结构特征强化的高效马尔可夫随机场社团发现方法

金 弟 尤心心 刘岳森 何东晓
(天津大学智能与计算学部 天津 300350)

摘 要 社团发现是非常重要的网络数据分析任务.统计模型类社团发现方法由于具有坚实的理论基础和优越的性能,因此越来越被人们关注.然而,已有社团发现模型一般都基于有向概率图模型,作为无向概率图模型的马尔可夫随机场极少被用于社团发现领域.2018 年我们提出了一个网络导向的马尔可夫随机场模型 NetMRF,该模型虽具有良好的性能,但仍存在如下问题:(1)NetMRF 的能量函数不够完整,缺少往往在 MRF 中起主导作用的单点势函数,仅采用了常被视为起辅助作用的成对势函数对社团进行描述;(2)也正因为如此,为了使成对势函数能有效建模网络中不规则的拓扑信息,NetMRF 采用了复杂的三层全连接马尔可夫随机场结构,这虽会增强其描述能力,却给推断算法带来了 $O(n^3)$级时间复杂度,n 为网络节点数.本文针对上述问题对 NetMRF 进行改进.首先基于网络嵌入方法,结合吉布斯分布设计有效的单点势函数,解决了 NetMRF 能量函数不完整的缺陷;进而通过对成对势函数结构的有效稀疏化,缓解了其效率不高的问题;从而构建了一个高精度、近线性的马尔可夫随机场新模型 iMRF.本文采用"最大化-加和"版本的信念传播算法对 iMRF 进行推断,通过最大化联合后验概率获得最优的社团配置.在两组人工网络和 20 个真实网络上,我们将 iMRF 与 6 个统计模型类社团发现方法(包含 NetMRF)进行比较,结果显示 iMRF 的平均精度高于对比算法 2.6%~12.9%;iMRF 的平均运行速度在对比算法中也名列前茅,尤其是对于大规模网络具有更强的处理能力.

关键词 社交网络;社团发现;网络嵌入;马尔可夫随机场;信念传播
中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2019.02821

Structural Feature Enhanced Markov Random Field for Community Detection in Large-Scale Networks

JIN Di YOU Xin-Xin LIU Yue-Sen HE Dong-Xiao
(*College of Intelligence and Computing, Tianjin University, Tianjin* 300350)

Abstract Networks can model interactions and relationships between units of various complex systems. They are powerful representations which can be used for analyzing the nature and function of complex systems. One of the most important properties of complex networks is their community structure, in which nodes are connected more densely within clusters than across clusters. Discovering communities is useful in many real applications. In recent years, discovering community structure in complex networks has attracted a great number of scholars from various research fields. A series of methods based upon different theories and techniques have been proposed. Among these methods, statistical model based methods have a solid theoretical basis and reasonably good performance, and hence have been broadly adopted. However, existing community detection methods are typically based on directed probabilistic graphical models, while Markov Random Field(MRF) which is a type of undirected probabilistic graphical model is rarely used in community detection. MRF is a general and potent statistical modeling technique. It can well represent the

收稿日期:2018-05-16;在线出版日期:2019-05-13.本课题得到国家自然科学基金(61772361,61876128)资助.**金 弟**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为复杂网络分析、社团检测.E-mail:jindi@tju.edu.cn.**尤心心**,硕士,主要研究方向为复杂网络分析、社团检测.**刘岳森**,本科生,主要研究方向为网络社团检测.**何东晓**(通信作者),博士,副教授,主要研究方向为数据挖掘、网络社团检测.E-mail:hedongxiao@tju.edu.cn.

structural relationships of task-specific properties and constraints underlying complex problems. In 2018, we presented a network-specific MRF approach (NetMRF) for community detection. Though this model owns a good performance, it still has the following problems. (1) The energy function of NetMRF is incomplete, i. e., it lacks the unary potentials which often play a dominant role in MRF. Only the pairwise potentials that play an assistant function in MRF are used to describe the community structure. (2) Due to this reason, in order to enable pairwise potentials to effectively model the irregular topology information in the network, NetMRF defined a three-layer and fully-connected structure for MRF. Although it enhances the description capability of NetMRF, the time complexity of NetMRF reaches at $O(n^3)$, where n is the number of nodes. To solve these two problems, we propose a new MRF model, i. e., iMRF, which means improved MRF. First, to make the energy function complete, we design the unary potentials for iMRF based on network embedding and Gibbs distribution. Furthermore, through the sparse treatment of the structure of pairwise potentials, the problem of low efficiency is also alleviated. Based on these, a new and effective MRF model iMRF is constructed. The max-sum version of belief propagation (BP) can be used to find a configuration corresponding to the maximum of joint probability of MRF models, which only needs a similar time with that of finding a set of memberships via individually maximizing the marginal probabilities. Therefore, we employ the max-sum version of BP to maximize the joint probability distribution of iMRF, so as to derive better configuration of community memberships. The pairwise structure of our framework is the same as network topology, thus the complexity of our algorithm is finally $O(m)$, where m is the number of edges. We validated the new MRF-based method on two types of synthetic benchmarks as well as on twenty real-world networks, and compared it with six state-of-the-art methods (including NetMRF). The results show that the average accuracy of iMRF is 2.6% to 12.9% higher than the algorithms compared, and it has the stronger processing capability on large-scale networks.

Keywords social networks; community detection; network embedding; Markov random field; belief propagation

1 引 言

网络代表了复杂系统中不同个体间的交互关系,是用于分析复杂系统性质和功能的有力表征工具.复杂网络,如生物网络、通讯网络和社交网络,分别是生物系统、通讯系统和交互系统的抽象表达.社团结构是复杂网络的一个重要性质,即相同社团内节点的交互要比不同社团间节点的交互更为频繁[1-2].探测复杂网络中蕴含的社团结构可帮助人们更好的理解复杂系统的组织原理,探测其功能,并预测其未来趋势.譬如:在社交网络中可发现具有相似兴趣爱好的社交团体,在引文网络中可探测出具有相似主题的论文簇,在蛋白质交互网络中可发现具有相似生物功能的组织模块等.

近年来,复杂网络社团结构检测已吸引了许多来自不同研究领域研究者的关注.目前已提出了许多基于不同理论和技术的方法[3],它们包括谱聚类[4-5]、层次聚类[6]、启发式方法[7]、模块度优化[8-9]、动力学过程[10]和统计模型推断[11]等.基于统计模型的方法,由于具有坚实的理论基础和优越的性能,得到了广泛的研究与应用.目前基于统计模型的方法可分为三类:第一类主要是基于随机块模型及其扩展,采用似然最大化进行求解[12],譬如 Karrer 等人[13]基于"在模型中保持节点度分布"的思想,提出了一个度修正的社团发现随机块模型.第二类是将非负矩阵分解用于社团发现任务[14],譬如 Zhao 等人[15]提出了一个概率矩阵分解模型来建模符号网络,并采用期望最大化方法进行参数估计,以发现符号网络中的社团结构.第三类是基于深度学习的社团发现模型,譬如 Yang 等人[16]提出了一个基于深度自编码器的网络及社团表征模型,在学习出大规模网络非线性表征的同时揭示社团结构.

马尔可夫随机场(Markov Random Field,MRF)

是一类通用的统计建模技术，它能够很好地表征复杂情形下特定任务的性质和约束关系. MRF 主要包含两类函数：单点势函数(unary potential)和成对势函数(pairwise potential). 单点势函数主要用来刻画对象自身具有的特征，成对势函数主要用来捕捉对象之间的依赖或约束关系，它们共同作用构成完整的 MRF 能量函数. 由于 MRF 目标函数(能量函数)的定义不受概率约束，因此善于捕获数据中更多的模式信息；同时，MRF 的能量函数能够依据 Hammersley-Clifford 定理转化为联合后验概率的形式，并可采用具有坚实理论基础的概率推断方法进行严格推断[17]. 目前 MRF 模型已被成功用于图像分割等领域. 图像分割与社团发现在形式上颇为相似，因此 MRF 在社团发现领域也应很具潜力.

然而，目前基于统计模型的社团发现方法大都仅关注于有向概率图模型，而马尔可夫随机场却极少被用于社团发现领域. 将 MRF 应用于社团发现任务的主要挑战在于：(1) 用于图像分割的 MRF 是以像素点特征(如 RGB 值)为主导构建单点势函数(unary potential)，在图像分割中起核心作用；而在复杂网络中则截然不同：拓扑结构是其最基础的信息，而节点特征(即使有的话)一般也仅起辅助作用；(2) MRF 利用像素间规则的网格化结构(如四邻域或八邻域)构建成对势函数(pairwise potential)，在图像分割中起辅助性的微调作用；然而网络拓扑是不规则的，而我们正是需要利用这种不规则的网络拓扑来发现其中蕴含的社团结构. 针对以上问题，2018 年我们提出了一个网络导向的成对 MRF 模型 NetMRF，首次将 MRF 用于社团发现领域[18]. 该模型能够克服(或弱化)网络中没有节点个体特征的缺陷，并通过巧妙地将网络中仅有的不规则拓扑信息建模到模型的核心能量函数中去，来达到刻画社团结构的目的，并通过最小化能量函数以发现最佳社团结构.

NetMRF 虽具有较好的性能，但仍存在如下问题：(1) 该模型的能量函数不够完整，缺少刻画节点个体特征的单点势函数. 一般来说，单点势函数是在成对 MRF 中起主导作用(即获取近优解)的部分，它的缺失会弱化模型对社团的全局描述能力，从而降低社团发现性能；(2) 为了使成对势函数(在 MRF 能量函数中起微调作用的部分)能有效建模网络中不规则的拓扑信息，NetMRF 模型定义了一个复杂的全连接成对势函数结构. 这种全连接的方式虽然有效，却会使其模型推断的时间复杂度非常高，达到 $O(n^3)$级，n 为网络节点数. 尽管文献[18]中还提供了一个高效的近似推断版本，但这种复杂的近似过程却会降低模型的描述能力，尤其会受限于有效处理大规模复杂网络.

针对上述问题本文对 NetMRF 模型进行改进，提出了一个新的成对马尔可夫随机场模型 iMRF (improved MRF). 首先，既然网络嵌入(network embedding)可描述网络节点的内在分布表征[19-20]，我们可采用网络嵌入方法获得节点的拓扑特征，并通过吉布斯分布定义 MRF 单点势函数，以获得全局描述能力. 进而，我们发现：一条边的存在或是由于这对节点具有较高的结构相似性，或是由于其中一个节点的流行度非常高，从而吸引了另一节点与它产生边，并基于此在真实存在的链接上定义成对势函数，以获取高效的局部描述能力. 通过上述，新模型不仅可通过网络嵌入解决 NetMRF 势能函数不完整的缺陷，而且还可通过对成对势函数结构的天然稀疏化来缓解其效率不高的问题，从而构建出一个高精度、近线性的 MRF 新模型. 我们采用最大化-加和版本的信念传播算法(Belief Propagation, BP)对该模型进行推断，通过最大化联合后验概率获得最优的社团配置. 实验结果表明，本文提出的 iMRF 方法平均精度要高于对比算法 2.6%～12.9%；同时，iMRF 的运行速度也要领先于 NetMRF 算法的加速版本，甚至相比于其他统计模型类的社团发现算法也仍具效率优势.

本文第 2 节简要介绍已提出的基准模型 NetMRF；第 3 节对改进的新模型 iMRF 及其推断算法进行描述；第 4 节展示新方法在人工和真实网络上的性能，并与已有方法进行比较分析；第 5 节对社团发现的相关工作进行介绍，最后对文章进行总结，并阐述未来工作.

2　基准马尔可夫随机场模型 NetMRF

不妨设一个无向无权网络 G 包含 n 个节点和 m 条边，它可被表示为二值邻接矩阵 $\boldsymbol{A}=(a_{ij})_{n\times n}$，如果节点 i 和 j 之间有边，则 $a_{ij}=1$，否则 $a_{ij}=0$. 我们的目的是将网络中的 n 个节点划分到 K 个不相交的社团，即每个节点 i 具有一个唯一的社团标签 $c_i\in\{1,\cdots,K\}$，表明它属哪个社团.

2.1　NetMRF 模型介绍

这里我们对已提出的 NetMRF 模型进行简要介绍. 在该模型中，任意节点对$\langle i,j\rangle$都有一个如下

形式的成对势函数：

$$\theta_{ij}(c_i,c_j;a_{ij})=-(-1)^{\delta(c_i,c_j)}\left(\frac{d_i d_j}{2m}-a_{ij}\right) \quad (1)$$

其中，d_i 表示节点 i 的度；若 $c_i=c_j$，则 $\delta(c_i,c_j)$ 为 1，否则为 0；$d_i d_j/2m$ 是节点 i 与 j 之间期望的边数. 上面的成对势函数表明：如果节点对 i 和 j 之间存在的真实边数大于(小于)期望的边数，那么这对节点属于相同社团时会比属于不同社团时贡献更小(更大)的势能值. 它实现了“奖励具有稠密连接的节点对属于相同社团、奖励具有稀疏连接的节点对属于不同社团，惩罚具有稠密连接的节点对属于不同社团、惩罚具有稀疏连接的节点对属于相同社团”的簇结构特性. 这四个奖励与惩罚机制保证了能量函数对于网络中社团结构的刻画能力，使得社团划分结果越好，能量函数值越小.

在最终的能量函数中，所有节点对形成的成对势函数相互配合，共同捕获网络拓扑中的社团结构，以获得全局一致的社团检测结果，即

$$E(C;\boldsymbol{A})=\sum_{i\neq j}\theta_{ij}(c_i,c_j;a_{ij}) \quad (2)$$

基于上面定义的社团划分 C 所对应的能量函数，利用吉布斯分布(Gibbs distribution)$P(C|\boldsymbol{A})\propto\exp\{-\beta E(C;\boldsymbol{A})\}$($\beta$ 是温度的倒数)，即可得到在给定网络拓扑 $\boldsymbol{A}$ 条件下，社团划分 C 的后验概率：

$$P(C|\boldsymbol{A})=\frac{1}{Z(\boldsymbol{A})}\prod_{i\neq j}\exp\left\{(-1)^{\delta(c_i,c_j)}\beta\left(\frac{d_i d_j}{2m}-a_{ij}\right)\right\} \quad (3)$$

其中 $Z(\boldsymbol{A})$ 是归一化系数.

3 改进的马尔可夫随机场模型 iMRF

上述 NetMRF 存在如下问题：(1) 缺少往往在 MRF 中起主导作用(或全局搜索功能)的单点势函数，仅采用了常被视为起辅助作用(或局部搜索功能)的成对势函数对社团进行描述；(2) 也正因为如此，NetMRF 采用了复杂的三层全连接马尔可夫随机场结构来定义成对势函数，这虽会增强其描述能力，但却给推断算法带来了随节点数呈立方级增长的高计算复杂度；(3) 文中虽也给出了启发式的快速近似算法，但其精度又难以有理论保证.

针对上述 NetMRF 的缺陷，这里提出了一个新的马尔可夫随机场社团发现模型 iMRF. 我们首先利用网络嵌入从拓扑结构中提取可表征每个节点自身结构特征的低维向量，并在此基础上定义能量函数中的单点势函数；进而对网络中真正有边相连的节点对，依据存在一条边的两个基本假设(即或是这对节点具有较高的结构相似性，或是由于其中一个节点的流行度非常高、从而吸引了另一节点与它产生边)，来构建它们的成对势函数. 然后，基于吉布斯分布，将模型定义为社团划分结果的联合后验概率. 最后采用信念传播算法最大化 iMRF 的联合后验概率，以获得最佳的社团划分结果.

3.1 模型概观

iMRF 模型的核心能量函数由两部分组成. 第一部分是一组单点势函数，它以个体节点为定义单位，旨在捕捉该节点的个性化结构特征，在模型中起主要作用. 第二部分是一组成对势函数，它以节点对为定义单位，旨在刻画该节点对之间的依赖或约束关系，在模型中起到对初始近优解的微调作用. 能量函数中所有单点势函数和成对势函数互相补充、协同工作，实现探测网络中全局一致的社团结构之目标. iMRF 模型的能量函数被定义为

$$E(C;\boldsymbol{A},\boldsymbol{V})=\sum_i\theta_i(c_i;\boldsymbol{v}_i)+\sum_{\langle i,j\rangle\in\varepsilon}\theta_{ij}(c_i,c_j;\boldsymbol{v}_i,\boldsymbol{v}_j) \quad (4)$$

其中，$C=(c_1,c_2,\cdots,c_n)$ 表示网络中所有节点的社团划分，ε 表示网络中边的集合，$\boldsymbol{v}_i$ 表示节点 i 的低维向量表征，$\boldsymbol{V}=\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n\}$ 表示 n 个节点低维向量的集合，θ_i 表示单点势函数，θ_{ij} 表示成对势函数(具体定义将在 3.2 节和 3.4 节中详细介绍). 式(1)定义的能量函数应该具有与给定社团划分相对应的特性(即社团内部包含的边越多，能量函数值越小；社团间的边越少，能量函数值越小)，以实现最小化能量函数可对应最优社团划分的目标. 此外，能量函数也可作为社团划分的一个量化指标，即社团划分结果越好，它对应的能量函数值越小.

基于定义在社团划分 C 上的能量函数，我们可利用吉布斯分布，来计算给定网络拓扑 $\boldsymbol{A}$ 时社团划分 C 的后验概率：

$$P(C|\boldsymbol{A},\boldsymbol{V})=\frac{1}{Z}\mathrm{e}^{-\beta E(C;\boldsymbol{A},\boldsymbol{V})} \quad (5)$$

这里 Z 是归一化常数系数. 最后，最佳的社团划分结果 $\hat{C}$ 可通过如下公式获得：

$$\hat{C}=\arg\max_C P(C|\boldsymbol{A},\boldsymbol{V}) \quad (6)$$

3.2 定义单点势函数

能量函数中单点势函数的定义是为了刻画每个个体的个性化特征. 对于社团发现任务来说，单点势函数应反映出一个节点结合自身结构特征做出的属于各个社团倾向的判断. 然而，网络中的观测数据一般是邻接矩阵形式，它反映了节点对之间的约束关

系而不是单个节点的个体结构特征，所以无法直接使用它来定义单点势函数. 幸运的是，网络嵌入技术可从网络拓扑中提取每个节点的结构低维表征，这使得单点势函数的定义变得可能. 网络嵌入旨在将网络拓扑数据从原始拓扑空间映射到一个低维的特征空间，以学习出特征空间中每个节点稠密、连续的低维分布表征. 新的节点表征一般被认为可保持节点最本质的结构特性，同时去除了噪声和冗余信息. 这样的新型节点表征在网络分析任务中主要有如下 3 点优势：(1) 低计算复杂度；(2) 高并行性；(3) 对于传统机器学习算法的良好适用性[21-22].

因此我们需进一步解决的问题就成了：如何有效地将网络嵌入提取出的节点特征建模为单点势函数，从而反映出一个节点对自己所属社团的自我认知. 令 $\theta_i(c_i;\boldsymbol{v}_i)$ 表示以特征向量 $\boldsymbol{v}_i$ 为观测条件下，节点 i 的单点势函数，其中 c_i 取值从 1 到 K. $\theta_i(c_i;\boldsymbol{v}_i)$ 衡量了给节点 i 分配社团标签 c_i 的成本，譬如，如果节点 i 更倾向于属于第一个社团而非第二个社团，我们应该有 $\theta_i(1;\boldsymbol{v}_i)<\theta_i(2;\boldsymbol{v}_i)$. 我们需要将由网络嵌入提取到的个体结构特征建模到单点势函数中. 然而，个体结构特征是节点的低维向量形式，它仍与单点势函数的形式具有很大差别. 为有效解决该问题，我们逆用式(2)中吉布斯分布给出的能量函数和概率分布之间的关系. 具体地，我们使用 γ_{c_i} 来表示节点 i 属于社团 c_i 的概率，利用 γ_{c_i} 可以将单点势函数定义为如下：

$$\theta_i(c_i;\boldsymbol{v}_i)=\frac{-\log\gamma_{c_i}}{\beta} \tag{7}$$

这里 β 表示温度，通常被设置为 1；因为 γ_{c_i} 已经是节点 i 属于各个社团的概率分布，所以不需要再次对其进行归一化. 根据文献[17]，概率分布和能量函数呈现出反比关系，即能量函数值越小，其对应的概率值越大.

我们利用模糊 C 均值算法[23](Fuzzy C-Means，FCM)来近似所有节点属于不同社团的概率分布：

$$J(\boldsymbol{U},z_1,\cdots,z_K)=\sum_{c=1}^{K}\sum_{j=1}^{n}u_{c_j}^{m}d_{c_j}^{2} \tag{8}$$

$$z_i=\frac{\sum_{j=1}^{n}u_{c_j}^{m}\boldsymbol{v}_j}{\sum_{j=1}^{n}u_{c_j}^{m}} \tag{9}$$

其中 u_{c_j} 表示节点 j 对于社团 c 的隶属度，$\boldsymbol{U}$ 代表所有节点对于每个社团的隶属度矩阵，m 表示加权指数，z_i 表示第 i 个聚类中心，d_{c_j} 表示第 c 个聚类中心与第 j 个数据点之间的欧氏距离. FCM 算法首先随机初始化每个节点的特征到各个社团的隶属度，然后根据式(9)计算每个社团的聚类中心，接着更新隶属度矩阵. 当所有隶属度变化的绝对值都低于设定阈值时，算法停止迭代，对应的隶属度即为节点属于社团的概率分布.

通过结合网络嵌入与模糊 C 均值方法，我们从网络拓扑中提取出所有节点的结构特征，进而基于吉布斯分布将其建模为在能量函数中起主要作用的单点势函数.

3.3　定义成对势函数

在马尔可夫随机场模型的能量函数中，另一个重要的部分就是成对势函数，它通常基于成对对象来定义，并且能够揭示出对象之间的约束关系. 从社团发现的角度来说，每一个成对势函数应该捕捉到一个节点对之间的耦合关系.

令 $\theta_{ij}(c_i,c_j;\boldsymbol{v}_i,\boldsymbol{v}_j)$ 表示基于节点特征 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$，节点对 i 和 j 的成对势函数，其中 c_i 和 c_j 分别表示节点 i 和 j 属于的社团. $\theta_{ij}(c_i,c_j;\boldsymbol{v}_i,\boldsymbol{v}_j)$ 衡量了分别分配社团标签 c_i 和 c_j 给节点 i 和 j 的成本. 例如，如果特征向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 之间的欧式距离非常小，说明节点 i 和 j 具有高度的结构相似性，那么它们倾向于属于相同的社团而非不同社团，即 $\theta_{ij}(c_i,c_i;\boldsymbol{v}_i,\boldsymbol{v}_j)<\theta_{ij}(c_i,c_j;\boldsymbol{v}_i,\boldsymbol{v}_j)$，此时 $c_i\neq c_j$. 为了避免全连接模型带来的高复杂度，这里我们仅为网络中有边的节点对定义成对势函数，即

$$\theta_{ij}(c_i,c_j;\boldsymbol{v}_i,\boldsymbol{v}_j)=\mu(c_i,c_j)\times\left[\frac{\omega_1}{d(\boldsymbol{v}_i,\boldsymbol{v}_j)}+\omega_2|d_i-d_j|\right] \tag{10}$$

其中 ω_1 和 ω_2 是两个权重系数，均为非负；如果 $c_i=c_j$，函数 $\mu(c_i,c_j)$ 的值为 -1，否则为 1；$d(\boldsymbol{v}_i,\boldsymbol{v}_j)$ 表示特征向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 之间的欧式距离，d_i 表示节点 i 的度. 对于社团中一条边的存在，式(10)定义的成对势函数考虑了两种可能情况. 第一种情况是如果两个节点具有高度的结构相似性，即表征它们结构特征的低维向量之间具有非常小的欧式距离，那么这对节点之间倾向于存在一条边. 如果两个节点的结构特征不相似，但是它们之间却仍然有一条边存在，本文认为这属于第二种情况，即其中一个节点的流行度非常高，吸引了另一个节点与它产生一条边. 我们用节点的度 d_i 来表示节点 i 的流行度，度值越大表示流行度越高，$|d_i-d_j|$ 的值越大，代表节点 j 受节点 i 流行度的影响越大，它

们之间越倾向于存在一条边.

本文定义的成对势函数描述了表 1 中 8 种可能的情况,即如果社团内部节点间存在一条边,那么我们认为要么是这对节点具有高度的结构相似性、要么是其流行度差异非常大,所以只要满足其中一点我们就有理由对其进行鼓励(即给一个较低的势能值),若同时满足上述两点则给予强鼓励(给一个最低的势能值),若都不满足则给予惩罚(给一个较高的势能值);同理,如果处于不同社团的两个节点之间存在一条边,那么我们认为也是上述两种原因导致的,所以只要满足其中一种就给予惩罚,同时满足两种则给予强惩罚,若都不满足则给予鼓励.基于概率分布和能量函数之间的反比关系,较低的能量值对应一个较高的概率.我们定义的成对势函数鼓励社团内部边的存在,惩罚不同社团之间边的存在,这和社团发现的目标(即划分节点到社团中,使社团内部节点连接稠密、社团间连接稀疏)是一致的.同时,由于我们只给有边相连的节点对定义成对势函数,因此该马尔可夫随机场具有与原始网络拓扑相同的结构.这种稀疏化的结构定义为后面高效的模型推断奠定了基础.

表 1 成对势函数描述的 8 种可能情况

	结构相似度	流行度差异	属于相同社团的倾向
社团内有边	高	大	强鼓励
	高	小	鼓励
	低	大	鼓励
	低	小	惩罚
社团间有边	高	大	强惩罚
	高	小	惩罚
	低	大	惩罚
	低	小	鼓励

3.4 基于信念传播的推断算法

定义好完整的模型之后,我们需要通过最大化后验概率 $P(C|\boldsymbol{A},\boldsymbol{V})$ 得到最好的社团划分结果 C.最大化联合概率能够推断出对应于联合概率分布中所有社团成员的一组最优配置,而最大化边缘概率会将节点看成是相对独立的存在.联合概率较边缘概率能够更好地刻画节点间的约束关系,而正是这种关系将有助于确定节点所属的社团类簇.因此,一般认为优化联合概率可提供更好的解,这也是我们选择最大化联合概率的原因.表 2 给出了一个示例,它展示了两个二元变量的联合概率与边缘概率值.具体而言,当 $x_1=1$ 和 $x_2=0$ 时,联合概率分布达到最大值 0.4.然而,如果最大化边缘概率 $p(x_1)$ 和 $p(x_2)$,得到的结果是 $x_1=0$ 和 $x_2=0$,它对应的联合概率值是 0.3.故最大化联合概率一般更佳.

表 2 在两个二值变量上的联合概率和边缘概率取值示例

	$x_1=0$	$x_1=1$
$x_2=0$	0.3	0.4
$x_2=1$	0.3	0.0

信念传播算法是适用于马尔可夫随机场模型的最流行的推断方法之一,它具有坚实的理论基础,并且在许多应用中展示出了良好的性能.信念传播算法具有两个版本:"加和-乘积(sum-product)"和"最大化-加和(max-sum)".其中,"加和-乘积"版本旨在计算出 MRF 模型中每个节点的边缘概率分布,"最大化-加和"版本旨在发现一组对应于联合概率最大值的节点配置.针对于文中模型,我们在信念传播"最大化-加和"版本的框架下设计和推导新模型的概率推断算法如下.

信念传播的关键就是每个节点 i 发送一条"消息"给节点 j,其中 j 是在网络拓扑中和 i 直接相连的节点(也称节点 i 的邻居).该消息表明:当不考虑节点 j 的情况下,固定节点 i 属于社团 c_i 时所对应的最大负能量值[24].我们使用 $\Psi_{c_i}^{i\to j}$ 表示在节点 j 缺席的情况下,节点 i 属于社团 c_i 的最大负能量值,它是通过递归计算 i 接收的所有其他邻居节点的消息来获得的:

$$\Psi_{c_i}^{i\to j}\leftarrow\beta\log\gamma_{c_i}+\sum_{k\in N(i)/j}\left[\max_{c_k}\left[-\beta\times\mu(c_i,c_k)\times\left(\frac{\omega_1}{d(\boldsymbol{v}_i,\boldsymbol{v}_k)}+\omega_2|d_i-d_k|\right)\right]+\Psi_{c_k}^{k\to i}\right] \tag{11}$$

这里 $N(i)$ 表示节点 i 的邻居节点集合.因为在信念传播算法的最大化-加和版本中,消息不再是一个概率分布,所以消息的归一化变为将最小值平移至 0.

当算法收敛后,可计算每个节点的最大化信念 $\mu_i(c_i)$,它表示当固定节点 i 属于社团 c_i 时,得到的最大负能量值,具体计算公式如下:

$$\mu_i(c_i)\leftarrow\beta\log\gamma_{c_i}+\sum_{k\in N(i)}\left[\max_{c_k}\left[-\beta\times\mu(c_i,c_k)\times\left(\frac{\omega_1}{d(\boldsymbol{v}_i,\boldsymbol{v}_k)}+\omega_2|d_i-d_k|\right)\right]+\Psi_{c_k}^{k\to i}\right] \tag{12}$$

为了得到对应联合最大后验概率的节点标签,我们选择每个变量 c_i 对应最大化信念时的状态:

$$\hat{c}_i=\arg\max_{c_i\in\{1,\cdots,K\}}\mu_i(c_i) \tag{13}$$

即得到了最终的社团划分结果.

3.5 算法描述及复杂度分析

由于我们的马尔可夫随机场模型 iMRF 具有和原始网络相同的拓扑结构,因此我们的推断算法可

以直接获得近线性的复杂度，而不需要再通过复杂的近似过程一步步降低时间复杂度. 下面“算法 1”中给出了完整的最大化-加和版本的信念传播描述.

算法 1.　iMRF 的最大化-加和信念传播算法.

输入：网络拓扑矩阵 $\boldsymbol{A}$，节点的特征矩阵 $\boldsymbol{V}$，社团数目 K，收敛阈值 ε，最大迭代次数 T，权重参数 ω_1 和 ω_2

输出：所有节点的社团划分结果 $\hat{c}_i$

1. 对于每条边(i,j)，随机初始化 K 维归一化向量$\{\boldsymbol{\Psi}_{c_i}^{i\to j}\}$
2. $conv\leftarrow\varepsilon+10$；$t\leftarrow 0$
3. WHILE $conv>\varepsilon$ 并且 $t<T$
4.　DO $conv\leftarrow 0$；$t\leftarrow t+1$；
5.　FOR 以随机的顺序计算每条消息 $\boldsymbol{\Psi}_{c_i}^{i\to j}$
6.　DO 通过式(11)更新所有 K 个消息 $\boldsymbol{\Psi}_{c_i}^{i\to j}$
7.　$conv\leftarrow conv+|\boldsymbol{\Psi}_{\text{new}}^{i\to j}-\boldsymbol{\Psi}_{\text{old}}^{i\to j}|$
8.　END DO
9.　END FOR
10.　END DO
11. END WHILE
12. 对于每个节点 i，利用式(12)计算它的最大化信念 $\mu_i(c_i)$
13. 使用式(11)在最大化信念中找到合适的 $\hat{C}$

由于 iMRF 模型具有和网络拓扑相同的结构，因此每次算法迭代时，每个节点只需传递 K 个消息给它的邻居节点，这样一来，每次迭代所需更新的消息总数是 $2mK$，其中 m 是网络中边的个数，K 是网络中的社团总数. 此外，更新一条消息的复杂度可被视为常数，即 $O(1)$，因此，算法每次迭代的复杂度是 $O(2mK)$. 考虑到初始阶段所有的消息需要被初始化，这需要 $O(2mK)$ 的时间，同时算法将会在 T 次迭代后收敛到一个固定点，所以算法 1 的复杂度是 $O(2mK+2TmK)$. 由于本文用于提取节点特征矩阵的网络嵌入方法 Node2Vec 之时间复杂度为 $O(n)$，又有 $n<m$，所以这里只需考虑算法 1 的复杂度 $O(2mK+2TmK)$ 即可. 另，由于本文最大化-加和版本推断算法采用的是异步消息传递方式，每次传递的都是最新消息，因此它减少了传播延迟；同时，算法还引入了震荡系数，使其能更加快速收敛，从而缓解了算法针对大规模网络收敛速度慢的问题. 故，算法迭代次数 T 一般可视为常数. 因此，文中算法的复杂度亦可表示为 $O(mK)$，即对于大规模稀疏网络可达到近线性.

4　实　验

这里主要是验证文中提出的 iMRF 方法的性能，其是否比已有的 6 个统计模型类社团发现方法有所提升，尤其是对比 NetMRF 方法. 我们在两组被广泛使用的人工网络和 20 个真实网络上进行实验. 除了我们自己提出的马尔可夫随机场方法 NetMRF，其他对比算法包括：(1) 度修正随机块模型 Karrer[13]，(2) 矩阵分解方法 SNMF[25]，(3) BNMTF[26]，(4) MNDP[27]，(5) 深度学习方法 DNR[16]. 当网络中的社团结构已知时，我们利用标准化互信息 NMI(Normalized Mutual Information) 和准确率 AC(Accuracy)[28] 作为评价指标；当社团结构未知时，我们利用模块度函数(Modularity Q)[8] 作为评价指标. 为公平比较，我们使用作者提供的源代码和默认参数. 但对于深度学习方法 DNR，它的结果对于参数十分敏感，通常需要花费很长的时间来调参以获取最好的结果，因此，这里我们仅使用作者在原文中报道的实验结果[16]. 在实验中，我们采用了 Dell 台式机(Intel® Core™ i5 2.70 GHz 处理器、8 GB 内存)、Windows 10 操作系统和 Matlab 2016 运行环境.

此外，本文采用了一个经典的网络嵌入方法：Node2Vec，来提取网络中节点的结构向量表征，并将其作为模型的观测 V. Node2Vec 针对目前相关研究中对网络节点邻居的定义不够灵活之缺陷，提出并充分利用了“邻居节点是有效学习节点表征的关键”. Node2Vec 设计了一个有偏的随机游走过程，以捕捉包含丰富结构信息的邻居节点. 通过最大化保持每个节点网络邻居的似然函数，学习出节点从拓扑空间到低维特征空间的映射. 在实验中，我们使用了作者提供的源代码，并且按照默认值设置了随机游走的步长($l=80$)和窗口的大小($k=10$).

4.1　性能评价指标

在社团发现性能评估的应用中，互信息 MI(Mutual Information)通常被用来估量两个社团分布的相似度. 给定两个社团分布的集合 C(真实社团)和 $\hat{C}$(模型推断的社团结果)，它们的互信息 $MI(C,\hat{C})$ 被定义为

$$MI(C,\hat{C})=\sum_{c_i\in C,\hat{c}_j\in\hat{C}}P(c_i,\hat{c}_j)\cdot\log\frac{P(c_i,\hat{c}_j)}{P(c_i)\cdot P(\hat{c}_j)}\tag{14}$$

其中 $P(c_i)$ 和 $P(\hat{c}_j)$ 分别表示节点 i 属于社团 c_i 和节点 j 属于社团 $\hat{c}_j$ 的边缘概率，$P(c_i,\hat{c}_j)$ 表示它们的联合概率. $MI(C,\hat{C})$ 的取值范围是从 0 到 $\max(H(C),H(\hat{C}))$，这里的 $H(C)$ 和 $H(\hat{C})$ 分别是 C 和 $\hat{C}$ 的熵. 当两个社团划分完全相同时 $MI(C,\hat{C})$ 取最大值，反之，当它们完全独立时取值为 0. $MI(C,\hat{C})$ 的一个重

要的特性是对于不同标签排列其值不变.在实验中,我们使用归一化的 $MI(C,\hat{C})$ 度量,即 Normalized NMI(NMI),其取值范围为从 0 到 1.

准确率 AC(Accuracy)常用来估量标签的准确性.给定一个包含 n 个节点的网络,对于每个节点,$\hat{C}$ 是我们通过一个模型得到的社团标签,而 r_i 是标签已知的真实值,那么准确率定义为

$$AC=\frac{\sum_{i=1}^{n}\delta(r_i,\mathrm{map}(\hat{c}_i))}{n} \tag{15}$$

其中 $\delta(x,y)$ 仅当 $x=y$ 时为 1,否则为 0.$\mathrm{map}(\hat{c}_i)$ 是一个映射函数,将每个社团标签 $\hat{c}_i$ 映射到与其相对应的真实值.

另外,当真实的社团分布未知时,我们采用著名的模块度函数 Q 作为算法性能的度量标准[8,29].一个包含 K 个社团的划分是一个标签集合 $\{c\}$,其中 $\hat{c}_i\in\{1,\cdots,K\}$ 是节点 i 所属的簇.一个包含 n 个节点和 m 条边的网络划分 $\{c\}$ 的模块度函数定义如下:

$$Q(\{c\})=\frac{1}{m}\left(\sum_{\langle ij\rangle\in\varepsilon}\delta_{c_ic_j}-\sum_{\langle ij\rangle}\frac{d_id_j}{2m}\delta_{c_ic_j}\right) \tag{16}$$

其中 ε 是边的集合,度 d_i 是节点 i 的邻居节点数目;$\delta_{c_ic_j}$ 是 Kronecker 函数,其仅当 $c_i=c_j$ 时取值为 1,否则为 0.这时,模块度 Q 的物理含义即为:网络中同一社团内节点的边数占网络总边数的比例,减去相同节点数相同社团划分时其期望边数的比例.如果社团内的边数与随机值相同,则 $Q=0$;对于最强的社团结构可有 $Q=1$.

4.2 人工网络上的性能评估

我们在两种不同类型的人工网络上测试文中 iMRF 算法的性能,它们分别为 Girvan-Newman(GN)人工网络[6]和 LFR 人工网络[30],这两种网络的社团结构都是已知的,因此可采用精度来评价不同算法性能的优劣.

4.2.1 GN 人工网络

Girvan-Newman(GN)人工网络[6]是被随机生成的,每个网络中包括 128 个节点,被平均分到 4 个社团中.每个节点的平均度值为 16,即 $z_{\mathrm{in}}+z_{\mathrm{out}}=16$,其中 z_{in} 表示在同一社团内连接到目标节点的边数;而 z_{out} 表示其它社团中节点连接到目标节点的边数.在下面的实验中,随着平均外度 z_{out} 的范围从 0 变化到 12,网络的社团结构变得越来越模糊.尤其是,当 $z_{\mathrm{out}}=12$ 时,网络中没有任何社团结构.

图 1(a)和(b)分别展示了在 NMI 和 AC 两种评价指标下,iMRF 与 Karrer、SNMF、BNMTF、MNDP、DNR、NetMRF 算法的实验比较.正如我们看到的,在 GN 网络上,我们的 iMRF 算法相对于其他对比算法表现出了明显的优势.具体地,采用 NMI 度量,当 z_{out} 的值从 8 变化到 10,iMRF 高出第二名 NetMRF 约 5%;采用 AC 度量时也具有相似的趋势.整体来说,文中 iMRF 算法在所有对比算法中具有最高的精度,这也说明了 iMRF 具有良好的性能,更加适合于解决社团发现问题.

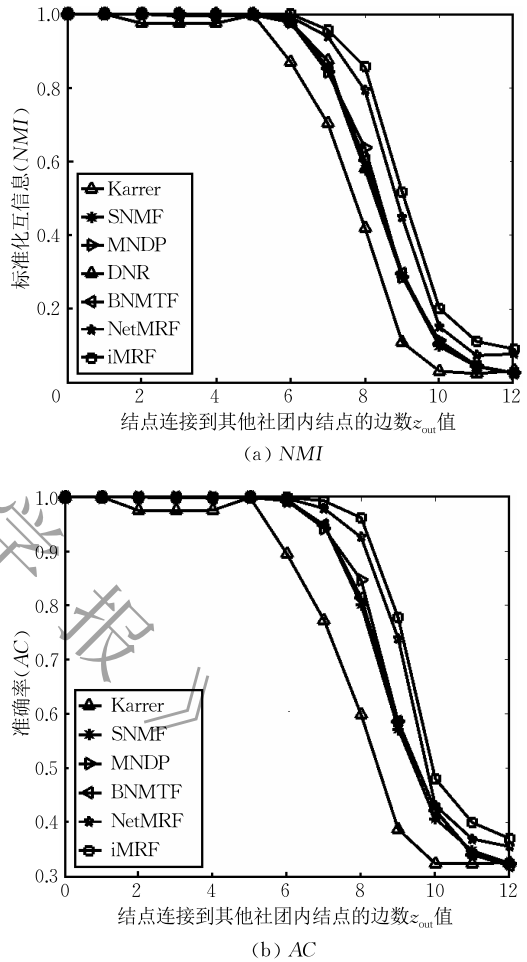


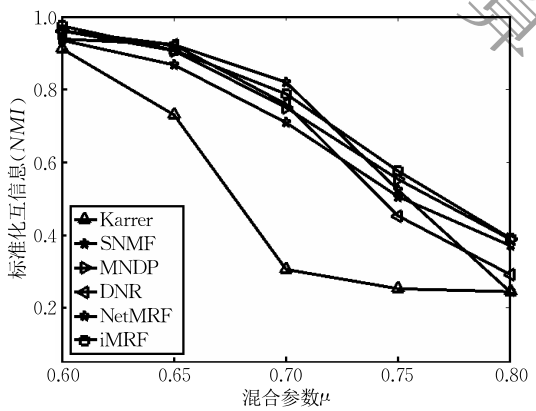
图 1 在 GN 人工网络上采用(a) NMI 和(b) AC 作为度量标准,不同社团发现方法的比较

4.2.2 LFR 人工网络

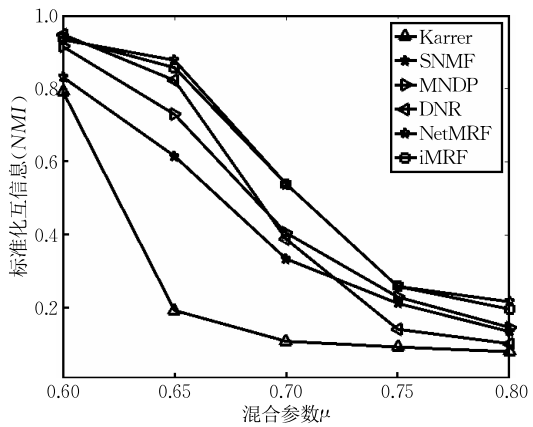
进一步,我们在 LRF 人工网络上比较不同社团发现方法的性能.LFR 模型由 Lancichinetti 等人提出[30],生成的网络具有无标度特征的节点度分布和社团规模分布,因此更接近现实生活中的真实网络.我们使用包含 1000 个节点、最小社团规模 $c_{\min}$ 分别为 10 和 20 的两组 LFR 网络.混合参数 μ 表示每个节点与其他社团中节点共享的边所占的比例,随着

μ 值的减小，网络的社团结构变得越来越模糊. 为了使实验结果更具判别性，我们令 μ 在 0.6 到 0.8 之间取值，每次增长 0.05. 剩余的参数值保持固定：节点的平均度值 $d=20$，最大节点度值 $d_{\max}=2.5\times d$，最大社团规模 $c_{\max}=5\times c_{\min}$，节点度值的幂律分布系数为 $\tau_1=-2$，社团规模的幂律分布系数为 $\tau_1=-1$. 由于采用 *AC* 和 *NMI* 作为精度度量具有相似的结果，这里我们仅展示出了采用 *NMI* 指标的结果. 另，由于 BNMTF 方法每次运行都不能在 100 h 内完成，因此我们没有对比它的实验结果.

图 2 展示出随着混合参数 μ 的变化，不同方法在 *NMI* 指标上的实验结果. 如图 2(a)所示，我们的 iMRF 方法在大多数情况下(混合参数 $\mu=0.6$、0.75 或 0.8 时)均呈现出了最优的 *NMI* 精度，其在剩余参数设置下($\mu=0.65$ 或 0.7)也表现出了很有竞争力的结果. 在图 2(b)中，我们的方法也体现出了相似的领先趋势. 另外，根据图 2 还可以看出，iMRF 方法比 NetMRF 方法的表现更加稳定.



(a) $n=1000$, $c_{\min}=10$, $c_{\max}=50$



(b) $n=1000$, $c_{\min}=20$, $c_{\max}=100$

图 2　在 LFR 人工网络上不同算法的性能比较((a)针对节点数 $n=1000$、最小社团规模 $c_{\min}=10$ 的 LFR 网络；(b)针对节点数 $n=1000$，最小社团规模 $c_{\min}=20$ 的 LFR 网络. 图中每个点都是 50 次实验结果的平均值)

4.3　真实网络上的性能评估

为了进一步评价 iMRF 模型的性能，我们继续在 20 个真实网络上开展对比实验，真实网络可能与人工网络具有不同的拓扑性质，因此可能在与人工网络上表现出不同的对比结果. 我们仍采用 Karrer、SNMF、BNMTF、MNDP、DNR 和 NetMRF 作为对比算法. 其中 NetMRF 方法包含了其未加速版本(记作 Original-NetMRF)和加速版本(记作 NetMRF)，它们的复杂度分别是节点数的立方级和近线性. 这 20 个真实网络中有的具有社团标签、有的无社团标签，因此我们对其分别开展实验.

4.3.1　已知社团结构的真实网络

首先，我们在已知社团结构的 12 个真实网络上评估 iMRF 模型的性能，这些网络的基本信息可参见表 3，更多细节①可在文献[31]中获得. 需要注意的是，表 3 中的“High school friendship6”和“High school friendship7”是两个具有相同拓扑结构的网络，但因其具有两种不同的社团结构而被作为两个网络来处理. 为了评价在上述真实网络上的对比实验结果，我们继续采用 *NMI* 和 *AC* 作为评价指标. 另外，DNR-L2 和 DNR-CE 是深度学习方法 DNR 的两个版本[16]，前者是采用 L2 范数的 DNR 方法，后者是采用交叉熵距离的 DNR 方法.

表 3　已知社团结构的真实网络介绍

数据集	节点数	边数	社团个数
Zachary's Karate club	34	78	2
Dolphin social network	62	160	2
High school friendship6	69	220	6
High school friendship7	69	220	7
Political books	105	441	3
American college football	115	613	12
Political blogs	1490	16717	2
Cora	2708	5429	7
UAI2010	3363	45006	19
Northeastern	13882	381935	7
PubMed diabetes	19729	44338	3
Maryland	20871	744862	7

表 4 和表 5 给出了对比实验结果. 我们观察到，采用 *NMI* 和 *AC* 两种度量指标，iMRF 在这 12 个网络中的 10 个网络上均得到了最好的结果. 具体的，采用 *NMI* 度量，iMRF 平均比 Karrer、SNMF、BNMTF、MNDP、DNR、Original-NetMRF 和 NetMRF 分别提升了 12.26%、8.97%、9.45%、6.95%、7.43%、5.50%、2.51% 和 6.50%；采用 *AC* 度量时也呈现出了相似的趋势. 这些实验结果表明了 iMRF 相比

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

其他统计模型在发现社团结构方面的优越性.同时,iMRF 比我们 2018 年提出的马尔可夫随机场方法(包括其原始版本 Original-NetMRF 和加速版 NetMRF)也具有更好的性能,这进一步验证了本文针对 NetMRF 模型能量函数不完整之缺陷的改进是有效的,以及单点势函数的引入增强了新模型对网络社团结构的捕捉能力,从而提高了模型精度.

表 4 在 12 个具有已知社团结构的真实网络上,采用 *NMI* 度量与 6 个代表性方法进行比较
("N/A"表示 DNR 方法未给出结果,"—"表示运行时间超过 100 h)

数据集/*NMI*(%)	方法								
	Karrer	SNMF	BNMTF	MNDP	DNR_{L2}	DNR_{CE}	Original	NetMRF	iMRF
Karate	83.72	**100.00**	**100.00**	**100.00**	**100.0**	**100.0**	**100.00**	**100.00**	**100.00**
Dolphin	88.88	81.41	81.41	88.88	88.9	81.8	75.32	54.44	**88.88**
Friend6	77.02	78.64	71.22	79.30	88.8	92.4	**96.07**	93.98	92.70
Friend7	85.10	82.11	84.30	84.26	90.7	93.2	**93.95**	93.24	**93.95**
Polbooks	54.20	56.48	51.18	53.01	55.2	58.2	**67.21**	56.88	64.68
Football	87.06	90.38	92.42	92.42	92.7	91.4	92.69	92.42	**92.77**
Polblogs	45.68	70.95	70.78	71.07	38.9	51.7	73.31	71.83	**74.21**
Cora	17.06	24.72	26.08	33.99	N/A	N/A	39.27	37.24	**42.80**
UAI2010	20.98	23.24	21.68	25.01	N/A	N/A	27.62	25.76	**34.15**
Northeastern	49.13	38.66	—	40.67	N/A	N/A	52.30	45.24	**52.90**
PubMed	12.28	13.80	—	14.96	N/A	N/A	17.69	16.89	**27.01**
Maryland	10.88	11.04	—	12.07	N/A	N/A	13.51	13.11	**15.03**

表 5 在 12 个具有已知社团结构的真实网络上,采用 *AC* 度量,与已有其他类型的方法进行比较
(这里没有将 DNR 方法作为对比算法,因为其原文中里没有记录这些结果)

数据集/*AC*(%)	方法						
	Karrer	SNMF	BNMTF	MNDP	Original	NetMRF	iMRF
Karate	97.06	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Dolphin	98.39	96.77	96.77	98.39	95.16	82.27	**98.39**
Friend6	81.16	78.26	60.87	78.26	**97.10**	95.35	95.10
Friend7	94.20	88.41	89.86	89.86	**95.65**	**95.65**	**95.65**
Polbooks	82.86	80.95	69.52	81.90	**88.57**	83.81	85.76
Football	84.35	87.83	91.30	91.30	90.43	91.30	93.24
Polblogs	87.18	94.69	94.61	94.69	95.34	95.01	**96.88**
Cora	37.70	42.25	40.95	44.39	60.22	58.05	**64.03**
UAI2010	27.78	28.52	25.51	28.92	33.25	31.14	**37.38**
Northeastern	66.36	58.87	—	56.40	68.87	65.11	**71.05**
PubMed	53.64	52.87	—	50.72	55.53	55.53	**66.50**
Maryland	50.24	46.49	—	58.23	58.23	53.64	**62.23**

4.3.2 社团结构未知的真实网络

在真实世界中,我们通常不知道网络中所包含的社团结构.因此这里我们进一步采用未知社团结构的真实网络,进一步评估 iMRF 方法的性能.这 8 个网络的基本信息见表 6,更详细的介绍①可参见文献[32].因为这里不知道真实的社团结构,而所有对比算法需要社团数目作为输入,因此这里我们使用 Louvain方法[33] 分别估计这些网络的社团个数,并将其用于所有对比算法.另外,我们使用模块度指标 Q 来评价不同算法在这些网络上的实验结果.这里没有将 DNR 方法作为对比算法,主要是由于其原文中没有记录这些结果.

表 6 未知社团结构的真实网络.

数据集	节点数	边数	社团个数
Les Miserables	77	254	6
Word adjacencies	112	425	7
Jazz musicians collaborations	198	2742	4
C. Elegans neural	297	2148	5
E. coli metabolic	453	2025	10
E-mail network URV	1133	5451	11
Power grid	4941	6594	39
Word association	5018	55234	12

表 7 给出了实验结果.根据模块度指标 Q,iMRF 方法在这 8 个网络中的 5 个网络上都展现出了最好的结果,在剩余的 3 个网络上均排名第二,并与最好结果也非常接近.采用模块度 Q,iMRF 方法平均比 Karrer、SNMF、BNMTF、MNDP、Original-NetMRF 和 NetMRF 分别提高了 0.2142、0.0487、0.0681、0.0447、0.0205 和 0.0043;又由于模块度 Q 的取值一般都在 0.3 到 0.8 范围内[8],因此 iMRF 方法较其他对比算法的性能改进是明显的.

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

表 7　在 8 个未知社团结构的真实网络上与其他方法的实验比较(“—”表示运行时间超过 100 h)

数据集/模块度 Q	方法						
	Karrer	SNMF	BNMTF	MNDP	Original	NetMRF	iMRF
Les Mis	0.4575	0.5453	0.5487	0.5434	0.5434	**0.5600**	**0.5600**
Adjnoun	−0.1041	0.2672	0.2634	0.2712	0.2712	0.2813	**0.2873**
Jazz	0.3696	0.4348	0.4347	0.4377	0.4377	0.4495	**0.4520**
Neural	0.2617	0.3701	0.3689	0.3811	0.3811	0.4120	**0.4213**
Metabolic	0.2656	0.3879	0.3834	0.3796	0.3796	**0.4579**	0.4417
E-mail	0.5126	0.5007	0.4685	0.5154	0.5154	0.5712	**0.5787**
Power	0.1796	0.8649	0.7212	0.8683	—	**0.9266**	0.9242
Word	**0.4595**	0.3546	—	0.3613	0.3613	0.4226	0.4500

4.4　运行效率比较

我们进一步以运行时间为单位来验证 iMRF 算法的效率. 首先,我们在上面被使用过的具有不同规模的 20 个真实网络上进行实验,并令 iMRF 与 Karrer、SNMF、BNMTF、MNDP、Original-NetMRF、NetMRF 方法进行比较. 对比算法中不包括 DNR 方法是因为原文中没有给出效率结果.

表 8 给出了实验结果. 首先分析 iMRF 与 NetMRF 模型的非加速版 Original-NetMRF 及加速版 NetMRF 的对比结果. 如表 8 所示,在大多数网络中,iMRF 的运行速度都领先于 Original-NetMRF 和 NetMRF,具有明显的效率优势. 这表明 iMRF 对于原马尔可夫随机场模型 NetMRF 的第二项改进具有显著效果,即新的成对势函数不仅对于社团结构具有良好的描述能力,而且使模型结构简化,便于高效推断. 我们进一步令 iMRF 与其他对比算法进行比较,采用粗体表示最好结果,同时粗斜体表示排名第二的结果. 正如我们看到的,尽管 iMRF 在小规模网络上会花费多一点儿的时间,但是在大规模网络上其均能够保持排名领先,相比于其他基于统计模型的社团发现算法仍具有效率优势.

表 8　在所有真实网络上与其他算法的运行时间比较(“—”表明运行时间超过 100 h)

数据集/方法	运行时间/s						
	Karrer	SNMF	BNMTF	MNDP	Original	NetMRF	iMRF
Karate	***0.006***	**0.004**	0.447	0.009	0.077	0.012	0.028
Dolphin	***0.010***	**0.006**	0.908	0.049	0.100	0.024	0.034
Friend6	0.061	**0.008**	3.108	0.089	0.894	0.062	***0.058***
Friend7	***0.068***	**0.009**	3.522	0.097	0.981	0.076	0.071
Polbooks	***0.039***	**0.010**	2.614	0.091	0.697	0.071	0.051
Football	0.510	**0.036**	11.28	0.242	5.796	0.222	***0.201***
Polblogs	3.458	**1.399**	1053	17.10	77.54	12.20	***2.847***
Cora	114.6	**8.019**	11.203	267.9	1666	56.40	***18.44***
UAI2010	891.9	**28.66**	64343	392.2	10192	133.7	***46.70***
Northeastern	3311	***197.7***	—	4641	25965	1962	**100.2**
Pubmed	2238	***666.7***	—	14793	25885	794.0	**119.9**
Maryland	25421	***8139***	—	24187	41778	18657	**9484**
LesMis	0.059	**0.018**	3.46	0.058	0.987	***0.045***	0.084
Adjnoun	0.352	**0.024**	6.74	0.390	1.592	0.117	***0.091***
Jazz	0.443	**0.075**	10.85	***0.220***	6.580	2.540	1.166
Neural	***0.806***	**0.098**	27.22	2.150	14.66	5.778	1.886
Metabolic	4.416	**0.639**	157.0	7.530	25.77	10.25	***3.254***
E-mail	58.32	**2.627**	1272	12.60	150.57	18.58	***8.912***
Power	3312	**34.58**	67757	258.6	7666	455.9	***249.8***
Word	1982	**98.03**	204196	378.3	45445	884.9	***71.51***

进一步,我们还利用人工网络,给出不同算法随网络规模变化时时间增长的趋势. 我们仍采用 Girvan-Newman(GN)网络生成器,但这里变化的是 GN 网络的规模,即节点数 n 被分别设置为 5000、10000、20000、40000 和 50000,依次增加;其余参数的设置与 4.2.1 节中生成 GN 网络的设置相同. 算法 iMRF、Karrer、SNMF、BNMTF、MNDP、Original-NetMRF 和 NetMRF 在上述网络上的运行时间如图 3 所示. 整体来说,我们算法 iMRF 的运行速度是最快的. 同时,iMRF 的运行时间与网络规模基本亦呈正比关系,这进一步表明其具有近线性时间复杂度,并适合于处理大规模网络.

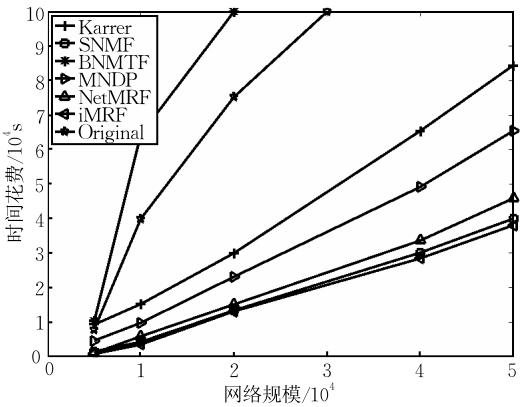


图 3 不同算法在不同规模 GN 网络上的运行时间比较

5 相关工作

近年来,复杂网络社团检测已吸引了许多来自不同研究领域研究者的关注.目前已提出了许多基于不同理论和技术的方法[3],它们包括谱聚类[4-5]、层次聚类[6]、启发式方法[7]、模块度优化[8-9]、动力学方法[34]和统计模型推断[11]等,一个较全面的介绍可参考文献[3].在这些相关工作中,基于统计模型的方法被认为是一类非常有前景的技术,并且正在被非常积极地研究.

目前,基于统计模型的方法大致可被分为三类.第一类主要是使用或者扩展著名的随机块模型、采用似然最大化进行模型优化的方法.例如,Karrer 等人[13]基于"在模型中保持节点度分布"的思想,提出了一个度修正的社团发现随机块模型.Jin 等人[35]提出了一个通过建模和排名节点的流行度来发现重叠社团的块模型.Ball 等人[36]通过扩展寻找节点社团的随机块模型来检测链接社团,并提出了一种新的期望最大化算法来学习模型参数.基于统计模型的第二类方法是利用非负矩阵分解框架来解决社团发现问题.例如,Wang 等人[25]采用平方损失函数,利用对称非负矩阵分解来最小化损失函数.Psorakis 等人[37]采用通用的 KL 散度作为损失函数,并利用贝叶斯非负矩阵分解模型来提取网络中的重叠社团.Zhang 等人[26]通过去掉"每个节点属于不同社团的概率之和必须为 1"的归一化约束,使得模型能够更好地探测重叠社团结构.第三类是基于深度学习的社团发现模型,譬如 Yang 等人[16]提出了一个基于深度自编码器的网络社团表征模型,在学习出大规模网络非线性表征的同时获取社团结构.

马尔可夫随机场 MRF 是一类有效的统计建模技术[17],它能够很好地表征复杂问题下特定任务的性质和约束关系.然而,目前基于统计模型的社团发现方法大都仅关注于有向概率图模型,而已被成功用于(与社团发现问题相相似的)图像分割问题的 MRF 却极少被用于社团发现领域.2018 年,我们提出了一个网络导向的成对 MRF 模型 NetMRF,将 MRF 有效用于社团发现领域[18].但它仍存在如下问题:(1)该模型的能量函数不够完整,缺少刻画节点个体特征的单点势函数;(2)该模型定义了一个十分复杂的全连接成对势函数结构,降低了其对于大规模网络的泛化能力.

为了将 MRF 更好地应用于社团发现任务,本文针对上述问题对 NetMRF 算法进行改进,提出了一个新的成对马尔可夫随机场模型 iMRF(improved MRF).新模型不仅可通过引入网络嵌入解决 NetMRF 势能函数不完整的缺陷,且可通过对成对势函数结构的稀疏化来缓解其效率不高的问题,从而构建了一个高精度、近线性的 MRF 新模型.

6 总结与展望

本文针对我们 2018 年提出的 NetMRF 模型[18]之缺陷进行了改进,提出了一个新的基于马尔可夫随机场的社团发现模型 iMRF.我们利用网络嵌入技术抽取节点的内在分布表征,并基于吉布斯分布将其定义为单点势函数,以获得模型对于社团结构的全局描述能力,在模型中起主导作用.进而,我们发现:一条边的存在或是由于一对节点具有高度的结构相似性,或是由于其中一个节点的流行度非常高,从而吸引了另一个节点.基于上述两点,我们为网络中有边相连的节点对精心定义成对势函数,以获得模型对社团结构的局部描述能力,在模型中起精细的微调作用.在两组人工网络和 20 个真实网络上的实验结果表明,iMRF 模型对于 NetMRF 模型的两点改进取得了显著效果;另外,和其他基于统计模型的社团发现方法相比较,本文提出的 iMRF 更适合于解决社团发现问题.

自动确定社团数目(即模型选择)是统计模型类社团发现方法所面临的一个开放性问题.对于未知社团个数的网络,文中模型 iMRF 可针对不同的社团个数 K 进行多次运行(例如,在 $K_{\min}$ 到 $K_{\max}$ 的范围内),然后寻找最小能量值所对应的 K 值,并将其作为最终社团数目.然而这种扫描式的模型选择方

法一般都比较费时.为了解决该问题,我们在实验中主要利用了著名的 Louvain 方法来确定社团个数.在未来的工作中,我们将进一步关注于如何基于模型本身快速获得社团个数,即如何将模型选择与社团发现同时建模到马尔可夫随机场框架之下.譬如,我们可首先设置一个较大的 K 值,然后利用贝叶斯模型选择思想,为候选社团添加合适的先验信息,使得在优化过程中无关的社团就能够被自动过滤掉,以实现快速自动确定社团个数的目的.

参　考　文　献

[1] Newman M E J, Peixoto T P. Generalized communities in networks. Physical Review Letters, 2015, 115(8): 088701

[2] Li Hui-Jia, Li Ai-Hua, Li Hui-Ying. Fast community detection algorithm via dynamical iteration. Chinese Journal of Computers, 2017, 40(4): 970-984(in Chinese)
(李慧嘉, 李爱华, 李慧颖. 社团结构迭代快速探测算法. 计算机学报, 2017, 40(4): 970-984)

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

[4] Li Y, He K, Bindel D, Hopcroft J E. Uncovering the small community structure in large networks: A local spectral approach//Proceedings of the 24th International World Wide Web Conference. New York, USA, 2015: 658-668

[5] Tsung C K, Ho H, Chou S, et al. A spectral clustering approach based on modularity maximization for community detection problem//Proceedings of the Computer Symposium (ICS). Chiayi, China, 2016: 12-17

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

[7] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007, 76(3): 036106

[8] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69: 026113

[9] Newman M E J. Equivalence between modularity optimization and maximum likelihood methods for community detection. Physical Review E, 2016, 94(5): 052315

[10] Li Hui-Jia, Yan Guan, Liu Zhi-Dong, et al. A linear community detection algorithm based on dynamical system in networks. Sciennia Sinica Mathematica, 2017, 47(2): 241-256(in Chinese)
(李慧嘉, 严冠, 刘志东等. 基于动态系统的网络社团线性探测算法. 中国科学: 数学, 2017, 47(2): 241-256)

[11] Martin T, Ball B, Newman M E J. Structural inference for uncertain networks. Physical Review E, 2016, 93(1): 012306

[12] He D X, Liu D Y, Jin D. A stochastic model for the detection of heterogeneous link communities in complex networks//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Texas, USA, 2015: 130-136

[13] Karrer B, Newman M E J. Stochastic block models and community structure in networks. Physical Review E, 2011, 83(1): 016107

[14] Yang J, Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York, USA, 2013: 587-596

[15] Zhao X Y, Yang B, Liu X, et al. Statistical inference for community detection in signed networks. Physical Review E, 2017, 95(4): 042313

[16] Yang L, Cao X, He D X, et al. Modularity based community detection with deep learning//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 2252-2258

[17] Bishop C M. Pattern Recognition and Machine Learning. New York, USA: Springer, 2006

[18] He D X, You X X, Feng Z Y, et al. A network-specific Markov random field approach to community detection//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. San Francisco, USA, 2018: 306-313

[19] Wang X, Cui P, Wang J, et al. Community preserving network embedding//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 203-209

[20] Cao S, Lu W, Xu Q. GraRep: Learning graph representations with global structural information//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York, USA, 2015: 891-900

[21] Cui P, Wang X, Pei J, et al. A survey on network embedding. IEEE Transactions on Knowledge & Data Engineering, 2019, 31(5): 833-852

[22] Grover A, Leskovec J. Node2Vec: Scalable feature learning for networks//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 855-864

[23] Zhang S, Wang R S, Zhang X S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications, 2007, 374(1): 483-490

[24] Zhang P, Moore C. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. Proceedings of the National Academy of Sciences, 2014: 111(51): 18144-18149

[25] Wang F, Li T, Wang X, et al. Community discovery using nonnegative matrix factorization. Data Mining and Knowledge Discovery, 2011, 22(3): 493-521

[26] Zhang Y, Yeung D Y. Overlapping community detection via bounded nonnegative matrix factorization//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 606-614

[27] Jin D, Chen Z, He D X, Zhang W X. Modeling with node degree preservation can accurately find communities// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, USA, 2015: 160-167

[28] Liu H, Wu Z, Li X, et al. Constrained nonnegative matrix factorization for image representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1299-1311

[29] Li Hui-Jia, Li Hui-Ying, Li Ai-Hua. Analysis of multi-scale stability in community structure. Chinese Journal of Computers, 2015, 38(2): 301-312(in Chinese)
(李慧嘉, 李慧颖, 李爱华. 多尺度的社团结构稳定性分析. 计算机学报, 2015, 38(2): 301-312)

[30] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4): 046110

[31] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks: The state of the art and comparative study. ACM Computing Surveys, 2013, 45(4): 43:1-43:35

[32] Nelson D L, McEvoy C L, Schreiber T A. The University of South Florida word association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, & Computers, 2004, 36(3): 402-407

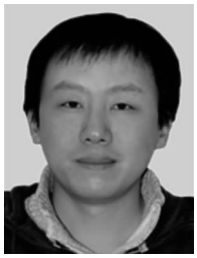
[33] Blondel V D, Guillaume J L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 1088(10): 1742-5468

[34] Li H J, Bu Z, Li A, et al. Fast and accurate mining the community structure: Integrating center locating and membership optimization. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(9): 2349-2362

[35] Jin D, Wang H C, Dang J W, et al. Detect overlapping communities via modeling and ranking node popularities// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Arizona, USA, 2016: 172-178

[36] Ball B, Karrer B, Newman M. Efficient and principled method for detecting communities in networks. Physical Review E, 2011, 84(3): 036103

[37] Psorakis I, Roberts S, Ebden M. Overlapping community detection using Bayesian nonnegative matrix factorization. Physical Review E, 2011, 83(6): 066114



JIN Di, Ph. D., associate professor. His current research interests include complex network analysis and network community detection.

YOU Xin-Xin, M. S. Her current research interests include complex network analysis and network community detection.

LIU Yue-Sen, undergraduate. His current research interest is network community detection.

HE Dong-Xiao, Ph. D., associate professor. Her current research interests include data mining and network community detection.

Background

Network which denotes the interaction relationships of units of various complex systems, is a powerful representation that can be used for analyzing the nature and function of complex systems. One of the most important properties of complex networks is community structure, in which nodes interact more densely within clusters than across clusters. Discovering communities is useful for many real applications, thus it has attracted a great number of scholars from various research fields. A wide variety of community detection algorithms have been proposed. They include hierarchical clustering, heuristic methods, modularity-based methods, spectral algorithms, dynamic algorithms, and statistical model based methods. Among these methods, statistical model based methods have a solid theoretical basis and reasonably good performance, and hence have been broadly adopted. Many statistical models have been explored and utilized to discover community structure, which include stochastic block model, nonnegative matrix factorization (NMF), and deep learning. However, most of the community detection methods which are based on statistical models only focus on directed probabilistic graphical model. Undirected probabilistic graphical model (e.g., Markov Random Field, MRF) which have been successfully used in image segmentation (similar to the community detection problem), is rarely used in community detection. In 2018, we have presented a network specific MRF approach for community detection (NetMRF). The

new method effectively encodes the structural properties of an irregular network in an energy function (the core of a MRF model) so that the minimization of the function gives rise to the best community structures.

However, NetMRF still has the following problems. (1) The energy function of NetMRF is incomplete. The unary potentials which can characterize the individual features of the node and play a dominant role in the energy function are not defined. (2) To effectively model the irregular topological information in the network, the pairwise potentials (another part in the energy function) are constructed on a complex fully-connected structure, which makes the complexity of model inference not satisfactory. To solve the above problems, we propose a new pairewise MRF model, i. e., iMRF, which means the improved MRF. We first employ network embedding methods to obtain the topological characteristics of each node, and use it to define the unary potentials using Gibbs distribution. Furthermore, we find that: the existence of an edge in a community is either due to the high structural similarity of this pair of nodes, or due to that the popularity of one node is high enough to attract another one well. Based on this, we define pairwise potentials for node pairs which are connected by links. Experimental results indicate that, iMRF can significantly improve the performance of NetMRF, and is also better than other community discovery methods compared which are based on statistical models.

This work is supported by the Natural Science Foundation of China with Nos. 61772361 and 61876128. The first is about "The Research on Accurate Semantic Community Detection in Large-Scale Complex Networks with Content" and the second is about "The Research on Community Detection in Large-Scale Networks Using Markov Random Field".