物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于平均场近似的BP算法求解随机块模型

马闯　杨晓龙　陈含爽　张海峰

A mean-field approximation based BP algorithm for solving the stochastic block model

Ma Chuang　Yang Xiao-Long　Chen Han-Shuang　Zhang Hai-Feng

引用信息 Citation: Acta Physica Sinica, 70, 228901 (2021)　DOI: 10.7498/aps.70.20210511
在线阅读 View online: https://doi.org/10.7498/aps.70.20210511
当期内容 View table of contents: http://wulixb.iphy.ac.cn

基于平均场近似的 BP 算法求解随机块模型*

马闯[1)] 杨晓龙[1)] 陈含爽[2)] 张海峰[3)†]

1) (安徽大学互联网学院, 合肥 230039)

2) (安徽大学物理与材料科学学院, 合肥 230601)

3) (安徽大学数学科学学院, 合肥 230601)

(2021 年 3 月 16 日收到; 2021 年 7 月 29 日收到修改稿)

置信传播 (BP) 算法作为推断概率图模型的主流算法是求解随机块模型中联合概率分布的重要方法之一. 但现有的方法要么在处理核边结构问题上存在精度不足问题, 要么在理论的推导上存在近似太多, 导致求解过程复杂且难以理解问题, 或两个问题均存在. 当然, 精度不足也是由近似多造成的. 导致理论近似多且推导复杂的主要原因, 是随机块模型推断过程中求解联合概率分布并不是直接套用 BP 算法, 即处理的图 (网络) 与概率图模型的图不统一. 因此, 本文利用平均场近似修正联合概率分布, 使其完全匹配 BP 算法的迭代公式, 这样使得在理论推导上简单易懂. 最后通过实验验证, 该方法是有效的.

关键词: 随机块模型, 置信传播算法, 联合概率分布, 平均场近似

PACS: 89.75.Hc, 89.75.Fb, 05.10.–a, 05.10.Ln **DOI:** 10.7498/aps.70.20210511

1 引 言

复杂系统可以用复杂网络进行建模, 而复杂网络主要由点与边组成. 在现实中, 点与边的信息是可以直接或间接[1]获取的, 如在航空网络系统中[2], 节点为机场, 如果机场之间具有航班, 就存在一条边; 科学家合作网络[3], 节点表示科学家, 如果两个科学家有合作就存在一条边等. 这些简单的点与连边关系可以帮助我们揭示复杂系统的隐藏信息, 如节点的属性, 可以对应着节点在系统中属于哪个功能块或组织[4]. 如航空网络中的核心与外围 (边缘) 机场[5], 科学家合作网络中的特定研究领域的小组[6]等. 因此, 如何通过节点的连边信息去探测这些中尺度结构, 如社团结构、核边结构 (核心-边缘), 是一个重要的科学问题, 这些问题得到了来自各个领域学者的研究.

社团结构的描述为社团内部紧密相连, 而社团之间连接稀疏[6]. 社团结构的探测方法已经有很多, 主要有基于模块度的方法[7–9]、谱划分的方法[4,10]、矩阵分解的方法[11,12]、基于动力学的方法[13,14]、随机块模型的方法[15,16]以及其他方法[17,18]; 核边结构的描述为核与核紧密相连, 核与边紧密相连, 边与边连接稀疏[19,20]. 相较于社团结构, 对核边结构探测方法的研究相对较少. 主要有给定某种中心性指标, 然后通过截断选取最重要的一部分节点作为核心节点[21,22]; 定义核边结构的指标, 然后利用最优化方法求解[23,24]; 通过随机块模型探测核边结构[25,26]以及其他方法[5,27–29].

可以看出, 随机块模型在解决社团结构与核边结构的探测问题上都是行之有效的[25]. 随机块模型是一种生成模型, 可以生成具有模块结构的网

* 国家自然科学基金 (批准号: 12005001, 61973001, 11875069),安徽省高校协同创新项目 (批准号: GXXT-2021-032) 和安徽省自然科学基金 (批准号: 2008085QF299) 资助的课题.

† 通信作者. E-mail: haifengzhang1978@gmail.com

络[15]. 即首先给定 n 个节点, 随机分配到 k 个模块中. 假设向量 $\boldsymbol{\gamma}$ 表示每个模块的规模, 即 γ_r 表示第 r 个模块的大小占总节点个数的比例, 则每个节点就以 γ_r 的概率分配到第 r 模块中. 然后给定一个 $k\times k$ 的关联矩阵 $\boldsymbol{p}$ (对称矩阵), 任意两个节点 (一个节点属于模块 s, 一个节点属于模块 r) 以 p_{rs} 的概率相连. 这样就可以得到一个以参数 $\boldsymbol{\gamma}$ 与 $\boldsymbol{p}$ 生成的随机网络. 如, 当 $k=2$ 时, 如果 $p_{11}>p_{12}$, $p_{22}>p_{12}$, 生成的网络具有社团结构; 如果 $p_{11}\geqslant p_{12}\geqslant p_{22}$, 生成的网络具有核边结构.

给定特定参数 $\boldsymbol{\gamma}$ 与 $\boldsymbol{p}$, 可以生成具有社团结构或核边结构的网络. 也就是说, 如果一个网络具有社团或核边结构, 那么可以通过学习该模型 (即模型参数), 来确定它是怎么生成的, 以此来探测哪些节点属于哪个社团, 或者哪些节点属于核, 哪些节点属于边. 因此, 随机块模型不仅可以用来探测社团结构, 还可以用来探测核心边缘结构. 其中 EM 算法是学习该模型的重要方法[16,25,26,30,31], 而通过联合概率公式, 确定边际概率又是 EM 算法的关键部分与难点. 基于马尔科夫链的蒙特卡罗方法 (MCMC) 是解决该难点的一种较好的方法[30], 但是由于采样空间非常大, 会造成时间复杂度过高、精度低的问题. 置信传播 (BP) 算法[32] 可以很好地解决上述问题[16,25,26,31], 因此被广泛用来解决社团或核边结构的探测问题. BP 算法用来推断马尔科夫随机场中的边际概率, 当马尔可夫网络是树状图时, 具有精确解; 当马尔可夫网络是稀疏的, 会得到近似精确解[33].

应用 BP 算法解决社团结构的探测问题时, 由于网络的稀疏性, 在公式的推导中关于节点邻居可以忽略不计的近似是合理的[16,31], 但是处理核边结构探测问题时, 这种近似却是致命的, 因为核节点具有天然的大度性质. 因此, 在另外一篇文章中, 我们对 BP 算法进行了稍微的改进 (IBP), 就可以很好地解决此问题[25]. 但是伴随而来的就是 IBP 算法的推导过程中, 需要进行大量的近似处理, 以及各种模糊的推导, 并不是可以套用 BP 算法公式[34]直接得到的. 同样的道理, BP 算法解决社团结构的探测问题时, 即使其中各种近似是合理的, 也会存在上述不能套用 BP 算法公式直接得到的情况. 这使得即使 BP 或 IBP 算法在社团划分或核边结构探测中能得到很好的结果, 但是推导过程却是复杂且难于理解, 或者说是不那么精确的. 因此, 本文从平均场近似[35] 的角度出发, 先应用平均场近似对联合概率进行处理, 得到可以完全匹配 BP 算法的形式, 然后再套用 BP 算法公式 (而不是以往的先套用公式, 然后在公式里近似处理), 在不需要任何修改、近似以及假设的情况下直接得到求解边际概率的迭代公式. 使得随机块模型解决社团结构或核边结构探测问题时, 既可以保证结果的精度, 又可以保证理论的简单明了、有理有据.

2 知识回顾

随机块模型最直接的用处就是按照某种需求生成具有某种模块结构的人工网络. 例如在社团探测中, 就可以用随机块模型生成具有社团结构的人工网络, 以此来评估社团挖掘算法的好坏. 但是本文将用到它的另外一个重要应用, 就是通过学习随机块模型中的参数使其最好地匹配给定网络, 从而揭示给定网络的结构. 显然, 这是一个参数估计问题, 可以采用最大似然估计的方法进行解决[16,25,26].

2.1 最大似然估计

最大似然估计研究的是给定什么样的参数可以使得似然函数的值最大, 也就是使得已经发生的事件概率最大. 假设给定一个无权无向网络的邻接矩阵 $\boldsymbol{A}$, 其中 $A_{ij}=1$ 表示第 i 个节点与第 j 个节点相连, $A_{ij}=0$ 表示不相连. 在随机块模型中, 当给定参数 $\boldsymbol{p}$ 与 $\boldsymbol{\gamma}$, 则似然函数就可以用 $P(\boldsymbol{A}|\boldsymbol{p},\boldsymbol{\gamma})$ 来表示. 因此, 需要最大化 $P(\boldsymbol{A}|\boldsymbol{p},\boldsymbol{\gamma})$, 以此求解 $\boldsymbol{p}$ 与 $\boldsymbol{\gamma}$, 从而来揭示网络的结构. 但是, 通过观察可以发现, 似然函数里包含一个隐形变量 $\boldsymbol{g}$, 其中 g_i 表示节点 i 所属的模块. 因此有

$$\begin{aligned} P(\boldsymbol{A}|\boldsymbol{p},\boldsymbol{\gamma}) &= \sum_{\boldsymbol{g}} P(\boldsymbol{A},\boldsymbol{g}|\boldsymbol{p},\boldsymbol{\gamma}) \\ &= \sum_{\boldsymbol{g}} P(\boldsymbol{A}|\boldsymbol{g},\boldsymbol{p},\boldsymbol{\gamma})P(\boldsymbol{g}|\boldsymbol{\gamma}) \\ &= \sum_{\boldsymbol{g}} \left[\prod_{i<j} p_{g_ig_j}^{A_{ij}}(1-p_{g_ig_j})^{1-A_{ij}} \prod_i \gamma_{g_i}\right]. \end{aligned} \tag{1}$$

可以看出, 最大化公式 (1) 是一个含有隐变量的最大似然估计问题, 因此可以应用 EM 算法进行求解[36].

2.2 EM 算法求解随机块模型

EM 算法在理论推导上是根据詹森不等式不断求解对数似然函数的下界函数最优值, 以此逐渐

收敛到原始函数的最优值. 因此, 对 (1) 式取对数, 然后根据詹森不等式有[25,26]:

$$\begin{aligned}&\log\sum_{\boldsymbol{g}}P\left(\boldsymbol{A},\boldsymbol{g}|\boldsymbol{p},\boldsymbol{\gamma}\right)\\&\geqslant\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\log\frac{P\left(\boldsymbol{A},\boldsymbol{g}|\boldsymbol{p},\boldsymbol{\gamma}\right)}{q\left(\boldsymbol{g}\right)}\triangleq L\left(\boldsymbol{p},\boldsymbol{\gamma}\right),\end{aligned}\tag{2}$$

其中

$$\begin{aligned}q\left(\boldsymbol{g}\right)&=\frac{P\left(\boldsymbol{A},\boldsymbol{g}|\boldsymbol{p},\boldsymbol{\gamma}\right)}{\sum\limits_{\boldsymbol{g}}P\left(\boldsymbol{A},\boldsymbol{g}|\boldsymbol{p},\boldsymbol{\gamma}\right)}\\&=\frac{\prod\limits_{i<j}p_{g_ig_j}^{A_{ij}}\left(1-p_{g_ig_j}\right)^{1-A_{ij}}\prod\limits_{i}\gamma_{g_i}}{\sum\limits_{\boldsymbol{g}}\prod\limits_{i<j}p_{g_ig_j}^{A_{ij}}\left(1-p_{g_ig_j}\right)^{1-A_{ij}}\prod\limits_{i}\gamma_{g_i}}.\end{aligned}\tag{3}$$

可以看出, $q(\boldsymbol{g})=P(\boldsymbol{g}|\boldsymbol{A},\boldsymbol{p},\boldsymbol{\gamma})$, 表示给定一个网络, 在参数 $\boldsymbol{p}$ 与 $\boldsymbol{\gamma}$ 给定的条件下, 该网络每个节点分配为 $\boldsymbol{g}$ 的概率, 这是一个联合概率.

当 (3) 式成立的情况下, (2) 式等号成立, 所以最大化公式 (1) 问题可以转化为最大化 $L\left(\boldsymbol{p},\boldsymbol{\gamma}\right)$ 问题. 对 $L\left(\boldsymbol{p},\boldsymbol{\gamma}\right)$ 进一步推导有:

$$\begin{aligned}&L\left(\boldsymbol{p},\boldsymbol{\gamma}\right)=\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\log\frac{P\left(\boldsymbol{A},\boldsymbol{g}|\boldsymbol{p},\boldsymbol{\gamma}\right)}{q\left(\boldsymbol{g}\right)}\\&=\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\log\left[\prod_{i<j}p_{g_ig_j}^{A_{ij}}\left(1-p_{g_ig_j}\right)^{1-A_{ij}}\prod_{i}\gamma_{g_i}\right]\\&\quad-\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\log q\left(\boldsymbol{g}\right)\\&=\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\left\{\sum_{i<j}\left[A_{ij}\log p_{g_ig_j}+\left(1-A_{ij}\right)\right.\right.\\&\quad\left.\left.\times\log\left(1-p_{g_ig_j}\right)\right]+\sum_{i}\log\gamma_{g_i}\right\}\\&\quad-\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\log q\left(\boldsymbol{g}\right)\\&=\frac{1}{2}\sum_{i\neq j}\sum_{\boldsymbol{rs}}\left[A_{ij}q_{\boldsymbol{rs}}^{ij}\log p_{\boldsymbol{rs}}+\left(1-A_{ij}\right)\right.\\&\quad\left.\times q_{\boldsymbol{rs}}^{ij}\log\left(1-p_{\boldsymbol{rs}}\right)\right]\\&\quad+\sum_{i\boldsymbol{r}}q_{\boldsymbol{r}}^{i}\log\gamma_{\boldsymbol{r}}-\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\log q\left(\boldsymbol{g}\right),\end{aligned}\tag{4}$$

其中 $q(\boldsymbol{g})$ 的边际概率 $q_{\boldsymbol{r}}^{i}$ 表示节点 i 属于群组 $\boldsymbol{r}$ 的概率:

$$q_{\boldsymbol{r}}^{i}=\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\delta_{g_i,\boldsymbol{r}},\tag{5}$$

$q(\boldsymbol{g})$ 的边际概率 $q_{\boldsymbol{rs}}^{ij}$ 表示节点 i 属于群组 $\boldsymbol{r}$ 的概率且节点 j 属于群组 $\boldsymbol{s}$ 的概率:

$$q_{\boldsymbol{rs}}^{ij}=\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\delta_{g_i,\boldsymbol{r}}\delta_{g_j,\boldsymbol{s}}.\tag{6}$$

如果 $g_i=\boldsymbol{r}$, 则 $\delta_{g_i,\boldsymbol{r}}=1$, 否则 $\delta_{g_i,\boldsymbol{r}}=0$.

对 (4) 式求解最大值 (满足约束条件 $\sum_{\boldsymbol{r}}\gamma_{\boldsymbol{r}}=1$), 应用拉格朗日乘子法可以得到:

$$p_{\boldsymbol{rs}}=\frac{\sum\limits_{i\neq j}A_{ij}q_{\boldsymbol{rs}}^{ij}}{\sum\limits_{i\neq j}q_{\boldsymbol{rs}}^{ij}},\tag{7}$$

$$\gamma_{\boldsymbol{r}}=\frac{1}{n}\sum_{i}q_{\boldsymbol{r}}^{i}.\tag{8}$$

根据上述推导过程, 可以得到一个迭代公式. 即首先初始化 $\boldsymbol{p}$ 与 $\boldsymbol{\gamma}$, 根据 (3) 式计算边际概率 $q_{\boldsymbol{r}}^{i}$ 与 $q_{\boldsymbol{rs}}^{ij}$, 然后根据 (7) 式与 (8) 式计算新的 $\boldsymbol{p}$ 与 $\boldsymbol{\gamma}$, 重复上述过程直到收敛为止. 最后可以通过收敛的边际概率 $q_{\boldsymbol{r}}^{i}$ 的值对每个节点进行划分. 其中, 当需要探测社团结构时, 只需初始化 $\boldsymbol{p}$ 为具有社团结构形式的关联矩阵, 如果需要探测核边结构时, 则需要初始化 $\boldsymbol{p}$ 为具有核边结构形式的关联矩阵.

需要注意的是, 在应用 (7) 式求解 $p_{\boldsymbol{rs}}$ 时, 需要计算任意节点对之间的 $q_{\boldsymbol{rs}}^{ij}$ 值, 这将花费大量的计算时间. 因此可以化简为

$$\begin{aligned}\sum_{ij}q_{\boldsymbol{rs}}^{ij}&=\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\sum_{i}\delta_{g_i,\boldsymbol{r}}\sum_{j}\delta_{g_j,\boldsymbol{s}}\\&=\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)n_{\boldsymbol{r}}n_{\boldsymbol{s}}=\left\langle n_{\boldsymbol{r}}n_{\boldsymbol{s}}\right\rangle\approx\left\langle n_{\boldsymbol{r}}\right\rangle\left\langle n_{\boldsymbol{s}}\right\rangle,\end{aligned}\tag{9}$$

其中 $n_{\boldsymbol{r}}=\sum_{i}\delta_{g_i,\boldsymbol{r}}$, $\langle\cdots\rangle$ 表示期望. 于是有

$$\begin{aligned}\left\langle n_{\boldsymbol{r}}\right\rangle&=\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\sum_{i}\delta_{g_i,\boldsymbol{r}}\\&=\sum_{i}\sum_{\boldsymbol{g}}q\left(\boldsymbol{g}\right)\delta_{g_i,\boldsymbol{r}}=\sum_{i}q_{r}^{i}.\end{aligned}\tag{10}$$

所以 (7) 式可以写成:

$$p_{\boldsymbol{rs}}=\frac{\sum\limits_{i\neq j}A_{ij}q_{\boldsymbol{rs}}^{ij}}{\sum\limits_{i}q_{\boldsymbol{r}}^{i}\sum\limits_{j}q_{\boldsymbol{s}}^{j}}.\tag{11}$$

应用 (11) 式代替 (7) 式只需要计算有边的节点对的两节点边际概率, 这将很大程度上减少计算量, 特别是在稀疏网络. 而且在 BP 算法中这种处理更显得尤为必要, 后面将进一步介绍.

到这里并没结束, 通过遍历所有构型 (k^n 种构型) 的方法根据 (3) 式求解全概率 $q(\boldsymbol{g})$ 以此求解边际概率 $q_{\boldsymbol{r}}^{i}$ 与 $q_{\boldsymbol{rs}}^{ij}$ 是显然不可行的. 因此如何根据

(3) 式求解边际概率 $q_{\boldsymbol{r}}^{i}$ 与 $q_{\boldsymbol{rs}}^{ij}$ 是 EM 算法的难点, 也是关键部分. 下面将介绍两种求解方法: MC 采样和 BP 算法, 并分析他们在处理社团与核边结构探测中存在的不足之处.

2.3 马尔科夫链蒙特卡罗方法 (MCMC)

这里应用 MCMC 模拟的方法按照联合概率公式 (3) 对网络的划分构型 (即 $g=[g_1,g_2,\cdots,g_n]^{\mathrm{T}}$) 进行采样. 本文将采用单分量 Metropolis-Hastings (M-H) 方法[37]. 设 $g^t=[g_1^t,\ g_2^t,\ \cdots,\ g_n^t]^{\mathrm{T}}$ 表示在马尔科夫链中第 t 轮迭代后的状态, 其中每一轮迭代都需要对每个分量的状态进行一次迭代更新.

例如, 在 $t+1$ 轮迭代中, 已经迭代好的分量 (第 i 个分量) 的状态记为 g_i^{t+1}. 假设要对第 j 个分量进行采样, 首先设 g_{-j}^{t+1} 表示在 $t+1$ 轮迭代中在对第 j 个分量进行处理时, 其他分量的状态. 特别地, 当对每一轮的各个分量进行迭代是按照顺序处理时, 有:

$$g_{-j}^{t+1}=[g_1^{t+1},\ g_2^{t+1},\ \cdots,\ g_{j-1}^{t+1},\ g_{j+1}^{t},\ \cdots,\ g_n^{t}]^{\mathrm{T}}. \tag{12}$$

把第 j 个分量的状态 g_j^t 随机变为一种状态 $\hat{g}_j^{t+1}$, 因为有 k 个模块, 所以建议转移概率 $q(g_j^t,\hat{g}_j^{t+1})=q(\hat{g}_j^{t+1},g_j^t)=1/k$. 根据 (3) 式, 按照 M-H 规则接收这种状态 (从状态 g_j^t 变为状态 $\hat{g}_j^{t+1}$) 的概率:

$$\alpha(g_j^t,\hat{g}_j^{t+1})=\min\left\{1,\frac{\left[\prod\limits_{i\neq j}p_{g_ig_j^t}^{A_{ij}}\left(1-p_{g_ig_j^t}\right)^{1-A_{ij}}\right]\gamma_{g_j^t}}{\left[\prod\limits_{i\neq j}p_{g_i\hat{g}_j^{t+1}}^{A_{ij}}\left(1-p_{g_i\hat{g}_j^{t+1}}\right)^{1-A_{ij}}\right]\gamma_{\hat{g}_j^{t+1}}}\right\}, \tag{13}$$

式中 $g_i(i\neq j)$ 为 g_{-j}^{t+1} 中分别对应的量. 在区间 $(0,1)$ 内按均匀分布取出一个数记为 μ, 如果 $\mu\leqslant\alpha(g_j^t,\hat{g}_j^{t+1})$, 则 $g_j^{t+1}=\hat{g}_j^{t+1}$, 否则 $g_j^{t+1}=g_j^t$. 当每一轮所有结点的状态更新完以后 (也可以不按顺序更新), 再进行下一轮的更新. 在更新的轮数足够大的情况下, 采样的分布收敛到 (3) 式的概率分布, 因此可以通过统计估计出边际概率 $q_{\boldsymbol{r}}^{i}$ 与 $q_{\boldsymbol{rs}}^{ij}$.

但是, 由于所有构型共用 k^n 种可能, 在网络较大的情况下, 采样空间巨大, 很难收敛到 (3) 式的概率分布, 从而造成计算结果产生误差, 并且不稳定. 下面介绍一种快速求解边际概率的方法, 即 BP 算法.

2.4 BP 算法

2.4.1 BP 算法简介

BP 算法可以用来推断马尔科夫随机场 (又称为概率图模型) 中的边际概率[33]. 在本文中, 考虑一种特殊的概率图模型——成对图模型. 成对图模型定义在一个无向图 $G(\boldsymbol{V},\boldsymbol{E})$ 上, 每一个顶点 $v_i\in\boldsymbol{V}$ 表示 $\boldsymbol{X}$ 中的一个随机变量 x_i, 每一条边 $(v_i,v_j)\in\boldsymbol{E}$ 表示随机变量 x_i 与 x_j 之间的某种依赖关系. 给定每个节点 v_i 一个点位势 $\phi_i(x_i)$, 以及每一条边 $(v_i,v_j)\in\boldsymbol{E}$ 一个边位势 $\psi_{ij}(x_i,x_j)$, 则在该成对马尔科夫网上的分布为

$$P(\boldsymbol{X})=\frac{1}{Z}\prod_{(i,j)\in\boldsymbol{E}}\psi_{ij}(x_i,x_j)\prod_{i\in\boldsymbol{V}}\phi_i(x_i), \tag{14}$$

其中 Z 为归一化常数, 通常称为配分函数.

求解 (14) 式这个联合概率的边际概率可以通过 BP 算法迭代公式求得, 迭代过程如下[34]:

$$v_{i\to j}(x_i)\propto\phi_i(x_i)\prod_{k\in N(i)/j}\sum_{x_k}\psi_{ik}(x_i,x_k)\,v_{k\to i}(x_k), \tag{15}$$

$$b_i(x_i)\propto\phi_i(x_i)\prod_{k\in N(i)}\sum_{x_k}\psi_{ik}(x_i,x_k)\,v_{k\to i}(x_k), \tag{16}$$

其中, $N(i)$ 表示节点 v_i 的邻居, $v_{i\to j}(x_i)$ 为概率图修改后 (即删除节点 v_j) x_i 的边际概率, $b_i(x_i)$ 为 x_i 的边际概率. 另外还可以得到一条边上两个节点的联合概率:

$$b_{ij}(x_i,x_j)\propto\psi_{ij}(x_i,x_j)\,v_{i\to j}(x_i)\,v_{j\to i}(x_j). \tag{17}$$

2.4.2 BP 算法求解随机块模型

对于联合概率公式 (3), 可以看成是一个马尔科夫网上的概率分布. 在这里, 马尔科夫网是一个全连通图, 且随机变量为 $g=[g_1,\ g_2,\ \cdots,\ g_n]^{\mathrm{T}}$. 对于一对节点 (v_i,v_j), 如果 $A_{ij}=1$, 其边位势为 $\psi_{ij}(g_i,g_j)=P_{g_ig_j}$, 如果 $A_{ij}=0$, 其边位势为 $\psi_{ij}(g_i,g_j)=1-P_{g_ig_j}$, 节点 v_i 的点位势为 $\phi_i(g_i)=\gamma_{g_i}$. 所以根据 BP 算法的迭代公式 (15), 当 j 节点是 i 节点的邻居时, 可以得到[26,31]:

$$\eta_{\boldsymbol{r}}^{i\to j}=\frac{\gamma_{\boldsymbol{r}}}{Z_{i\to j}}\prod_{k\in\boldsymbol{V}/N^*(i)}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}(1-p_{\boldsymbol{rs}})\prod_{k\in\boldsymbol{N}(i)/j}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}p_{\boldsymbol{rs}}, \tag{18}$$

其中“消息” $\eta_{\boldsymbol{r}}^{i\to j}$ 对应于 (15) 式中的 $v_{i\to j}(x_i)$, 在这里可以理解为移除节点 v_j, v_i 属于模块 $\boldsymbol{r}$ 的概率. $Z_{i\to j}$ 表示配分函数, $N(i)$ 表示节点 v_i 的邻居, $N^*(i)=N(i)\bigcup\{i\}$. 为了更方便更快捷地计算, 需要做两个假设, 会得到两个近似.

假设 1 如果第 j 个节点不是第 i 节点的邻居, 移除节点 v_j 对节点 v_i 属于模块 $\boldsymbol{r}$ 的概率没有影响, 即有 $\eta_{\boldsymbol{s}}^{k\to i}\approx q_{\boldsymbol{s}}^{k}$;

假设 2 当网络规模很大且很稀疏的时候, 对集合 $\boldsymbol{V}/N^*$ 中所有元素做运算近似于对集合 $\boldsymbol{V}$ 中所有元素做运算, 即有 $f(\boldsymbol{V}/N^*)\approx f(\boldsymbol{V})$, 其中 $f(*)$ 代表某种运算.

根据上述两个假设, 有:

$$\begin{aligned}&\prod_{k\in\boldsymbol{V}/N^*(i)}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\left(1-p_{\boldsymbol{rs}}\right)\\ \approx&\prod_{k\in\boldsymbol{V}/N^*(i)}\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}\left(1-p_{\boldsymbol{rs}}\right)\\ \approx&\prod_{k\in\boldsymbol{V}}\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}\left(1-p_{\boldsymbol{rs}}\right)\\ =&\prod_{k\in\boldsymbol{V}}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right),\end{aligned}\tag{19}$$

这是一个外场量. 所以 (18) 式可以写成:

$$\begin{aligned}\eta_{\boldsymbol{r}}^{i\to j}=&\frac{\gamma_{\boldsymbol{r}}}{Z_{i\to j}}\prod_{k}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\\ &\times\prod_{k\in N(i)/j}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}p_{\boldsymbol{rs}},\end{aligned}\tag{20}$$

其中配分函数为

$$\begin{aligned}Z_{i\to j}=&\sum_{\boldsymbol{r}}\gamma_{\boldsymbol{r}}\prod_{k}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\\ &\times\prod_{k\in N(i)/j}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}p_{\boldsymbol{rs}}.\end{aligned}\tag{21}$$

根据 (16) 式可以得到边际分布 $q_{\boldsymbol{r}}^{i}$:

$$q_{\boldsymbol{r}}^{i}=\frac{\gamma_{\boldsymbol{r}}}{Z_i}\prod_{k}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\prod_{k\in N(i)}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}p_{\boldsymbol{rs}},\tag{22}$$

其中配分函数为

$$Z_i=\sum_{\boldsymbol{r}}\gamma_{\boldsymbol{r}}\prod_{k}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\prod_{k\in N(i)}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}p_{\boldsymbol{rs}}.\tag{23}$$

根据 (17) 式可以得到一条边上两个点的联合概率:

$$q_{\boldsymbol{rs}}^{ij}=\frac{\eta_{\boldsymbol{r}}^{i\to j}\eta_{\boldsymbol{s}}^{j\to i}p_{\boldsymbol{rs}}}{\sum\limits_{\boldsymbol{rs}}\eta_{\boldsymbol{r}}^{i\to j}\eta_{\boldsymbol{s}}^{j\to i}p_{\boldsymbol{rs}}}.\tag{24}$$

通过不停迭代 (20) 式和 (22) 式, 直到收敛, 然后通过 (22) 式与 (24) 式即可计算出 $q_{\boldsymbol{r}}^{i}$ 与 $q_{\boldsymbol{rs}}^{ij}$.

上述算法在网络规模比较大且稀疏的时候是适用的, 但是应用此方法探测核边结构, 就会存在问题. 众所周知, 一个具有核边结构的网络的核节点不仅和核节点紧密相连, 而且和边节点紧密相连, 也就是说核节点与整个网络中的节点都是紧密相连的, 具有天然的大度性质. 所以在假设 2 中, 对集合 $\boldsymbol{V}/N^*$ 中所有元素做运算近似于对集合 $\boldsymbol{V}$ 中所有元素做运算, 这种假设在核边结构的探测中是不合理的, 有可能是错误的. 针对于此, 我们在另外一篇文章中对 BP 算法进行了修正 [25], 可以很好地解决这个问题.

2.4.3 修正后的 BP 算法 (IBP)

首先对联合概率分布公式 (3) 稍作处理, 可得:

$$\begin{aligned}q\left(\boldsymbol{g}\right)=&\frac{\prod\limits_{i<j}p_{g_ig_j}^{A_{ij}}\left(1-p_{g_ig_j}\right)^{1-A_{ij}}\prod\limits_{i}\gamma_{g_i}}{\sum\limits_{\boldsymbol{g}}\prod\limits_{i<j}p_{g_ig_j}^{A_{ij}}\left(1-p_{g_ig_j}\right)^{1-A_{ij}}\prod\limits_{i}\gamma_{g_i}}\\ =&\frac{\prod\limits_{i<j}\left(\frac{p_{g_ig_j}}{1-p_{g_ig_j}}\right)^{A_{ij}}\left(1-p_{g_ig_j}\right)\prod\limits_{i}\gamma_{g_i}}{\sum\limits_{\boldsymbol{g}}\prod\limits_{i<j}\left(\frac{p_{g_ig_j}}{1-p_{g_ig_j}}\right)^{A_{ij}}\left(1-p_{g_ig_j}\right)\prod\limits_{i}\gamma_{g_i}},\end{aligned}\tag{25}$$

则每条边的信使为 [25,38]

$$\begin{aligned}\eta_{\boldsymbol{r}}^{i\to j}=&\frac{\gamma_r}{Z_{i\to j}}\prod_{k\in\boldsymbol{V}/\{i,j\}}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\left(1-p_{\boldsymbol{rs}}\right)\\ &\times\prod_{k\in N(i)/j}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}.\end{aligned}\tag{26}$$

给定一个假设:

假设 3 当网络很大且很稀疏的时候, 对集合 $\boldsymbol{V}/\{i,j\}$ 中所有元素做运算近似于对集合 $\boldsymbol{V}$ 中所有元素做运算. 显然, 这个假设对结果的影响可以忽略不计.

根据假设 1 与假设 3 可以得到下面近似:

$$\begin{aligned}&\prod_{k\in\boldsymbol{V}/\{i,j\}}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\left(1-p_{\boldsymbol{rs}}\right)\approx\\ &\prod_{k\in\boldsymbol{V}/\{i,j\}}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\approx\prod_{k\in\boldsymbol{V}}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right),\end{aligned}\tag{27}$$

这是一个与 (19) 式一样的外场量, 但是这里的近似与 (19) 式的近似相比, 也就是假设 3 与假设 2 相比不会依赖节点度的大小, 所以更适用于具有核边结构的网络. (26) 式可以写成:

$$\eta_{\boldsymbol{r}}^{i\to j}=\frac{\gamma_{\boldsymbol{r}}}{Z_{i\to j}}\prod_{k}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\times\prod_{k\in N(i)/j}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{rs}},\tag{28}$$

其中配分函数为

$$Z_{i\to j}=\sum_{\boldsymbol{r}}\gamma_{\boldsymbol{r}}\prod_{k}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\times\prod_{k\in N(i)/j}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}},\tag{29}$$

则边际分布 $q_{\boldsymbol{r}}^{i}$ 为 [25,38]

$$q_{\boldsymbol{r}}^{i}=\frac{\gamma_{r}}{Z_{i}}\prod_{k}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\times\prod_{k\in N(i)}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{rs}},\tag{30}$$

其中

$$Z_{i}=\sum_{\boldsymbol{r}}\gamma_{\boldsymbol{r}}\prod_{k}\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^{k}p_{\boldsymbol{rs}}\right)\times\prod_{k\in N(i)}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}.\tag{31}$$

根据 (17) 式可以得到一条边上两个点的联合概率:

$$q_{\boldsymbol{rs}}^{ij}=\frac{\eta_{\boldsymbol{r}}^{i\to j}\eta_{\boldsymbol{s}}^{j\to i}\dfrac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}}{\displaystyle\sum_{\boldsymbol{rs}}\eta_{\boldsymbol{r}}^{i\to j}\eta_{\boldsymbol{s}}^{j\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}}.\tag{32}$$

通过 (28) 式, (30) 式, (32) 式与 (20) 式, (22) 式, (24) 相比, 可以发现算法改进后, 最终的结果仅仅是 $\dfrac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}$ 代替了 $p_{\boldsymbol{rs}}$, 没有增加任何多余的计算复杂度, 但是修改后的方法已经验证了更适用于含有大度节点的网络 [25], 如核边结构网络.

改进后的算法虽然在结果上表现出来强大的优势, 但是在理论推导上存在一定的问题. 根据 BP 算法的计算信使公式 (15), 从联合概率公式 (25) 并不能直接推导得到的信使迭代公式 (26)(这种直接写法是参考文献 [35]). 这是因为概率分布公式 (25) 相比 (3) 式虽然做了变形处理, 但是两个公式在数值上依然相等, 即在把一个全连通图当做一个马尔科夫网的情况下, 对于一对节点 (v_i,v_j), 如果 $A_{ij}=1$, 其边位势还是 $\psi_{ij}(g_i,g_j)=P_{g_ig_j}$, 如果 $A_{ij}=0$, 其边位势也还是 $\psi_{ij}(g_i,g_j)=1-P_{g_ig_j}$. 所以说这里得到的信使迭代公式应该还是 (18) 式而不是 (26) 式.

在这里如果想从概率分布公式 (25) 得到相比信使迭代公式 (18) 更好的 (26) 式, 换句话说, 如果想得到比 BP 算法更好的 IBP 算法, 可以按照下面的思路去理解. 这里的马尔科夫网不仅仅是一个全连通图, 还需要在全连通图上再加上 M (原网络上边的个数) 条额外边. 对于一对节点 (v_i,v_j), 如果 $A_{ij}=1$, 马尔科夫网上的边就要看成两条边, 其边位势就分解成 $\psi_{ij}(g_i,g_j)=1-P_{g_ig_j}$ 与 $\psi_{ij}(g_i,g_j)=\dfrac{P_{g_ig_j}}{1-P_{g_ig_j}}$, 如果 $A_{ij}=0$, 其边位势还是 $\psi_{ij}(g_i,g_j)=1-P_{g_ig_j}$. 这样理解以后, 就可以按照 BP 算法公式 (15) 得到信使迭代公式 (26). 即使这样, 根据假设 3 与假设 2 还是得不到最终的信使迭代公式 (28), 在这里还应满足相比假设 2更严格的假设, 即: 在马尔科夫网中所有边位势为 $\psi_{ij}(g_i,g_j)=1-P_{g_ig_j}$ 的节点对 (v_i,v_j)(包括原网络中不存在边以及额外增加的存在边), 有 $q_{\boldsymbol{r}}^{i}\approx\eta_{\boldsymbol{r}}^{i\to j}$. 意思就是存在一些少部分节点对 (v_i,v_j), 即使 v_j 是 v_i 的邻居也要近似处理, 即移除节点 v_j 对节点 v_i 属于模块 $\boldsymbol{r}$ 的概率的影响忽略不计. 总的来说, 虽然改进后的 BP 算法在实验结果上会表现出很好的效果, 但是存在理论推导上比较复杂、近似处理的过程太多、不易理解等问题.

2.4.4 小 结

BP 算法可以对联合概率公式 (3) 与 (25) 进行求解, 以此得到两种迭代公式 (前者用 BP 算法表示, 后者用 IBP 算法表示) 求解边际概率. 但是都存在一些问题: 1) 在 BP 算法中, 要满足假设 1 与假设 2, 因此不适用核边结构的探测; 2) 在 IBP 算法中, 虽然在实验中已经验证了比 BP 算法有着更好的性能, 但是在理论上处理得太过复杂, 还存在一些在理论上不易理解的假设.

而且, 通过观察可以看出, 无论是 BP 算法还是 IBP 算法所处理的马尔科夫网 (至少也是全连

通图) 都是与要处理的原始网络不一致的. 虽然原始网络是稀疏的, 但是马尔科夫网络却是全连通的, 这也造成了即使 BP 算法与 IBP 算法表现得效果好, 却有所疑惑. 这是因为我们都知道: 当马尔科夫网是树状网络, BP 算法是精确的, 当马尔科夫网是稀疏的时候, BP 算法是近似精确的.

所以, 下面将从另外一个角度出发, 即首先对联合概率公式 (3) 与 (25) 进行近似处理 (而不是在 BP 算法的推理中近似处理), 得到一个新的概率公式, 这个概率公式要满足两个条件: 1) 由这个公式导出的马尔科夫随机网与原始网络一致; 2) 这个公式在形式上与 (14) 式完全一致. 前者可以保证马尔科夫网的稀疏性, 后者可以保证在应用 BP 算法的时候简洁明了且便于理解.

3 基于平均场近似的 BP 算法

首先对联合概率公式 (3) 进行平均场近似处理. 概率分布 (3) 式可以表示为

$$q(\boldsymbol{g}) \propto \prod_{i<j} p_{g_ig_j}^{A_{ij}} \left(1-p_{g_ig_j}\right)^{1-A_{ij}} \prod_i \gamma_{g_i} = \left[\exp\left(\sum_{i<j}(1-A_{ij})\ln(1-p_{g_ig_j})\right)\right] \prod_{i<j} p_{g_ig_j}^{A_{ij}} \prod_i \gamma_{g_i}. \tag{33}$$

令 $h_i = \sum_{j\neq i}(1-A_{ij})\ln\left(1-p_{g_ig_j}\right)$, 通过平均场近似有

$$h_i^{\mathrm{MF}} = \sum_{j\neq i}\sum_{\boldsymbol{s}} q_{\boldsymbol{s}}^j (1-A_{ij}) \ln\left(1-p_{g_i\boldsymbol{s}}\right) \approx \sum_{\boldsymbol{s}} \left[\ln\left(1-p_{g_i\boldsymbol{s}}\right) \sum_j q_{\boldsymbol{s}}^j (1-A_{ij})\right], \tag{34}$$

再应用一次平均场近似, 即

$$\sum_j q_{\boldsymbol{s}}^j A_{ij} \approx \frac{d_i}{n} \sum_j q_{\boldsymbol{s}}^j, \tag{35}$$

其中 $d_i = \sum_j A_{ij}$, 表示节点 v_i 的度. 则 (34) 式可以写成:

$$h_i^{\mathrm{MF}} = \sum_{\boldsymbol{s}} \left[\ln\left(1-p_{g_i\boldsymbol{s}}\right)\left(1-\frac{d_i}{n}\right)\sum_j q_{\boldsymbol{s}}^j\right]. \tag{36}$$

因此联合概率公式 (3) 经过两次平均场近似可以写成:

$$q(\boldsymbol{g}) \propto \left[\exp\left(\sum_i h_i^{\mathrm{MF}}\right)\right] \prod_{i<j} p_{g_ig_j}^{A_{ij}} \prod_i \gamma_{g_i} = \prod_{i<j} p_{g_ig_j}^{A_{ij}} \prod_i \left\{\exp\left[\sum_{\boldsymbol{s}} \ln\left(1-p_{g_i\boldsymbol{s}}\right) \times \left(1-\frac{d_i}{n}\right)\sum_j q_{\boldsymbol{s}}^j\right]\gamma_{g_i}\right\}. \tag{37}$$

把此联合概率分布定义为一个成对马尔科夫网上的概率分布, 而且马尔科夫网与原始网络一致. 其中, 对于一条边 (v_i, v_j), 其边位势为 $\psi_{ij}(g_i, g_j) = P_{g_ig_j}$, 对于一个节点 v_i, 其点位势为 $\phi_i(g_i) = \exp\left[\sum_{\boldsymbol{s}} \ln\left(1-p_{g_i\boldsymbol{s}}\right)\left(1-\frac{d_i}{n}\right)\sum_j q_{\boldsymbol{s}}^j\right]\gamma_{g_i}$. 所以直接根据 BP 算法的公式 (15)—(17) 式, 可以不经过任何近似地得到:

$$\eta_{\boldsymbol{r}}^{i\to j} = \frac{\gamma_{\boldsymbol{r}}}{Z_{i\to j}} \exp\left(\sum_{\boldsymbol{s}} \ln\left(1-p_{\boldsymbol{rs}}\right)\left(1-\frac{d_i}{n}\right) \times \sum_k q_{\boldsymbol{s}}^k\right) \prod_{k\in N(i)/j} \sum_{\boldsymbol{s}} \eta_{\boldsymbol{s}}^{k\to i} p_{\boldsymbol{rs}}, \tag{38}$$

$$\eta_{\boldsymbol{r}}^{i\to j} = \frac{\gamma_{\boldsymbol{r}}}{Z_{i\to j}} \exp\left(\sum_{\boldsymbol{s}} \ln\left(1-p_{\boldsymbol{rs}}\right)\left(1-\frac{d_i}{n}\right) \times \sum_k q_{\boldsymbol{s}}^k\right) \prod_{k\in N(i)/j} \sum_{\boldsymbol{s}} \eta_{\boldsymbol{s}}^{k\to i} p_{\boldsymbol{rs}}, \tag{39}$$

$$q_{\boldsymbol{rs}}^{ij} = \frac{\eta_{\boldsymbol{r}}^{i\to j}\eta_{\boldsymbol{s}}^{j\to i} p_{\boldsymbol{rs}}}{\sum_{\boldsymbol{rs}} \eta_{\boldsymbol{r}}^{i\to j}\eta_{\boldsymbol{s}}^{j\to i} p_{\boldsymbol{rs}}}, \tag{40}$$

其中配分函数 $Z_{i\to j}$ 与 Z_i 可以分别表示为

$$Z_{i\to j} = \sum_{\boldsymbol{r}} \gamma_{\boldsymbol{r}} \exp\left[\sum_{\boldsymbol{s}} \ln\left(1-p_{\boldsymbol{rs}}\right)\left(1-\frac{d_i}{n}\right)\sum_k q_{\boldsymbol{s}}^k\right] \times \prod_{k\in N(i)/j} \sum_{\boldsymbol{s}} \eta_{\boldsymbol{s}}^{k\to i} p_{\boldsymbol{rs}}, \tag{41}$$

$$Z_i = \sum_{\boldsymbol{r}} \gamma_{\boldsymbol{r}} \exp\left[\sum_{\boldsymbol{s}} \ln\left(1-p_{\boldsymbol{rs}}\right)\left(1-\frac{d_i}{n}\right)\sum_k q_{\boldsymbol{s}}^k\right] \times \prod_{k\in N(i)} \sum_{\boldsymbol{s}} \eta_{\boldsymbol{s}}^{k\to i} p_{\boldsymbol{rs}}. \tag{42}$$

上述迭代过程我们记为 MFBP 算法.

类似地, 针对 IBP 算法处理过程, 也可以对联合概率公式 (3) 先进行变形、再平均场近似处理.

即概率分布 (3) 式可以表示为

$$q(\boldsymbol{g}) \propto \prod_{i<j}\left(\frac{p_{g_ig_j}}{1-p_{g_ig_j}}\right)^{A_{ij}}\left(1-p_{g_ig_j}\right)\prod_i \gamma_{g_i}$$
$$= \left[\exp\left(\sum_{i<j}\ln\left(1-p_{g_ig_j}\right)\right)\right] \times \prod_{i<j}\left(\frac{p_{g_ig_j}}{1-p_{g_ig_j}}\right)^{A_{ij}}\prod_i \gamma_{g_i}, \tag{43}$$

令 $\hat{h}_i = \sum_{j\neq i}\ln\left(1-p_{g_ig_j}\right)$, 通过平均场近似有

$$\hat{h}_i^{\mathrm{MF}} = \sum_{j\neq i}\sum_{\boldsymbol{s}} q_{\boldsymbol{s}}^j \ln\left(1-p_{g_i\boldsymbol{s}}\right) \approx \sum\nolimits_{\boldsymbol{s}}\left[\ln\left(1-p_{g_i\boldsymbol{s}}\right)\sum\nolimits_j q_{\boldsymbol{s}}^j\right]. \tag{44}$$

因此联合概率公式 (3) 经过一次平均场近似可以写成

$$q(\boldsymbol{g}) \propto \left[\exp\left(\sum_i \tilde{h}_i^{\mathrm{MF}}\right)\right]\prod_{i<j} p_{g_ig_j}^{A_{ij}}\prod_i \gamma_{g_i}$$
$$= \prod_{i<j}\left(\frac{p_{g_ig_j}}{1-p_{g_ig_j}}\right)^{A_{ij}} \times \prod_i\left\{\exp\left[\sum_s \ln\left(1-p_{g_is}\sum_j q_s^j\right)\gamma_{g_i}\right\}\right. . \tag{45}$$

把此联合概率分布定义为一个成对马尔科夫网上的概率分布, 而且马尔科夫网与原始网络一致. 其中, 对于一条边 (v_i, v_j), 其边位势为 $\psi_{ij}(g_i, g_j) = \frac{p_{g_ig_j}}{1-p_{g_ig_j}}$, 对于一个节点 v_i, 其点位势为 $\phi_i(g_i) = \exp\left[\sum_{\boldsymbol{s}}\ln\left(1-p_{g_i\boldsymbol{s}}\right)\sum_j q_{\boldsymbol{s}}^j\right]\gamma_{g_i}$. 所以直接根据 BP 算法的公式 (15)—(17) 式, 可以得到:

$$\eta_{\boldsymbol{r}}^{i\to j} = \frac{\gamma_{\boldsymbol{r}}}{Z_{i\to j}}\exp\left(\sum_{\boldsymbol{s}}\ln\left(1-p_{\boldsymbol{rs}}\right)\sum_k q_{\boldsymbol{s}}^k\right) \times \prod_{k\in N(i)/j}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}, \tag{46}$$

$$q_{\boldsymbol{r}}^i = \frac{\gamma_{\boldsymbol{r}}}{Z_i}\exp\left(\sum_{\boldsymbol{s}}\ln\left(1-p_{\boldsymbol{rs}}\right)\sum_k q_{\boldsymbol{s}}^k\right) \times \prod_{k\in N(i)}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}, \tag{47}$$

$$q_{\boldsymbol{rs}}^{ij} = \frac{\eta_{\boldsymbol{r}}^{i\to j}\eta_{\boldsymbol{s}}^{j\to i}\dfrac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}}{\sum\limits_{\boldsymbol{rs}}\eta_{\boldsymbol{r}}^{i\to j}\eta_{\boldsymbol{s}}^{j\to i}\dfrac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}}, \tag{48}$$

其中配分函数 $Z_{i\to j}$ 与 Z_i 可以分别表示为

$$Z_{i\to j} = \sum_{\boldsymbol{r}}\gamma_{\boldsymbol{r}}\exp\left[\sum_{\boldsymbol{s}}\ln\left(1-p_{\boldsymbol{rs}}\right)\sum_k q_{\boldsymbol{s}}^k\right] \times \prod_{k\in N(i)/j}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}, \tag{49}$$

$$Z_i = \sum_{\boldsymbol{r}}\gamma_{\boldsymbol{r}}\exp\left[\sum_{\boldsymbol{s}}\ln\left(1-p_{\boldsymbol{rs}}\right)\sum_k q_{\boldsymbol{s}}^k\right] \times \prod_{k\in N(i)}\sum_{\boldsymbol{s}}\eta_{\boldsymbol{s}}^{k\to i}\frac{p_{\boldsymbol{rs}}}{1-p_{\boldsymbol{rs}}}. \tag{50}$$

上述迭代过程记为 MFIBP 算法.

对比 BP 算法与 MFBP 算法、IBP 算法与 MFIBP 算法的最终迭代公式, 发现不同的只是外场量. 在节点个数足够多, 且网络比较稀疏的情况下, BP 与 IBP 算法的外场都可以表示为

$$\prod_{k\in\boldsymbol{V}}\left[1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^k p_{\boldsymbol{rs}}\right] = \exp\left[\sum_k\ln\left(1-\sum_{\boldsymbol{s}}q_{\boldsymbol{s}}^k p_{\boldsymbol{rs}}\right)\right] = \exp\left(\sum_{\boldsymbol{s}}\sum_k q_{\boldsymbol{s}}^k p_{\boldsymbol{rs}}\right).$$

又因为, $d_i/n\to 0$ 且 $\ln\left(1-p_{\boldsymbol{rs}}\right)\to p_{\boldsymbol{rs}}$, 则 MFBP 算法的外场量可以近似为

$$\exp\left[\sum_{\boldsymbol{s}}\ln\left(1-p_{\boldsymbol{rs}}\right)\left(1-\frac{d_i}{n}\right)\sum_k q_{\boldsymbol{s}}^k\right] \to \exp\left(\sum_{\boldsymbol{s}}\sum_k q_{\boldsymbol{s}}^k p_{\boldsymbol{rs}}\right),$$

MFIBP 算法的外场量可以近似为

$$\exp\left[\sum_{\boldsymbol{s}}\ln\left(1-p_{\boldsymbol{rs}}\right)\sum_k q_{\boldsymbol{s}}^k\right] \to \exp\left(\sum_{\boldsymbol{s}}\sum_k q_{\boldsymbol{s}}^k p_{\boldsymbol{rs}}\right).$$

可以看出: 理论上, 经过平均场近似后得到的 MFBP 算法和 MFIBP 算法在一定条件下分别与原有的 BP 算法和 IBP 算法是等价的. 但是在理论推导上, 平均场近似的方法要更简单合理且便于理解. MFIBP 算法与 MFBP 算法相比较, 少了一次平均场近似.

下面将对比 BP 算法和 MFBP 算法, 以及 IBP 算法和 MFIBP 算法实验上的效果.

4 中尺度结构的探测

下面将在社团结构及核边结构探测上进行实验验证. 在社团结构探测中, 为了验证不同算法探测结果的好坏, 将采用 NMI 指标进行衡量, 即 [39]:

$$\mathrm{NMI}(X,Y)=\frac{2I(X,Y)}{H(X)+H(Y)}, \tag{51}$$

其中, X 与 Y 分别为算法探测的结果以及真实的结果; $I(X,Y)$表示 X 与 Y 之间的互信息; $H(X)$与 $H(Y)$分别表示 X 与 Y 的信息熵.

对于核边结构的探测结果的评价, NMI 并不是一个合理的指标, 这是因为核边结构的核与边的地位并不是等价的. 所以将采用 $F1$指标进行衡量, 即[40]:

$$F1=2PR/(P+R), \tag{52}$$

其中, P 表示预测结果中, 预测为正 (本文中为核) 的样本中预测正确的概率, R 表示数据样本中, 正样本中预测正确的概率.

4.1 社团结构探测

首先采用随机块模型生成具有社团结构的基准网络, 以此来验证 MFBP 算法和 MFIBP 算法分别与 BP 算法和 IBP 算法的区别. 假设有 k 个模块, 则每个节点就以 $\gamma_{\boldsymbol{r}}$ 的概率分配到第 $\boldsymbol{r}$ 模块中. 社团之间节点的连边概率定义为 $p_{\boldsymbol{rs}}=c_{\mathrm{in}}/n(\boldsymbol{r}=\boldsymbol{s})$ 和 $p_{\boldsymbol{rs}}=c_{\mathrm{out}}/n(\boldsymbol{r}\neq\boldsymbol{s})$, 因此网络的平均度为 $c=\sum_{\boldsymbol{r}<\boldsymbol{s}}2\gamma_{\boldsymbol{r}}\gamma_{\boldsymbol{s}}c_{\mathrm{out}}+\sum_{\boldsymbol{r}}\gamma_{\boldsymbol{r}}^2c_{\mathrm{in}}$.

如果每个模块的大小一致, 即 $\gamma_{\boldsymbol{r}}=1/k$, 则网络的平均度为 $c=[c_{\mathrm{in}}+q(c_{\mathrm{out}}-1)]/q$. 定义模块外部与社团内部的连接概率比为 $\varepsilon=c_{\mathrm{out}}/c_{\mathrm{in}}$. 所以, 当 $\varepsilon<1$时生成网络具有社团结构, 且越小社团结构越明显; 当 $\varepsilon=1$时生成网络为 ER 随机图. 当 $n\to\infty$时, 网络检测社团结构会存在一个阈值[38]:

$$\varepsilon^*=(\sqrt{c}-1)/(\sqrt{c}-1+k), \tag{53}$$

即, 如果 $\varepsilon>\varepsilon^*$时, 各种社团算法都无法检测出社团结构; 当 $\varepsilon<\varepsilon^*$时, BP 算法可以探测出社团结构.

从理论上, 前面的分析可以看出, 当网络节点个数足够多, 网络连接比较稀疏时, BP 算法、MFBP 算法、IBP 算法以及 MFIBP 算法是近似等价的, 通过图 1(a) 可以验证, 所有结果为 10 次结果的平均值 (下同). 通过图 1(b)—图 1(d) 可以发现, 当网络连接稀疏且规模较小时, BP 算法、MFBP 算法、IBP 算法以及 MFIBP 算法的结果也是一致的. 但是, 在网络连接稠密的情况下 (如图 2), 可以发现 MFBP 算法要优于 BP 算法, 且与 MFIBP 算法、IBP 算法一致.

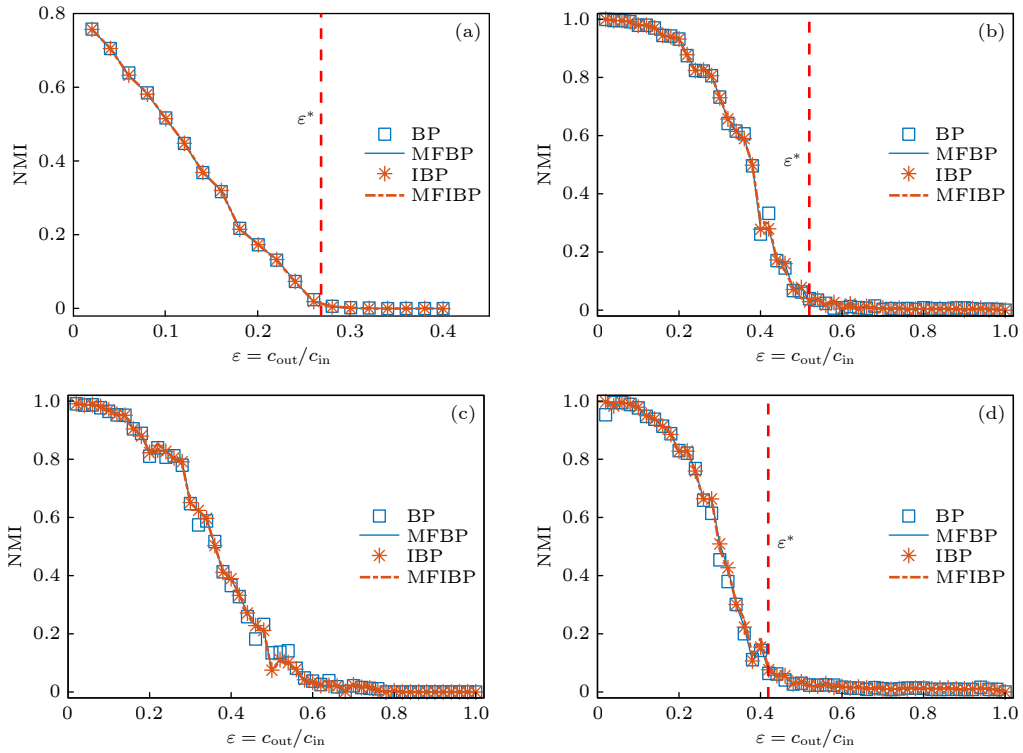

图 1 稀疏基准网络社团结构探测 (a) $n=10000$, $c=3$, $\gamma=[0.5,\ 0.5]^{\mathrm{T}}$, $\varepsilon^*=0.27$; (b) $n=200$, $c=10$, $\gamma=[0.5,\ 0.5]^{\mathrm{T}}$, $\varepsilon^*=0.520$; (c) $n=200$, $c=10$, $\gamma=[0.3,\ 0.7]^{\mathrm{T}}$; (d) $n=200$, $c=10$, $\gamma=[1/3,\ 1/3,\ 1/3]^{\mathrm{T}}$, $\varepsilon^*=0.419$

Fig. 1. Community structure detection in sparse benchmark networks: (a) $n=10000$, $c=3$, $\gamma=[0.5,\ 0.5]^{\mathrm{T}}$, $\varepsilon^*=0.27$; (b) $n=200$, $c=10$, $\gamma=[0.5,\ 0.5]^{\mathrm{T}}$, $\varepsilon^*=0.520$; (c) $n=200$, $c=10$, $\gamma=[0.3,\ 0.7]^{\mathrm{T}}$; (d) $n=200$, $c=10$, $\gamma=[1/3,\ 1/3,\ 1/3]^{\mathrm{T}}$, $\varepsilon^*=0.419$.

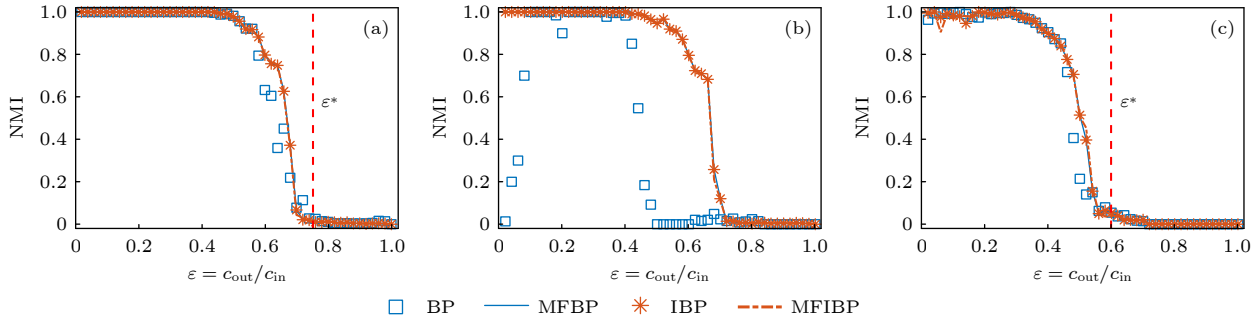

图 2 稠密基准网络社团结构探测 (a) $n=200$, $c=50$, $\gamma=[0.5,\ 0.5]^{\mathrm{T}}$, $\varepsilon^*=0.75$; (b) $n=200$, $c=50$, $\gamma=[0.3,\ 0.7]^{\mathrm{T}}$, (c) $n=200$, $c=30$, $\gamma=[1/3,\ 1/3,\ 1/3]^{\mathrm{T}}$, $\varepsilon^*=0.599$

Fig. 2. Community structure detection in dense benchmark networks: (a) $n=200$, $c=50$, $\gamma=[0.5,\ 0.5]^{\mathrm{T}}$, $\varepsilon^*=0.75$; (b) $n=200$, $c=50$, $\gamma=[0.3,\ 0.7]^{\mathrm{T}}$; (c) $n=200$, $c=30$, $\gamma=[1/3,\ 1/3,\ 1/3]^{\mathrm{T}}$, $\varepsilon^*=0.599$.

4.2 核边结构探测

生成具有核边结构特性的基准网络来验证 MFBP 算法和 MFIBP 算法分别与 BP 算法和 IBP 算法的区别. 给定 3 个参数: 核与核的连边概率 P_{CC}, 核与边的连边概率 P_{CP}, 以及边与边的连边概率 P_{PP}, 当设置 $P_{\text{CC}}>P_{\text{CP}}>P_{\text{PP}}$ 时, 可以生成一个具有核边结构的网络 [28]. 在本文中, 设 $P_{\text{PP}}=0.05$, $P_{\text{CC}}=\theta$, $P_{\text{CP}}=0.6\theta$. 网络的大小 $n=200$, 其中核节点的个数为 50, 边节点的个数为 150. 图 3 所示为 4 种算法在不同参数基准网络的核边结构的探测情况, 其中实验结果为 10 次结果的平均值. 同样可以发现, MFBP 算法要优于 BP 算法, 且与 MFIBP 算法、IBP 算法一致. 在 BP 算法中, 当核边结构足够明显时, 却表现出了较差的结果, 这是因为核边结构明显, 意味着网络结构稠密, 特别是核节点的度很大, 这会导致 (19) 式关于用所有节点集合近似邻居的补集变得极其不合理, 而 MFBP 算法、IBP 算法以及 MFIBP 算法没有这方面的缺陷.

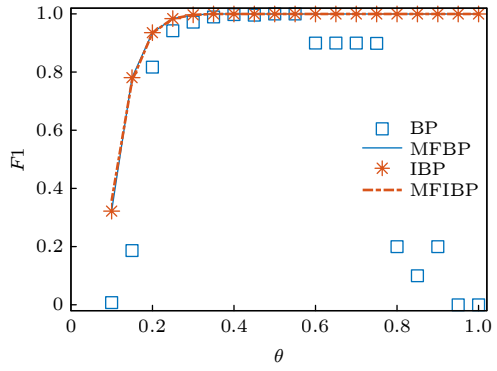

图 3 基准网络的核边结构探测

Fig. 3. CP structure detection in benchmark networks.

接着, 在美国航空网络上验证本文中的算法, 其中每个节点代表一个机场, 共有 332 个节点; 每条边表示两个机场具有航班, 共有 2126 条边 [5]. 应用 BP 算法, 求解结果有 27 个核; 应用 MFBP 算法, 求解结果有 34 个核; 应用 IBP 算法, 求解结果有 47 个核; 应用 MFIBP 算法, 求解结果有 47 个核. 其中对应的参数结果见 (54) 式—(57) 式. 为了验证上述 4 个算法求得结果的精确性, 采用 MCMC 采样求解 EM 算法, 共采样 1000 轮 (每轮 n 次), 得到的结果共有 47 核, 具体参数如 (58) 式. 如果把 MCMC 结果作为真实结果, 则 BP 算法、MFBP 算法、IBP 算法以及 MFIBP 算法结果的 $F1$ 值分别为 0.730, 0.840, 1.000 以及 1.000. 从上述结果可以看出, 在美国航空网络的核边结构实验中, MFBP 算法结果要优于 BP 算法, MFIBP 算法结果与 IBP 算法一致, 且接近精确解. 而且 MFIBP 算法要优于 MFBP 算法, 这是因为 MFBP 算法比 MFIBP 算法多一次平均场近似.

$$\gamma_{\text{BP}}=\begin{bmatrix}0.081\\0.919\end{bmatrix},\ p_{\text{BP}}=\begin{bmatrix}0.873 & 0.151\\0.151 & 0.012\end{bmatrix}, \tag{54}$$

$$\gamma_{\text{MFBP}}=\begin{bmatrix}0.102\\0.898\end{bmatrix},\ p_{\text{MFBP}}=\begin{bmatrix}0.836 & 0.120\\0.120 & 0.010\end{bmatrix}, \tag{55}$$

$$\gamma_{\text{IBP}}=\begin{bmatrix}0.142\\0.858\end{bmatrix},\ p_{\text{IBP}}=\begin{bmatrix}0.711 & 0.074\\0.074 & 0.008\end{bmatrix}, \tag{56}$$

$$\gamma_{\text{MFIBP}}=\begin{bmatrix}0.142\\0.858\end{bmatrix},\ p_{\text{MFIBP}}=\begin{bmatrix}0.713 & 0.074\\0.074 & 0.008\end{bmatrix}, \tag{57}$$

$$\gamma_{\text{MCMC}}=\begin{bmatrix}0.142\\0.858\end{bmatrix},\ p_{\text{MCMC}}=\begin{bmatrix}0.709 & 0.074\\0.074 & 0.008\end{bmatrix}. \tag{58}$$

5 结 论

本文详细描述了随机块模型探测社团结构、核边结构的理论, 以及 MCMC 和 BP 算法求解该理论的方法, 并分析其优缺点. 通过理论分析以得到: 1) 原始的 BP 算法不适用于核边结构的探测; 2) 修正后得到的 IBP 算法, 虽然适用于核边结构, 但在理论上推导近似多且不明朗; 3) 无论是 BP 算法还是 IBP 算法, 所处理的马尔科夫网 (联合概率分布), 既不是原始网络, 也不满足网络稀疏条件. 最后一条是造成随机块模型中 BP 算法和 IBP 算法理论难以理解的主要原因. 针对于此, 本文首先对联合概率分布进行平均场近似, 得到一个好的联合分布函数, 然后可以直接套用 BP 算法理论, 且在这一步不需要任何近似, 使得 BP 算法求解随机块模型这一套方法, 不仅在结果上效果显著, 而且在理论上通俗易懂. 这种先通过平均场近似再套用 BP 算法还适用于利用最小化哈密顿量 (关于模块度的函数) 探测社团结构的方法中[38], 和本文一样, 将使得理该论推导部分变得简单易懂.

本文的计算工作得到了安徽大学高性能计算平台的支持.

参考文献

[1] Zhang H F, Wang W X 2020 *Acta Phys. Sin.* **69** 088906 (in Chinese) [张海峰, 王文旭 2020 物理学报 **69** 088906]
[2] Guimerà R, Mossa S, Turtschi A, Amaral L A N 2005 *PNAS* **102** 7794
[3] Newman M E J 2006 *Phys. Rev. E* **74** 36104
[4] Benson A R, Gleich D F, Leskovec J 2016 *Science* **353** 163
[5] Xiang B B, Bao Z K, Ma C, Zhang X, Chen H S, Zhang H F 2018 *Chaos* **28** 13122
[6] Newman M E J 2003 *SIAM Rev.* **45** 167
[7] Leicht E A, Newman M E J 2008 *Phys. Rev. Lett.* **100** 118703
[8] Newman M E J 2006 *Proc. Natl Acad. Sci. U.S.A.* **103** 8577
[9] Newman M E J 2012 *Nat. Phys.* **8** 25
[10] Zhang X, Newman M E J 2015 *Phys. Rev. E* **92** 52808
[11] Lee D D, Seung H S 1999 *Nature* **401** 788
[12] Chang Z C, Chen H C, Liu Y, Yu H T, Huang R Y 2015 *Acta Phys. Sin.* **64** 218901 (in Chinese) [常振超, 陈鸿昶, 刘阳, 于洪涛, 黄瑞阳 2015 物理学报 **64** 218901]
[13] Shao J, Han Z, Yang Q, Zhou T 2015 *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Sydney NSW, Australia, August 10–13, 2015 p1075
[14] Gregory S 2010 *New J. Phys.* **12** 103018
[15] Karrer B, Newman M E J 2011 *Phys. Rev. E* **83** 16107
[16] Decelle A, Krzakala F, Moore C, ZdeborováL 2011 *Phys. Rev. Lett.* **107** 65701
[17] Ledwith M 2020 *Community development: A critical approach* (Bristol: Policy Press) pp1–252
[18] Wang X Y, Zhao Z X 2014 *Acta Phys. Sin.* **63** 178901 (in Chinese) [王兴元, 赵仲祥 2014 物理学报 **63** 178901]
[19] Everett M G, Borgatti S P 2000 *Social Networks* **21** 397
[20] Verma T, Russmann F, Araújo N A M, Nagler J, Herrmann H J 2016 *Nat. Commun.* **7** 10441
[21] Lee S H, Cucuringu M, Porter M A 2014 *Phys. Rev. E* **89** 32810
[22] Rombach P, Porter M A, Fowler J H, Mucha P J 2017 *SIAM Rev.* **59** 619
[23] Kojaku S, Masuda N 2017 *Phys. Rev. E* **96** 52313
[24] Kojaku S, Masuda N 2018 *New J. Phys.* **20** 43012
[25] Ma C, Xiang B B, Chen H S, Zhang H F 2020 *Chaos* **30** 23112
[26] Zhang X, Martin T, Newman M E J 2015 *Phys. Rev. E* **91** 32803
[27] Della Rossa F, Dercole F, Piccardi C 2013 *Sci. Rep.* **3** 1467
[28] Ma C, Xiang B B, Chen H S, Small M, Zhang H F 2018 *Chaos* **28** 53121
[29] Kang L, Xiang B B, Zhai S L, Bao Z K, Zhang H F 2018 *Acta Phys. Sin.* **67** 198901 (in Chinese) [康玲, 项冰冰, 翟素兰, 鲍中奎, 张海峰 2018 物理学报 **67** 198901]
[30] Ball B, Karrer B, Newman M E J 2011 *Phys. Rev. E* **84** 36103
[31] Decelle A, Krzakala F, Moore C, ZdeborováL 2011 *Phys. Rev. E* **84** 66106
[32] Mugisha S, Zhou H J 2016 *Phys. Rev. E* **94** 12305
[33] Yedidia J S, Freeman W T, Weiss Y 2005 *IEEE Trans. Inf. Theory* **51** 2282
[34] Mezard M, Montanari A 2009 *Information, Physics, and Computation* (Oxford: Oxford University Press) pp304–305
[35] Perotti J I, Tessone C J, Clauset A, Caldarelli G 2018 arXiv: 1806.07005 v1[soc-ph]
[36] Dempster A P, Laird N M, Rubin D B 1977 *Journal of the Royal Statistical Society: Series B (Methodological)* **39** 1
[37] Tiago P, Peixoto 2019 *Advances in Network Clustering and Blockmodeling* (New York:Wiley) pp289–332
[38] Zhang P, Moore C 2014 *Proc. Natl. Acad. Sci. U.S.A.* **111** 18144
[39] Gerlof B 2009 Proceedings of the 21th Biennial GSCL Conference Potsdam, Germany, September 30–October 2 2009 p31
[40] Sokolova M, Laxpalme G 2009 *Inf. Process. Manage.* **45** 427

A mean-field approximation based BP algorithm for solving the stochastic block model*

Ma Chuang [1)] Yang Xiao-Long [1)] Chen Han-Shuang [2)] Zhang Hai-Feng [3)†]

1) (*School of Internet*, *Anhui University*, *Hefei* 230039, *China*)

2) (*School of Physics and Material Science*, *Anhui University*, *Hefei* 230601, *China*)

3) (*School of Mathematical Science*, *Anhui University*, *Hefei* 230601, *China*)

(Received 16 March 2021; revised manuscript received 29 July 2021)

Abstract

As a mainstream algorithm for inferring probabilistic graphical models, belief propagation (BP) algorithm is one of the most important methods to solve the joint probability distribution in the stochastic block model. However, existing methods either lead to low accuracy in dealing with the core-periphery structure problem, or the theoretical derivation is difficult to understand due to a large number of approximation, or both exist. Of course, the reason for low accuracy comes from too many approximations. The main reason for many approximations and complex theoretical derivation is that the joint probability distribution in the inference process of the stochastic block model is not directly solved by the BP algorithm, that is, the graph (network) being processed is not consistent with the graph considered in the probabilistic graph model. Therefore, in this paper, a mean-field approximation is developed to modify the joint probability distribution to make the BP algorithm match perfectly, which makes the theoretical derivation easy to understand. Finally, the effectiveness of the proposed method is validated by the experimental results.

Keywords: stochastic block model, belief propagation algorithm, joint probability distribution, mean-field approximation

PACS: 89.75.Hc, 89.75.Fb, 05.10.–a, 05.10.Ln **DOI:** 10.7498/aps.70.20210511

* Project supported by the National Natural Science Foundation of China (Grant Nos. 12005001, 61973001, 11875069),the University Synergy Innovation Program of Anhui Province, China (Grant No. GXXT-2021-032), and the Natural Science Foundation of Anhui Province, China (Grant No. 2008085QF299).

† Corresponding author. E-mail: haifengzhang1978@gmail.com