

面向属性网络社团检测的度修正广义随机块模型^{*}

王笑, 戴芳, 郭文艳, 王军锋

(西安理工大学 理学院, 陕西 西安 710054)

通信作者: 戴芳, E-mail: daifang@xaut.edu.cn



摘要: 随机块模型可以拟合各种网络的生成, 挖掘网络的隐含结构与潜在联系, 在社团检测中具有明显的优势. 广义随机块模型 GSB 是基于链接社团的思想发现广义社团的, 但其仅适用于有向无属性网络. 针对无向属性网络, 对网络拓扑信息建模的同时对节点属性进行建模, 提出一种度修正的属性网络广义随机块模型 DCGSB (degree corrected general stochastic block model). 在该模型中, 假设网络拓扑信息和属性信息的生成过程都服从幂函数形式的分布, 并且引入节点的度来刻画网络的无标度特性, 可以更好地拟合真实网络的生成. 利用期望最大化算法对 DCGSB 模型的参数进行估计, 通过硬划分处理, 得到节点隶属度, 进而完成社团检测. 在 3 个含有不同结构的真实属性网络数据集上进行实验, 并与 10 种社团检测算法进行对比, 结果表明 DCGSB 模型不仅继承了 GSB 模型的优点, 能发现广义社团, 而且由于属性信息和节点度的引入, 使其社团检测能力优于其他 10 种比较算法.

关键词: 随机块模型; 节点度; 属性网络; 社团检测

中图法分类号: TP311

中文引用格式: 王笑, 戴芳, 郭文艳, 王军锋. 面向属性网络社团检测的度修正广义随机块模型. 软件学报, 2025, 36(5): 2308–2320. <http://www.jos.org.cn/1000-9825/7197.htm>

英文引用格式: Wang X, Dai F, Guo WY, Wang JF. Degree Corrected General Stochastic Block Model for Community Detection in Attributed Network. Ruan Jian Xue Bao/Journal of Software, 2025, 36(5): 2308–2320 (in Chinese). <http://www.jos.org.cn/1000-9825/7197.htm>

Degree Corrected General Stochastic Block Model for Community Detection in Attributed Network

WANG Xiao, DAI Fang, GUO Wen-Yan, WANG Jun-Feng

(School of Science, Xi'an University of Technology, Xi'an 710054, China)

Abstract: Stochastic block models can fit the generation of various networks, mining implicit structures and potential connections within these networks. Thus, they have significant advantages in community detection. General stochastic block (GSB) models discover general communities based on link communities, but they are only applicable to directed non-attributed networks. This study proposes a degree corrected general stochastic block (DCGSB) model for undirected attributed networks which models both network topology information and node attributes. In the DCGSB model, it is assumed that the generation of network topology information and node attributes follows a distribution in the form of power functions. Node degrees are introduced to characterize the scale-free property of networks, which allows the model to better fit the generation of real networks. The expectation-maximization algorithm is employed to estimate the parameters of the DCGSB model, and node-community memberships are obtained by hard partition to complete community detection. Experiments are conducted on three real attributed network datasets containing different network structures, and the proposed model is compared with ten existing community detection algorithms. Results show that the DCGSB model not only inherits the advantages of GSB models in identifying general communities but also outperforms the ten algorithms in community detection due to the introduction of attribute information and node degrees.

Key words: stochastic block model (SBM); node degree; attributed network; community detection

* 收稿时间: 2023-09-23; 修改时间: 2024-01-12; 采用时间: 2024-03-28; jos 在线出版时间: 2024-12-11
CNKI 网络首发时间: 2024-12-12

近年来,在许多学科中,将结构化数据建模成复杂网络的趋势日益增长.复杂网络作为一种描述实体之间关系的抽象概念,已成为数据科学研究的热点.网络分析也日益受到数学、计算机科学以及生物学等领域的关注^[1,2].社团检测是网络分析的一项重要任务,其目的是将网络划分为多个社团,使得同一社团内的节点链接比较紧密,不同社团间的节点链接比较稀疏^[3].

社团检测可看作网络节点的聚类问题.经典的启发式聚类算法主要包括基于层次聚类的凝聚式算法(FN)^[4]和分裂式算法(GN)^[5]等.网络结构和离散优化的学习方式对启发式算法的影响较大,导致其很难描述比较复杂的网络数据^[6].生成模型在启发式算法的基础上利用概率分布对网络进行建模,把社团检测问题转化成统计推断问题.

随机块模型(stochastic block model, SBM)^[7]是基于随机等价思想建立的一类概率生成模型,它不仅可以生成各种网络,反过来,还可以探测网络的社团结构. SBM 利用概率分布对网络进行建模,把社团检测问题转化为带有约束条件的参数估计问题,具有很强的理论支撑性.同时, SBM 不需要提前假设网络中都存在哪种网络结构,在不知道网络结构的情况下可以准确地识别真实结构,能识别的结构包括社团结构、多分结构以及混合结构等.具有代表性的 SBM 有: 2008 年, Airoldi 等人^[8]提出的混合隶属度随机块模型; 2011 年, Karrer 等人^[9]提出的度修正随机块模型; 2019 年, Lu 等人^[10]提出的正则化随机块模型等. 这些模型在社团检测的过程中可以借助优化算法对其参数进行估计,如 MCMC (Markov chain Monte Carlo) 算法、置信度算法、期望最大化算法与变分推断算法等.

广义随机块模型(general stochastic block model, GSB)^[11]是基于链接社团的思想提出来的,它不仅可以检测传统的社团结构,还可以检测多分结构、混合结构等多种网络结构.在该模型中,假设边的产生仅与两个端点所在社团有关,将节点看作是无差别的,也没有考虑属性信息对社团检测精度的影响.后来, Jiang 等人^[12]虽然将其扩展到了符号网络,但依然没有考虑属性信息对社团检测精度的影响.因此,本文在 GSB 模型的基础上,针对无向属性网络,综合考虑了网络拓扑信息和属性信息对社团检测精度的影响,在对网络拓扑信息建模的同时对节点属性进行建模,并且引入节点的度刻画真实网络的无标度特性,提出了一种度修正的属性网络广义随机块模型 DCGSB.值得注意的是, DCGSB 模型是对 GSB 模型的一种扩展,两者的区别在于 DCGSB 模型不仅对节点属性进行建模,而且引入了节点的度这个观测变量,可以更好地发现网络中的社团结构.

本文第 1 节介绍和分析 SBM 的相关工作.第 2 节给出 DCGSB 模型的建模过程.第 3 节介绍 DCGSB 模型的参数估计及时间复杂度分析.第 4 节介绍基于 DCGSB 模型的社团检测过程.第 5 节给出实验结果与分析.总结在第 6 节中给出.

1 相关工作

随机块模型是一种经典的网络分析模型,可以描述不同类型的网络结构并还原网络的生成. SBM 按照概率分布可分为 3 类: 基于 Bernoulli 分布的随机块模型、基于 Poisson 分布的随机块模型以及基于幂函数形式分布的随机块模型.

1.1 基于 Bernoulli 分布的随机块模型

2008 年, Airoldi 等人^[8]假设两点间的连边服从 Bernoulli 分布,提出了混合隶属度随机块模型. Xing 等人^[13]将混合隶属度随机块模型扩展到动态网络,提出了动态混合隶属度随机块模型. Aicher 等人^[14]假设边的权重服从指数族分布,并将其引入 SBM 中,提出了加权随机块模型,将 SBM 扩展到了加权网络. Li 等人^[15]对网络结构的异质性进行建模,提出了度修正的重参数化随机块模型,可以表征具有异构块和度分布的各种网络结构. Qiao 等人^[16]通过引入度衰减变量对所有节点上的不同度分布进行编码,提出了幂律度随机块模型,在真实网络中近似地保持了无标度特征. Wu 等人^[17]结合节点流行度刻画网络的无标度特性,提出了动态网络社团检测随机块模型,解决了动态网络中的社团检测和演化追踪问题. Xu 等人^[18]通过对网络中噪声节点建模,提出了鲁棒随机块模型,缓解了网络中的噪声问题. Liu 等人^[19]通过神经网络将节点属性引入到 SBM 中,提出了属性网络生成模型 ANGM (attributed network generative model),可以检测同配网络和异配网络.

1.2 基于 Poisson 分布的随机块模型

2011 年, Karrer 等人^[9]假设两点间的连边服从 Poisson 分布并且融合节点度的幂律分布, 提出了度修正的随机块模型 DCSBM. Chang 等人^[20]假设网络的拓扑信息和属性信息都服从 Poisson 分布, 提出了融合网络拓扑信息和节点属性的随机块模型 PSB_PG. 鉴于 PSB_PG 模型没有考虑节点度对社团检测精度的影响, 郑忆美等人^[21]通过引入节点度来刻画网络的无标度特性, 提出了度修正的属性网络随机块模型 DPSB_PG. Liu 等人^[22]受 Newman 混合模型^[23]思想的启发, 在拓扑模型中引入了一组表征节点链接行为的参数, 并假设网络拓扑信息和属性信息都服从 Poisson 分布, 提出了融合拓扑信息和节点属性信息的 GNAN 模型. Zhu 等人^[24]在不考虑自环的情况下, 提出了无自环随机块模型, 可以检测网络中的反社团 (anti-community) 结构. Wu 等人^[25]通过对原始网络和模体网络 (motif network) 进行统一建模, 提出了混合阶随机块模型, 避免了现有社团检测方法中高阶和低阶连接模式的分离和碎片化问题. Abossedgh 等人^[26]在 DCSBM 的基础上增加每条边标签出现的概率, 提出了带标签的随机块模型, 改善了对个体行为的分析.

1.3 基于幂函数形式分布的随机块模型

2011 年, Shen 等人^[11]假设两节点间的连边服从幂函数形式的分布, 提出了广义随机块模型 GSB. Yang 等人^[27]认为两节点间的连边不仅与节点所在社团有关, 而且与节点的流行度有关, 通过引入节点流行度对网络中边的生成进行建模, 提出了基于流行度的条件链接 (popularity-based conditional link, PCL) 模型, 并且利用内容判别模型 DC 对节点内容进行建模, 提出了 PCL_DC 模型. Chai 等人^[28]不仅考虑了属性信息和节点度的幂律分布, 而且继承了 PPL 模型 (popularity and productivity link model)^[29]和 GSB 模型的优点, 提出了 PPSB 模型 (popularity-productivity stochastic block model). Jiang 等人^[12]将 GSB 模型扩展到符号网络, 提出了符号广义随机块模型. Chen 等人^[30]将节点属性以概率的形式融入 GSB 模型中, 提出了子空间的随机块模型, 在节点之间连边生成的过程中融合拓扑信息和属性信息构建了隐网络, 避免了负采样策略. 虽然有研究人员对 GSB 模型进行了扩展, 但是没有考虑节点度对网络链接的影响, 也没有对属性信息进行建模. 因此, 本文针对无向属性网络, 在 GSB 模型研究的基础上, 对网络拓扑信息和属性信息同时建模, 并且引入节点的度刻画网络的无标度特性, 提出一种度修正的属性网络广义随机块模型 DCGSB, 使其达到更好的社团检测精度.

2 DCGSB 模型

设 $G(V, E, X)$ 是一个无向属性网络, 其中集合 $V = \{1, 2, 3, \dots, n\}$ 由 n 个节点组成, 集合 $E = \{e_1, e_2, \dots, e_m\}$ 由 m 条边组成, $X = (x_{ik})_{n \times K}$ 表示所有节点的属性矩阵. 若节点 i 具有属性 k , 则 $x_{ik} = 1$; 否则 $x_{ik} = 0$. $A = (a_{ij})_{n \times n}$ 为网络的邻接矩阵, 表示网络中节点之间的连边. 若节点对 (i, j) 之间存在连边, 则 $a_{ij} = 1$; 否则 $a_{ij} = 0$. 假设网络 G 被分成 c 个不同的社团, 用 $V_1, V_2, \dots, V_c$ 表示, 则有 $G = \bigcup_{r=1}^c V_r$.

2.1 基于拓扑信息的生成模型

在标准的随机块模型^[7]中, 概率矩阵 $H = (h_{rs})_{c \times c}$ 控制了网络中所有节点的连边概率, 当 $i \in V_r$ 且 $j \in V_s$ 时, 节点 i 与节点 j 之间的连边概率只与其所在社团有关, 即与 h_{rs} 有关. 在 GSB 模型^[11]中, Shen 等人通过引入节点隶属度矩阵 $B = (b_{ir})_{n \times c}$ 放松了这个限制, 其中 b_{ir} 表示节点 i 在社团 V_r 中的概率. 但是 GSB 模型只适用于有向无属性网络, 且将社团内的节点看作是无差别的, 忽略了节点度的异质性, 没有建模真实网络中节点度的幂律分布. 所以, 本文针对无向属性网络, 在对网络拓扑信息建模的同时对节点属性进行建模, 并且引入节点的度刻画网络的无标度特性, 更好地拟合了真实网络.

节点隶属度矩阵 B , 概率矩阵 H 以及节点度 $D = (d_i)_{1 \times n}$ 对网络中节点 i 与节点 j 之间存在连边的可能性有很大的影响. 一般地, 节点间存在连边的可能性随着节点度的增大而增大, 同时影响网络链接的生成. 假设节点 i 与节点 j 之间边的生成服从幂函数形式的分布, 则社团 V_r 中的节点 i 和社团 V_s 中的节点 j 存在连边的可能性为:

$$A_{ij}^{rs} = d_i b_{ir} h_{rs} d_j b_{js} \quad (1)$$

考虑所有社团, 节点对 (i, j) 生成连边的可能性为:

$$\hat{A}_{ij} = \sum_{r,s=1}^c d_i b_{ir} h_{rs} d_j b_{js} \quad (2)$$

给定模型参数 B, H 和观测变量节点的度 $D = (d_i)_{1 \times n}$ 后, 生成网络 G 的概率 $P(A|B, H, D)$ 为:

$$P(A|B, H, D) = \prod_{i,j=1}^n (\hat{A}_{ij})^{a_{ij}} \quad (3)$$

其中, 概率矩阵 H 是对称的且满足约束条件 $\sum_{r,s=1}^c h_{rs} = 1$, 节点隶属度矩阵 B 满足约束条件 $\sum_{i=1}^n d_i b_{ir} = \sum_{i=1}^n \tilde{b}_{ir} = 1$. 由于概率矩阵 $H = (h_{rs})_{c \times c}$ 的引入, 继承了随机块模型的优点, 可以检测多种网络结构, 比如概率矩阵 H 对角线元素大于非对角线元素时, 可以检测传统的社团结构; 对角线元素小于非对角线元素时, 可以检测网络的异配结构; 其他情况可以检测更多的网络结构.

2.2 基于属性信息的生成模型

在属性网络中, 节点属性通常用高维向量表示. 如果一组高维向量中具有相同分量的个数越多, 其对应节点构成的网络拓扑结构一致的可能性越大, 越容易形成社团结构. 假设网络中节点属性的生成也服从幂函数形式的分布, θ_{rk} 为社团 V_r 具有属性 k 的概率, $\Theta = (\theta_{rk})_{c \times K}$ 是社团相关属性矩阵. 一般地, 对于不同类型的网络, 节点度与节点属性的关系可能呈现多样性. 但在一些网络中, 节点度与节点属性具有相关性. 基于此, 社团 V_r 中节点 i 具有属性 k 与节点隶属度 b_{ir} 、节点的度 d_i 以及社团相关属性 θ_{rk} 有关, 社团 V_r 中节点 i 具有属性 k 的可能性为:

$$\hat{X}_{ik}^r = d_i b_{ir} \theta_{rk} \quad (4)$$

考虑所有社团, 节点 i 具有属性 k 的可能性为:

$$\hat{X}_{ik} = \sum_{r=1}^c d_i b_{ir} \theta_{rk} \quad (5)$$

其中, $\sum_{i=1}^n d_i b_{ir} = \sum_{i=1}^n \tilde{b}_{ir} = 1, \sum_{k=1}^K \theta_{rk} = 1$ 满足归一化约束条件.

给定模型参数 B, Θ 以及观测变量节点的度 D 后, 网络中生成节点属性的概率 $P(X|B, \Theta, D)$ 为:

$$P(X|B, \Theta, D) = \prod_{i=1}^n \prod_{k=1}^K (\hat{X}_{ik})^{x_{ik}} \quad (6)$$

2.3 融合拓扑信息和属性信息的度修正广义随机块模型

假设网络拓扑信息和节点属性的生成过程相互独立, 则度修正的属性网络广义随机块模型 DCGSB 的联合概率为:

$$\begin{aligned} P(A, X|B, H, \Theta, D) &= P(A|B, H, D) \times P(X|B, \Theta, D) \\ &= \prod_{i,j=1}^n \left(\sum_{r,s=1}^c d_i b_{ir} h_{rs} d_j b_{js} \right)^{a_{ij}} \times \prod_{i=1}^n \prod_{k=1}^K \left(\sum_{r=1}^c d_i b_{ir} \theta_{rk} \right)^{x_{ik}} \end{aligned} \quad (7)$$

如图 1 所示, 与 GSB 模型相比, DCGSB 模型进一步假设一条边的生成概率还依赖于其两端节点的度, 并且针对无向属性网络, 在对网络拓扑信息建模的同时对节点属性进行建模, 其节点属性的生成也服从幂函数形式的分布. 图 1 中阴影圆表示观测变量, 图 1(a) 中的无阴影圆表示隐变量, 虚线表示两个变量之间的关系, 实线表示观测到的连边; 图 1(b) 中的无阴影圆表示模型参数. DCGSB 算法的建模过程如图 1(b) 所示. 根据建模过程, 网络中节点间连边的生成和节点属性的生成步骤可以总结如下.

- (1) 从社团 V_r 和 V_s 中分别以概率 b_{ir} 和 b_{js} 抽取节点 i 和 j .
- (2) 在节点 i 和 j 之间以概率 $\hat{A}_{ij}$ 生成连边, 其中 $\hat{A}_{ij} = \sum_{r,s=1}^c d_i b_{ir} h_{rs} d_j b_{js}$.
- (3) 从社团 V_r 中以概率 θ_{rk} 选择一种属性 k .
- (4) 给节点 i 以概率 $\hat{X}_{ik}$ 选择一种属性, 其中 $\hat{X}_{ik} = \sum_{r=1}^c d_i b_{ir} \theta_{rk}$.

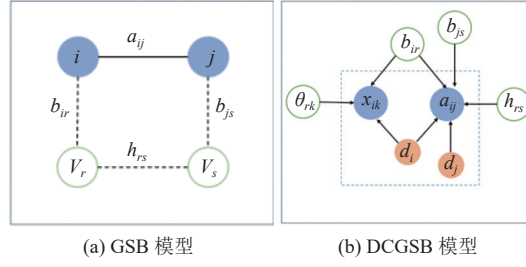


图1 GSB 与 DCGSB 建模过程示意图

3 DCGSB 模型的参数估计及时间复杂度分析

3.1 DCGSB 模型参数的 EM 算法估计

DCGSB 模型总共包含 3 种变量, 分别为: 观测变量 A 、 X 和 D ; 隐变量 $Q = (q_{ij}^{rs})_{n \times n}$ 和 $\Upsilon = (\gamma_{ik}^r)_{n \times K}$; 模型参数 $B = (b_{ir})_{n \times c}$ 、 $H = (h_{rs})_{c \times c}$ 和 $\Theta = (\theta_{rk})_{c \times K}$. 由于模型中存在隐变量, 无法直接对似然函数进行求解, 期望最大化 (expectation-maximum, EM) 算法^[31]是常见的隐变量估计方法, 能很好地处理此类问题. 所以, 本文采用 EM 算法对模型进行参数估计. 推断过程如下.

对联合概率模型公式 (7) 取对数得到对数似然为:

$$L(B, H, \Theta) = \sum_{i,j=1}^n a_{ij} \ln \left(\sum_{r,s=1}^c d_i b_{ir} h_{rs} d_j b_{js} \right) + \sum_{i=1}^n \sum_{k=1}^K x_{ik} \ln \left(\sum_{r=1}^c d_i b_{ir} \theta_{rk} \right) \quad (8)$$

在期望 E (expectation) 步, 参数 B, H, Θ 固定后, 利用 Jensen 不等式得到对数似然的下界为:

$$\bar{L}(B, H, \Theta) = \sum_{i,j=1}^n \sum_{r,s=1}^c \left[a_{ij} q_{ij}^{rs} \ln \left(\frac{d_i b_{ir} h_{rs} d_j b_{js}}{q_{ij}^{rs}} \right) \right] + \sum_{i=1}^n \sum_{k=1}^K \sum_{r=1}^c \left[x_{ik} \gamma_{ik}^r \ln \left(\frac{d_i b_{ir} \theta_{rk}}{\gamma_{ik}^r} \right) \right] \quad (9)$$

其中,

$$q_{ij}^{rs} = \frac{d_i b_{ir} h_{rs} d_j b_{js}}{\sum_{r,s=1}^c d_i b_{ir} h_{rs} d_j b_{js}}, \quad \gamma_{ik}^r = \frac{d_i b_{ir} \theta_{rk}}{\sum_{r=1}^c d_i b_{ir} \theta_{rk}} \quad (10)$$

其中, q_{ij}^{rs} 表示节点 i, j 分别在社团 V_r 和 V_s 中且节点对 (i, j) 之间存在连边的概率; γ_{ik}^r 表示节点 i 在社团 V_r 中且具有属性 k 的概率.

在最大化 M (maximum) 步, 隐变量 $q_{ij}^{rs}, \gamma_{ik}^r$ 固定, 根据 Lagrange 乘子法可以得到 $b_{ir}, h_{rs}, \theta_{rk}$ 这 3 个参数的估计, 分别为:

$$b_{ir} = \frac{\sum_{j=1}^n \sum_{s=1}^c a_{ij} q_{ij}^{rs} + \sum_{k=1}^K x_{ik} \gamma_{ik}^r}{d_i \times \left(\sum_{j=1}^n \sum_{s=1}^c a_{ij} q_{ij}^{rs} + \sum_{k=1}^K x_{ik} \gamma_{ik}^r \right)} \quad (11)$$

$$h_{rs} = \frac{\sum_{i,j=1}^n a_{ij} q_{ij}^{rs}}{\sum_{i,j=1}^n \sum_{r,s=1}^c a_{ij} q_{ij}^{rs}}, \quad \theta_{rk} = \frac{\sum_{i=1}^n x_{ik} \gamma_{ik}^r}{\sum_{i=1}^n \sum_{k=1}^K x_{ik} \gamma_{ik}^r} \quad (12)$$

参数估计方法的具体步骤如算法 1 所示.

算法 1. DCGSB 的参数推断算法.输入: $A, X, c, I_T, \varepsilon$;输出: B, H, Θ .

1. 计算节点的度 $D = (d_i)_{1 \times n}$;
2. 初始化 $B^{(0)}, H^{(0)}, \Theta^{(0)}$;
3. 根据公式 (8) 计算目标函数 $L(B^{(0)}, H^{(0)}, \Theta^{(0)})$;
4. **for** $t = 1 : I_T$ **do**
5. E-step: 根据公式 (10) 计算 $q_{ij}^{rs}, \gamma_{ik}^r$;
6. M-step: 根据公式 (11)、公式 (12) 计算 $B^{(t)}, H^{(t)}, \Theta^{(t)}$;
7. 根据公式 (8) 计算目标函数 $L(B^{(t)}, H^{(t)}, \Theta^{(t)})$;
8. **if** $|L(B^{(t)}, H^{(t)}, \Theta^{(t)}) - L(B^{(t-1)}, H^{(t-1)}, \Theta^{(t-1)})| < \varepsilon$ 或者 $t = I_T$ **then**
9. $B = B^{(t)}, H = H^{(t)}, \Theta = \Theta^{(t)}$; **break**;
10. **end if**
11. **end for**

3.2 初值设置

在算法 DCGSB 中, 概率矩阵 H 的初始化对其收敛速度有着较大的影响. 当 H 的初始化所产生的网络结构与真实网络结构一致时, 算法就很容易收敛; 但当 H 初始化产生的网络结构与真实网络结构不一致时, 算法收敛速度就会很慢. 为了让算法尽可能少的迭代次数就达到收敛, 在实验中, 利用最大熵^[32]与最大似然的思想对 H 的初始值进行设置, 分为以下 3 种情况: (1) 当 H 对角线元素大于非对角线元素时, 对应网络的同配结构; (2) 当 H 对角线元素小于非对角线元素时, 对应网络的异配结构; (3) 当 H 中的元素在某个值附近浮动时 (比如 0.5), 对应网络的其他结构. 针对每个网络, 我们首先将这 3 种情况分别执行若干次 (比如 10 次), 然后计算每种情况下这若干次 (10 次) 似然函数的平均值, 最后将平均值最大的那种情况对应的 H 初始化形式作为 DCGSB 算法的初始化.

3.3 时间复杂度分析

DCGSB 模型的时间复杂度主要由 EM 算法的 E 步 (公式 (10)) 和 M 步 (公式 (11)、公式 (12)) 决定. 在每一次迭代过程中, E 步的时间复杂度为 $O(mc^2 + nKc)$; M 步的时间复杂度为 $O(nc + c^2 + nKc)$. 由于社团数 c 远远小于网络节点数 n , 即 $c \ll n$, 因此 M 步的时间复杂度可以写成 $O(nKc)$, 并且算法的最大迭代次数为 I_T , 故整个算法的时间复杂度为 $O(I_T(mc^2 + nKc))$.

4 基于 DCGSB 模型的社团检测

由于节点隶属度矩阵 B 刻画了网络中节点在社团上的分布情况, 我们的主要目标是求出节点隶属度矩阵 $B = (b_{ir})_{n \times c}$, 即每个节点属于任意一个社团 V_r ($r = 1, 2, \dots, c$) 的概率. 本文利用硬划分的方式来推断网络中节点的社团隶属度, 即利用 b_{ir} 的最大值限定一个节点只能属于一个社团. 由于 EM 算法求解模型参数时引入了隐变量, 不能直接对 b_{ir} 进行处理. 为了得到硬划分, 本文对参数 b_{ir} 与 h_{rs} 设置了一种运算, 即:

$$M_{ir} = \frac{\sum_{s=1}^c b_{ir} h_{rs}}{\sum_{r,s=1}^c b_{ir} h_{rs}} \quad (13)$$

DCGSB 算法可通过优化 $r^* = \arg \max_r M_{ir}$ 得到节点 i 最终属于社团 r^* .

5 实验结果及分析

5.1 数据集

本文选取 3 个不同规模的真实属性网络数据集 WebKB (Cornell, Texas, Washington, Wisconsin), Cora 和 Citeseer 来检验 DCGSB 模型的划分效果. 属性网络数据集的基本特征如表 1 所示, 其中 n 和 m 分别为节点数和边数; K 为属性种类; c 为社团个数.

表 1 属性网络数据集特征

Datasets		n	m	K	c	Structure
WebKB	Cornell	195	304	1 703	5	disassortative
	Texas	187	328	1 703	5	disassortative
	Washington	230	446	1 703	5	disassortative
	Wisconsin	265	530	1 703	5	disassortative
Cora	—	2 708	5 429	1 433	7	assortative
Citeseer	—	3 312	4 723	3 703	6	assortative

1) WebKB 数据集^[33]. WebKB 是一种引文网络, 由康奈尔大学 (Cornell), 得克萨斯大学 (Texas), 华盛顿大学 (Washington) 以及威斯康星大学 (Wisconsin) 这 4 所大学的网页及网页间的链接组成, 共有 877 个节点, 表示所有的网页; 1 608 条边, 表示网页之间的超链接; 网络中的网页 (即节点) 被分成学院主页、课程主页、学生主页、员工主页以及研究项目主页 5 种类型. 每个节点都由一个 1 703 维的属性向量构成. 此外, WebKB 数据集包含的 4 个网络均为混合结构.

2) Cora 数据集^[34]. Cora 是由 2 708 个节点和 5 429 条边组成的科技文献引文网络. 网络中所有文献 (即节点) 被分成神经网络、强化学习、规则学习、基于案例的推理、概率方法、遗传算法以及理论知识这 7 种类型.

3) Citeseer 数据集^[35]. 该数据集是学术引用网络, 包含 3 312 个节点, 表示所有的学术论文; 4 723 条边, 表示各论文之间的引用关系; 网络中所有的论文 (即节点) 被分成以下 6 种类型, 即代理人 (agents)、人工智能、数据库、人机交互、信息检索以及机器学习.

5.2 评价指标

为了评估 DCGSB 模型在真实网络上的检测效果, 本文采用社团检测常用的评价指标标准化互信息 (normalized mutual information, NMI)^[36]、 F 测度 (pairwise F -measure, PWF)^[37] 以及准确率 (accuracy, ACC)^[38] 对其进行评估.

(1) 标准化互信息 (NMI)

文献 [36] 提出的 NMI 是根据混淆矩阵判断社团划分后信息保留的完整程度. 其定义如公式 (14) 所示:

$$NMI = \frac{-2 \sum_{r=1}^{c_1} \sum_{s=1}^{c_2} N_{rs} \log \left(\frac{N_{rs} N}{N_r N_s} \right)}{\sum_{r=1}^{c_1} N_r \log \left(\frac{N_r}{N} \right) + \sum_{s=1}^{c_2} N_s \log \left(\frac{N_s}{N} \right)} \quad (14)$$

其中, c_1 表示真实社团数, c_2 表示社团检测算法划分的社团数, N 是网络节点总数, N_r 、 N_s 分别表示社团 r 、社团 s 的节点数, N_{rs} 表示本该属于社团 r 的节点却被错误地划分到社团 s 的节点数. NMI 的取值范围为 $[0, 1]$. 值越大, 表示算法社团划分效果越好.

(2) F 测度 (PWF)

文献 [37] 提出的 PWF 是将精确度 (Precision) 和召回率 (Recall) 的概念归到单个评估中, 其定义如公式 (15) 所示:

$$PWF = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (15)$$

其中, $Precision = |S \cap T|/|S|$ 是社团检测算法检测结果的精确度, $Recall = |S \cap T|/|T|$ 是社团检测算法检测结果的召

回率; S 是社团检测算法检测的社团中含有相同标签的节点集合, T 是真实网络社团中含有相同标签的节点集合, $|\cdot|$ 表示集合中元素的数量. PWF 的取值范围也是 $[0, 1]$. 值越大, 说明社团检测算法划分的效果越好.

(3) 准确率 (ACC)

文献 [38] 提出的 ACC 是用来度量正确划分的节点百分比的. 假设一个网络有 n 个节点, 对于每一个节点, l_i 表示社团检测算法检测到的社团标签, r_i 是真实的社团标签, 其定义如公式 (16) 所示:

$$ACC = \frac{1}{n} \sum_{i=1}^n I(r_i, \text{map}(l_i)) \quad (16)$$

其中, $I(x, y)$ 是一个指示函数, 如果 $x = y$, 则 $I(x, y) = 1$; 否则 $I(x, y) = 0$. $\text{map}(l_i)$ 是一个映射函数, 它将每个社团标签 l_i 映射为等价的真实社团标签. ACC 的取值范围依然是 $[0, 1]$. 值越大, 说明社团检测算法划分的效果越好.

5.3 社团数 c 对社团检测精度的影响分析

利用 DCGSB 模型进行社团检测时, 社团数 c 的不同会影响社团检测精度. 因此, 本节分别取 $c = 2, 3, 4, 5, 6, 7, 8, 9, 10$, 在数据集 Cornell, Texas, Washington, Wisconsin, Cora 以及 Citeseer 上进行了 10 次实验, 分别计算这 10 次实验的 NMI 、 PWF 以及 ACC 的平均值, 并绘制成折线图, 如图 2 所示.

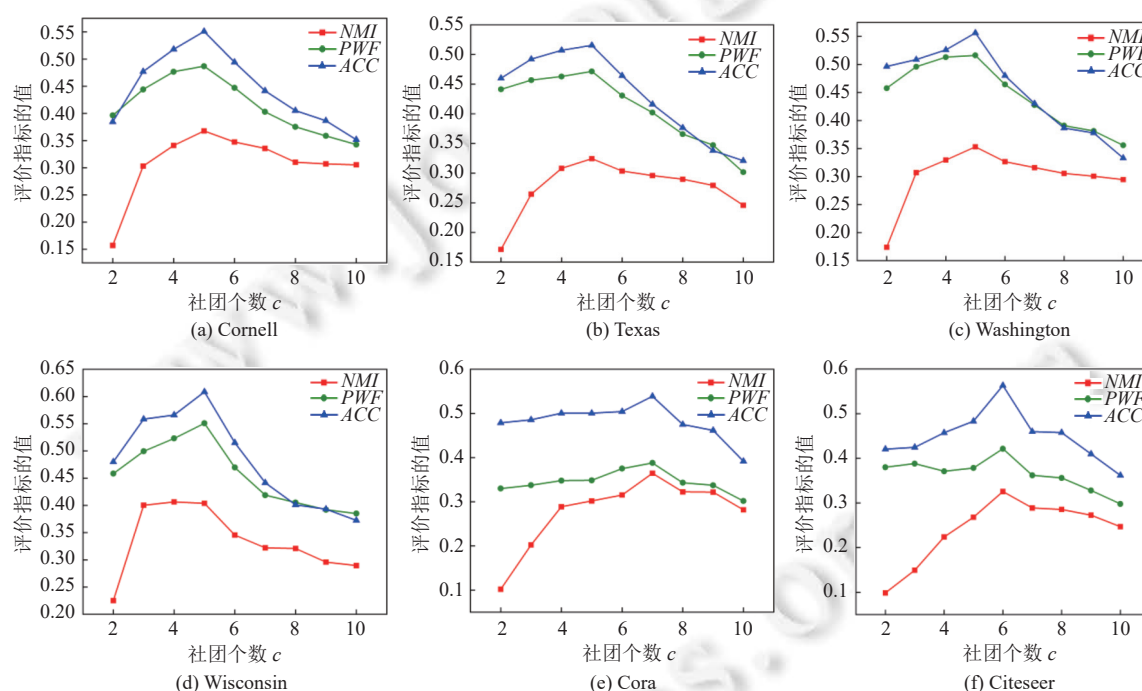


图2 参数 c 对不同网络社团检测精度的影响

由图 2 可以看出, 在数据集 Cornell, Texas, Washington, Wisconsin, Cora 以及 Citeseer 上, 随着社团数量 c 的增加, DCGSB 模型社团检测的评价指标 NMI 、 PWF 和 ACC 都呈现“先上升后下降”的趋势, 并且每个网络在真实社团数量处取到最大值. 因此, 在接下来的实验中, 我们取社团个数 c 等于真实网络社团数量.

如果不知道真实社团数量时, c 从 2 开始, 每次增加 1, 分别计算给定网络上的 NMI 、 PWF 以及 ACC , 直到评价指标出现“拐点”, 此时所对应的 c 值, 就是社团个数. 也可以采用其他方法确定社团数 c .

5.4 节点度对社团检测精度的影响分析

为了验证 DCGSB 模型中节点度对社团检测精度的影响, 本文对节点度做了消融实验. 在表 1 所示的 3 个网络数据集上, 分别对网络拓扑信息和节点属性建模时, 没有引入节点度的属性网络广义随机块模型 CGSB 与加了

节点度的 DCGSB 模型进行了实验. 表 2 显示了两个模型分别在 3 个网络数据集上 30 次实验结果的 *NMI*、*PWF* 和 *ACC* 的平均值和标准差.

表 2 节点度对社团检测精度的影响 (均值±标准差)

评价指标	模型	Cornell	Texas	Washington	Wisconsin	Cora	Citeseer
<i>NMI</i>	CGSB	0.2952±0.0608	0.3077±0.0628	0.3271±0.0472	0.3899±0.0774	0.3350±0.0291	0.2936±0.0237
	DCGSB	0.3678±0.0369	0.3244±0.0625	0.3531±0.0283	0.4038±0.0294	0.3645±0.0250	0.3255±0.0253
<i>PWF</i>	CGSB	0.4860±0.0749	0.4213±0.0456	0.5142±0.0408	0.5576±0.0408	0.3735±0.0277	0.3936±0.0250
	DCGSB	0.4869±0.0493	0.4715±0.0338	0.5162±0.0221	0.5508±0.0322	0.3879±0.0105	0.4214±0.0221
<i>ACC</i>	CGSB	0.5427±0.0781	0.4708±0.0590	0.5321±0.0630	0.5887±0.0543	0.5201±0.0354	0.5119±0.0323
	DCGSB	0.5508±0.0458	0.5155±0.0658	0.5563±0.0280	0.6086±0.0296	0.5389±0.0302	0.5629±0.0287

由表 2 可以看出,除了在网络 Wisconsin 上,没有加节点度的 CGSB 模型的 *PWF* 指标比加了节点度的 DCGSB 模型稍微高了一点外,其余网络上的 *NMI*、*PWF* 以及 *ACC* 都是加了节点度的 DCGSB 模型比没有加节点度的 CGSB 模型高,并且在网络 Cornell, Texas, Washington, Wisconsin, Cora 以及 Citeseer 上,DCGSB 模型的 *NMI* 指标比 CGSB 模型的 *NMI* 指标分别提高了 24.59%, 5.43%, 7.95%, 3.57%, 8.81%, 10.87%, 说明引入节点度刻画网络的无标度特征对社团检测精度是有积极影响的.

为了验证节点度与节点属性之间的关系,本文在对网络拓扑信息和节点属性建模时,只在网络拓扑信息中引入节点度的模型 TGSB 与对两者都引入节点度的模型 DCGSB 进行了对比. 表 3 是两个模型在表 1 所示的 3 个网络数据集上进行 30 次实验结果的 *NMI*、*PWF* 以及 *ACC* 的平均值和标准差.

表 3 节点度与节点属性之间的关系 (均值±标准差)

评价指标	模型	Cornell	Texas	Washington	Wisconsin	Cora	Citeseer
<i>NMI</i>	TGSB	0.3027±0.0487	0.2377±0.0595	0.2875±0.0277	0.3886±0.0582	0.2353±0.0387	0.2791±0.0290
	DCGSB	0.3678±0.0369	0.3244±0.0625	0.3531±0.0283	0.4038±0.0294	0.3645±0.0250	0.3255±0.0253
<i>PWF</i>	TGSB	0.4489±0.0866	0.4028±0.0349	0.4779±0.0484	0.5436±0.0604	0.3111±0.0342	0.3756±0.0256
	DCGSB	0.4869±0.0493	0.4715±0.0338	0.5162±0.0221	0.5508±0.0322	0.3879±0.0105	0.4214±0.0221
<i>ACC</i>	TGSB	0.4862±0.0619	0.4219±0.0515	0.5400±0.0624	0.5962±0.0729	0.4510±0.0607	0.5200±0.0222
	DCGSB	0.5508±0.0458	0.5155±0.0658	0.5563±0.0280	0.6086±0.0296	0.5389±0.0302	0.5629±0.0287

由表 3 可以看出,对网络拓扑信息和节点属性建模时,两者都引入节点度的 DCGSB 模型各个评价指标都比只在网络拓扑信息中引入节点度的 TGSB 模型高,其中在网络 Cornell, Texas, Washington, Wisconsin, Cora 以及 Citeseer 上,DCGSB 模型的 *NMI* 指标比只在网络拓扑信息中引入节点度的 TGSB 模型分别提高了 21.51%, 36.47%, 22.82%, 3.91%, 54.91%, 16.62 %, 说明在这些网络中节点度与节点属性具有一定的相关性.

5.5 社团检测结果与分析

为了评估模型 DCGSB 的检测精度,本文在表 1 所示的 3 个网络数据集上进行社团检测实验,与现有生成模型 PCL_DC^[27]、DCSBM^[9]、PPSB^[28]、PSB_PG^[20]、DPSB_PG^[21]以及 ANG^[19]进行了对比,并且与经典的社团检测算法 SA-Cluster^[39]、In-Custer^[1]、BAGC^[40]以及 BANE^[41]进行了比较. 为了保持实验的公平性,对比算法的参数都采用原文献中的最优参数,实验结果如表 4、表 5 和表 6 所示,其中红色加粗数值表示 11 个算法中最好的 *NMI*、*PWF* 和 *ACC*, 黑色加粗表示次好的; BANE 和 ANG^[19]的数据来自文献 [19]; SA-Cluster、In-Custer、PCL_DC、PPSB 以及 BAGC 的数据来自文献 [42]; DCSBM 的数据来自文献 [43]. 表 4 是 11 种算法的 *NMI* 指标, 表 5 是 11 种算法的 *PWF* 指标, 表 6 是 11 种算法的 *ACC* 指标. 为了使实验结果更稳定,我们对每个算法进行了 30 次实验,给出了 3 个指标 30 次的平均值和标准差.

由实验结果可以看出,与其他 10 种算法相比,本文提出的 DCGSB 模型不仅适用于多种网络结构检测,而且表现出更好的检测精度和性能. 在含有混合结构的异配网络 (Cornell, Texas, Washington, Wisconsin) 上, *NMI* 的值都是最优的. 在网络 Cornell 和 Texas 上, DCGSB 模型的 *NMI* 值比表现最好的对比算法 DPSB_PG 分别提高了

13.31% 和 6.92%, DPSB_PG 也考虑了节点的度, 但是检测效果却没有本文提出的度修正的属性网络广义随机块模型 DCGSB 好, 这说明在 DCGSB 模型中拓扑信息和属性信息均服从幂函数形式的分布具有一定的优势. 在异配网络 Washington 和 Wisconsin 上, DCGSB 的 *NMI* 值比表现最好的对比算法 ANGEM 分别提高了 1.64% 和 0.70%. ANGEM 没有考虑网络中节点的度, 而本文的模型 DCGSB 考虑了节点的度, 这说明加入节点的度对提高网络社团检测精度具有很大的帮助. 在含有社团结构的同配网络 Citeseer 上, DCGSB 模型的 *NMI* 值比表现最好的对比算法 ANGEM 提高了 12.09%, 但在同样含有社团结构的同配网络 Cora 上, DCGSB 的 *NMI* 值对比算法 PCL_BC 低了 12.38%, 这是因为 Cora 中的社团结构没有 Citeseer 中的社团结构明显.

表 4 DCGSB 模型和对比算法在属性网络上的 *NMI* (均值±标准差)

算法	Cornell	Texas	Washington	Wisconsin	Cora	Citeseer
PCL_DC	0.073±0.010	0.061±0.011	0.092±0.015	0.060±0.000	0.416±0.003	0.170±0.003
SA-Cluster	0.064±0.000	0.082±0.000	0.077±0.000	0.101±0.000	0.117±0.000	0.047±0.000
In-Custer	0.038±0.000	0.106±0.000	0.063±0.000	0.089±0.000	0.112±0.000	0.043±0.000
DCSBM	0.0969	0.1665	0.0987	0.0314	0.1707	0.0413
BAGC	0.040±0.006	0.052±0.007	0.053±0.006	0.034±0.015	0.008±0.005	0.017±0.000
PPSB	0.068±0.001	0.111±0.015	0.112±0.006	0.078±0.013	0.068±0.035	0.033±0.001
BANE	0.1257	0.1767	0.1682	0.2000	—	0.0924
PSB_PG	0.3131±0.0476	0.2965±0.0579	0.3235±0.0374	0.3736±0.0550	0.3012	0.2533±0.0348
DPSB_PG	0.3246±0.0385	0.3034±0.0346	0.3263±0.0247	0.3774±0.0368	0.3143±0.0205	0.2658±0.0326
ANGM	0.2909	0.2864	0.3474	0.4010	—	0.2904
DCGSB	0.3678±0.0369	0.3244±0.0625	0.3531±0.0283	0.4038±0.0294	0.3645±0.0560	0.3255±0.0253

表 5 DCGSB 模型和对比算法在属性网络上的 *PWF* (均值±标准差)

算法	Cornell	Texas	Washington	Wisconsin	Cora	Citeseer
PCL_DC	0.281±0.000	0.316±0.018	0.326±0.034	0.274±0.002	0.441±0.002	0.299±0.002
SA-Cluster	0.386±0.000	0.383±0.000	0.474±0.000	0.398±0.000	0.282±0.000	0.298±0.000
In-Custer	0.401±0.000	0.399±0.000	0.472±0.000	0.426±0.000	0.284±0.000	0.299±0.000
BAGC	0.342±0.034	0.546±0.003	0.480±0.002	0.479±0.005	0.299±0.006	0.298±0.000
PPSB	0.308±0.007	0.467±0.005	0.358±0.009	0.328±0.006	0.190±0.010	0.190±0.005
PSB_PG	0.4378±0.0627	0.4117±0.0365	0.4879±0.0413	0.5290±0.0502	0.3554	0.3569±0.0366
DPSB_PG	0.4498±0.0476	0.4293±0.0561	0.4891±0.0175	0.5364±0.0437	0.3629	0.3854±0.0359
DCGSB	0.4869±0.0493	0.4715±0.0338	0.5162±0.0221	0.5508±0.0322	0.3879±0.0105	0.4214±0.0221

表 6 DCGSB 模型和对比算法在属性网络上的 *ACC* (均值±标准差)

算法	Cornell	Texas	Washington	Wisconsin	Cora	Citeseer
PCL_DC	0.329±0.014	0.348±0.015	0.380±0.039	0.336±0.000	0.564±0.009	0.412±0.016
SA-Cluster	0.415±0.000	0.401±0.000	0.491±0.000	0.404±0.000	0.264±0.000	0.233±0.000
In-Custer	0.405±0.000	0.423±0.000	0.465±0.000	0.464±0.000	0.267±0.000	0.230±0.000
DCSBM	0.3794	0.4809	0.3180	0.3282	0.3848	0.2657
BAGC	0.439±0.003	0.563±0.003	0.464±0.003	0.474±0.011	0.301±0.007	0.222±0.000
PPSB	0.362±0.026	0.506±0.012	0.402±0.021	0.385±0.032	0.263±0.006	0.263±0.006
BANE	0.3805	0.4556	0.4291	0.4362	—	0.3334
PSB_PG	0.4780±0.085	0.4396±0.056	0.5348±0.0399	0.5849±0.0380	0.4494±0.671	0.5148±0.063
DPSB_PG	0.5199±0.0400	0.4438±0.0630	0.5330±0.0330	0.5884±0.0466	0.4339±0.0249	0.5233±0.0360
ANGM	0.4421	0.5059	0.5435	0.5660	—	0.5498
DCGSB	0.5508±0.0458	0.5155±0.0658	0.5563±0.0280	0.6086±0.0296	0.5389±0.0302	0.5629±0.0287

与考虑节点度并且服从 Poisson 分布的 DPSB_PG 模型相比, 在 3 个属性网络数据集上, 不论是 *NMI* 指标, 还是 *PWF* 和 *ACC* 指标, 模型 DCGSB 都比 DPSB_PG 的值高, 进一步说明了本文提出的模型 DCGSB 中, 让拓扑信息和属性信息均服从幂函数形式的分布对于多种网络社团检测均表现出良好的性能.

综上所述, 本文提出的 DCGSB 模型综合性能优于其他 10 种对比算法, 并且让网络的拓扑信息和属性信息的生成均服从幂函数形式的分布且引入节点度刻画网络无标度特性的随机块模型能较好地识别属性网络的同配结构和异配结构。

6 结 论

本文针对无向属性网络, 提出了一种度修正的属性网络广义随机块模型 DCGSB. 该模型联合了网络拓扑信息和属性信息, 在对网络拓扑信息建模的同时对节点属性进行建模, 并且引入节点的度刻画网络的无标度特性, 更好地拟合了真实网络, 进而提高了社团检测精度. 统一的幂函数形式的分布有利于模型参数的估计, 利用 EM 算法具有较好的算法收敛性. 实验结果显示: 本文的模型 DCGSB 能够发现网络的多种结构, 相较于比较的现有其他算法在性能上有不同程度的提升, 这说明网络拓扑信息和属性信息的生成均服从幂函数形式的分布并且引入节点的度可以更好地拟合真实网络, 对社团检测精度也有提升.

本文主要研究了广义随机块模型在无向属性网络中的社团检测问题, 下一步可以考虑将广义随机块模型 GSB 扩展到其他类型的网络. 同时, 本文在模型参数估计时采用了 EM 算法. 但是 EM 算法随着网络规模的增大, 会增加计算量. 在未来工作中, 可以考虑其他参数估计方法.

References:

- [1] Zhou Y, Cheng H, Yu JX. Clustering large attributed graphs: An efficient incremental approach. In: Proc. of the 2010 IEEE Int'l Conf. on Data Mining. Sydney: IEEE, 2010. 689–698. [doi: 10.1109/ICDM.2010.41]
- [2] Clauset A, Moore C, Newman MEJ. Hierarchical structure and the prediction of missing links in networks. *Nature*, 2008, 453(7191): 98–101. [doi: 10.1038/nature06830]
- [3] Newman MEJ. Detecting community structure in networks. *The European Physical Journal B*, 2004, 38(2): 321–330. [doi: 10.1140/epjb/e2004-00124-y]
- [4] Newman MEJ. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, 69(6): 066133. [doi: 10.1103/PhysRevE.69.066133]
- [5] Girvan M, Newman MEJ. Community structure in social and biological networks. *Proc. of the National Academy of Sciences of the United States of America*, 2002, 99(12): 7821–7826. [doi: 10.1073/pnas.122653799]
- [6] Wang HR, Ding DZ, Zhang M. Gradient-based overlapping hierarchical community detection. *Computer Systems & Applications*, 2021, 30(8): 207–212 (in Chinese with English abstract). [doi: 10.15888/j.cnki.csa.008016]
- [7] 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]
- [8] Airoldi EM, Blei DM, Fienberg SE, Xing EP. Mixed membership stochastic blockmodels. In: Proc. of the 21st Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2008. 33–40.
- [9] Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. *Physical Review E*, 2011, 83(1): 016107. [doi: 10.1103/PhysRevE.83.016107]
- [10] Lu XY, Szymanski BK. A regularized stochastic block model for the robust community detection in complex networks. *Scientific Reports*, 2019, 9(1): 13247. [doi: 10.1038/s41598-019-49580-5]
- [11] Shen HW, Cheng XQ, Guo JF. Exploring the structural regularities in networks. *Physical Review E*, 2011, 84(5): 056111. [doi: 10.1103/PhysRevE.84.056111]
- [12] Jiang JQ. Stochastic block model and exploratory analysis in signed networks. *Physical Review E*, 2015, 91(6): 062805. [doi: 10.1103/PhysRevE.91.062805]
- [13] Xing EP, Fu WJ, Song L. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 2010, 4(2): 535–566. [doi: 10.1214/09-AOAS311]
- [14] Aicher C, Jacobs AZ, Clauset A. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 2015, 3(2): 221–248. [doi: 10.1093/comnet/cnu026]
- [15] Li Y, Chen HC, Yang B. Reparameterized stochastic block model adaptive to heterogeneous degree and block distributions. *IEEE Access*, 2018, 6: 37615–37626. [doi: 10.1109/ACCESS.2018.2853115]
- [16] Qiao MY, Yu J, Bian W, Li Q, Tao DC. Adapting stochastic block models to power-law degree distributions. *IEEE Trans. on*

- Cybernetics, 2019, 49(2): 626–637. [doi: [10.1109/TCYB.2017.2783325](https://doi.org/10.1109/TCYB.2017.2783325)]
- [17] Wu XX, Jiao PF, Wang YP, Li TP, Wang WJ, Wang B. Dynamic stochastic block model with scale-free characteristic for temporal complex networks. In: Proc. of the 24th Int'l Conf. on Database Systems for Advanced Applications. Chiang Mai: Springer, 2019. 502–518. [doi: [10.1007/978-3-030-18579-4_30](https://doi.org/10.1007/978-3-030-18579-4_30)]
 - [18] Xu ZJ, Liu XY, Cui XJ, Li XM, Yang B. Robust stochastic block model. Neurocomputing, 2020, 379: 398–412. [doi: [10.1016/j.neucom.2019.10.069](https://doi.org/10.1016/j.neucom.2019.10.069)]
 - [19] Liu XY, Yang B, Song WZ, Musial K, Zuo WL, Chen HX, Yin HZ. A block-based generative model for attributed network embedding. World Wide Web, 2021, 24(5): 1439–1464. [doi: [10.1007/s11280-021-00918-y](https://doi.org/10.1007/s11280-021-00918-y)]
 - [20] Chang ZH, Jia CY, Yin XJ, Zheng YM. A generative model for exploring structure regularities in attributed networks. Information Sciences, 2019, 505: 252–264. [doi: [10.1016/j.ins.2019.07.084](https://doi.org/10.1016/j.ins.2019.07.084)]
 - [21] Zheng YM, Jia CY, Chang ZH, Li XY. A degree corrected stochastic block model for attributed networks. Journal of Computer Research and Development, 2020, 57(8): 1650–1662 (in Chinese with English abstract). [doi: [10.7544/issn1000-1239.2020.20200158](https://doi.org/10.7544/issn1000-1239.2020.20200158)]
 - [22] Liu W, Chang ZH, Jia CY, Zheng YM. A generative node-attribute network model for detecting generalized structure and semantics. Physica A: Statistical Mechanics and its Applications, 2022, 588: 126557. [doi: [10.1016/j.physa.2021.126557](https://doi.org/10.1016/j.physa.2021.126557)]
 - [23] Newman MEJ, Leicht EA. Mixture models and exploratory analysis in networks. Proc. of the National Academy of Sciences of the United States of America, 2007, 104(23): 9564–9569. [doi: [10.1073/pnas.0610537104](https://doi.org/10.1073/pnas.0610537104)]
 - [24] Zhu JJ, Liu YG, Wu H, Chen Z, Zhang Y, Yang SM, Yang CH, Yang W, Wu XD. A no self-edge stochastic block model and a heuristic algorithm for balanced anti-community detection in networks. Information Sciences, 2020, 518: 95–112. [doi: [10.1016/j.ins.2020.01.005](https://doi.org/10.1016/j.ins.2020.01.005)]
 - [25] Wu XX, Wang CD, Jiao PF. Hybrid-order stochastic block model. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 4470–4477. [doi: [10.1609/aaai.v35i5.16574](https://doi.org/10.1609/aaai.v35i5.16574)]
 - [26] Abossedgh S, Saghaei A, Amiri A. Monitoring a labeled degree-corrected stochastic block model. Quality and Reliability Engineering International, 2023, 39(1): 99–112. [doi: [10.1002/qre.3221](https://doi.org/10.1002/qre.3221)]
 - [27] 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. Paris: ACM, 2009. 927–936. [doi: [10.1145/1557019.1557120](https://doi.org/10.1145/1557019.1557120)]
 - [28] 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(1): 012807. [doi: [10.1103/PhysRevE.88.012807](https://doi.org/10.1103/PhysRevE.88.012807)]
 - [29] Yang TB, Chi Y, Zhu SH, Gong YH, Jin R. Directed network community detection: A popularity and productivity link model. In: Proc. of the 10th SIAM Int'l Conf. on Data Mining. Columbus: Society for Industrial and Applied Mathematics, 2010. 742–753. [doi: [10.1137/1.9781611972801.6](https://doi.org/10.1137/1.9781611972801.6)]
 - [30] Chen HR, Yu ZJ, Yang QL, Shao JM. Attributed graph clustering with subspace stochastic block model. Information Sciences, 2020, 535: 130–141. [doi: [10.1016/j.ins.2020.05.044](https://doi.org/10.1016/j.ins.2020.05.044)]
 - [31] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1–22. [doi: [10.1111/j.2517-6161.1977.tb01600.x](https://doi.org/10.1111/j.2517-6161.1977.tb01600.x)]
 - [32] Xuan GR, Shi YQ, Chai PQ, Sutthiwan P. An enhanced EM algorithm using maximum entropy distribution as initial condition. In: Proc. of the 21st Int'l Conf. on Pattern Recognition (ICPR 2012). Tsukuba: IEEE, 2012: 849–852.
 - [33] Sen P, Namata G, Bilgic M, Getoor L, Gallagher B, Eliassi-Rad T. Collective classification in network data. AI Magazine, 2008, 29(3): 93–106. [doi: [10.1609/aimag.v29i3.2157](https://doi.org/10.1609/aimag.v29i3.2157)]
 - [34] Rostami M, Oussalah M. A novel attributed community detection by integration of feature weighting and node centrality. Online Social Networks and Media, 2022, 30: 100219. [doi: [10.1016/j.osnem.2022.100219](https://doi.org/10.1016/j.osnem.2022.100219)]
 - [35] Giles CL, Bollacker KD, Lawrence S. Citeseer: An automatic citation indexing system. In: Proc. of the 3rd ACM Conf. on Digital Libraries. Pittsburgh: ACM, 1998. 89–98. [doi: [10.1145/276675.276685](https://doi.org/10.1145/276675.276685)]
 - [36] Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(9): P09008. [doi: [10.1088/1742-5468/2005/09/P09008](https://doi.org/10.1088/1742-5468/2005/09/P09008)]
 - [37] Shang RH, Wang S, Zhang WT, Feng J, Jiao LC, Stolkin R. Evolutionary multi-objective overlapping community detection based on fusion of internal and external connectivity and correction of node intimacy. Applied Soft Computing, 2024, 154: 111414. [doi: [10.1016/j.asoc.2024.111414](https://doi.org/10.1016/j.asoc.2024.111414)]
 - [38] Ma HP, Liu ZJ, Zhang XY, Zhang L, Jiang H. Balancing topology structure and node attribute in evolutionary multi-objective community detection for attributed networks. Knowledge-based Systems, 2021, 227: 107169. [doi: [10.1016/j.knosys.2021.107169](https://doi.org/10.1016/j.knosys.2021.107169)]
 - [39] Zhou Y, Cheng H, Yu JX. Graph clustering based on structural/attribute similarities. Proc. of the VLDB Endowment, 2009, 2(1): 718–729. [doi: [10.14778/1687627.1687709](https://doi.org/10.14778/1687627.1687709)]

- [40] Xu ZQ, Ke YP, Wang Y, Cheng H, Cheng J. A model-based approach to attributed graph clustering. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. Scottsdale: ACM, 2012. 505–516. [doi: [10.1145/2213836.2213894](https://doi.org/10.1145/2213836.2213894)]
- [41] Yang S, Yang B. Enhanced network embedding with text information. In: Proc. of the 24th Int'l Conf. on Pattern Recognition (ICPR). Beijing: IEEE, 2018. 326–331. [doi: [10.1109/ICPR.2018.8545577](https://doi.org/10.1109/ICPR.2018.8545577)]
- [42] Berahmand K, Haghani S, Rostami M, Li YF. A new attributed graph clustering by using label propagation in complex networks. Journal of King Saud University-computer and Information Sciences, 2022, 34(5): 1869–1883. [doi: [10.1016/j.jksuci.2020.08.013](https://doi.org/10.1016/j.jksuci.2020.08.013)]
- [43] Shang RH, Zhang WT, Li ZY, Wang C, Jiao LC. Attribute community detection based on latent representation learning and graph regularized non-negative matrix factorization. Applied Soft Computing, 2023, 133: 109932. [doi: [10.1016/j.asoc.2022.109932](https://doi.org/10.1016/j.asoc.2022.109932)]

附中文参考文献:

- [6] 王寒蕊, 丁岱宗, 张謐. 基于梯度的重叠式层次社区检测. 计算机系统应用, 2021, 30(8): 207–212. [doi: [10.15888/j.cnki.csa.008016](https://doi.org/10.15888/j.cnki.csa.008016)]
- [21] 郑忆美, 贾彩燕, 常振海, 李轩涯. 一种度修正的属性网络随机块模型. 计算机研究与发展, 2020, 57(8): 1650–1662. [doi: [10.7544/issn1000-1239.2020.20200158](https://doi.org/10.7544/issn1000-1239.2020.20200158)]



王笑(1994—), 女, 硕士生, 主要研究领域为图像处理, 计算机视觉, 复杂网络.



郭文艳(1973—), 女, 博士, 副教授, 主要研究领域为智能计算, 网络优化.



戴芳(1966—), 女, 博士, 教授, 主要研究领域为图像处理, 计算机视觉, 复杂网络.



王军锋(1969—), 男, 博士, 副教授, 主要研究领域为神经网络, 图像处理.