了解决 PCL 模型中只考虑一种类型边的情况,2010 年 Yang 在其模型基础上针对有向网络引入节点流行度和生成度,即同时考虑了节点被其他节点连接构成边以及连接其他节点进而构成边 2 种情况,提出了生成模型 PPL(popularity and productivity link model)[26].该方法更加充分地考虑了网络拓扑结构,但同样只适合于检测网络中的传统社区结构.

之后,Chai 等人继承 PPL 模型和 GSB 模型的优点,结合 DC 模型构建了 PPSB_DC 模型[18].该模型不仅考虑了节点度的幂律分布,也融合了节点的属性信息.为解决 PCL_DC 只能识别传统社区结构的弊端,模型采用了 GSB 中的块结构假设,可以识别网络中广义结构.在模型参数求解中,PPSB_DC 同样使用了与 PCL_DC 类似的两阶段投影算法.但近期研究发现[20]:这种两阶段算法不能保证收敛.

Chen 等人 2016 年在贝叶斯框架下提出了融合节点链接与属性的概率模型 BNPA(Bayesian nonparametric attribute model)[27],以解决社区数量的定义问题.该模型在 NMM 模型的基础上引入社区中节点含有属性的概率,与节点连边共同构成生成式模型.其继承了 NMM 模型的优点,因此能发现网络中存在的广义社区结构.并且,该模型利用非参数贝叶斯方法引入先验信息 CRP(Chinese restaurant process)[28]对社区数量进行推断,在一定程度上解决了大多数方法需要预先定义社区数量的问题.但其可能对网络社区数量推断不准确,进而影响广义结构识别的精度.

近两年,He 等人为进一步探究社区的语义关系,将网络拓扑与节点内容联合建模来刻画社区及社区相关的属性语义,构建了 NEMBP(nested EM algorithm with belief propagation)模型[19].该模型假设社区和属性主题共享潜在的类结构,对二者之间的关系进行刻画,同时结合度修正以及多项式分布对边和属性的生成过程进行建模,更好地拟合了真实网络.最后,利用 BP(belief propagation)[29]算法和嵌套 EM 算法对模型参数进行求解.相对单层 EM 算法,NEMBP 模型求解复杂度较高.同样,该模型可以识别网络中的广义结构,但是由于模型对属性主题的刻画,其需要预先给定社区和内容主题个数,并在实现过程中假设二者一致,而这往往与真实网络不符.

2 DPSB_PG 模型

在对网络进行社区检测的过程中融入节点属性信息可以在一定程度上改善对网络的理解.近期我们针对属性网络提出了 PSB_PG 模型[20],从融合节点属性与链接的角度出发,根据二者之间的潜在关系构建生成式模型.然而,由于真实网络具有无标度特性,其中节点的度分布符合幂律分布.因此各节点的链接情况(节点度数)分布不均匀,一般地,节点的度越大,越倾向于与其他节点相连接.近年来,相关社区检测算法也逐渐将节点度相关的变量引入网络生成过程中,以更加准确地拟合真实网络.

本文主要考虑了节点度对节点间连边的影响,以进一步提高模型对观测网络的拟合能力,进而构建了属性网络中含有节点度信息的度修正随机块模型 DPSB_PG.该模型中网络结构与节点属性的生成均服从泊松分布,且二者相互独立.值得注意的是,本文的 DPSB_PG 模型是对 PSB_PG 模型的扩展,其主要差别在于是否引入节点度这一可观测变量,并在实验过程中对模型的性能进一步分析.本节我们直接介绍 DPSB_PG 模型,第 3 节我们将介绍其 EM 算法求解过程,并在第 4 节对实验进行完善,不仅证明了此改进模型度修正信息引入的有效性,

同时对不同的网络(网络中是否考虑节点属性信息)进一步比较,证明了该模型的可扩展性.

2.1 问题定义

设 $G(V,E,W)$ 表示一个无向无权属性网络,其中 $V=\{1,2,\cdots,n\}$,$|V|=n$ 表示网络中 n 个节点的集合,$E=\{e_1,e_2,\cdots,e_m\}$,$|E|=m$ 表示网络中 m 条边的集合.通常网络中节点之间的连边用邻接矩阵 $\boldsymbol{A}=(A_{ij})_{n\times n}$ 表示,若节点对 (i,j) 之间存在一条边,则 $A_{ij}=1$,否则 $A_{ij}=0$.设每个节点的属性用 K 维属性向量表示,则所有节点对应属性矩阵 $\boldsymbol{W}=(W_{ik})_{n\times k}$,若节点 i 具有属性 k,则 $W_{ik}=1$,否则 $W_{ik}=0$.现假设网络可以被划分为 c 个不同的社区,用 $V_r(r=1,2,\cdots,c)$ 表示,则有:$V=\bigcup_{r=1}^{c} V_r$.

为清晰理解模型中各符号的意义,表 1 展示了模型中的主要参数.

Table 1 Symbol Description of the DPSB_PG Model

表 1 模型部分符号说明

Symbol	Description
n,m,K	Number of nodes, number of edges and dimension of the node attributes
$\boldsymbol{A},\boldsymbol{W}$	Adjacency matrix and attribute matrix
c	Number of communities
$\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi}$	Model parameters: community membership matrix, stochastic block matrix, community related attribute matrix
A_{ij}	$A_{ij}=1$ if a link exists between node i and j, or 0 otherwise
W_{ik}	$W_{ik}=1$ if node i has the k-th attribute, or 0 otherwise
x_i	The degree of node i
d_{ir}	The probability that a node i belongs to the r-th community
θ_{rs}	The connecting probability between community r and s
ϕ_{rk}	The probability that the r-th community has the k-th attribute
q_{ij}^{rs}	The probability between node i in the r-th community and node j in the s-th community to be linked
γ_{ik}^{r}	The probability of node i in the r-th community possessing the k-th attribute

2.2 节点链接生成过程

我们引入一个节点隶属度矩阵 $\boldsymbol{D}=(d_{ir})_{n\times c}$ 对网络中节点所在社区分布进行软约束,$d_{ir}\in[0,1]$ 表示节点 i 在社区 V_r 中的概率,即每个节点以一定的概率存在于社区中,且满足归一化约束 $\sum_{i=1}^{n} d_{ir}=1$.当对节点所在社区进行估计后,节点对 (i,j) 之间存在边的概率与社区间连边概率 θ_{rs} 有关,由随机块矩阵 $\boldsymbol{\Theta}=(\theta_{rs})_{c\times c}$ 表示,同时满足约束 $\sum_{r,s=1}^{c}\theta_{rs}=1$.由于真实网络节点的度分布符合幂律分布,其中每个节点 i 与其他节点之间存在边的概率并不相同,我们用 $x_i(i=1,2,\cdots,n)$ 表示节点 i 真实的度大小.一般来说,节点的度越大,与其他节点相连的可能性越大,进而影响网络中链接的生成,影响社区的分布.

因此,网络中任意节点对 (i,j) 之间是否存在边不仅与节点的社区隶属度 $\boldsymbol{D}$、社区间连接概率矩阵 $\boldsymbol{\Theta}$ 有关,同时受节点度的影响.通过节点度对网络生成过程进行控制,可以更好地拟合真实情况.这里,假设节点对 (i,j) 间边的生成是独立的且服从泊松分布,可以得到节点对 (i,j) 之间生成一条边的期望为

$$\hat{A}_{ij}=\sum_{r,s=1}^{c} x_i d_{ir}\theta_{rs} x_j d_{js},$$

其中:

$$\sum_{i=1}^{n} x_i d_{ir}=\sum_{i=1}^{n}\overline{d_{ir}}=1,$$

满足归一化约束.

与度修正随机块模型 DCSBM[12] 相对应,我们允许生成网络中包含多边和自环现象.因此,根据泊松分布过程,当给定参数 $\boldsymbol{D},\boldsymbol{\Theta}$ 及观测变量 x_i 后,生成网络 G 的概率 $P(\boldsymbol{A}|\boldsymbol{D},\boldsymbol{\Theta})$ 为

$$\begin{aligned}P(\boldsymbol{A}\mid\boldsymbol{D},\boldsymbol{\Theta})=&\prod_{i,j=1,i<j}^{n}\frac{\left(\sum_{r,s=1}^{c}x_i d_{ir}\theta_{rs}x_j d_{js}\right)^{A_{ij}}}{(A_{ij})!}\times\\&\exp\left(-\sum_{r,s=1}^{c}x_i d_{ir}\theta_{rs}x_j d_{js}\right)\times\\&\prod_{i=1}^{n}\frac{\left(\frac{1}{2}\sum_{r=1}^{c}x_i d_{ir}\theta_{rr}x_i d_{ir}\right)^{\frac{1}{2}A_{ii}}}{(A_{ii}/2)!}\times\\&\exp\left(-\frac{1}{2}\sum_{r=1}^{c}x_i d_{ir}\theta_{rr}x_i d_{ir}\right).\end{aligned}\tag{1}$$

以上网络链接的生成过程中结合节点度信息进一步利用了网络的拓扑结构,对真实网络结构充分拟合.其中,随机块矩阵 $\boldsymbol{\Theta}$ 可以对网络中隐含的结构类型给出直观的解释.当其主对角线元素大且非对角线元素小时,网络呈现传统社区结构;当主对角线元素小且非对角线元素大时,对应社区内连接相对稀疏,社区间连接紧密的二部图结构;其他一般情况可对应于混合社区结构.因此,本模型能在生成过程中对网络中隐含的结构类型进行判断,具有识别网络中广义结构的能力.同时,网络中节点可能属于一个或多个社区,而隶属度矩阵 $\boldsymbol{D}$ 可以反映出节点

社区的分布情况，可用于识别重叠社区（隶属度以 $\boldsymbol{d}_{ir}$ 为阈值指派）或非重叠社区（隶属度最大指派），PSB_PG 中对重叠社区检测能力进行了比较[20]，本文重点针对非重叠社区进行实验.

2.3 节点属性生成过程

针对属性网络，通常网络中每个节点对应的属性往往是高维的，而对于一个社区而言，其中的节点是否含有公共属性具有稀疏性.若一个社区中节点的属性高度相关，也会与网络拓扑结构一致或互补，进而促进社区的构成.因此，我们假设社区中节点属性的生成同样服从泊松分布.

设 ϕ_{rk} 表示社区 V_r 含有属性 k 的概率，并构成社区相关属性矩阵 $\boldsymbol{\Phi}=(\phi_{rk})_{c\times K}$.显然，社区内所含有的属性信息可以为相关社区提供丰富的语义解释，挖掘其中的主题分布.类似地，网络中节点 i 具有属性 k 与节点度 x_i、社区隶属度 $\boldsymbol{D}$ 以及社区相关属性 $\boldsymbol{\Phi}$ 有关，则期望为

$$\hat{W}_{ik}=\sum_{r=1}^{c}x_i d_{ir}\phi_{rk},$$

且满足归一化约束：

$$\sum_{k=1}^{K}\phi_{rk}=1.$$

因此根据泊松分布过程，在给定参数矩阵 $\boldsymbol{D},\boldsymbol{\Phi}$ 后，网络中生成节点属性的概率 $P(\boldsymbol{W}|\boldsymbol{D},\boldsymbol{\Phi})$ 为

$$P(\boldsymbol{W}\mid\boldsymbol{D},\boldsymbol{\Phi})=\prod_{i=1}^{n}\prod_{k=1}^{K}\frac{\left(\sum_{r=1}^{c}x_i d_{ir}\phi_{rk}\right)^{W_{ik}}}{(W_{ik})!}\exp\left(-\sum_{r=1}^{c}x_i d_{ir}\phi_{rk}\right).\quad(2)$$

2.4 融合节点属性的网络生成过程

设社区属性隶属度与节点社区隶属度共享网络中的潜在社区，以便对二者之间隐含联系进行刻画.假设网络中节点邻接矩阵 $\boldsymbol{A}$ 与属性矩阵 $\boldsymbol{W}$ 生成过程彼此独立，则融合度修正信息的属性网络随机块模型 DPSB_PG 的联合分布形式表示为

$$P(\boldsymbol{A},\boldsymbol{W}|\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi})=P(\boldsymbol{A}|\boldsymbol{D},\boldsymbol{\Theta})\times P(\boldsymbol{W}|\boldsymbol{D},\boldsymbol{\Phi})=$$

$$\prod_{i,j=1,i<j}^{n}\frac{\left(\sum_{r,s=1}^{c}x_i d_{ir}\theta_{rs}x_j d_{js}\right)^{A_{ij}}}{(A_{ij})!}\times$$

$$\exp\left(-\sum_{r,s=1}^{c}x_i d_{ir}\theta_{rs}x_j d_{js}\right)\times$$

$$\prod_{i=1}^{n}\frac{\left(\frac{1}{2}\sum_{r=1}^{c}x_i d_{ir}\theta_{rr}x_i d_{ir}\right)^{\frac{1}{2}A_{ii}}}{(A_{ii}/2)!}\times$$

$$\exp\left(-\frac{1}{2}\sum_{r=1}^{c}x_i d_{ir}\theta_{rr}x_i d_{ir}\right)\times$$

$$\prod_{i=1}^{n}\prod_{k=1}^{K}\frac{\left(\sum_{r=1}^{c}x_i d_{ir}\phi_{rk}\right)^{W_{ik}}}{(W_{ik})!}\exp\left(-\sum_{r=1}^{c}x_i d_{ir}\phi_{rk}\right).\quad(3)$$

相应地，网络中的链接和节点属性的生成过程可总结为：

1）以概率 d_{ir} 从社区 V_r 中抽取一个节点 i 且以概率 d_{js} 从社区 V_s 中抽取一个节点 j；

2）以概率为节点对 (i,j) 生成一条边：

$$A_{ij}\sim Poisson(\hat{A}_{ij})=Poisson\left(\sum_{r,s=1}^{c}x_i d_{ir}\theta_{rs}x_j d_{js}\right);$$

3）以概率 ϕ_{rk} 在社区 V_r 中选择一个属性 k；

4）以概率为节点 i 选择一个属性 k：

$$W_{ik}\sim Poisson(\hat{W}_{ik})=Poisson\left(\sum_{r=1}^{c}x_i d_{ir}\phi_{rk}\right).$$

对应概率图模型如图 1 所示：

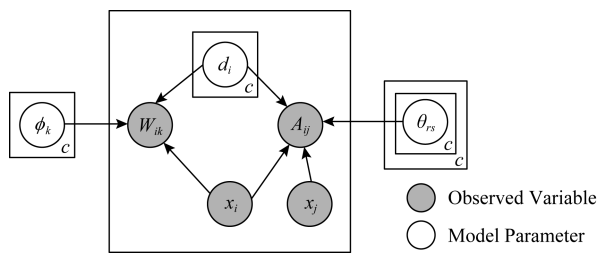

Fig. 1 The probabilistic graph model for DPSB_PG

图 1 DPSB_PG 概率图模型

3 DPSB_PG 模型 EM 求解算法

本文改进后的模型中共包括 3 种变量：网络中的观测变量 $\boldsymbol{A},\boldsymbol{W}$ 以及节点度 x_i；潜变量 q_{ij}^{rs} 及 γ_{ik}^{r}；模型参数 $\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi}$.我们期望构造的生成模型能够很好的拟合真实网络的拓扑结构和属性分布，因此目标是对式(3)求对数，最大化对数似然函数：

$$L(\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi})=\ln(P(\boldsymbol{A},\boldsymbol{W}|\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi}))=$$

$$\sum_{i,j=1}^{n}\left[\frac{1}{2}A_{ij}\ln\left(\sum_{r,s=1}^{c}x_i d_{ir}\theta_{rs}x_j d_{js}\right)-\frac{1}{2}\sum_{r,s=1}^{c}x_i d_{ir}\theta_{rs}x_j d_{js}\right]+$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\left[W_{ik}\ln\left(\sum_{r=1}^{c}x_i d_{ir}\phi_{rk}\right)-\sum_{r=1}^{c}x_i d_{ir}\phi_{rk}\right].\quad(4)$$

然而，由于式(4)难以直接计算，因此本文利用处理这类问题常用的 EM 算法[21]求解.EM 算法能保证收敛到模型的局部最大值，从而获得模型最优参数.

根据 Jensen 不等式 $E[f(X)]\geqslant f(EX)$（f 为

凸函数,当且仅当 X 为常数时等式成立)可得到对数似然的下界:

$$L(\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi})\geqslant\bar{L}(\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi})=\frac{1}{2}\sum_{i,j=1}^{n}\sum_{r,s=1}^{c}\left[A_{ij}q_{ij}^{rs}\ln\left(\frac{x_i d_{ir}\theta_{rs}x_j d_{js}}{q_{ij}^{rs}}\right)-x_i d_{ir}\theta_{rs}x_j d_{js}\right]+\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{c}\left[W_{ik}\gamma_{ik}^{r}\ln\left(\frac{x_i d_{ir}\phi_{rk}}{\gamma_{ik}^{r}}\right)-x_i d_{ir}\phi_{rk}\right],\tag{5}$$

其中,q_{ij}^{rs} 表示节点 i 在社区 V_r 中,节点 j 在社区 V_s 中且节点对(i,j)间有边相连的概率;γ_{ik}^{r} 表示节点 i 在社区 V_r 中并具有属性 k 的概率,二者满足约束:

$$\sum_{r,s=1}^{c}q_{ij}^{rs}=1,\sum_{r=1}^{c}\gamma_{ik}^{r}=1.$$

E 步:根据模型参数和观测变量 $\boldsymbol{A},\boldsymbol{W}$ 及 x_i,估计潜变量 q_{ij}^{rs} 和 γ_{ik}^{r} 的期望值:

$$q_{ij}^{rs}=\frac{x_i d_{ir}\theta_{rs}x_j d_{js}}{\sum_{r,s=1}^{c}x_i d_{ir}\theta_{rs}x_j d_{js}},\gamma_{ik}^{r}=\frac{x_i d_{ir}\phi_{rk}}{\sum_{r=1}^{c}x_i d_{ir}\phi_{rk}}.\tag{6}$$

M 步:固定变量 q_{ij}^{rs} 和 γ_{ik}^{r},根据上述约束运用拉格朗日乘子法最大化式(5),可得到模型参数估计:

$$d_{ir}=\frac{\sum_{j=1}^{n}\sum_{s=1}^{c}A_{ij}q_{ij}^{rs}+\sum_{k=1}^{K}W_{ik}\gamma_{ik}^{r}}{x_i\left(\sum_{i,j=1}^{n}\sum_{s=1}^{c}A_{ij}q_{ij}^{rs}+\sum_{i=1}^{n}\sum_{k=1}^{K}W_{ik}\gamma_{ik}^{r}\right)},\tag{7}$$

$$\theta_{rs}=\frac{\sum_{i,j=1}^{n}A_{ij}q_{ij}^{rs}}{\sum_{i,j=1}^{n}\sum_{r,s=1}^{c}A_{ij}q_{ij}^{rs}},\tag{8}$$

$$\phi_{rk}=\frac{\sum_{i=1}^{n}W_{ik}\gamma_{ik}^{r}}{\sum_{i=1}^{n}\sum_{k=1}^{K}W_{ik}\gamma_{ik}^{r}}.\tag{9}$$

首先,随机初始化参数 $\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi}$,然后重复迭代 EM 算法,对式(6)和式(7)~(9)不断更新直至目标函数收敛或达到最大迭代次数,输出对应的模型参数 $\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi}$,对应最优的网络拟合效果,具体算法过程见算法 1.因此,可以通过该算法挖掘网络中节点的社区分布情况,社区对应的属性分布以及网络中广义社区结构.由于 EM 算法求解过程,上述参数的形式看起来与 PSB_PG 中参数结果相差不多,这也表明概率生成模型数学形式的优美.但实质上,节点度 x_i 的引入直观影响了节点所在的社区分布,同时由于潜变量的变化,会间接影响社区属性以及网络结构的推断,因此相较于原 PSB_PG 算法仍具有一定程度的变化.

算法 1. DPSB_PG 模型算法.

输入:邻接矩阵 $\boldsymbol{A}$、属性矩阵 $\boldsymbol{W}$、社区数量 c、最大迭代次数 I_T 和阈值 ε;

输出:模型参数 $\boldsymbol{D},\boldsymbol{\Theta},\boldsymbol{\Phi}$.

① 由邻接矩阵 $\boldsymbol{A}$ 计算每个节点 i 的度 x_i;

② 随机初始化模型参数 $\boldsymbol{D}^{(0)},\boldsymbol{\Theta}^{(0)},\boldsymbol{\Phi}^{(0)}$;

③ for $t=1:I_T$ do

④ E 步:根据式(6)计算 $q_{ij}^{rs},\gamma_{ik}^{r}$;

⑤ M 步:根据式(7)~(9)分别更新模型参数 $\boldsymbol{D}^{(t)},\boldsymbol{\Theta}^{(t)},\boldsymbol{\Phi}^{(t)}$;

⑥ if $t=I_T$ 或者 $|L^{(t)}-L^{(t-1)}|<\varepsilon$ then

⑦ $\boldsymbol{D}=\boldsymbol{D}^{(t)},\boldsymbol{\Theta}=\boldsymbol{\Theta}^{(t)},\boldsymbol{\Phi}=\Phi^{(t)}$;

⑧ break;

⑨ end if

⑩ end for

由于社区隶属度矩阵 $\boldsymbol{D}$ 刻画了每个节点 i 在所有 c 个社区上的分布,基于此可以推断网络中节点的社区归属度,可以是软划分(即通过限定 d_{ir} 的阈值,超过此阈值时一个节点可以归属于对应的多个社区),也可以是硬划分(即只利用 d_{ir} 的最大值,限定一个节点只能归属于一个社区),分别对重叠社区或非重叠社区检测,本文主要针对非重叠社区检测进行.随机块矩阵 $\boldsymbol{\Theta}$ 反映了网络中存在的结构类型,可以根据其对角分布推断其是经典社区结构、还是二部图结构或是混合结构等.社区相关属性矩阵 $\boldsymbol{\Phi}$ 描述了每个社区相关的属性特征,可以通过属性类别分析当前社区的语义,增强所识别社区的可解释性.并且,由于 EM 算法自身的收敛性质保证了 DPSB_PG 求解算法的收敛性.

算法 1 的时间复杂度主要取决于 EM 算法对参数的更新迭代过程.其中,执行 E 步的时间复杂度为 $O(mc^2+nKc)$,执行 M 步的时间复杂度为 $O(nc+c^2+nck)$,又由于实际网络中社区数量 c 常小于节点个数 n 且算法的最大迭代次数是 I_T,故算法 1 整体的时间复杂度可为 $O(I_T(mc^2+nKc))$.

4 实验结果及分析

4.1 评价指标

本文采用常用的标准化互信息(normalized mutual information, NMI)[30] 以及 F 测度(pairwise F-measure, PWF)[16] 2 种评价指标度量本文模型与

现有方法在真实网络上的广义结构识别效果.其分别表示为

$$NMI(C^*,C)=\frac{H(C^*)+H(C)-H(C^*,C)}{(H(C^*)+H(C))/2},$$

其中，$C^*=(C_1^*,C_2^*,\cdots,C_k^*)$表示已知的真实社区分布，$C=(C_1,C_2,\cdots,C_k)$表示利用算法检测出的社区分布，$|C^*|=|C|=k$ 表示社区的数量.真实社区划分 C^* 的熵 $H(C^*)$、模型推断社区划分 C 的熵 $H(C)$以及 C^* 与 C 的联合熵 $H(C^*,C)$分别表示为

$$H(C^*)=-\sum_{i=1}^{|C^*|}P_{C^*}(i)\ \mathrm{lb}(P_{C^*}(i)),$$

$$H(C)=-\sum_{j=1}^{|C|}P_C(j)\ \mathrm{lb}(P_C(j)),$$

$$H(C^*,C)=-\sum_{i,j=1}^{k}P_{C^*,C}(i,j)\ \mathrm{lb}(P_{C^*,C}(i,j)),$$

$$PWF=\frac{2\times precision\times recall}{precision+recall},$$

其中，$precision=|S\cap T|/|S|$，$recall=|S\cap T|/|T|$分别表示算法划分结果的精确度和召回率，S 表示算法识别社区中具有相同类标的节点集合，T 表示对应的真实网络中具有相同类标的节点集合，$|\cdot|$表示集合中元素数量.

利用 NMI，PWF 这 2 种指标对社区划分的质量进行评价可较为直观地表示算法检测到的社区划分与真实划分的差异.其中二者的值介于 0 和 1 之间，值越大，表示所对应的社区划分效果越好.

4.2 参数设置

由于节点对(i,j)所对应社区随机块矩阵 $\boldsymbol{\Theta}$ 反映了网络中隐含广义结构的类型.因此其初始值的设置会严重影响算法的收敛速度与社区检测的精度.当 $\boldsymbol{\Theta}$ 的初始值对应的结构与真实的网络结构一致时，算法收敛速度快，检测结果也相对更优；相反当二者完全相反时(如真实结构是二部图，但初始值为传统社区结构)，算法收敛速度很慢，且容易陷入局部最小.因此，在实验过程中，我们利用最大熵以及最大似然的思想对 $\boldsymbol{\Theta}$ 的初始值进行预判.

在矩阵 $\boldsymbol{\Theta}=(\theta_{rs})_{c\times c}$ 的初始值选择中，我们考虑 3 种情况：主对角线元素大、非对角线元素大以及所有元素值随机但近乎相等.这 3 种情况分别对应：社区结构、二部图结构以及混合结构.因此针对每个网络我们首先对 3 种初值迭代固定的少量几步(如 10 步)EM 并反复几次，根据式(5)计算多次的平均似然值.最后选择 3 种情形中平均似然最大时对应的 $\boldsymbol{\Theta}$ 情况作为 $\boldsymbol{\Theta}$ 的初始设置.

在实验过程中，我们设置社区数量 c 与已知的实际网络社区数量相等，同时不限定真实网络的结构类型，对输入的网络通过上述初始化步骤进行网络结构的预判，进而执行对网络社区结构的检测.

4.3 数据集描述

我们将本模型在真实属性网络与非属性网络上进行了实验.表 2、表 3 分别描述了本文所选用的 4 个真实属性网络数据集和 6 个非属性网络数据集的相关统计特征.

Table 2 Feature of the Attributed Networks

表 2 属性网络数据集特征

Datasets	n	m	K	c	Structure
Cornell	195	304	1 703	5	Mixture
Texas	187	328	1 703	5	Mixture
Washington	230	446	1 703	5	Mixture
Wisconsin	265	530	1 626	5	Mixture
Facebook	1 018	26 717	576	10	Community
Cora	2 708	5 429	1 433	7	Community
Citeseer	3 312	4 723	3 703	6	Community

Table 3 Feature of the Non-attributed Networks

表 3 非属性网络数据集特征

Datasets	n	m	c	Structure
Karate	34	78	2	Community
Dolphins	62	159	2	Community
Football	115	613	12	Community
Adjnoun	112	425	2	Bipartite
Polbooks	105	441	3	Community
Polblogs	1 490	19 079	2	Community

4 个属性网络数据集包括：

1) WebKB 数据集①.该数据集由 Cornell，Texas，Washington，Wisconsin 四所大学的网页以及网页间的链接关系构成，共包含 877 个节点，1 608 条边，均划分为 5 个社区.同时，每个节点都关联一个高维的二值属性特征，由各网页中出现的不同词构成.该数据集的 4 个网络含有的结构均为混合结构.

2) Facebook 数据集②.该数据集记录了社交网站 Facebook 上从 2005 年起的社交关系数据.这里

① http://www.cs.cmu.edu/~webkb/

② http://snap.stanford.edu/data/

选取其中一个数据子集,包含 1 018 个节点用户,26 717 条边表示了用户之间的朋友关系,同时每个节点由一个 576 维的属性特征向量描述.

3) Cora 数据集[31].该数据集为科技文献引文网络的一个子集,其中包含 2 708 个节点,代表所有的文献,5 429 条边代表文献之间的引用关系,共包括 7 种文献类别构成的相应社区,同时每个节点都关联一个 1 433 维的属性向量.

4) Citeseer 数据集[32].该数据集也是科技文献引文网络的一个子集,含有 3 312 个节点,4 732 条引用关系,包括 6 个文献类别构成的相应社区,同时每个节点由一个 3 703 维的属性向量构成.

6 个非属性网络数据集包括:

1) Karate 数据集[33].该数据集是一所美国大学空手道俱乐部成员构成的关系网络,包含 34 个节点和 78 条边,由于内部矛盾分裂为 2 个社区.

2) Dolphins 数据集[34].该数据集由新西兰某海峡 62 只海豚群体的交流情况得到,海豚之间的频繁交流构成了网络中 159 条边,分为 2 个社区.

3) Football 数据集[1].该数据集是由美国大学生足球联赛构建的网络,包含 115 个节点和 613 条边,其中节点表示各个足球队,边表示两球队之间进行过比赛,所有代表队分为 12 个联盟.

4) Adjnoun 数据集[35].该数据集表示了某英文文章中形容词与名词的连接情况.共包含名词、形容词构成的 112 个顶点和表示二者相邻情况的 425 条边,由于形容词与名词大体上总是相邻出现而很少有名词与名词、形容词与形容词相邻的情况,因此构成了典型的二部图网络结构.

5) Polbooks 数据集①:该数据集由关于美国政治的图书在线购买情况构成,包含 105 个在线商家销售的书籍构成的节点,节点之间共 441 条连边,表示同一买家频繁购买的书籍之间的关系,构成 3 个社区.

6) Polblogs 数据集[36]:该数据集是由政治博客构成的网络,包含 1 490 个节点和 19 079 条边.节点代表网页博客,每条边代表网页间超链接情况,原数据中节点包含了网页政治倾向,代表了 2 个社区.

4.4 实验结果对比

本节我们通过与相关工作中的提到的部分融合节点属性的属性网络社区检测方法以及只利用网络结构的社区检测方法分别在属性网络以及非属性网络上进行对比实验,以验证本文方法的有效性.

4.4.1 属性网络结果分析

在真实属性网络数据集上,我们将本文的模型 DPSB_PG 与现有融合节点链接与属性的生成式模型 NEMBP,BNPA,PPSB_DC,PCL_DC 以及未改进的 PSB_PG 模型进行了对比实验.为了保持公平,所有比较算法均采用原文中提及的最优参数设置.表 4、表 5 展示了对应的 NMI,PWF 结果.由于 EM

Table 4 Comparison of *NMI* for Attributed Networks

表 4 属性网络 *NMI* 结果对比

Model	Value	Cornell	Texas	Washington	Wisconsin	Facebook	Cora	Citeseer
DPSB_PG	mean	**0.325 6**	**0.297 6**	0.293 9	**0.384 8**	**0.283 1**	0.317 6	0.265 3
	max	**0.470 2**	**0.422 1**	0.374 3	**0.448 7**	0.313 3	0.418 3	0.339 4
PSB_PG	mean	0.319 6	0.293 2	**0.315 4**	0.376 4	0.279 0	0.301 2	0.250 1
	max	0.435 8	0.411 0	**0.387 4**	0.430 0	0.304 5	0.369 9	0.334 7
NEMBP	mean	0.151 0	0.296 5	0.193 8	0.232 2	0.276 0	0.403 3	0.200 3
	max	0.279 3	0.420 2	0.310 7	0.407 5	**0.315 5**	0.475 7	0.291 1
BNPA	mean	0.077 2	0.226 5	0.246 9	0.321 2	0.273 9	**0.439 1**	0.170 0
	max	0.093 3	0.269 4	0.270 1	0.341 3	0.286 6	0.502 2	0.319 6
PPSB_DC	mean	0.112 8	0.208 5	0.172 6	0.131 5	0.094 6	0.182 0	0.133 5
	max	0.250 3	0.366 3	0.369 0	0.240 9	0.218 9	**0.522 1**	0.380 5
PCL_DC	mean	0.071 5	0.071 2	0.072 3	0.076 5	0.243 9	0.436 7	**0.359 2**
	max	0.114 0	0.096 5	0.128 6	0.126 3	0.291 5	0.496 1	**0.409 9**

Note: The best performance are highlighted with bold, and the suboptimal performance are highlighted with underline.

① http://www-personal.umich.edu/~mejn/netdata/

Table 5 Comparison of *PWF* for Attributed Networks

表 5 属性网络 *PWF* 结果对比

Model	Value	Cornell	Texas	Washington	Wisconsin	Facebook	Cora	Citeseer
DPSB_PG	mean	**0.4468**	0.4215	0.4344	**0.5333**	0.2413	0.3629	0.3676
	max	**0.5980**	0.5013	0.4790	**0.6275**	0.2593	0.4495	0.4408
PSB_PG	mean	0.4346	0.4155	0.4461	0.5296	0.2401	0.3554	0.3528
	max	0.5550	0.4837	0.5125	0.5878	0.2566	0.4228	0.4334
NEMBP	mean	0.3646	**0.5622**	0.4034	0.3867	0.3197	0.4337	0.3236
	max	0.4722	**0.7002**	0.6064	0.5384	0.3962	0.5203	0.3840
BNPA	mean	0.3446	0.5084	0.3851	0.4818	0.2148	**0.4809**	0.3548
	max	0.3528	0.5257	0.3921	0.4967	0.2218	**0.5423**	0.3984
PPSB_DC	mean	0.3789	0.5610	**0.4751**	0.3900	**0.3393**	0.2835	0.2733
	max	0.5025	0.6753	**0.6118**	0.4839	**0.5448**	0.3592	0.4768
PCL_DC	mean	0.3000	0.3364	0.3209	0.2929	0.3268	0.4562	**0.4679**
	max	0.3238	0.3844	0.3791	0.3319	0.3930	0.5225	**0.5313**

Note: The best performance are highlighted with bold, and the suboptimal performance are highlighted with underline.

算法迭代过程会根据初始值的不同收敛到局部最优解，我们对每个模型进行了 30 轮实验，并报告了 2 个指标 30 轮的均值以及最大值.

由表 4 和表 5 可知：本文提出的 DPSB_PG 模型可以适用于广义结构检测，且对于含有混合结构的真实属性网络（WebKB 数据集的 4 个网络上）的检测效果有明显改善.相较于没有考虑节点度信息的 PSB_PG 模型，除在属性网络 Washington 上表现略差外，我们改进的模型在其他网络中表现更好.说明融入节点的度信息对网络社区结构的检测具有积极影响，使得 DPSB_PG 模型对数据具有较好的拟合能力.对 Washington 网络，我们分析发现：该网络规模相对较小，而其中存在多个孤立节点，因此这些节点的度分布可能会影响本模型的整体表现，相较于未引入度信息的 PSB_PG 模型表现略差.

BNPA 模型对 Cora 数据集中社区结构的检测得到了最好的效果，但由于其引入了先验信息，需要对先验参数进行调参，且其检测精度依赖于社区数量估计的准确性，在其他网络中表现效果不佳. PCL_DC 模型在具有社区结构的 Cora，Citeseer 网络上取得了较好的效果，但对非传统社区结构适应性较差.进一步，我们的模型在 Citeseer 中表现仅次于 PCL_DC，由指标 *NMI* 可知，在 Facebook 数据中表现最优，这也说明本模型在混合结构、社区结构网络中均表现出良好的性能，而 NEMBP，PPSB_DC 模型在各个网络中的表现具有不稳定性，NEMBP 仅在 Texas 网络中取得了较好的效果，PPSB_DC 由于初值及收敛性的影响使得表现效果差异较大.因此，本文的 DPSB_PG 模型综合性能优于其他相关算法，说明利用度修正思想以及泊松分布随机块模型能较好地识别属性网络的广义结构.

4.4.2 非属性网络结果分析

原 PSB_PG 模型中重点针对属性网络，对节点属性进行了充分分析，但对非属性网络中模型的表现缺少进一步比较.因此，对模型改进后，在非属性网络数据集中，我们对比了 DPSB_PG 与其他模型 PSB_PG，BNPA-link，PPSB 和 PCL 在只利用网络拓扑信息时性能上的差异（同样所有比较算法均采用原文中的最优参数设置）.除此之外，我们的模型在生成过程中引入了节点度修正的思想，与 DCSBM[12] 模型类似.而二者差异之处在于：DPSB_PG 模型采用 EM 算法求解模型参数，DCSBM 模型基于贪心算法利用标签交换思想最大化似然函数求解模型参数.因此，我们也将改进的模型与 DCSBM 模型进行了对比，以验证本文算法的性能.同样，我们将对比的前 4 个模型执行 30 轮并报告了 *NMI*，*PWF* 指标的均值以及最大值.为了与原文保持一致，对于 DCSBM 模型，我们基于贪心策略执行 20 次 KL 算法以实现标签转换，并报告其最大值.为保持公平，我们在与 DCSBM 对比时只将其与其他模型的最大值进行对比.表 6、表 7 中展示了相应的实验结果.

由表 6 和表 7 中结果可知，改进后的 DPSB_PG 模型在其中 4 个非属性网络中取得了第 1 或第 2 的

表现效果,进一步表明我们的模型在仅利用网络拓扑信息时同样能较好地挖掘网络的多种类型结构.在 PSB_PG 的基础上引入节点度信息进行修正后,除在 Adjnoun 数据集上 DPSB_PG 效果略差外,在其他真实非属性网络中的效果均优于 PSB_PG 模型,可见度修正的引入起到了积极作用,进一步对网络拓扑结构进行了拟合.对 Adjnoun 网络,我们同样分析发现:在其网络规模下,孤立节点占据较大比重,因此其度分布可能会影响网络链接生成过程,相较于未引入度修正信息的 PSB_PG 模型性能略有下降.

Table 6 Comparison of *NMI* for Non-attributed Networks

表 6 非属性网络 *NMI* 结果对比

Model	Value	Karate	Dolphin	Football	Adjnoun	Polbooks	Polblogs
DPSB_PG	mean	**1.000 0**	0.800 5	0.874 9	0.459 6	0.530 0	**0.509 5**
	max	**1.000 0**	0.890 4	0.924 2	0.525 8	**0.604 0**	**0.547 3**
PSB_PG	mean	**1.000 0**	0.781 4	0.874 2	0.466 4	0.526 2	0.497 2
	max	**1.000 0**	0.890 4	0.924 2	0.525 8	**0.604 0**	0.543 6
BNPA-link	mean	**1.000 0**	**0.952 5**	0.846 6	**0.531 7**	**0.548 7**	0.204 9
	max	**1.000 0**	**1.000 0**	0.898 4	**0.631 1**	0.553 8	0.231 9
PPSB	mean	0.994 5	0.764 3	0.843 6	0.004 9	0.511 8	0.496 2
	max	**1.000 0**	0.890 4	0.924 2	0.030 1	0.582 6	**0.547 3**
PCL	mean	0.895 4	0.746 8	**0.901 9**	0.003 6	0.502 8	0.487 8
	max	**1.000 0**	**1.000 0**	**0.926 9**	0.022 7	0.593 4	**0.547 3**
DCSBM	max	0.837 2	**1.000 0**	0.870 6	0.446 6	0.542 0	0.489 3

Note: The best performance are highlighted with bold, and the suboptimal performance are highlighted with underline.

Table 7 Comparison of *PWF* for Non-Attributed Networks

表 7 非属性网络 *PWF* 结果对比

Model	Value	Karate	Dolphin	Football	Adjnoun	Polbooks	Polblogs
DPSB_PG	mean	**1.000 0**	0.917 6	0.784 0	0.758 2	0.722 3	**0.794 9**
	max	**1.000 0**	0.970 1	**0.904 9**	0.805 8	0.785 6	**0.822 0**
PSB_PG	mean	**1.000 0**	0.906 4	0.782 4	0.761 5	0.711 4	0.786 4
	max	**1.000 0**	0.970 1	**0.904 9**	0.805 8	0.785 3	0.819 9
BNPA-link	mean	**1.000 0**	**0.987 0**	0.724 4	**0.815 8**	**0.786 9**	0.382 1
	max	**1.000 0**	**1.000 0**	0.857 9	**0.865 4**	0.787 4	0.484 9
PPSB	mean	0.964 1	0.896 0	0.746 2	0.495 9	0.713 3	0.784 9
	max	**1.000 0**	0.970 1	**0.904 9**	0.541 7	0.757 9	**0.822 0**
PCL	mean	0.920 8	0.894 9	**0.842 6**	0.498 2	0.733 0	0.778 9
	max	**1.000 0**	**1.000 0**	**0.904 9**	0.516 8	**0.814 2**	**0.822 0**
DCSBM	max	0.939 4	**1.000 0**	0.833 5	0.767 0	0.771 2	0.795 8

Note: The best performance are highlighted with bold, and the suboptimal performance are highlighted with underline.

同样,由于 BNPA 模型在生成过程中引入了先验,通过对参数的调整可得到相对更优的社区检测结果,因此在 Karate, Dolphin, Adjnoun, Polbooks 数据集中都取得了最好的效果.但由文献[37]可知,BNPA 模型在自动判断社区数量时,与真实网络社区数量差异明显.更一般的情况下,我们的模型效果明显好于 PPSB 以及 PCL 模型(除 Football 网络数据集外).因 PCL 模型只适用于传统的社区结构,在明显的二部图结构网络 Adjnoun 中该模型并不适用.对 DCSBM 模型,其贪心优化策略正确发现了 Dolphin 网络中社区结构,但对其他网络的检测效果相较于我们的算法较差,在这里标签交换思想没

有额外的优势.因此通过对比可以反映出：我们的模型在仅利用节点链接信息时同样对网络中社区结构、二部图结构的检测具有较好的性能.

5 结　　论

融合网络中节点之间的链接关系以及节点固有的属性信息挖掘网络中潜在的结构，并利用属性信息增强所识别社区的可解释性，进而揭示网络系统的功能逐渐被人们所重视.本文在 PSB_PG 模型的基础上引入节点的度信息对模型进行改进，提出了一种属性网络中链接和节点属性产生过程都服从 Poisson 分布的度修正随机块模型 DPSB_PG，以充分拟合节点度在真实世界中的幂律分布特征，并基于随机块特性识别网络中广义社区结构.并且，DPSB_PG 模型易于通过 EM 算法进行求解，具有较好的算法收敛性.通过与现有方法在真实属性网络、非属性网络上的实验比较研究：我们的模型能够发现网络中的广义结构，如传统的社区结构、二部图结构以及混合结构等多种网络结构；节点度的引入更好地拟合了真实网络，因此识别精度优于基础模型 PSB_PG，且相较于现有的其他同类算法在性能上也有不同程度的改善.另外，该模型保持了 PSB_PG 模型灵活的社区语义关系学习特性以及重叠社区检测能力.

由于 EM 算法在模型参数求解过程中根据网络规模的不同，可能需要较大的计算量，进而影响模型的效率.因此，在今后的工作中，我们拟引入变分推理或随机变分推理等其他策略加快模型参数的求解过程，在保证社区检测精度的基础上提高算法的计算效率，以适应更加大型的复杂网络，更快更好地挖掘网络中的“社区”结构.

参 考 文 献

[1] Girvan M, Newman M E. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826

[2] Fortunato S, Hric D. Community detection in networks: A user guide [J]. Physics Reports, 2016, 659: 1-44

[3] Newman M E. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6): No.066133

[4] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): No.P10008

[5] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49(2): 291-307

[6] Shi Jianbo, Malik J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[7] Yang J, Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach [C] //Proc of the 6th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2013: 587-596

[8] Rosvall M, Bergstrom C T. Maps of information flow reveal community structure in complex networks [J]. Proceedings of the National Academy of Sciences, 2008, 105(4): 191-218

[9] Li He, Yin Ying, Li Yuan, et al. Large-scale dynamic network community detection by multi-objective evolutionary clustering [J]. Journal of Computer Research and Development, 2019, 56(2): 281-292 (in Chinese)
(李赫，印莹，李源，等. 基于多目标演化聚类的大规模动态网络社区检测[J]. 计算机研究与发展，2019，56(2)：281-292)

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

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

[12] Karrer B, Newman M E. Stochastic blockmodels and community structure in networks [J]. Physical Review E, 2011, 83(1): No.016107

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

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

[15] Jia Caiyan, Li Yafang, Carson M B, et al. Node attribute-enhanced community detection in complex networks [J]. Scientific Reports, 2017, 7(1): No.2626

[16] Yang Tianbao, Jin Rong, Chi Yun, et al. Combining link and content for community detection: A discriminative approach [C] //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 927-936

[17] Yang J, McAuley J, Leskovec J. Community detection in networks with node attributes [C] //Proc of the 13th IEEE Int Conf on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2013: 1151-1156

[18] Chai Bianfang, Yu Jian, Jia Caiyan, et al. Combining a popularity-productivity stochastic block model with a discriminative content model for general structure detection [J]. Physical Review E, 2013, 88(1): No.012807

[19] He Dongxiao, Feng Zhiyong, Jin Di, et al. Joint identification of network communities and semantics via integrative modeling of network topologies and node contents [C] //Proc of the 31st AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2017: 116-124

[20] Chang Zhenhai, Jia Caiyan, Yin Xianjun, et al. A generative model for exploring structure regularities in attributed networks [J]. Information Sciences, 2019, 505: 252-264

[21] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22

[22] Airoldi E M, Blei D M, Fienberg S E, et al. Mixed membership stochastic blockmodels [J]. Journal of Machine Learning Research, 2008, 9(5): 1981-2014

[23] Shen Huawei, Cheng Xueqi, Guo Jiafeng. Exploring the structural regularities in networks [J]. Physical Review E, 2011, 84(5): No.056111

[24] Barron A, Rissanen J, Yu B. The minimum description length principle in coding and modeling [J]. IEEE Transactions on Information Theory, 1998, 44(6): 2743-2760

[25] Newman M E J, Clauset A. Structure and inference in annotated networks [J]. Nature Communications, 2016, 7(1): No.11863

[26] Yang Tianbao, Chi Yun, Zhu Shenghuo, et al. Directed network community detection: A popularity and productivity link model [C] //Proc of the 10th SIAM Int Conf on Data Mining. Philadelphia: SIAM, 2010: 742-753

[27] Chen Yi, Wang Xiaolong, Bu Junzhao, et al. Network structure exploration in networks with node attributes [J]. Physica A: Statistical Mechanics and its Applications, 2016, 449: 240-253

[28] Aldous D J. Exchangeability and related topics [M]. Berlin: Springer, 1985: 84-96

[29] Decelle A, Krzakala F, Moore C, et al. Inference and phase transitions in the detection of modules in sparse networks [J]. Physical Review Letters, 2011, 107(6): No.065701

[30] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification [J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(9): No. P09008

[31] Mccallum A, Nigam K, Rennie J D, et al. Automating the construction of internet portals with machine learning [J]. Information Retrieval, 2000, 3(2): 127-163

[32] Giles C L, Bollacker K D, Lawrence S, et al. CiteSeer: An automatic citation indexing system [C] //Proc of the 3rd ACM Int Conf on Digital Libraries. New York: ACM, 1998: 89-98

[33] Zachary W W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1976, 33(4): 452-473

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

[35] Newman M E. Finding community structure in networks using the eigenvectors of matrices [J]. Physical Review E, 2006, 74(3): No.036104

[36] Adamic L A, Glance N. The political blogosphere and the 2004 US election: Divided they blog [C] //Proc of the 3rd Int Workshop on Link Discovery. New York: ACM, 2005: 36-43

[37] Chang Zhenhai, Yin Xianjun, Jia Caiyan, et al. Mixture models with entropy regularization for community detection in networks [J]. Physica A: Statistical Mechanics and Its Applications, 2018, 496: 339-350

Zheng Yimei, born in 1996. PhD candidate. Her main research interests include machine learning and complex network analysis.

Jia Caiyan, born in 1976. PhD, professor. Her main research interests include machine learning, complex network analysis and social computing.

Chang Zhenhai, born in 1979. PhD. His main research interests include statistical analysis and complex network analysis.

Li Xuanya, born in 1987. PhD. Director of Baidu Campus and the executive member of CCF. His main research interests include internet of things and artificial intelligence.