基于统计推理的社区发现模型综述

柴变芳[1,2] 贾彩燕[1] 于 剑[1]
(北京交通大学计算机与信息技术学院 北京 100044)[1] (石家庄经济学院信息工程系 石家庄 050031)[2]

摘 要 社区有助于揭示复杂网络结构和个体间的关系。研究人员从不同视角提出很多社区发现方法,用来识别团内紧密、团间稀疏的网络结构。自2006年以来,提出了一些基于统计推理的社区发现方法,它们可识别实际网络中更多的潜在结构,并以其可靠的理论基础和优越的结构识别能力成为当前的主流。该类方法的主要目标是建立符合实际网络的生成模型以拟合观测网络,将社区发现问题转化为贝叶斯推理问题。首先给出社区发现中生成模型的相关定义;其次按照模型中社区组成元素将已有统计推理模型分为节点社区推理模型和链接社区推理模型,并深入探讨各种模型的设计思想及实现算法;再次,总结各模型适用的网络类型及规模、发现的社区结构、算法复杂度等,给出一种选择已有基于统计推理的社区发现模型的方法,并利用基准数据集对已有典型统计推理模型进行验证及分析;最后探讨了基于统计推理模型的社区发现存在的主要问题和未来发展的方向。

关键词 社区发现,概率模型,随机块模型,统计推理,混合隶属度

Overview of Community Detection Models on Statistical Inference

CHAI Bian-fang[1,2] JIA Cai-yan[1] YU Jian[1]
(Institute of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)[1]
(Department of Information Engineering, Shijiazhuang University of Economic, Shijiazhuang 050031, China)[2]

Abstract Community detection can identify salient structure and relations among individuals from the complex network. Researchers put forward many different methods, which are mainly used to detect the groups with dense connections within groups but sparser connections between them. To detect more latent structures in reality networks, various models on statistical inference have been proposed since 2006, which are on sound theoretical principles and have better performances identifying structures, and have become the state-of-the-art models. These models' aims are to define a generative process to fit the observed network, and transfer the community detecting problem to Bayesian inference. First, the concepts on generation model were defined. Then, the article divided the generation models on community detection into vertex community and link community based on composition in community, and discussed design ideas and algorithms of each model in detail. What these models adapt to was also summarized from aspects of network type and scale, community structure, complexity etc, and then a method was given on how to select an existed statistical model. The existing classical models were tested and analyzed on the popular benchmark datasets. In the end, main problems on these models were highlighted, as well as the future progress.

Keywords Community detection, Probabilistic model, Stochastic block model, Statistical inference, Mixed membership

许多复杂网络系统,如社会网络、生物网络、引用网络、新陈代谢网络、信息网络、万维网等,蕴含着一些潜在的社区结构,这些社区内的节点具有相似的特性,在网络中扮演相似的角色,以其为单位的粗粒度网络描述,可简化对整个网络的功能、交互及其演化的研究,因此识别社区结构有助于我们更深入地了解网络的本质,认识网络结构与其功能之间的关系。社区发现的关键是社区的定义,不同社区结构的定义导致不同的社区发现方法。通常社区定义为团内节点连接稠密、团间节点连接稀疏的结构,下面将此类社区结构称为传统社区。实际网络中还存在这样的节点集,它们具有相似的链接模式,但它们之间是否存在稠密链接不确定,下面将此类有相似链接模式的结构称为广义社区(兼容传统社区)。社区发现的目标是利用图拓扑结构所蕴藏的信息,甚至网络节点的属性信息,从复杂网络中解析出模块化的社区结构,这是网络分析的关键。目前它在社会学、生物学、物理学和计算机科学等领域都有广泛的应用,对人们准确理解复杂系统的特性有十分重要的意义。一个好的社区发现算法要求满足这样几个标准:(1)能发现网络的各种结构规律;(2)能处理各种类型的网络,包括有向的、无向的或加权网络;(3)时间复杂度和空间复杂度不因网络规模过大而无法控制;(4)有可靠的理论基础,不

到稿日期:2011-10-02 返修日期:2012-01-15 本文受国家自然科学基金项目(61033013),北京市自然科学基金(4112046),河北省自然科学基金项目(F2008000204)资助。

柴变芳(1979—),女,博士生,讲师,主要研究领域为复杂网络分析,E-mail:11112088@bjtu.edu.cn;贾彩燕(1976—),女,博士,副教授,主要研究领域为数据挖掘、生物信息学、复杂网络分析;于 剑(1969—),男,博士,教授,主要研究领域为机器学习、图像分割等。

能仅是凭经验的启发式方法。

按照目标函数设定的依据将社区发现算法分为基于启发式度量的方法和基于统计推理的方法。前者采用启发式的度量衡量传统社区结构的优劣，缺乏严格的理论基础。为了识别网络中的广义社区结构，近来出现了一些基于统计推理的社区发现方法，该类方法可识别网络中结构对等(structural equivalence)和规律对等(regular equivalence)的结构，利用生成模型拟合观测到的网络来获得节点的划分及网络的结构，具有完备的概率理论基础和解释，能更好地满足社区发现算法的标准。

已有的基于统计推理的社区发现模型有混合模型、随机块模型 SBM(Stochastic Block Model)和混合隶属度随机块模型(mixed membership model)等。简单的 SBM 在社会科学和计算机科学中有很长的研究历史[1-3]，是当前流行的一种社区发现生成模型，但在实际网络应用中有很多局限性，不能很好地拟合模型和实际观测网络。为了克服此问题，将其扩展到指数随机图模型，但此类模型随数据规模增长引起的复杂度增长使其存在不可计算性。近年来，人们设计了一些扩展的 SBM，使其能生成更符合实际的网络。2006 年，Hastings[4] 采用的物理种植分区模型 PPM 是一种特殊的 SBM，将社区划分问题转化为统计推理问题；2008 年，Hoffman 等人[5] 采用贝叶斯方法处理该模型。2007 年，Newman 和 Leicht[6] 用混合概率模型发现网络的结构，不仅能识别传统意义的协调匹配(assortative mixing)社区，还能发现有相似链接模式的非协调匹配(disassortative mixing)社区；但 Newman 的模型限制发现的社区必须有出度，因此不能正确识别实际网络中没有出度的社区；Ramasco 等人[7] 于 2008 年解决了此问题，并定义信息熵来确定最优的社区个数。2008 年，Clauset 等[8] 提出的层次结构(Hierarchical Structure)模型使用树状随机图来表示网络的层次结构，建立一个似然模型来求解能够最佳表达网络层次结构的随机图。2008 年，Airoldi 等人[9,10] 针对关系数据的依赖性特点，将随机块模型和混合隶属度模型结合，建立混合隶属度随机块模型。2009 年，任伟等[11] 基于链接社区的思想识别边的社区隶属度，进而由边的社区隶属度计算两个端点的社区隶属度。2011 年，Karrer 等[12] 在随机块模型中融入节点度信息，得到了更好的社区结构，也证明了其比相似的模块社区发现算法有更可靠的解释、更优的结果。同年，Ball 等人[13] 基于链接社区的思想设计了一个融入边分布的随机块模型。为克服已有统计推理模型只能处理无向网络和只能进行社区硬划分等局限，2011 年底，中科院沈华伟等[14] 设计了一个扩展的随机块模型，其不仅能发现广义社区，还能获取广义社区间的关系矩阵及各社区节点的中心度。2012 年，华中科技大学段东圣等人[15] 在社区发现中考虑链接和内容属性，其用不同的变量对基于拓扑属性的网络社区和基于内容属性的文档主题建模。

自 2006 年 Hastings 提出种植分区模型以来，国内外基于统计推理的社区发现模型吸引了物理学、生物学、机器学习等多领域研究者的关注，以其坚实的理论基础和识别能力成为目前的主流方法。但该类模型的研究还处于起步阶段，许多模型只能处理中小型网络，运行效率和性能都有待提高。此前，Fortunato[16] 于 2010 年综述了统计物理学领域的各类社区发现算法，包括 2009 年之前物理学领域的几个典型的统计推理模型；国内社区发现学者[17-19] 主要对传统聚类方法、模块度方法、谱方法进行详细综述。从 2009 年至今，物理学、生物学和机器学习领域出现了许多基于统计推理的社区发现模型，但未见专门针对此类模型的综述。本文综述各领域基于统计推理的社区发现模型的研究现状以及目前面临的主要问题，详细分析各模型的原理及应用场景，并在基准数据集上通过实验比较分析各种典型模型，最后探讨基于统计推理的社区发现方法的未来发展前景及存在问题。

1 相关定义

社区发现目标是根据网络拓扑结构识别网络中功能或性质对等的潜在结构，最近流行的一种方法是将社区发现转化为统计推理问题。统计推理目标是根据观测数据和模型假设，推理数据集的特征。如果数据集是网络，建立生成模型来拟合观测网络。随机块模型是常见的一种社区发现生成模型。

定义 1(社区发现的生成模型) 生成模型是无监督学习常用的技术，它利用专家知识对输出数据做预测，将专家的知识编码到概率模型中，利用贝叶斯推理学习模型的参数来预测无标记数据。社区发现的生成模型对观测数据建立一系列假设，用假设中的潜变量和参数的配置表示潜在未观测数据，对观测数据的生成过程建模。社区发现的生成模型包括 3 部分：观测变量集 $A(\{A_{ij}\})$、潜在变量集 $q(\{q_i\})$(缺失的网络节点的指派)；基于变量 A 和 q 的联合分布及其生成式分解；与联合分布分解一致的独立假设和变量的分布假设，其会涉及一些参数$\{\theta\}$。

定义 2(简单随机块模型) 简单随机块模型认为网络 G 中每个节点 i 隶属一个社区 $g_i(1,\cdots,K)$，不同社区 g_i 和 g_j 中的节点 i 和 j 的链接概率相同，社区与社区的链接关系用 $K*K$(K 表示社区个数)维概率矩阵 W 描述，矩阵元素 Wg_ig_j 表示在 g_i 中的节点 i 与 g_j 中的节点 j 间生成边的概率。建立参数为$\{W,g,K\}$的生成模型，利用贝叶斯推理拟合观测网络，求得节点的社区指派 g 及社区的关系矩阵 W。根据 W 的不同分布假设，随机块模型能够生成不同的网络结构。对角概率矩阵可以生成独立的社区；小的非对角元素可生成传统社区结构；其他的矩阵可生成星型结构、层次结构、多分图结构等各种广义社区。

2 基于统计推理的社区发现模型

已有的基于统计推理的社区发现方法对图聚类问题建立生成模型，根据生成模型中社区构成元素的不同，将其分为节点社区(vertex community)和链接社区[20,21](link community)。节点社区模型将节点指派到各个社区。基于链接社区的模型将观测网络的边指派到各个社区。由于边的顶点可被指派到不同的社区，链接社区的思想很容易解释重叠社区现象。下面从各个生成模型的建模思想、网络生成过程、处理的网络的特征(是否有向、是否重叠)、问题求解的复杂度等多方面进行详细讨论。

2.1 基于节点社区的统计推理模型

大部分基于统计推理的社区发现模型采用节点社区的思想来划分社区，主要包括：种植分区模型 PPM(planted partition model)、混合模型 NMM(Newman's mixture model)、混

合隶属度模型 MMM(mixed membership model)、混合隶属度随机块模型 MMSB(mixed membership stochastic block model)、融合节点度的随机块模型 DCSBM(degree-corrected stochastic block model)。

2.1.1 种植分区模型

种植分区模型[22]是用来生成基准测试网络的模型，属于特殊的随机块模型，随机块矩阵的对角线元素值为 p_{in}，非对角元素值为 p_{out}，分别表示社区内节点链接概率和社区间节点链接概率。Hastings[4]设计了一种基于种植分区模型的社区发现模型；该模型假设社区内节点以 p_{in} 生成边，社区间节点以 p_{out} 生成边。依据假设生成观测网络 A 的过程如下：

(1)为网络的每个节点 i 指派社区 $q_i \in \{1,\cdots,K\}$；

(2)判断网络中的每对节点 (i,j) 是否存在边：若 $q_i = q_j$，则以概率 p_{in} 生成边，否则生成边概率为 p_{out}。

根据生成过程计算给定观测网络 A 下社区指派 $\{q_i\}$ 的概率为：

$$p(\{q_i\}) \propto \{\exp[-\sum_{(i,j)} J\delta_{q_i q_j} - \sum_{i \neq j} J'\delta_{q_i q_j}/2]\}^{-1} \tag{1}$$

该模型将社区发现问题表示为该函数的统计推理问题，用信念传播近似求解 $p(\{q_i\})$ 最大时的社区划分 $\{q_i\}$)，可用来发现无向图中非重叠的传统社区，复杂度较高，但在稀疏图上速度较快，复杂度为 $O(n\log^{\alpha} n)$，α 需要在运行过程中人为估计。

Hastings 模型需要输入参数 p_{in} 和 p_{out}，这两个参数在实际应用中不易获得。2008 年，Hofman 和 Wiggins[5]提出了一种贝叶斯框架来解决此类问题，即通过为社区分布和这两个参数设定先验分布来避免输入一些参数。该框架假设节点社区指派 π 服从 Dirichlet 分布，节点对间生成边的概率 p_{in} 和 p_{out} 服从 beta 分布。将社区发现问题转化为求 $K^* = \arg\max_k p(K|A)$ 及最大化后验分布 $p(\pi, p_{in}, p_{out}|A)$、$p(\{q_i\}|A)$ 时参数 p_{in}、p_{out} 的分布和社区分布 $\{q_i\}$。实际上对 K 的先验很少，求 $K^* = \arg\max_k p(K|A)$ 转化为求 $K^* = \arg\max_k p(A|K)$，根据生成过程和先验分布的假设写出 $p(A|K)$ 的生成式分解，利用平均场理论求解参数。该模型可以在给定范围内找到最好的 K 值，参数 p_{in} 和 p_{out} 不用输入，Hofman 的方法能自动确定网络最可能的社区个数，用贝叶斯的框架改进了 Hastings 的模型，时间复杂度为 $O(n\alpha)$($\alpha = 1.44$)。该类模型只能识别无向图中的传统社区，限制了其广泛应用。

2.1.2 混合模型

已有许多社区发现方法，需要指定社区的结构特征，如通过团内节点形成边的和大于其与团外节点形成边的和这一特征来描述社区。这些方法的各种假设都是基于传统社区的假设，致力于发现团内顶点链接紧密的“协调混合”结构。实际网络中存在一些结构对等的“非协调混合”模式，而用已有的方法发现不了这种结构。Newman 的混合模型(NMM)用来发现具有相似链接模式的节点形成的社区，能识别传统的社区，也能识别具有相似链接模式的“非协调混合”结构。该模型假设社区中的节点 i 与节点 j 有相似的链接，不关心 j 是否与 i 在同一社区。其思想类似于随机块模型，不同之处在于其没有清晰描述社区间的链接概率关系，只描述社区与节点间的关系。为描述该模型生成观测网络的过程，该模型假设：社区的节点比例服从 π 上的多项式分布，θ_{ij} 表示社区 r 中节点链接到节点 j 的概率，限制条件为 $\sum_{r=1}^{K}\pi_r = 1$，$\sum_{r=1}^{N} q_{ir} = 1$。该模型生成观测网络 A 的过程如下：

(1)为网络中的每个节点指派社区 g_i($i = 1,\cdots,N$)，以概率 π_r 指派 g_i 为社区 r($r = 1,\cdots,K$)；

(2)对每条边 (i,j)，以概率 $\pi_r\theta_{rj}$ 生成。

根据模型生成过程可得生成观测网络的似然值为 $\prod_i \pi_{g_i} \prod_j \theta_{g_{ij}}^{A_{ij}}$。利用最大似然估计可得每个节点属于各社区的概率 $\{q_{ir}\}$ 及参数 $\{\pi_r\}$ 和 $\{\theta_{rj}\}$。计算公式如下：

$$q_{ir} = \frac{\pi_r \prod_j \theta_{rj}^{A_{ij}}}{\sum_s \pi_s \prod_j \theta_{sj}^{A_{ij}}} \tag{2}$$

$$\pi_r = \frac{1}{n}\sum_i q_{ir}, \theta_{rj} = \frac{\sum_i A_{ij} q_{ir}}{\sum_i k_i q_{ir}} \tag{3}$$

式中，k_i 表示节点 i 的度。用 EM 算法迭代求得这些参数，根据参数 q 可知节点关于每个社区的隶属度，根据 θ 可知节点在各社区中的中心度。该模型可用在中小型有向、无向、加权网络的重叠社区发现中，模型设计初衷是解决非重叠社区发现，但该类算法参数 q 可用来识别重叠节点，通过 $g_i = \max_{r=1,\cdots,K}\{q_{ir}\}$ 计算每个节点 i 的隶属社区。

Newman 的模型参数 θ 的限制条件 $\sum_{r=1}^{N}\theta_{ri} = 1$，隐含每个社区要至少指向一个节点，即社区的出度不能为 0，因此不能正确识别一个节点集指向另一个节点集的二分图结构，Ramasco 等人于 2008 年[7]对其进行改进。该文献将参数 θ_{ri} 扩展为 3 种情况：社区 r 中节点指向节点 i 的概率、节点 i 指向社区 r 中任一节点的概率、社区 r 中节点与节点 i 存在双向链接的概率；并用分类平均熵推出最优的分类个数。文献[23]证明利用 Newman 的模型，可将节点对其它节点的影响度进行排序，该结果可用来识别对团结构和稳定性起作用的节点。文献[24]应用该技术解决人口层次化的问题，并设计标准确定最优的聚类数，该作者又在文献[25]中用变分贝叶斯来获取更好的分类结果。该类方法能够发现传统社区和“非协调混合”结构，但不能说明识别的是哪种结构，识别的结构不能清晰地描述网络的结构规律。其时间复杂度和空间复杂度为 $O(KN)$，可用来发现中等规模的网络社区。

2.1.3 混合隶属度模型

自 2003 年 Blei 等人的 LDA 混合隶属度模型提出后，其在文本分析、图像处理、人口学等多领域都广泛应用。在文本处理领域，如 PLSA、LDA，其依据文档的内容属性发现具有相同或相似主题的文档聚类，这里的主题类似社会网络或物理领域的社区，其是网络中相同或相似语义的单元。该类模型在链接分析或社会网络分析的应用中，可用网络拓扑属性来处理社区发现。关于文本分析和社区发现之间的联系可以参考文献[26-34]。

该类模型假设每个节点隶属多个类，用一个概率隶属度向量描述节点属于各个类的概率，各向量独立同分布，向量的各维值表示该数据隶属类的概率。观测对象隶属于多个类，相比混合模型、简单随机块模型，其更接近现实。但上述几种方法假设相同主题的节点间才能生成链接，即只能生成传统社区，而实际网络中不同社区间也可以生成链接；另外其假设各节点的混合隶属度向量相互独立，该假设不符合实际的关系数据间的交互事实。该类模型主要处理有向网络的链接分析问题，已有研究表明与节点内容融合起来可提高聚类结果，时间复杂度为 $O(N^2K)$。目前的模型只能处理中等规模的稀

疏网络的社区发现问题。

2.1.4 混合隶属度随机块模型

种植分区模型及混合模型假设网络中各节点独立、属于单个社区，而实际上许多关系网络中节点以不同的角色与其它节点交互，一种角色使其隶属一个社区。PPM，MMM，NMM 模型假设太强，不能处理实际关系网络的节点相互依赖和节点混合隶属多个社区的问题。混合隶属度模型中的数据具有潜在类混合隶属特性，使其有描述关系数据间的多种角色交互的能力，但数据独立假设不满足关系数据相互依赖的特征。文献[9,10]将混合隶属度模型和随机块模型结合，建立了混合隶属度随机块模型(mixed membership stochastic block model)，该模型将全局参数(块链接矩阵)和局部参数(链接中节点的混合隶属度)结合，以便解决成对节点的功能隶属度问题。其对有向关系数据形成的网络 $G=(N,Y)$ 建模，其中 N 为节点集合，$Y(p,q)$ 为节点对的值，即边的权重，文献中只考虑 $Y(p,q)\in\{0,1\}$ 的二值情况。该模型的参数包括：每个节点 i 隶属多个社区，其社区隶属度向量 $\theta_i\sim \mathrm{Dir}(\alpha)$；$K$ 个组间交互随机块矩阵 B；分别从交互对 (p,q) 两个端点隶属度向量抽取社区 $Z_{p->q}$ 和 $Z_{p<-q}$，若 B 中这两个社区链接概率不为 0，则生成链接。生成网络过程如下：

(1)对每个节点 $p\in N$，从 $Dir(\alpha)$ 上为 p 取样一个 K 维混合隶属度向量 θ_p；

(2)对节点 $(p,q)\in N*N$：

从多项式分布 $\mathrm{Mult}(\theta_p)$ 上为 p 选择组 $Z_{p->q}$；

从多项式分布 $\mathrm{Mult}(\theta_q)$ 上为 q 选择组 $Z_{p<-q}$；

从贝努力分布 $\mathrm{Bernoulli}((Z_{p->q})^{\mathrm{T}}BZ_{p<-q})$ 为 $Y(p,q)$ 取样，确定节点间的交互值。

根据上述假设可写出观测网络和潜在变量的联合分布概率 $P(Y,\theta,Z_{->},Z_{<-}|\alpha,B)$。

$$p(Y,\theta,z_{i\to j},z_{i\leftarrow j}|\alpha,B)=\prod_{i=1}^{N}p(\theta_i|\alpha)*\prod_{j=1}^{N}p(z_{i\to j}|\theta_i)*p(z_{i\leftarrow j}|\theta_j)*p(Y_{ij}|z_{i\to j},z_{i\leftarrow j},B) \tag{4}$$

MMSB 关于节点隶属多个社区、节点社区隶属度向量通过社区关系矩阵建立联系的假设更接近现实，在算法实现中用变分贝叶斯可快速获得节点的隶属度向量和社区的交互矩阵。该模型假设每个节点隶属多个社区，可用来识别重叠社区。缺点是为节点指派社区的思想不易扩展为层次模型；另外网络的两个节点隶属度相似度越大越易生成链接，其暗含重叠区域的边的密度高于非重叠区域边的，这在许多情况下不能反映现实网络的特征。该类模型时间复杂度为 $O(KN^2)$，适合对小规模有向网络建模。

2.1.5 融合节点度的随机块模型

已有的随机块模型忽略节点度的变化，生成的网络的度较分散，不符合实际网络度的幂律分布。文献[12]证明没有考虑度的随机块模型容易将节点度总和大的和总和小的社区归到其他社区，提出的度修正的随机块模型考虑了节点度对生成网络的影响，能够识别网络的真正结构。为了模型的简单化，假设网络包含多重边和自循环边，该假设在大的稀疏图中几乎不影响结果，但给计算带来很大便利；还假设两点间生成边的概率与边的期望值相等。令 A 表示无向多重图 G 的邻接矩阵，矩阵中的元素 A_{ij} 在 $i\neq j$ 时为两点间的实际边数，对角元素表示对应节点 i 的自循环边数的 2 倍。i 属于组 r，j 属于组 s，组间链接概率矩阵 w 的元素 w_{rs} 表示 r 和 s 间生成边 (i,j) 的概率；节点 i 期望度为 θ_i，节点 j 期望度为 θ_j，则 i 和 j 生成的边数服从期望为 $\theta_i\theta_j w_{rs}$ 的泊松分布，约束条件 $\sum_{i=1}^{K}\theta_i\delta_{g_i r}=1$。给定块矩阵 w 和节点组指派 $\{g\}$ 下生成图 G 的概率为：

$$P(G|w,g)=\prod_{i<j}\frac{(\theta_i\theta_j w_{g_ig_j})^{A_{ij}}}{A_{ij}!}\exp(-\theta_i\theta_j w_{g_ig_j})\prod_i\frac{(\frac{1}{2}\theta_i^2 w_{g_ig_i})^{A_{ij}}}{(A_{ii}/2)!}\exp(-\frac{1}{2}\theta_i^2 w_{g_ig_i}) \tag{5}$$

化简该概率分布，通过求导可得最大似然下的参数值如下：

$$\hat{\theta}=k_i/k_{g_i},\ \hat{w}_{rs}=m_{rs} \tag{6}$$

式中，k_i 为节点 i 的度，k_{g_i} 为 i 所属社区的度，m_{rs} 为社区 r 和 s 间的边数。将其代入似然函数化简得下式：

$$L(G|g)=\sum_{rs}m_{rs}\log\frac{m_{rs}}{k_rk_s} \tag{7}$$

Karrer 等人还设计一种快速的蒙特卡洛迭代算法，其每次只需复杂度为 $O(K^2)$ 的时间计算变化的似然函数值，最终收敛到似然函数最大时的指派 $\{g\}$。该模型可处理大型无向多重网络，属于非重叠社区发现算法，将其融合到重叠社区发现模型和混合隶属度模型中可以提高原模型的性能。该模型缺点是可能生成不现实的度序列，而且不能表示多尺度的社区结构，在 K 值选择上也没有提供一种好的方法。

2.2 基于链接社区的统计推理模型

链接社区在物理学和机器学习领域已有一些研究和应用，思想源于：社会网络中的节点通常属于多个社区，在不同社区其与其它节点之间存在不同类型的链接，不同的链接扮演不同的角色；同一个节点参与到网络的不同社区，整个网络就是通过这些重叠的节点联合起来的。主要的基于链接社区的统计推理模型有：对称联合链接模型 SPAEM、基于链接社区的随机块模型 SBMLC(stochastic block model for link community)、通用随机块模型 GSBM(general stochastic block model)。

2.2.1 对称联合链接模型

在文本处理领域的 PLSA[35] 是对称联合内容模型，用于发现文档集合主题的生成模型。将其扩展到链接处理上出现了 PHITS[29]，该模型认为文档间存在链接关系源于二者包含相同的主题，利用最大似然法求生成所有链接对的似然概率最大时各文档的主题概率分布参数。任伟在文献[11]中提出模型 SPAEM 是一个特殊的对称联合链接模型，认为网络中节点存在多种类型的链接，相同类型的链接存在一个社区。其采用对称的 PLSA 模型对网络数据建模。生成观测网络过程如下：

(1)以概率 π_r 选择一个社区 $r(1,\cdots,K)$，限制条件为 $\sum_{r=1}^{K}\pi_r=1$；

(2)在给定社区 π_r 条件下，为网络中的每对边 (i,j) 以概率 β_{ri} 选择节点 i，以概率 β_{rj} 选择节点 j，限制条件为 $\sum_{i=1}^{N}\beta_{ri}=1$。

根据生成过程可得生成网络的概率分布如下：

$$\mathrm{Prob}(A|\pi,\beta)=\prod_{(i,j)}(\sum_r\pi_r\beta_{ri}\beta_{rj})^{A_{ij}} \tag{8}$$

利用最大似然参数估计边的社区隶属度 $q_{ij,r}$、社区的分布 π_r、社区 r 指向节点 i 的概率 β_{ri}，参数迭代公式如下：

$$q_{ij,r}=\frac{\pi_r\beta_{ri}\beta_{rj}}{\sum_s \pi_r\beta_{ri}\beta_{rj}} \tag{9}$$

$$\pi_r=\frac{\sum_{ij} q_{ij,r}}{\sum_{ij}\sum_s q_{ij,r}},\beta_{ri}=\frac{\sum_{j:j\in N(i)} q_{ij,r}}{\sum_{k=1}^{n}\sum_{j:j\in N(i)} q_{kj,r}} \tag{10}$$

用 EM 算法迭代直到收敛，E 步求解 $q_{ij,r}$，M 步求解 π_r，β_{ri}。该方法的时间复杂度为 $O(KL)$（K 为社区个数，L 为边的个数），当边很多时较耗时，可用来处理小型的无向网络和加权网络。相比模块化社区发现方法[9]，其能更好地识别不同大小、不同度序列的非对称网络中的社区；相比 Newman 的混合模型，其发现传统社区的效果更优。SPAEM 中还提供了最优类个数选择的解决方法，用最小描述长度方案[36]能在最大似然和类个数间取得折中。

2.2.2 基于链接社区的随机块模型

链接社区的思想可以很好地处理重叠社区发现问题，全局社区发现方法又能捕获较大的社区。Ball 等人[13]提出一个基于链接社区的全局重叠社区发现统计方法。不同于已有的启发式链接社区发现方法，其不仅为链接指派社区，还通过生成模型拟合观测数据的方法，克服已有链接社区不能发现不同尺度、不同结构的社区的缺点。该模型假设网络为无向多重网络，边用 K 种颜色之一着色，每种颜色对应一种社区（在社会网络中代表节点的一种角色）。参数 θ_{iz} 表示节点 i 有 z 颜色边的倾向，两个顶点有 z 颜色边的概率越大，生成 z 社区边的概率就越大，$\theta_{iz}\theta_{jz}$ 表示 i 和 j 间 z 颜色边的期望值，自循环边的期望值为 $1/2\theta_{iz}\theta_{jz}$，实际边数服从期望值为 $\theta_{iz}\theta_{jz}$ 的泊松分布。假设独立的泊松分布的随机变量的和还服从泊松分布，则两个节点间所有颜色边的数量为 $\sum_z\theta_{iz}\theta_{jz}$，实际的边数量服从以其为均值的泊松分布。生成有邻接矩阵为 A 的图 G 的概率为：

$$P(G|\theta)=\prod_{i<j}\frac{(\sum_z\theta_{iz}\theta_{jz})^{A_{ij}}}{A_{ij}!}\exp(-\sum_z\theta_{iz}\theta_{jz})\prod_i\frac{(\sum_z 1/2\theta_{iz}\theta_{iz})^{A_{ij}}}{(A_{ij}/2)!}\exp(-1/2\sum_z\theta_{iz}\theta_{jz}) \tag{11}$$

用最大似然估计对上式进行化简，求得参数如下：

$$q_{ij}(z)=\frac{\theta_{iz}\theta_{jz}}{\sum_z\theta_{iz}\theta_{jz}},\theta_{iz}=\frac{\sum_j A_{ij}q_{ij}(z)}{\sqrt{\sum_{ij}A_{ij}q_{ij}(z)}} \tag{12}$$

用 EM 算法迭代求得参数 q 和 θ，对于大型网络使用该迭代公式会造成空间复杂度和时间复杂度过大。为了适合百万级的网络，对算法进行了改进。令 $k_{iz}=\sum_j A_{ij}q_{ij}(z)$，表示节点 i 的 z 颜色边数；$k_z=\sum_i k_{iz}$，$D=\sum_z\theta_{iz}\theta_{jz}=\sum_z(k_{iz}k_{jz})/k_z$，利用 k_{iz} 和 D 带入参数 $q_{ij}(z)$、θ_{iz} 可得：

$$\theta_{iz}=\frac{k_{iz}}{\sqrt{k_z}},q_{ij}(z)=\frac{k_{iz}k_{jz}}{Dk_z} \tag{13}$$

这样我们只需存储上次和本次的 k_{iz} 和 k_{jz} 就可以计算 E 步的 $q_{ij}(z)$ 和 M 步的 θ_{iz}，时间复杂度为 $O(2nK)$，空间复杂度为 $O(mK)$。由于大部分 k_{iz} 趋于 0，导致 $q_{ij}(z)$ 和 θ_{iz} 也趋于 0，通过两种策略来减少不必要的 0 计算：(1) 当 k_{iz} 小于阈值 δ 时将其设置为 0，相应的 $q_{ij}(z)$ 也为 0，只需计算两个端点 θ_{iz} 都不为 0 的 $q_{ij}(z)$；(2) 如果一个节点只有一个 k_{iz} 不为 0，则节点会固定为该社区，一条边的两个端点的 k_{iz} 都不改变，这条边收敛为社区 z，其对计算不再有影响，可以从计算中裁剪该边的所有变量。通过上述裁剪可以将此算法应用在大型的无向网络上。该模型通过 θ_{iz} 可计算节点属于各个社区的概率；也可用来发现非重叠社区，即选择每个节点具有最大 θ_{iz} 值的社区作为其所属社区。文献也证明了该启发式方法与融合节点度的随机块模型结果一致。

2.2.3 通用随机块模型

已有社区发现的统计推理模型的假设使其不能灵活发现不同类型网络中的社区规律，MMSB、SPAEM、SBMLC 没考虑边的方向，NMM 不能清晰捕获社区间的链接关系，DCSBM 只能识别非重叠社区。文献[15]建立了一个通用的随机块模型，在对网络规律类型没有任何先验的情况下探测网络的结构规律，能发现有向、无向网络的各种内在结构规律。其假设网络节点分为 K 个组，任意两个组 r 和 s 中的所有节点有相似的链接模式。该模型将组指派表示为隐含变量 $\{g\}$，组间的关系建模为块矩阵 W，用模型生成的网络拟合观测的网络数据 A。根据假设生成观测网络的过程如下：

(1) 对网络中的每条有向边 $\langle i,j\rangle$，以概率 W_{rs} 选择 $g_{i->j}=r$ 社区中的尾节点、$g_{j->i}=s$ 社区中的头节点；

(2) 对社区 r 中的每个节点以概率 θ_{ri} 选择节点 i，对社区 s 中的每个节点以概率 ϕ_{sj} 选择节点 j。

限制条件为 $\sum_{r=1}^{K}\sum_{s=1}^{K}w_{rs}=1$，$\sum_{i=1}^{n}\theta_{ri}=1$，$\sum_{j=1}^{n}\phi_{sj}=1$。根据生成过程得观测网络的似然值：

$$\text{Prob}(A|w,\theta,\phi)=\prod_{ij}(\sum_{rs}w_{rs}\theta_{ri}\phi_{sj})^{A_{ij}} \tag{14}$$

用最大似然法求得参数，用 EM 算法迭代求解，E 步计算 q_{ijrs}，M 步计算 w_{rs}、θ_{ri}、Φ_{sj}。公式如下：

$$q_{ijrs}=\frac{w_{rs}\theta_{ri}\phi_{sj}}{\sum_{rs}w_{rs}\theta_{ri}\phi_{sj}} \tag{15}$$

$$w_{rs}=\frac{\sum_{ij}A_{ij}q_{ijrs}}{\sum_{ijrs}A_{ij}q_{ijrs}},\theta_{ri}=\frac{\sum_{js}A_{ij}q_{ijrs}}{\sum_{ijrs}A_{ij}q_{ijrs}},\phi_{sj}=\frac{\sum_{ir}A_{ij}q_{ijrs}}{\sum_{ijrs}A_{ij}q_{ijrs}} \tag{16}$$

θ_{ri} 从出边的角度描述组 r 中 i 的中心度，Φ_{sj} 从入边的角度描述组 s 中 j 的中心度，w 描述了不同组间的交互模式，即结构规律。根据模型的参数从出边和入边的角度定义组隶属度参数 α_{ir} 和 β_{js}，前者表示作为尾节点的 i 关于组 r 的隶属度，后者表示作为头节点的 j 关于组 s 的概率，α_{ir} 和 β_{js} 计算公式如下：

$$\alpha_{ir}=\frac{\sum_s w_{rs}\theta_{ri}}{\sum_{rs}w_{rs}\theta_{ri}},\ \beta_{js}=\frac{\sum_r w_{rs}\phi_{ri}}{\sum_{rs}w_{rs}\phi_{ri}} \tag{17}$$

令节点 i 属于社区 $r=\text{argmax}_s\{\alpha_{is},s=1,2,\cdots,K\}$，从入边角度可将该模型扩展为非重叠社区发现模型，从出边角度可做类似的划分；令 θ 和 Φ 相等可将该模型扩展到无向图。该模型的时间复杂度为 $O(cTmK^2)$，c 为更好地得到近似全局最优值选择初始点的次数，T 为求解参数的迭代次数，m 为边数，K 为类数，该算法只能处理小规模网络。

3 基于统计推理的社区发现模型分析

3.1 存在的问题

基于统计推理的社区发现模型的优点包括：可用来实现重叠的和非重叠的社区发现；能发现广义社区和社区间的关系矩阵；考虑多种因素对社区结构的影响。但该类模型还处于研究阶段，存在许多问题需要解决，分析如下：

(1) 没有针对广义社区的标准网络数据集：统计推理模型的优势是能识别实际网络中有重叠的广义社区，已有基准数

据集中的实际网络主要针对传统社区发现方法设定，不存在非传统社区结构，无法用来测试能发现多种结构的统计推理模型。

(2)重叠社区评价方法有待完善：统计推理模型获取的是边或节点关于社区的隶属度，虽然已有的 NMI 和模块度函数已扩展到重叠社区度量上，但不能有效度量节点混合隶属度的准确率。

(3)没有有效地将链接社区思想与网络潜在的层次结构融合：链接社区可以有效处理和解释社区中的重叠问题，层次结构是网络中实际存在的现象，已有的统计推理模型大多数不考虑网络的层次结构，不能发现多尺度的网络结构。

(4)算法复杂度高：目前的生成模型采用 EM 算法进行参数学习，迭代次数多，每次迭代计算量大；并且 EM 算法要重复多次才能克服随机选择初值导致的局部最优问题，运行效率极低。

(5)模型的类个数选择问题：社区发现也属于图聚类问题，已有的 SPAEM、GSB 都采用最小描述长度方法来选择适当的类个数，但还需要用户指定类个数的范围。

3.2 模型定性比较

为了便于用户选择基于统计推理的社区发现算法，下面选择各类模型的几个代表算法，从多方面对其进行对比，对比结果见表 1。“时间复杂度”中的一些符号说明：N 表示网络节点个数，L 表示网络的边个数，K 表示社团个数。为了更直观地比较各模型的 E 步和 M 步的复杂度，表中的复杂度只记这两步的复杂度，实际复杂度需乘上初始点选择次数及 EM 迭代次数。

表 1 基于统计推理的社区发现典型模型比较

类别	模型	社区结构	类型	大规模	重叠	时间复杂度
节点社区	PPM	传统	无向	否	否	$O(N^a)$
	NMM	广义	全部	否	兼	$O(KL)$
	MMM	传统	有向	否	是	$O(KN^2)$
	MMSB	传统	有向	否	是	$O(KN^2)$
	DCSBM	广义	无向	是	否	$O(NK^2)$
链接社区	SPAEM	传统	无向加权	否	是	$O(KL)$
	SBMLC	广义	无向	是	兼	$O(NK)$
	GSB	广义	有向无向	否	兼	$O(LK^2)$

表 1 中各种模型的比对可辅助用户在处理一个网络的社区发现问题时，更准确地选择适当的模型。处理社区发现问题时首先要考虑发现社区结构特征，如果已有社区结构为传统社区，可以使用其它的流行方法；实际网络中不可能只包含传统的社区，因此能发现广义社区的模型应用范围更广。其次，社区发现算法要同时识别能重叠社区和非重叠社区问题，基于链接社区的思想能更好地解释重叠社区问题，非重叠社区是重叠社区的特例，因此该类模型是最佳的选择。另外，处理的网络要兼容无向网络、有向网络甚至加权网络。GSB 是同时满足上述 3 个条件的模型，其基于链接社区的思想发现有向网络(无向网络、加权网络是其的特例)中的不同结构，但没有考虑节点不同角色对生成网络的不同程度的影响；且算法的时间空间复杂度也有待提高，以适应处理大规模的网络；SBMLC 能够用来处理大型的无向图的重叠社区划分问题，将其与 GSB 的优点结合并考虑节点度可以设计更适合实际网络的社区发现算法。

3.3 模型定量对比

为了验证统计推理模型在发现传统社区和广义社团上的有效性，在实际小网络上将统计推理中的典型模型：NMM、SPAEM、GSB、PCL 和已有典型社团发现方法进行定量比对。实际网络包括：(1)无向网络：Zachary 的空手道俱乐部网络(34 个节点)、Lusseau 的海豚社会关系网络、美国大学足球队比赛关系网络(115 个节点)；(2)有向网络：美国的政治博客网络(1490 个节点)、Cora 文档引用数据集合(2708 个节点)、citeseer 文档引用数据集合(3312 个节点)；(3)二分图无向网络：词邻接网络(112 个节点)。表 2 和表 3 给出典型的统计推理模型与流行的社团发现算法划分准确率的平均值。

表 2 典型模型在无向图上的准确率比对

	kara	dolphins	football
k-means	1	1	0.9043
GN	0.9706	1	0.9043
OSLOM	0.9714	0.873	0.913
NMM	1	0.9355	0.65
SPAEM	1	1	**0.913**
GSB	1	1	0.7826

表 3 统计模型在有向图上的准确率比对

	polblogs	cora	citeseer
NMM	0.8255	0.2381	0.2052
GSB	**0.8264**	—	—
PCL	0.8205	**0.3018**	**0.2483**

根据表 2、表 3 的实验结果做如下分析：

(1)在小的无向网络 kara 和 dolphins 上，NMM、SPAEM、GSB 可得到和流行社区发现算法一样好的结果，这表明该类算法可有效发现传统社区。

(2)在 football 上 NMM、GSB 效果不如 SPAEM、k-means、GN，原因在于 SPAEM 主要用来发现传统社区，与已有方法有同样好的结果。而 NMM 和 GSB 发现的社区为广义社区，其发现的社区特点是团内节点与其他节点或其它社区有相似的链接模式，当网络结构混杂(如节点同时存在多分结构和传统社区结构中)时，NMM 和 GSB 更倾向于将同类链接模式的节点分到一个社区，传统社区发现方法更易将链接稠密的节点归到一个社区。

(3)在有向网络上的比较方面，已有的典型社区发现模型主要识别社区结构较明显的网络社区，对稀疏的网络 cora 和 citeseer 的识别准确率极低，与统计模型没有可比性，因此只对统计推理模型进行比较。NMM、GSB 可用来识别有向网络的广义社区，PCL 只能识别有向网络中的传统社区。实验结果表明这些模型在这些稀疏的实际网络上的准确率都偏低，PCL 相对好些，但也有待继续改进。

(4)NMM 和 GSB 还可用来识别网络中的“非协调混合”结构，如多部图结构。由于该类算法在稍大点的网络上的运行效率极低，只测试了在二分的词邻接网络上的社区发现准确率，测试结果可达 89%准确率，而其他传统社区发现方法只能达到 51%。

结束语 现实中许多复杂网络存在各种各样的复杂结构，挖掘这些结构能帮助我们更好地理解这些复杂系统。基于统计推理的社区发现模型试图利用网络“潜在”结构生成观测网络，利用贝叶斯推理解决社区发现问题，避免在没有任何先验知识下对网络结构做错误的假设，其以可靠的理论基础辅助社区发现的问题求解。另外，基于随机块模型扩展的方法不仅能判别节点社区指派，还能发现社区间交互规律，该特

点使我们可以通过社区交互矩阵从全局把握整个网络的结构。该类模型的研究也涉及吉布斯算法、信念传播算法、变分推理等近似求解方法。研究基于统计推理的社区发现模型近似推理不仅对社区发现有重要意义，对近似求解方法的理论研究和应用都有积极的意义。

本文总结了2006年至2012年物理学、机器学习、生物学等领域各种社区发现的统计推理模型，将其分为两大类，从多方面详细分析比较各类典型模型，并给出选择基于统计推理的社区发现模型的方法，最后在基准数据集上对典型模型的准确率做定量比较。但当前的统计推理模型仍处于研究阶段。下面给出统计推理的社区发现算法未来的研究方向：

(1)研究该类模型的参数学习算法的改进策略以提高运行效率，如用并行的方法提高运行效率，增加先验约束对原算法进行裁减以减少计算量。算法运行效率决定该类模型能否在实际中广泛应用。

(2)如何识别网络的聚类个数。已有的统计推理模型采用最小描述长度选择某个范围内最好的类个数，但该范围如何给定还需要依靠先验。非参数贝叶斯方法狄利克雷过程是一种可利用的工具。

(3)针对大型有向网络，考虑网络节点度的异质性，设计基于链接社区的统计推理模型。目前存在许多大型有向网络，网络中边的两个端点有不同角色，各节点有多重角色，已有的模型没有完全考虑这些特征，将这些特征都考虑到网络的生成模型中来提高模型的准确率。

(4)融合节点内容属性到物理学社区发现的统计推理模型中，在准确率和运行效率上达到双赢。在文本处理领域，最近也出现一些融合链接网络的内容属性和链接属性的混合隶属度模型，其假设节点独立同分布，且只能发现传统的社区，这导致其不能很好地处理网络关系数据。MMB针对这个问题做了一些工作，但只考虑网络的链接，该模型复杂度极高。将混合隶属度模型与随机块模型有效结合，并在模型中考虑节点的内容属性以提高运行效率和准确率。

(5)将链接社区思想、随机块模型、网络的层次结构有效结合，设计能识别层次结构的重叠社区发现算法，并在该类模型参数求解过程中应用和发展参数贝叶斯和非参数贝叶斯的近似求解理论。

参 考 文 献

[1] Hollandkathryn B P W, Leinhardt S. Stochastic blockmodels: First steps [J]. Social Networks, 1983, 5(2): 109-137

[2] Snijders T A B, Nowicki K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure [J]. Journal of Classification, 1997, 14(1): 75-100

[3] Nowicki K, Snijders T A B. Estimation and prediction for stochastic blockstructures [J]. Journal of the American Statistical Association, 2001, 96(455): 1077-1087

[4] Hastings M B. Community detection as an inference problem [J]. Physical Review E, 2006, 74(3): 035102

[5] Hofman J M, Wiggins C H. Bayesian approach to network modularity [J]. Physical review letters, 2008, 100(25): 258701

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

[7] Ramasco J J, Mungan M. Inversion method for content-based networks [J]. Physical Review E, 2008, 77(3): 036122

[8] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks [J]. Nature, 2008, 453(7191): 98-101

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

[10] Airoldi E M, Blei D M, Fienberg S E, et al. Mixed membership stochastic block models for relational data with application to protein-protein interactions [C] // Proceedings of the International Biometric Society. 2006

[11] Ren W, Yan G, Liao X, et al. Simple probabilistic algorithm for detecting community structure [J]. Physical Review E, 2009, 79 (3): 036111

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

[13] Karrer B B B, Newman M E J. Aan efficient and principled method for detecting communities in networks [J]. Physics Review E, 2011, 84(3): 036103

[14] Shen H W, Cheng X Q, Guo J F. Exploring the structural regularities in networks [J]. Physical Review E, 2011, 84(5): 056111

[15] Duan D, Li Y, Li R, et al. MEI: mutual enhanced infinite generative model for simultaneous community and topic detection [C]// Discovery Science. 2011: 91-106

[16] Fortunato S. Community detection in graphs [J]. Physics Reports, 2010, 486(3-5): 75-174

[17] 程学旗，沈华伟. 复杂网络的社区结构[J]. 复杂系统与复杂性科学，2011，8(1)：57-70

[18] 骆志刚，丁凡，蒋晓舟，等. 复杂网络社团发现算法研究新进展[J]. 国防科技大学学报，2011，33(1)：48-52

[19] 杨博，刘大有，Liu Ji-ming, et al. 复杂网络聚类方法 [J]. 软件学报，2009，20(1)：54-66

[20] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks [J]. Nature, 2010, 466 (7307): 761-764

[21] Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities [J]. Physical Review E, 2009, 80 (1): 016105

[22] Condon A, Karp R M. Algorithms for graph partitioning on the planted partition model [J]. Random Structures and Algorithms, 2001, 18(2): 116-140

[23] Mungan M, Ramasco J J. Stability of Maximum likelihood based clustering methods: exploring the backbone of classifications [J]. Journal of Statistical Mechanics, 2010, 4: 04028

[24] Vazquez A. Population stratification using a statistical model on hypergraphs [J]. Physical Review E, 2008, 77(6): 066106

[25] Vazquez A. Finding hypergraph communities: a Bayesian approach and variational solution [J]. Journal of Statistical Mechanics: Theory and Experiment, 2009: 07006

[26] Psorakis I, Roberts S, Sheldon B. Soft Partitioning in Networks via Bayesian Non-negative Matrix Factorization [C] // the 24th Annual Conference on NIPS. 2010

[27] Parkkinen J, Sinkkonen J, Gyenge A, et al. A block model suitable for sparse graphs [C]// Proceedings of the 7th International Workshop on Mining and Learning with Graphs. 2009

[28] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993-1022

(下转第30页)

案，由于其基于的是一般群中的指数运算，因此同上规定方法，令 S^* 和 M^* 分别表示在一般群中的单次幂运算及多次幂运算，在一般群中的幂运算代价要远大于在双线性群中的运算代价[5]，与一次配对运算代价相当。PBA-BB+方案及其它属性证明方案的计算代价如表1所列。

表1 PBA-BB+与其他方案的性能比较

方案 / 算法	PBA-CL[4]	PBA-BM[5]	PBA-BB
Issue	2M*	3S+1M+4P	4S+2P
Sign	2S* +3M*	9S+3M+4P	7S+4M+2P
Verify	2M*	4M+7P	7M+2P
Revoke	(k+3)S* + (3k+4)M*	M	2S
总体代价	(k+5)S* + (3k+11)M*	12S+9M+15P	13S+11M+6P

其次比较各方案的签名长度。定义各属性证明方案的签名长度。对于基于双线性配对的方案，系统一般选择的是椭圆曲线 $E(Fq)(|q|=170)$，所以在群 G 和群 G_T 中的元素大小分别为171bits和1020bits，因此其安全参数为：$l_p(170)$，$l_H(160)$，$l_\phi(80)$。对于本方案，PBA签名中包含一个TPM签名 σ_M、4个 G_2 中元素、一个哈希值、7个 Z_p^* 中元素。TPM签名长度一般为512bit，所以签名总长度为2717bit；同理对于PBA-BM方案，其签名长度为3397bit；对于PBA-CL方案，其签名长度[4]为9210bit。

综上可以看出，PBA-BB+在运算代价及签名长度方面均小于以上的其他方案。

结束语 本文提出了一个全新的属性证明方案，给出了具体的构造，并在随机预言模型下证明了其安全性。通过与其他属性证明方案的比较可以明显地看到，本文提出的PBA-BB方案计算代价明显小于其他的属性证明方案。在未来的工作中，将进一步提高方案的性能，使其更符合实际应用的需求。

参考文献

[1] Trusted Computing Group. TPM Main Part 1, Design Principles Specification, Version 1.2 Revision 62[EB/OL]. https://www.trustedcomputinggroup.org/home

[2] Jaeger T, Sailer R, Shankar U. PRIMA: policy-reduced integrity measurement architecture[C]// Proc. of the 11th ACM Symposium on Access Control Models and Technologies. New York, 2006: 19-28

[3] Sadeghi A, Stuble C. Property-based attestation for computing platforms: caring about properties, not mechanisms [C]// Proc. of the 2004 Workshop on New Security Paradigms. Nova Scotia: ACM, 2004: 67-77

[4] Chen Li-qun, Landfermann R, Lohr H, et al. A protocol for property-based attestation[C]// Proc. of the first ACM workshop on Scalable trusted computing. New York: ACM, 2006: 7-16

[5] Feng Deng-guo, Qin Yu. A property-based attestation protocol for TCM [J]. Science China (Information Sciences), 2010, 53(3): 454-464

[6] Boneh D, Boyen X. Short signatures without random oracles[C]// Cachin C, Camenisch J L, eds. EUROCRYPT 2004. LNCS, vol. 3027, Heidelberg, Springer Press, 2004: 56-78

[7] Ateniese G, Camenisch J, Hohenberger S, et al. Practical Group Signatures without Random Oracles [EB/OL]. Cryptology ePrint Archive, Report 2005/385, 2005, http://eprint.iacr.org/

[8] Boneh D, Boyen X, Shacham H. Short Group Signatures[C]// Franklin M, ed. Proc. of CRYPTO 2004. LNCS, vol 3152, Heidelberg, Springer Press, 2004: 41-55

[9] Pedersen T P. Non-interactive and information-theoretic secure verifiable secret sharing[C]// Feigenbaum J, ed. Proc. of CRYPTO'91. LNCS, vol 576, Berlin: Springer-Verlag, 1992: 129-140

[10] Jens G. Simulation-sound NIZK proofs for a practical language and constant size group signatures[C]// Proc. of ASIACRYPT' 2006. Shanghai, 2006: 444-459

[11] Neal K, Alfred M. Pairing-based cryptography at high security levels[C]// Proc. of the 10th IMA International Conference on Cryptography and Coding. LNCS, vol 3796, Berlin: Springer, 2005: 13-36

[12] Mao Wen-bo. Modern Cryptography: Theory and Practice [M]. New Jersey: Prentice Hall Press, 2003

[13] Poritz J, Herreweghen V, et al. Property Attestation- Scalable and Privacy-friendly Security Assessment of Peer Computers [R]. RZ 3548. IBM Press, 2004

[14] Fiat A, Shamir A. How To Prove Yourself: Practical Solutions to Identification and Signatrue Problems[C]// Proc. of CRYPTO'86. LNCS, vol 263, Berlin: Springer-Verlag, 1986: 186-194

[15] 秦宇，冯登国. 基于组件属性的远程证明[J]. 软件学报，2009，20(6)：1625-1641

(上接第7页)

[29] Cohn D, Chang H. Learning to probabilistically identify authoritative documents [C]// Citeseer. 2000: 167-174

[30] Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101: 5220-5233

[31] Nallapati R M, Ahmed A, Xing E P, et al. Joint latent topic models for text and citations [C]// ACM. 2008: 542-550

[32] Yang T, Chi Y, Zhu S, et al. Directed network community detection: A popularity and productivity link model [C]// SDM. 2010: 742-753

[33] Yang T, Jin R, Chi Y, et al. Combining link and content for community detection: a discriminative approach [C]// KDD. 2009: 927-936

[34] Yang T, Jin R, Chi Y, et al. A Bayesian framework for community detection integrating content and link [C]// BUAI. 2009: 615-622

[35] Hofmann T. Probabilistic latent semantic indexing [C]// ACM. 1999: 50-57

[36] Rissanen J. Modeling by shortest data description [J]. Automatica, 1978, 14(5): 465-471