基于连接行为驱动的合作网络模型与实证分析①

马英红，刘志远，王文倩
（山东师范大学管理科学与工程学院，济南 250014）

摘要：合作网络是由特定的群体为解决某些特定问题而构成的一种社会网络. 在合作网络中，网络成员之间建立的合作关系被称为他们之间建立了连接. 在科学研究领域常见的合作网络是由论文作者、成果完成者、专利开发者等构成的. 合作者之间建立的连接可能是通过完全随机、择优机制或者两者兼有的方式建立起来的. 构成合作关系的不同连接方式被称为连接行为. 用GR-QC数据，估算了网络中随机连接、择优连接行为的可能性，构建了基于连接行为的网络演化模型，并探讨了该模型的统计性质. 为了更好的证明网络模型的科学性，收集并分析了从2000年到2016年中国管理科学与工程学会部分核心成员发表的论文以及合作者，发现作者之间的合作网络演化符合提出的合作网络模型，并且具有与GR-QC网络相似的连接行为. 研究还发现，科学家合作网络中无论是老成员之间还是新成员与老成员之间，他们在建立连接时既有随机连接又有择优选择的行为. 数据实证分析表明，论文作者之间择优连接行为远远超过随机连接，而且老成员之间的新连接远低于新成员加入网络而产生的新连接；连接行为的差异可能是促使网络出现巨大连通分支和社团生成的影响因素.

关键词：合作网络；连接行为；优先机制；随机连接
中图分类号：N032；C931.1 **文献标识码**：A **文章编号**：1007-9807(2018)08-0083-15

0 引 言

社会网络的应用和大数据研究的快速发展使得对各种网络用户行为特征的记录和挖掘成为可能. 冯芷艳[1]等人指出，未来不仅要重视社会化网络环境中的行为机理研究，更要关注基于大数据的网络行为机理识别等问题的探讨. 杨善林[2]等人也指出，大数据可以用于社会网络分析. 社会网络是对现实世界中诸多社会成员之间关系行为的结构化的一种表示形式. 通常社会成员被作为研究对象，称为网络的节点. 研究对象之间的关系由网络中节点及节点之间的连接（或者边）来表示. 例如，创建一个新网页的时候，设计者经常在网页中建立与之相关的网页链接. 那么每个网页就是节点，网页之间的指向链接就是一条边. 随着时间的演进，已有的两个网页之间可能被重新链接或者断开链接，那么原来网络的结构就会发生变化. 随着时间的累积，网络中的这些局部连接的变化最终可能导致网络中某些宏观结构的涌现[3]. 如网络中的巨大分支的出现、网络中节点度的幂律行为等.

网络成员之间的交互作用而建立关系的连接行为被称为网络的连接行为. 无论完全随机、择优连接，还是随机和择优混合的连接方式都存在于网络连接的形成过程中. 许多研究者致力于网络中节点之间连接机制的研究. 早在1999年，Barabási和Albert提出了一类具有无标度和择优连接的网络[4]. Newman[5]等人都对网络中的择

① 收稿日期：2017-07-20；修订日期：2018-03-02.
基金项目：国家自然科学基金资助项目(71471106).
作者简介：马英红(1971—)，女，山东淄博人，博士，教授，博士生导师. Email：yinghongma71@163.com

优连接行为进行了细致的研究. 而李鹏翔等人发现只有当网络处于极端同质、极端异质情况下时，才会涌现无标度行为. 现实社会网络的度分布根本不可能与无标度的度分布或累积度分布完全吻合[6]. Kleinberg 和 Mendes 以及 Jackson 等人相继对既有择优连接又有随机连接行为的网络模型进行了研究[7-9]. 他们发现网络无标度幂律行为是一种极其少见的情况，实际的社会网络不可能只是一种择优连接或者是随机连接行为. David[10] 等人还指出，网络中的强连接能够促进网络中节点间的合作. 在这些探索中，研究者们发现无论是在理论上还是在现实的网络中，社会网络并不只是以择优连接作为增长机制，而是伴随着随机连接行为. 科学家合作网络是一类比较有代表性的社会网络. Newman 发现了科学家合作网的小世界特征和无标度性质，并且研究了科学家合作模式[11]. 也有诸多国内学者对某些类的科学家合作网络的演化、网络的统计特征进行了研究[12,13].

上述研究大多是对网络宏观层面的统计分析，而从网络中个体间连接行为的变化与网络结构之间的关系研究相对较少，这种对微观层面的探讨有助于揭示网络结构的某些差异性. 如，合作网络中巨大连通分支出现的影响因素，网络中部分节点成长为拥有很多邻居的成因等.

从 GR-QC 数据分析入手，估算了该网络中新老成员之间的连接的均值，得到了该网络中成员的度分布. 构建了基于连接行为的合作网络模型. 为验证模型的科学性和可行性，对模型的统计性质进行了研究，并以此模型与 GR-QC 网络进行了比较. 利用中国管理科学与工程领域部分学者自 2000 年到 2016 年间 17 年的论文作者数据，构建了合作网络，拟合发现了学者的度分布指数完全符合模型指数的分析. 同时发现，中国管理科学与工程领域的作者合作网络的平均度、连接行为的选择与 GR-QC 的网络极其相似. 通过这两个不同领域的现实数据与模型的拟合，揭示了合作网络中成员的连接行为具有一定的共性.

1 GR-QC 网络

GR-QC 网络取自 SNAP(Stanford network analysis project) Datasets Large Network，是 General Relativity and Quantum Cosmology Collaboration Network 的简写，称为广义相对论和量子宇宙学合作网络. 该数据描述的是广义相对论和量子宇宙学类杂志中发表的论文作者之间的合作关系. 若作者 i 和作者 j 共同合作发表论文，则在二者间连一条边；否则，两者之间没有边. 由此构造的网络是一个 0 - 1 网络. Leskovec, Kleinberg 和 Faloutsos 对该网络直径的收缩引起的网络结构的变化进行了研究[14]，但没有对其中的作者之间的连接行为进行研究.

1.1 GR-QC 网络统计特征

GR-QC 合作网络的基本拓扑特征如表 1 所示. 网络中的平均度是 5.53，即在平均意义下，网络中的每个成员的合作者是 5.53 人. 该网络具有较高的平均聚集系数和较短的平均路径，符合小世界网络的特征. 该网络的可视化见附录中的图 1 至图 4.

表 1 GR-QC 网络基本结构性质

Table 1 GR-QC network basic structural information

指标	节点总数	总边数	平均聚集系数	网络直径	平均距离
Size	5 242	14 496	0.687	17	6.049

附录的图 1 是 GR-QC 的网络的整体结构. 网络的中间部分是网络的巨大连通分支，周围散落的是一些小的合作团体. 巨大连通分支的涌现符合文献[2]中描述的社会网络演化性质. 附录的图 2 是图 1 中连通分支的中心节点的部分连接. 图中圆圈的大小表示节点的连接的边数的多少. 节点连接的边越多，圆圈越大，也就是节点的度越大. 附录的图 3 是网络中一个由 23 个节点构成的近于完全图的子图. 其中 22 个节点通过一个关键节点与整个巨大分支进行连接，该分支的合作关系是 23 个节点构成的紧密的研究团队. 该结构有助于理解合作网络的演化过程中巨大连通

分支形成过程. 附录的图 4 是图 1 中部分小合作团队的结构. 这种小的团队数量多，研究领域相对孤立，但是所有的小团队的成员的总和只占整体网络成员的 30%左右.

进一步对 GR-QC 合作网络的节点的度分布情况进行统计分析，该网络节点度分布如图 1 所示. 在这个有 5 242 位研究者的群体中，拥有 10 个以上合作者的科学家大约占总数的 10^{-4}，具有 20 个合作者的科学家大约占总数的 10^{-6}，具有 80 个合作者的科学家近于 10^{-9}. 由此，可以较为直观地发现，研究者的合作者数量 d 基本符合幂律分布.

Barabási 和 Albert[4] 证明了网络中节点的连接择优行为导致网络幂律现象的出现. 图 1 所示，网络中的节点的度分布基本呈幂律分布，但是其中有一部分数值异常. 如右图双对数图中的横轴[3,4]区间，其对应的左图中度分布大约(20,50)区间，偏离幂律. 这说明合作者的频数的分布不是完全遵循幂律，预示着网络中的连接不仅仅是择优连接，还可能有其他连接行为存在.

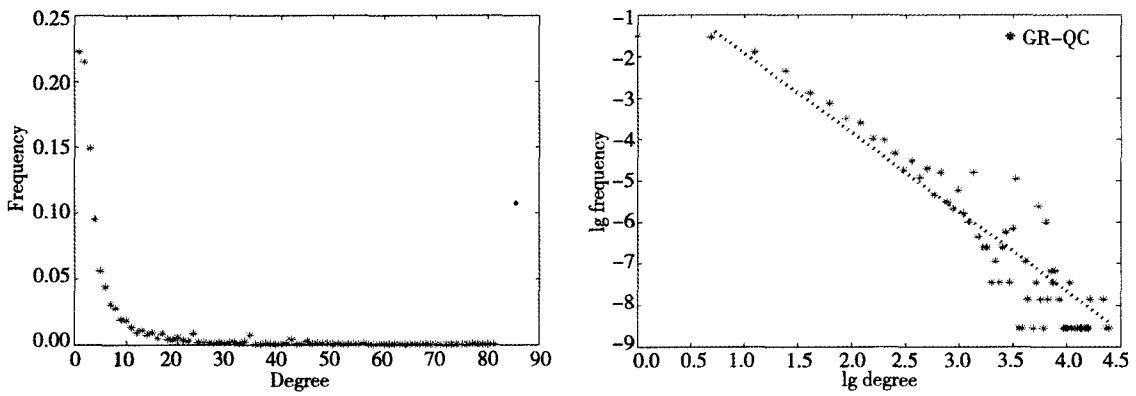

图 1　GR-QC 合作网络中节点的度分布

Fig. 1 The node degree distribution of GR-QC network

1.2　GR-QC 网络中连接行为

用参数 α 和 $1-\alpha$ 来描述 GR-QC 网络中新节点加入时的随机连接和择优连接的程度，$\alpha \in [0,1]$. 当 $\alpha \to 1$，新作者倾向于完全随机选择科学家并进行合作，$\alpha \to 0$，新作者则倾向于择优连接的方式选择大科学家进行合作. 根据 α 的取值情况，讨论不同的连接方式引起的网络节点度分布情况.

当 $\alpha = 1$ 时，即新节点加入时是完全随机连接行为. Erdos 和 Renyi[15] 提出的 ER 模型被认为是典型的完全随机网络模型. 引入时间序列，用 ER 模型描述 GR-QC 网络的连接行为：设在时间序列 $t \in \{0,1,2,\cdots,m\}$，每个时刻产生一个新节点，生成了 $m+1$ 个节点且两两连接的完全图. 当 $t \geqslant m+1$，每个新加入节点都随机选择网络中 m 个节点建立连接. 在节点 i 产生后的任何时刻 $t,(t > i)$，i 节点度的增加量是 $\frac{\Delta d_i(t)}{\Delta t} \approx \frac{m}{t}$. 由初始值 $d_i(i) = m$，得到 t 时刻节点 i 的度为 $d_i(t) = m + \ln(\frac{t}{i})$. 对于任意 d 及 t，只要得到使 $d_{i(d)}(t) = d$ 的 $i(d)$，便可得到网络中节点的累积度分布 $F_t(d) = 1 - \frac{i(d)}{t}$. 于是，在 t 时刻 $\frac{i(d)}{t} = e^{-\frac{d-m}{m}}$. 故 GR-QC 网络的 ER 模型表示的节点累积度分布 $F_t(d) = 1 - e^{-\frac{d-m}{m}}$.

当 $\alpha = 0$ 时，新节点加入建立 m 条连接时是完全择优连接的 BA 模型[3]，即新节点与老节点与之连接的概率与老节点的度成正比. 因此，对每个时刻 $t > i$，新加入网络中的节点导致的老节点 i 的度随 t 的变化率为 $\frac{\Delta d_i(t)}{\Delta t} \approx \frac{d_i(t)}{2t}$. 利用初始值 $d_i(i) = m$，得到 $\frac{i(d)}{t} = \left(\frac{m}{d}\right)^2$. 于是，GR-QC网络的 BA 模型表示的节点累积度分布为 $F_t(d) = 1 - m^2 d^{-2}$.

当 $0 < \alpha < 1$ 时，即每个新节点加入网络时既有随机连接又有择优连接行为. 由 α 的定义知道，α 是 m 个老节点进行随机选择连接的概率，$1-\alpha$ 是与 m 个老节点以择优连接的概率. 于是，网络中节点 i 的度变化率为

$$\frac{\Delta d_i(t)}{\Delta t} \approx \frac{\alpha m}{t} + \frac{(1-\alpha)md_i(t)}{2mt} = \frac{\alpha m}{t} + \frac{(1-\alpha)d_i(t)}{2t}$$

将初始值 $d_i(i)=m$ 代入上式，得到 $d_i(t) = (m+\frac{2\alpha m}{1-\alpha})(\frac{t}{i})^{(\frac{1-\alpha}{2})} - \frac{2\alpha m}{1-\alpha}$. 于是，网络中节点的累积度分布为

$$F_t(d) = 1 - (\frac{m+\gamma\alpha m}{d+\gamma\alpha m})^{\gamma}, \gamma = \frac{2}{1-\alpha} \quad (1)$$

公式(1)中，当 $\alpha=1$ 时，度分布为 $F_t(d) = 1 - e^{-\frac{d-m}{m}}$ 的完全随机连接，当 $\alpha=0$ 时，度分布为 $F_t(d) = 1 - m^2d^{-2}$ 的择优连接.

1.3 GR-QC 网络度分布函数的参数估计

由于公式(1)由参变量 α 和 m 共同影响，若是 m 确定，那么只需要考察 α 的值对公式(1)的影响，也就是 GR-QC 网络中，新节点有确定的新连接数时更倾向于以哪种连接行为与网络中老节点建立连接. 当 $\alpha=1$ 和 $\alpha=0$ 时分别是 ER 和 BA 模型形式，所以，只需要考虑 $0<\alpha<1$ 时. 由网络演化过程和欧拉定理知道，在 t 时刻，网络中的总边数是 tm，总度数是 $2mt$. 由此，得到网络中节点的平均度是 $2m$. 由 1.1 节知 GR-QC 网络的平均度 $2m=5.54$，故 $m=2.77$. 为估算 α 值，将公式(1)取自然对数，得到

$$\ln(1-F(d)) = \frac{2}{1-\alpha}\ln(m+\frac{2\alpha m}{1-\alpha}) - \frac{2}{1-\alpha}\ln(d+\frac{2\alpha m}{1-\alpha}) \quad (2)$$

取 α 的估计值 α_0，$\alpha_0 \in (0,1)$，(如表 2 中的第一行所示)以及 $m=2.77$ 代入公式(2)右端拟合 $F(d)$. 同时，将 GR-QC 网络中节点度、$m=2.77$ 代入公式(2)计算得到 α 的近似值 α_1 (如表 2 所示第二行). 当 $\alpha_0=\alpha_1$ 时公式(2)的拟合 $F(d)$ 与 GR-QC 网络的 $F(d)$ 一致.

表 2　α 值迭代估算

Table 2 Iterating and examining values for α

α_0	0.10	0.20	0.30	0.390	0.40	0.410	0.50	0.60	0.70	0.80
α_1	0.26	0.31	0.35	0.396	0.40	0.407	0.46	0.52	0.58	0.67

由表 2，$\alpha_0=\alpha_1=0.4$ 时，公式(2)的近似值与真实值相等. 即该网络中新节点以 $\alpha=0.4$ 的概率与老节点建立随机连接，以 0.6 的概率择优连接. 于是，公式(1)中 $\alpha=0.4$ 和 $m=2.77$. 数值模拟公式(1)，如图 2 所示. $\alpha=0.4$ 表明新研究者以较大的随机性选择与老研究者合作，显示了合作研究的盲目性；同时，新成员以 0.6 的概率进行择优连接，证明了现实中“富者越富”现象在科学家合作中也是普遍存在的. 这在一定程度上能解释 GR-QC 网络在图 1 中的偏离现象.

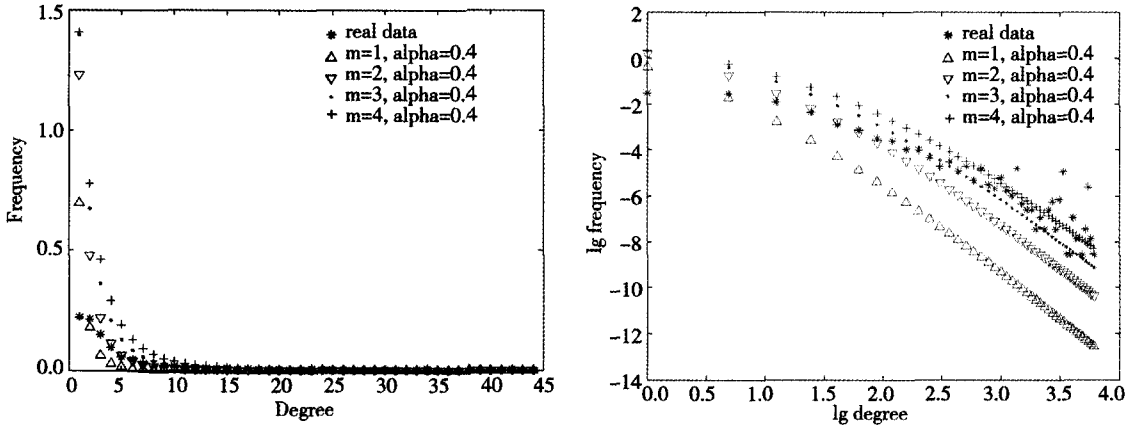

图 2　$\alpha=0.4$，分别取 m=1,2,3,4 时，公式(1) 与实际数据的吻合程度

Fig. 2 The numerical simulations of equation (1) when $\alpha=0.4, m=1,2,3,4$

从图 2 的拟合中发现，当 m=2,3 时，公式(1)与实际数据吻合程度较好，说明该公式(1)拟合实际网络有一定的合理性. 但是在度值较大的区域中，拟合曲线与实际数据还是差距较大，

说明公式(1)还有不尽合理之处.

2　连接行为的网络模型

在 GR-QC 网络中，只考虑了新节点加入合作网络时的连接行为就得到公式(1). 而 GR-QC 合作网络中有巨大分支的产生(如附件中图 1 所示)，巨大的连通分支的形成不仅是新节点与老节点间建立连接，也有老节点间建立新连接，这些连接的建立使得网络的连通分支规模不断扩大而形成了巨大分支. 利用新节点与老节点的连接行为和老节点间的新连接两种连接行为综合描述合作网络演化，进一步优化公式(1)中的连接行为.

2.1　模型的生成

假设新节点和新边在加入网络时服从 Possion 流，即网络时间划分的足够细时，使得每个单位时间内“一个新节点加入网络”和“两个老节点间建立连接”这两个事件同时发生的概率趋近 0. 也就是说，在一个单位时间内，“一个新节点加入网络”和“两个老节点间建立连接”最多有一种情况发生. GR-QC 网络是在一个具体的研究领域的合作行为，因此，在以下模型构造中，假设连接行为是全局性行为.

模型初始状态　假设 $t = m$ 时刻，网络已经有 $m+1$ 个节点，并且所有节点之间彼此有连接，构成一个完全图.

为方便，用 i 标记节点，并且令 i 是时刻 $t \geqslant i$ 时新生成的节点，$i \in \{0,1,2,\cdots\}$. 用 $d_i(t)$ 表示节点 i 在 t 时刻的度，于是 $d_i(t)>0$ 当且仅当 $t \geqslant i$.

新成员加入　在 $t > m > 0$ 时刻，一个新产生的节点与老节点建立 m 条连接. 以 $0 < \alpha < 1$ 的概率与 m 个不同的老节点随机连接，以 $1-\alpha$ 的概率与 m 个不同的老节点进行择优连接.

老成员间的新连接　当 $t>m+2n$，在老节点间建立 n 条不重复的新连接. 以 $0<\beta<1$ 的概率随机选择老节点建立 n 条不重复新连接，以 $1-\beta$ 的概率进行择优选择老节点建立 n 条新连接，$n > 0$.

2.2　模型的度分布

在上述网络生成中，设 m,n,α,β 都是给定常参数. 当 $t > m + 0.5\left((2m+1)^2+4n\right)^{\frac{1}{2}}$，网络中节点数量近似为 t，总边数近似的是 $t(m+n)$，网络的平均度为 $m+n$. 下面就 α,β 取值进行分类讨论节点的度分布函数的数学表达.

情况 1　当 $\alpha = 1,\beta = 1$ 时，即新节点加入时与老节点的连接、老节点之间的新连接都是以完全随机行为建立连接. 根据平均场理论，i 节点的度随时间的变化估计值可表示为 $\dfrac{\Delta d_i(t)}{\Delta t} \approx \dfrac{m+2n}{t}$. 结合初始值 $d_i(i) = m$，得到 $d_i(t) = m + (m+2n)\ln\left(\dfrac{t}{i}\right)$. 故 $i/t = \mathrm{e}^{\frac{m-d}{m+2n}}$. 于是，在 t 时刻，节点的累加度分布

$$F_t(d) = 1 - \mathrm{e}^{-\left(\frac{d-m}{m+2n}\right)} \tag{3}$$

情况 2　当 $\alpha = 0,\beta = 0$ 时，即新节点加入时与老节点的连接、老节点之间的新连接都是择优连接. 此时，i 节点的度随时间的变化为 $\dfrac{\Delta d_i(t)}{\Delta t} \approx \dfrac{md_i(t)}{2(m+n)t} + \dfrac{2nd_i(t)}{2(m+n)t} = \dfrac{(m+2n)d_i(t)}{2(m+n)t}$.

将初始解 $d_i(i) = m$ 代入上式，得 $d_i(t) = m \times \left(\dfrac{t}{i}\right)^{\gamma_1}$，其中 $\gamma_1 = \dfrac{m+2n}{2(m+n)}$. 得 $\dfrac{i}{t} = \left(\dfrac{d}{m}\right)^{-1/\gamma_1}$. 于是，在 t 时刻，节点的累加度分布为

$$F_t(d) = 1 - \left(\frac{d}{m}\right)^{-1/\gamma_1} \tag{4}$$

其中 $\gamma_1 = \dfrac{m+2n}{2(m+n)}$. 故新连接模式模型下，节点度密度函数 $p(d) \propto d^{-1-\frac{1}{\lambda}}$，由于 $\dfrac{1}{2} \leqslant \gamma_1 = \dfrac{m+2n}{2(m+n)} \leqslant 1$，故 $-1-\dfrac{1}{\gamma_1} \in [-3,-2]$. 此时，节点度密度函数 $p(d)$ 具有经典的 BA 模型的 Scale-free 性质[4].

情况 3　当 $\alpha = 0,\beta = 1$ 时，即新节点加入时与老节点的连接行为是择优连接，而老节点之间的新连接都是随机连接. 此时，i 节点的度随时间的变化估计值为 $\dfrac{\Delta d_i(t)}{\Delta t} \approx \dfrac{md_i(t)}{2(m+n)t} + \dfrac{2n}{t}$，将初始解 $d_i(i) = m$ 代入，得到

$$d_i(t) = \left(m + \frac{4n(m+n)}{m}\right)\left(\frac{t}{i}\right)^{\frac{m}{2(m+n)}} - \frac{4n(m+n)}{m}.$$

记$\frac{2(m+n)}{m}=\gamma_2,\gamma_2\geqslant 2$. 于是，节点的累积度分布函数为

$$F_t(d)=1-\left(\frac{m+2n\gamma_2}{d+2n\gamma_2}\right)^{\gamma_2}=1-\left(\frac{d+2n\gamma_2}{m+2n\gamma_2}\right)^{-\gamma_2}.$$

情况 4 当$\alpha=1,\beta=0$时，即新节点加入时与老节点的连接都是完全随机连接，而老节点之间的新连接都是择优连接. 类似于 $\alpha=0,\beta=1$ 的情形，得到 t 时刻节点累积度分布

$$F_t(d)=1-\left(\frac{m+m\gamma_3}{d+m\gamma_3}\right)^{\gamma_3},\text{ 其中}\frac{(m+n)}{n}=\gamma_3.$$

情况 5 对于一般的 α,β，即 $0<\alpha<1$，$0<\beta<1$. 此时，新节点加入时与老节点的连接、老节点之间的新连接都既有随机连接又有择优连接. 此时，i 节点的度随时间 t 的变化为

$$\begin{aligned}\frac{\Delta d_i(t)}{\Delta t}&\approx\frac{\alpha m}{t}+\frac{(1-\alpha)md_i(t)}{2(m+n)t}+\\&\quad\frac{2n\beta}{t}+\frac{(1-\beta)2nd_i(t)}{2(m+n)t}\\&=\frac{A_1}{t}+\frac{d_i(t)A_2}{t}\end{aligned}\tag{5}$$

其中 $A_1=m\alpha+2n\beta,A_2=\frac{m+2n-m\alpha-2n\beta}{2(m+n)}$.

将初始值 $d_i(i)=m$ 代入公式(5)，得到在 t 时刻 $(t>i)$ 节点 i 的度

$$\begin{aligned}d_i(t)&=\left(m+\frac{2B(m+n)}{A-B}\right)\left(\frac{t}{i}\right)^{A-B}-\\&\quad\frac{2(m+n)B}{A-B}\\&=(m+C)\left(\frac{t}{i}\right)^{A-B}-C,\end{aligned}$$

其中 $A=\frac{m+2n}{2(m+n)}$，$B=\frac{m\alpha+2n\beta}{2(m+n)}$，$C=\frac{2(m+n)B}{A-B}$. 于是，$t$ 时刻 $(t>i)$ 网络中节点的累积度分布

$$F_t(d)=1-\left(\frac{m+C}{d+C}\right)^{\gamma_4},\gamma_4=\frac{1}{A-B}>1\tag{6}$$

由于 m,n,α,β 都是给定的常参数，因此，A,B,C 的值也是常参数. 由公式(6)得到网络中节点的度分布密度函数是

$$p(d)=\frac{\gamma_4}{(m+C)^{\gamma_4+1}}(d+C)^{-\gamma_4-1}\propto d^{-1-\gamma_4}\tag{7}$$

其中 公式(7) 中的参数 γ_4,m,C 都是常数. 因此，公式(7)中的指数 $-1-\gamma_4\leqslant -2$ 且 γ_4 与网络的规模无关.

可以推断：在一般情况下，网络中节点的累积度分布函数(6)具有幂律、无标度的统计性质. 情况 3、情况 4 和情况 5 的累积度分布都是情况 2 的漂移幂律的形式. 公式(6)较公式(1)增加了描述老节点之间的新连接行为参数，比公式(1)更准确地反映网络中节点的度的变化行为.

2.3 模型的数值拟合

根据 2.2 节中模型的表达式(6)或者表达式(7)中的参数 A,B,C 仅与输入参数 m,n,α,β 有关. 根据公式(6)计算出 α,β 值，也就是知道了新节点与老节点、老节点与老节点之间建立新连接时的不同连接行为的可能性. 由于公式(6)中涉及两个参变量 α,β，在 1.2 节计算的方法已失效，因此，利用数据拟合的方式描述模型的数学表达式.

由于 GR-QC 网络数据为累积数据，在其数据中没有时间序列，无从得知其新成员加入网络建立新连接、老节点之间建立新连接的真实数量. 根据平均场理论，当演化时间 t 足够长时，网络中节点的平均度 $m+n$ 是常数. 根据 m 和 n 的非负整数的性质，可以取满足 $m+n$ 是给定常数的 m 和 n，估算不同 α 值，β 值时公式(6). 因此，在 GR-QC 网络中，网络的度序列 d 以及对应的频率值 $F(d)$ 是已知的. 由公式(6)计算 α 和 β 的值，也就是知道了新节点与老节点、老节点与老节点之间建立新连接时的不同连接行为的可能性.

图 3 是 $m+n=5$，α,β 取不同值时公式(7)与真实数据的拟合. 左图显示 $m=2,n=3,\beta=0.5$ 时模拟数值与真实值的吻合程度高. 右图是 $m=1,n=4,\beta=0.25$ 的两条曲线与真实值吻合较好. 图 3 拟合真实数据都比图 2 的要更准确. 这就说明网络中不仅有新节点与老节点的连接，也有老节点之间的新连接产生.

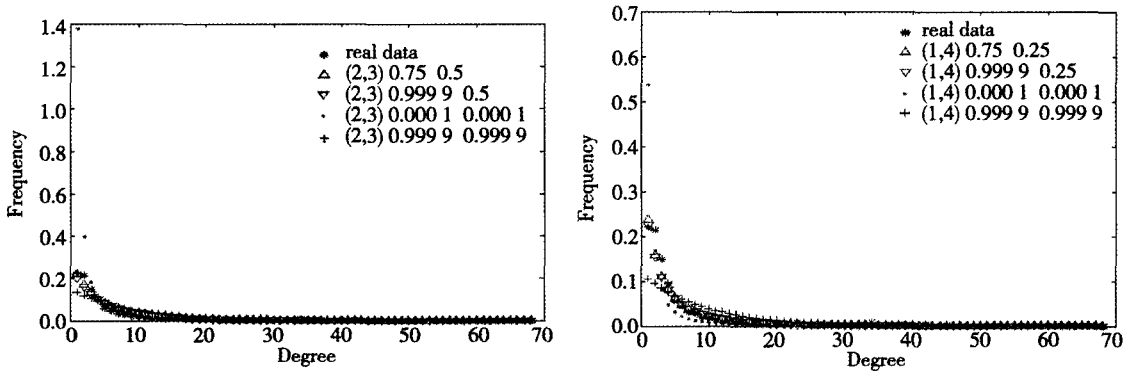

图 3　当 $m+n=5$ 时，公式(6)的模拟

Fig. 3 The numerical simulations of equation (6) when $m+n=5$

进一步地，将模型与实际 GR-QC 网络中节点的连接数值模拟进行对比，通过选择满足 $m+n$ 的不同 m 和 n 值以及 α 与 β 可能的取值，对公式(7)进行数值拟合. 如附录中图 5(a)—5(h)所示的 8 幅图片都是在双对数坐标系中，分别取 $m+n=5, m+n=6$ 时不同的 α 与 β 取值，公式(7)与实际数据的拟合的情况. 根据 1.2 节，计算得到的 GR-QC 网络中节点的平均度是 5.54，即 $m+n\approx 5.54$. 前四幅是 $m+n=5$ 时的拟合情况. 在双对数坐标系中，横坐标 $\ln 5\approx 1.6$. 附录图 5 中(a)是 $[m,n]=[1,4]$. $m=1$，说明新节点的度值是 1. n=4 是指的老节点中添加了 4 条连接. 图示中曲线 ($\alpha=0.999\ 9, \beta=0.25$)，($\alpha=0.75, \beta=0.25$) 与实际值几乎重合，此时，$\alpha$ 相对于 β 取值大，说明新节点加入网络时，以较大的概率进行随机连接，而老节点是有较大概率的择优连接行为；节点的度值从 2 到 5 时，对应的图(a)中 [0.7, 1.6] 之间，此时，α 的值趋近 0.999 9，说明新节点加入时几乎是完全随机连接. β 的值先是减小后增大(表现为与圆线重合后又偏离，在横轴 1.6 处与 $\beta=0.999\ 9$ 的曲线重合). 这说明度值在区间 [2,5] 的老节点以大概率选择随机连接行为；节点的度从 6 到 20，对应的(a)中的区间 (1.6, 3] 时，曲线 ($\alpha=0.999\ 9, \beta=0.25$)，($\alpha=0.75, \beta=0.25$) 与真实值拟合好后介于曲线 ($\alpha=0.999\ 9, \beta=0.25$) 与曲线 ($\alpha=0.000\ 1, \beta=0.000\ 1$) 之间. 此时，$\alpha$ 从 0.999 9降低为 0.000 1，说明新节点加入时是随着老节点的度值的增大而由完全随机连接趋向于择优连接. β 的取值由 0.25 趋于 0.000 1，说明老节点在选择老节点建立新连接的时候，趋向于择优连接；度大于 20 的节点，对应(a)图中的(3, 4.5)区间，该区间无论哪条曲线都和实际数据有交集也有偏离. 其中 ($\alpha=0.999\ 9, \beta=0.999\ 9$) 与实际值偏离程度大，说明新节点与老节点建立新连接或者老节点之间的新连接行为不都是完全随机连接. 曲线 ($\alpha=0.000\ 1, \beta=0.000\ 1$) 在 [3.3, 4.5] 上与实际数据吻合好，说明新节点与老节点建立新连接或者老节点之间的新连接行为趋向于择优连接. 同样，可以分析附录中图 2(b)、图 2 (c)和图 2 (d). 通过对图 2 (a)—图 2 (d)的分析，发现节点的度在 4 或者 5 的时候，区间两侧的连接行为变化较大. 在节点的度超过 20，网络中的连接进入一个择优和随机混杂区间.

由图 1 的左图知道，网络中度不超过 5 的节点占网络总数的 75%以上，度数居于 5 到 20 左右的节点，约占网络节点总数的 15%，网络中不足 10%的大度节点(节点度超过 20 的节点). 结合附录图 5(a)—图 5 (d)可以推断：GR-QC 网络中的大部分新节点加入时，随机连接行为占主导地位. 大部分的度小的老节点间建立新连接时，随机连接行为占主导；度值居于 5 到 20 左右的节点，无论是新节点加入时选择老节点还是老节点选择老节点都是择优连接行为占主导；而网络中大度节点(度大于 20 的节点)中新边的加入时既有随机连接又有择优连接. 附录图 5 (e)—图 5(h)表示的是 $m+n=6$ 时的拟合情况. 讨论与图图 5 (a)—图 5 (d)类似.

GR-QC 网络仅在一个时间点窗口的数据，不

能很好地证明模型的演化特征. 为了进一步说明模型的参数变化与网络演化的关系,搜集了具有时间序列的真实数据,并进行拟合.

3 模型的实证分析——中国管理科学与工程领域部分论文作者数据

3.1 数据收集与整理

选取数据是管理科学与工程学会(www. glkx-ygc. cn)第二届理事会的名誉理事长李京文、理事长高自友,以及副理事长马庆国、李一军、黄海军、谭跃进、陈国青、齐二石、李垣、党延忠、徐玖平11 位学者署名的学术论文,把论文的所有作者作为研究对象. 选取这些数据的主要原因是管理科学与工程学科理事会成员在某些方面代表学科的研究发展方向和核心领域. 在一定程度上与GR-QC 网络的固定领域具有共性. 数据时间和范围是自 2000 年 1 月到 2016 年 12 月期间发表的所有 SCI(科学引文索引)、SSCI(社会科学引文索引)、CSSCI(中文社会科学引文索引)检索的论文. 论文来自于 Web of Science 的 SCI-Expanded (1900 年至今)及 SSCI(1900 年至今)数据库,以及 CNKI 平台的 CSSCI 数据库,检索相应的中英文论文. 既包含了作者高水平的英文论文,也兼顾了高质量的中文论文. 既搜集了自然科学方面的成果,也获取了社会科学方面研究.

数据收集过程中,对数据进行了多次清洗. 由于外文期刊分别以 11 位教授的汉语全拼姓名进行收集,结合论文作者的研究领域、工作单位、学习和工作经历等多重属性消除了重名的作者,同时解决了英文的名前姓后、姓前名后、姓名简写等问题可能导致论文收集不全面等问题. 经过多次数据整理,截至 2017 年 3 月 14 日共获得了 2000 年—2016 年 1 510 篇论文,包含 1 033 位论文作者.

3.2 数据的统计分析

从论文中抽取论文的所有作者作为研究的网络节点,同一篇论文的所有作者之间都建立连接(不考虑独立作者的论文),构成论文合作网络. 两个作者之间多篇论文预示着两个节点之间边的重数定义为边的权重(见附录图 6). 包含 11 位管理科学与工程理事会理事长和副理事长的加权网络中,截至 2017 年 3 月,共有 1 033 个节点,边权总值 5 371(附录图 6 是 2016 年的加权网络). 如果只考虑作者之间是否建立连接,就得到了表示作者之间是否存在合作关系的合作关系网络,其连接数为 2 554(附录图 6 的 2016 年图). 网络节点度和权值累积到 2016 年的度分布和权值分布函数如图 4 所示. 从图中发现这两个分布函数具有相似的的幂律性质. 在表 3 中的"节点的度分布指数"和"节点权重分布指数"的 17 个值也证明了这一点.

附录图 6 展示了 2000 年、2010 年和 2016 年的论文作者合作网络. 从图中很容易看出作者之间的连接、以及围绕核心成员的那个团队的成员的增长的快慢. 对这 17 年的数据进行统计分析,基本参数如表 3 所示. 在作者连接行为构成的合作关系网络中,节点的度分布指数都大于 2,平均值是 2.377. 作者构成的加权合作网络中,节点的权重分布函数的指数也都大于 2,平均值 2.198. 符合模型的度分布函数公式(7)的指数区间.

由表 3 知道,自 2000 年到 2016 年中,作者总共增加了 1 003 人,连接数增加了 2 511 条边. 根据平均场理论,网络中平均每月增加 5.22 个作者,节点度的月平均增量是 26.14. 网络中已有的老作者(上一个时刻已经在网络中的节点)之间的新合作增量月平均是 13.56,而新连接的增量如表 3 中的"老节点间的新连接"并没有随着网络节点总数增加而出现较大的增长. 老作者之间建立的新连接月平均增量是 1.59;同时,由表 3 的"新节点加入产生的新连接"(当前时刻),即新加入的作者与老作者之间的新连接的增量月平均值是 14.14. 平均意义之下,每增加 1 个新作者而增加的新连接数是 2.77. 即 $m = 1.59$, $n = 2.77$. 根据 1.2 节,GR-QC 网络中节点的平均度是 $m + n \approx 5.54$. 管理科学与工程的论文作者之间的平均度 $m + n = 1.59 + 2.77 = 4.36$ 接近 GR-QC网络. 对比附录图5(b)发现,$\alpha = \beta = 0.000\,1$. 并且对应的公式(7)中的 $\gamma = 2.142$ 时,与实际数据的吻合程度最高. 事实上,在管理科学工程的这些核心成员之间的连接很少,表 3 中的老节点之间的新连接几乎全部是名气不大的作者或者

是核心成员的学生与其之间建立的连接；另一方面，新节点加入产生的新连接几乎全部发生在新作者与某位核心成员之间。即模型中老节点之间、新节点与老节点之间的随机连接行为少，择优选择行为多. 表 3 中的数据进一步证明了，促使网络演化的是新节点的加入以及择优连接行为.

表 3　11 位作者形成的合作网络部分统计参数.

Table 3 The statistical of the cooperation network with 11 authors

年度	节点数（作者人数）	连接数（是否有合作）	节点平均度	加权连接数（合作的次数）	老节点间的新连接	新节点加入产生的新连接	节点的度分布指数	节点权重分布指数	网络平均聚集系数
2000	30	43	2.867	53	0	0	2.413 2	2.033 2	0.315
2001	58	78	2.69	98	0	35	2.619	2.510 8	0.283
2002	95	136	2.863	178	0	58	2.506 8	2.490 1	0.275
2003	125	185	2.96	275	3	46	2.515 1	2.348 2	0.29
2004	165	253	3.067	409	3	65	2.460 8	2.270 2	0.299
2005	201	319	3.174	534	1	65	2.413	2.218 7	0.33
2006	262	445	3.397	749	6	120	2.354 1	2.200 4	0.339
2007	318	568	3.572	1 004	8	115	2.344 7	2.177 5	0.35
2008	391	709	3.627	1 300	7	134	2.296 3	2.153 3	0.349
2009	425	793	0.732	1 498	16	68	2.296 5	2.134 9	0.344
2010	488	946	3.877	1 835	19	234	2.304 5	2.124 6	0.344
2011	573	1 175	4.101	2 281	35	194	2.298 3	2.119 4	0.351
2012	668	1 422	4.257	2 765	22	225	2.315 6	2.112 3	0.365
2013	761	1 702	4.473	3 324	51	229	2.310 7	2.117 5	0.369
2014	844	1 954	4.63	3 950	37	215	2.322 5	2.117 1	0.375
2015	948	2 287	4.852	4 671	64	269	2.32	2.120 6	0.38
2016	1 033	2 554	4.945	5 371	34	233	2.323 9	2.112 2	0.384

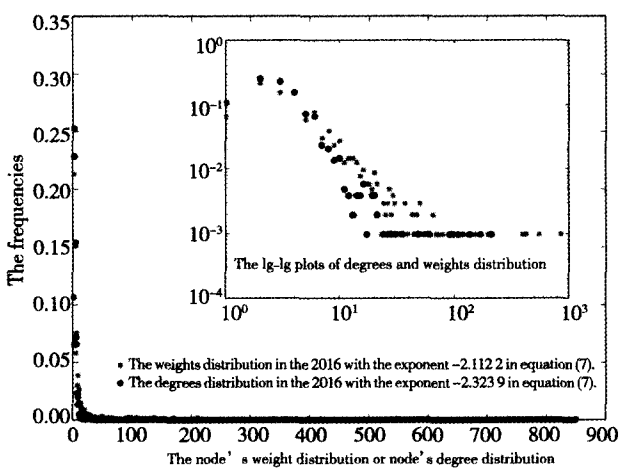

图 4　11 位核心理事成员在 2016 年构成的合作网络的节点权值分布和节点度分布

Fig. 4 The node degrees and weights distribution with 11 key councilors in 2016

为了更好地描述新节点加入网络中时择优选择和随机选择概率的大小，将图 5 中边数的增量进一步细化，分别考虑新节点加入网络时，选择连接行为的不同对网络的新连接的贡献. 由 2.1 节中模型的演化机制中，定义的新成员加入时，分别以 α 和 $1-\alpha$ 的概率与 m 个不同的老节点进行随机选择和择优选择建立连接. 因此，$(1-\alpha)/\alpha$ 就是择优选择与随机选择的比例.

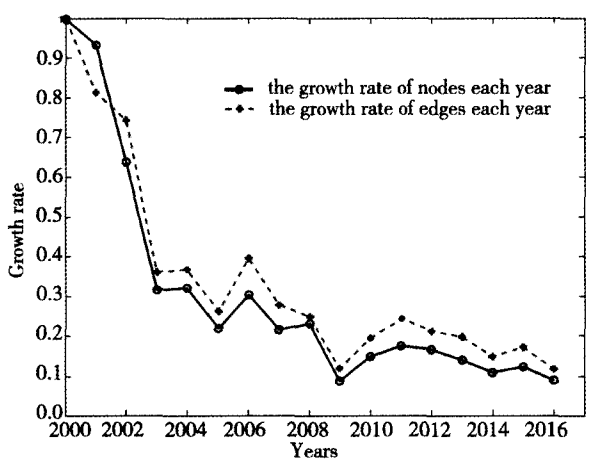

图 5　合作网络中作者和新连接的增加率

Fig. 5 The growth trends of nodes/edges

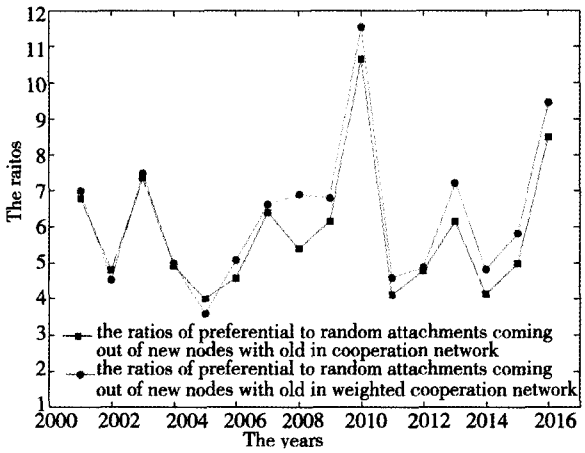

图 6　择优连接和随机连接的比率

Fig. 6 The ratios of preferential to random attachments

通过计算,图 6 的两条曲线分别表示连接关系网络和加权连接关系网络中 $(1-\alpha)/\alpha$ 的比值,也就是新节点加入时择优连接数和随机连接数的比值. 两条曲线的平均值分别是 5.42 和 5.99. 也就是说,在平均意义之下,网络和加权连接网络中新节点加入时,随机连接的概率和择优连接的概率分别是 0.155 8、0.844 2 和 0.143 1、0.856 9. 即使是比值最大情况下,也就是说随机连接概率最大的可能是 0.20 和 0.217 9,对应的是 2005 年度值. 这充分说明了,在新节点加入网络中时,以较大的概率选择大度节点进行连接.

为了更直观的解释合作网络模型的演化与节点连接行为是密切相关的,因此, 定义了另外的一个比值:年贡献率, 即,老节点对网络的贡献率和新节点对网络的贡献率. 老节点的贡献率定义为老节点之间产生的新连接数与老节点数的比值. 新节点的贡献率是新节点与老节点之间产生的新连接数与新节点数的比值. 利用表 3 中的数据计算老节点的贡献率和新节点的贡献率,分别如图 7 的右图和左图所示. 老节点的贡献率最大在 2014 年和 2015 年,不足 8%. 而新节点对网络的贡献率几乎是逐年增长的,到 2016 年的时候,年贡献率接近 300%. 由此,进一步表明网络的演化受到新节点连接行为的直接影响,新节点的连接行为是网络结构增长的动力.

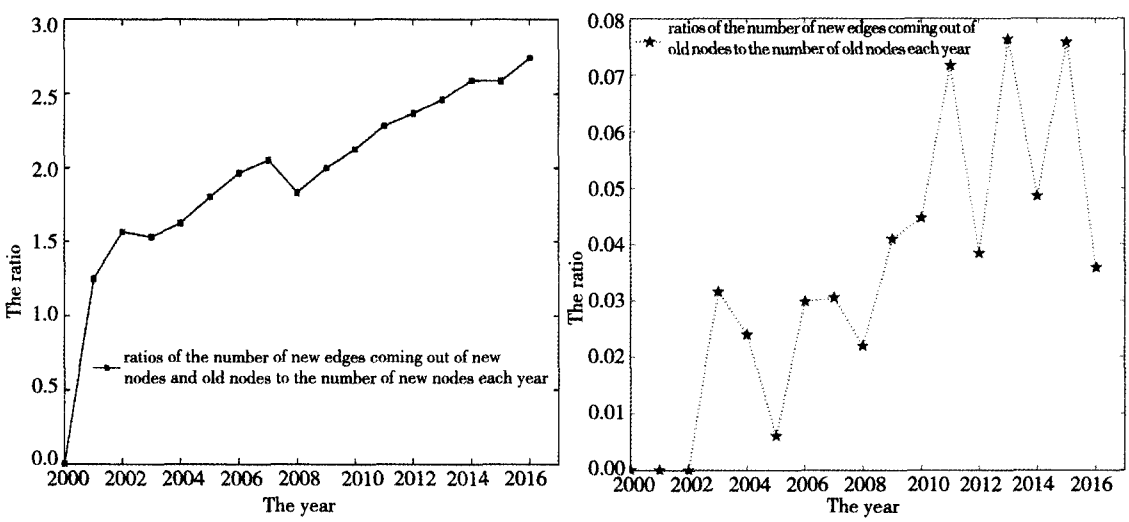

图 7　左图是新节点对网络的贡献率,右图是老节点对网络的贡献率

Fig. 7 The ratios of new edges to the number of nodes

4　结束语

通过网络中新成员的加入和老成员之间的新连接行为对节点度分布的刻画, 引入两个参数 α 与 β 并结合对 GR-QC 网络连接行为的测算,以及应用中国管理科学与工程部分学者的不同时期的论文数据作为实证, 构建了基于连接行为的合作网络演化模型. 为验证模型的科学性和可行性,对模型的统计性质进行了研究,并以此模型与 GR-QC 网络进行了比较. 利用中国管理科学与工程领域部分著名学者自 2000 年到 2016 年间 17 年的论文作者数据, 拟合发现了作者合作数的分布指数完全符合模型指数的分析. 同时, 发现中国管

理科学与工程领域的作者合作网络的平均度、连接行为的选择等都与 GR-QC 的网络相似. 两个不同领域的现实数据与模型的拟合，揭示了合作网络中成员的连接行为是合作网络共有的属性.

表 3 显示，随着网络规模的增大，网络中新成员规模也是增大的；新成员与老成员连接的增加是促使网络增大的主要因素；网络中核心成员之间的合作较少；不同社团之间的“桥梁”往往是一些开始在网络中并不重要的节点联系起来的. 当这些“桥梁”节点在网络中变得重要的时候，两个社团就凝聚为一个社团，网络也因此逐步发展为一个巨大的连通分支.

附录图 5 中的 8 幅图的数值拟合表明，网络中节点的连接行为与网络中节点平均度的大小息息相关. 网络中大部分的新节点加入网络时与老节点间建立新连接是择优连接行为占主导. 而且，新节点加入时对网络增长的贡献率远远超过网络中老节点之间的新连接. 网络中的巨大连通分支的形成不是网络中的核心成员之的连接，而是由于网络中的新节点与老节点间、度小的老节点与核心成员之间建立的连接，并不断的将独立的分支联系起来. 于是，就出现了网络中的小分支被大分支不断吞并，从而使得网络连通块的规模不断扩大而形成了巨大连通分支. 这种现象与现实社会关系中的“富人俱乐部”有异曲同工之妙：新人加入某种组织或者圈子的时候更倾向与网络中的“富人”先建立联系，而“圈子”中的老成员之间以随机或者择优选择等行为与他人建立联系，在演化过程中不断长成一张越来越大的“网”.

上述探讨都是基于平均场理论下的网络演化模型与实证，从网络结构方面讨论两种随机连接和择优连接行为对网络的演化的影响，没有考虑节点连接行为的内在驱动因素、节点的实际地理位置、以及节点所在的资源条件等影响因素. 进一步的研究可以结合上述问题进行更深入探索.

参 考 文 献：

[1]冯芷艳，郭迅华，曾大军，等. 大数据背景下商务管理研究若干前沿课题[J]. 管理科学学报，2013，16(1)：1－9.
Feng Zhiyan，Guo Xunhua，Zeng Dajun，et al. On the research frontiers of business management in the context of Big Data [J]. Journal of Management Sciences in China，2013，16(1)：1－9. (in Chinese)

[2]杨善林，周开乐. 大数据中的管理问题：基于大数据的资源观[J]. 管理科学学报，2015，18(5)：1－8.
Yang Shanlin，Zhou Kaile. Management issues in Big Data：The resource-based view of Big Data[J]. Journal of Management Sciences in China，2015，18(5)：1－8. (in Chinese)

[3]胡海波. 在线社会网络的结构、演化及动力学研究[J]. 系统工程学报，2014，29(3)：3－10.
Hu Haibo. Preferential linking in the growth of online social networks[J]. Journal of Systems Engineering，2014，29(3)：3－10. (in Chinese)

[4]Barabási A，Albert R. Emergence of scaling in random networks[J]. Science，1999，286：509－512.

[5]Newman M E J. The structure and function of complex networks[J]. SIAM Review，2003，45：167－256.

[6]李鹏翔，张萌物，席酉民，等. 组织网络中的无标度行为：极端情形的结果[J]. 管理科学学报，2009，12(4)：42－50.
Li Pengxiang，Zhang Mengwu，Xi Youmin，et al. Scale-free behavior in organizational networks：Consequences in extreme situations[J]. Journal of Management Sciences in China，2009，12(4)：42－50. (in Chinese)

[7]Kleinberg J M. The small-world phenomenon：An algorithmic perspective[C]. Proceedings of the 32nd ACM Symposium on Theory of Computing Annual ACM Symposium on Theory of Computing archive，Portland，Oregon，2006，163－170.

[8]Dorogovtsev S N，Mendes J F F. Scaling properties of scale-free evolving networks：Continuous approach[J]. Physical Review Letters，2007，63：56－125.

[9]Jackson M O, Rogers B W. Meeting strangers and friends of friends: How random are social networks[J]. American Economic Review, 2007, 5207 - 5211.

[10]David M, Brent S. Strong ties promote the evolution of cooperation in dynamic networks[J]. Social Networks, 2016, 45: 32 - 44.

[11]Newman M E J. Co-authorship networks and patterns of scientific collaboration[J]. PNAS, 2004, 101(1): 5200 - 5205.

[12]李志宏, 马 倩, 周广刚. 国内管理科学领域高校间学术论文合著网络的时间演化分析[J]. 管理工程学报, 2013, 27(4): 126 - 135.
Li Zhihong, Ma Qian, Zhou Guanggang. Analysis on the time evolution of cross-university academic papers co-author network in the filed of management science[J]. Journal of Industrial Engineering and Engineering Management, 2013, 27(4): 126 - 135. (in Chinese)

[13]李倩倩, 顾基发. 用户行为驱动的在线社交网络建模[J]. 系统工程学报, 2015, 30(1): 9 - 15.
Li Qianqian, Gu Jifa. Activity driven modelling of online social network[J]. Journal of Systems Engineering, 2015, 30(1): 9 - 15. (in Chinese)

[14]Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations [C]. Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005, 5: 21 - 24.

[15]Erdos P, Renyi A. On the evolution of random graphs[J]. Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 1960, 5: 17 - 61.

Cooperation network driven by attachment behaviors: Model and empirical analysis

MA Ying-hong, LIU Zhi-yuan, WANG Wen-qian

School of Management Science and Engineering, Shandong Normal University, Jinan 250014, China

Abstract: Cooperation networks are social networks consisting of a lot of collaborators to achieve better researches in science, engineering or other related fields. In cooperation networks, there are three probable attachments: random, preferential, and the mixed attachments. This paper analyzes the data of GR-QC network, and evaluates the attachment probability of the random, the preferential and the mixed. A novel collaborative network model driven by the attachment behaviors is presented. Compared with data of GR-QC, the attachment behaviors of scientists of management science and engineering in China have similar trends. Numerical simulations show that the attachment behaviors are affected by the average degree of the network, and the attachment behaviors evolves with the structure of the network. Discussing the numerical simulations and real data, the paper finds that the diversity and the homobium of attachment behaviors may cause the giant component to emerge or the distinct communities' to become cooperation networks.

Key words: cooperation network; attachment behavior; preferential mechanism; random attachment

附录

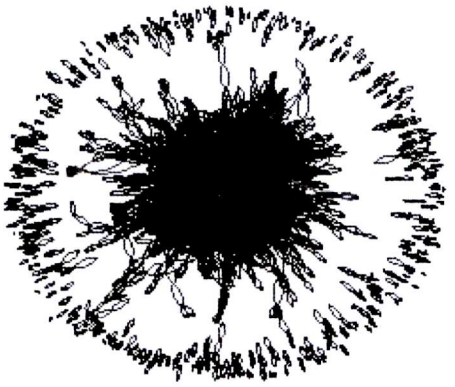

图 1　GR-QC 的可视化视图

Fig. 1 The global structure of GR-QC

图中以环形展示的形式把网络中的节点和连接展现出来. 图中间是一个巨大的连通分支，周围散落的是相对独立的合作团体. 中间巨大分支中的节点占据整个网络的接近 70%.

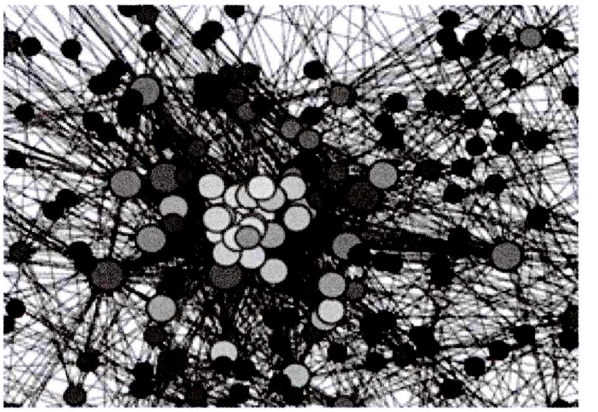

图 2　图 1 中的巨大分支核心节点簇

Fig. 2 The core node-clusters in huge component in Fig. 1

图中圆圈越大表示节点度越大，可见该核心节点簇中节点间合作关系非常紧密. 与其周围节点的合作关系亦较为紧密. 对核心节点簇成员信息进行提取可以用于该科学家合作领域中主要研究力量的分析.

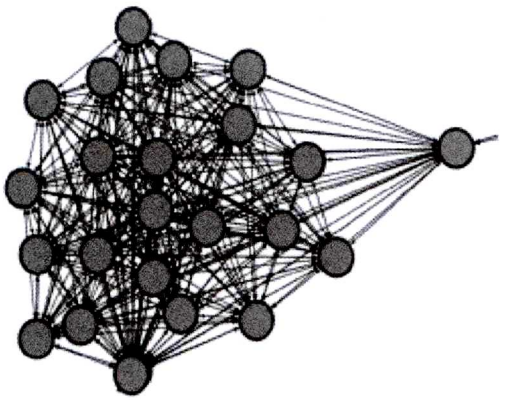

图 3　图 1 巨大分支中的一个节点簇

Fig. 3 A subgraph took from the huge component in Fig. 1

从网络巨大分支中截取的较为密集的节点簇：该簇中有 23 节点，近乎是全连接，并通过一个节点与巨大分支连接.

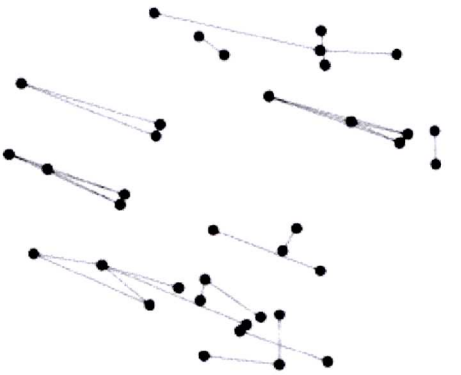

图 4　散布在图 1 巨大分支周围的合作团体

Fig. 4 Small groups scattered around the huge component in Fig. 1

网络巨大分支周围散布的团体里截取的部分小合作团体. 这样的独立合作团体在网络中较多.

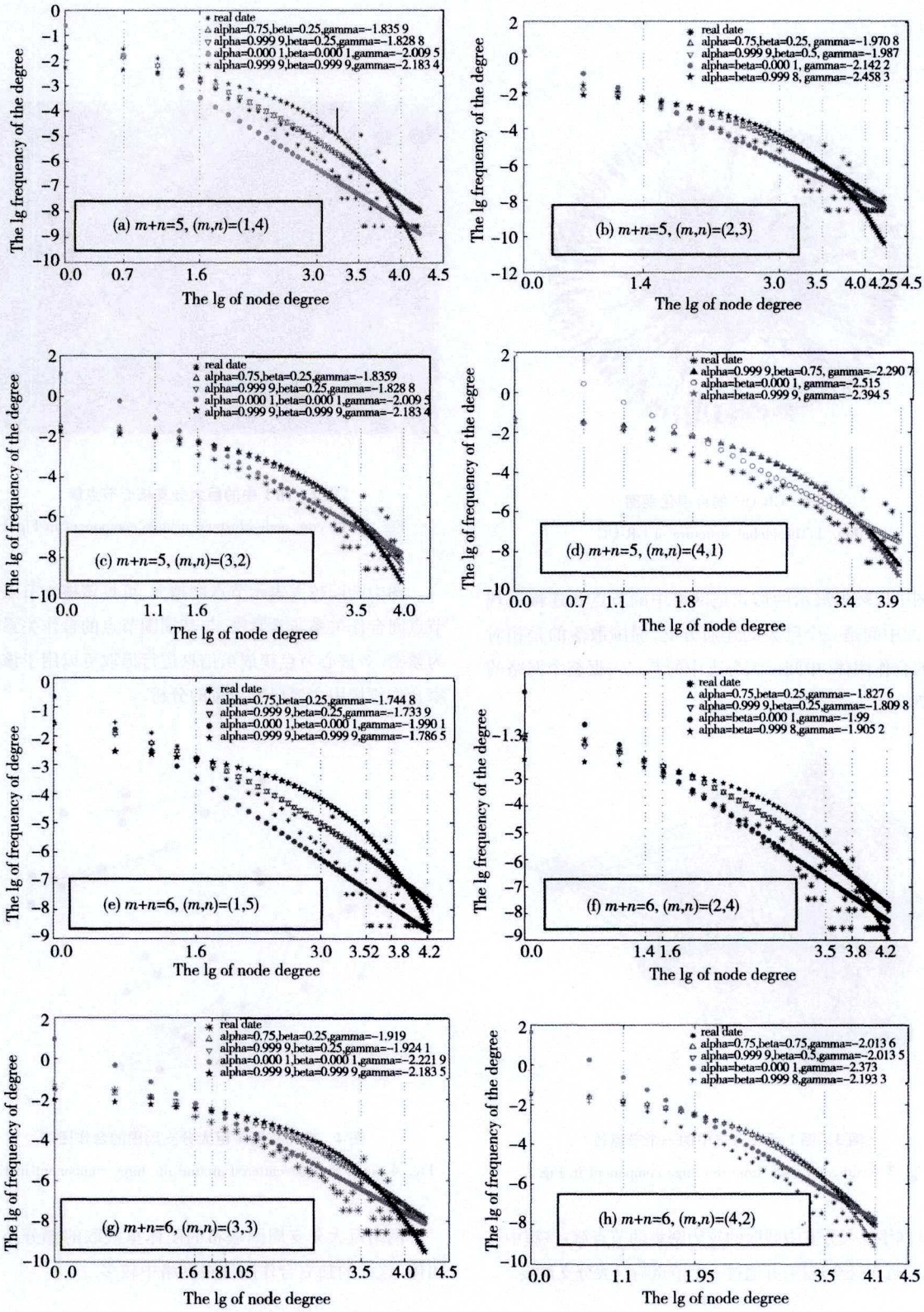

图 5 α、β 取值在[0,1]区间时，公式(7)的数值模拟与真实值的比较

Fig. 5 The comparing the real data with the numerical simulations when $0 \leqslant \alpha, \beta \leqslant 1$

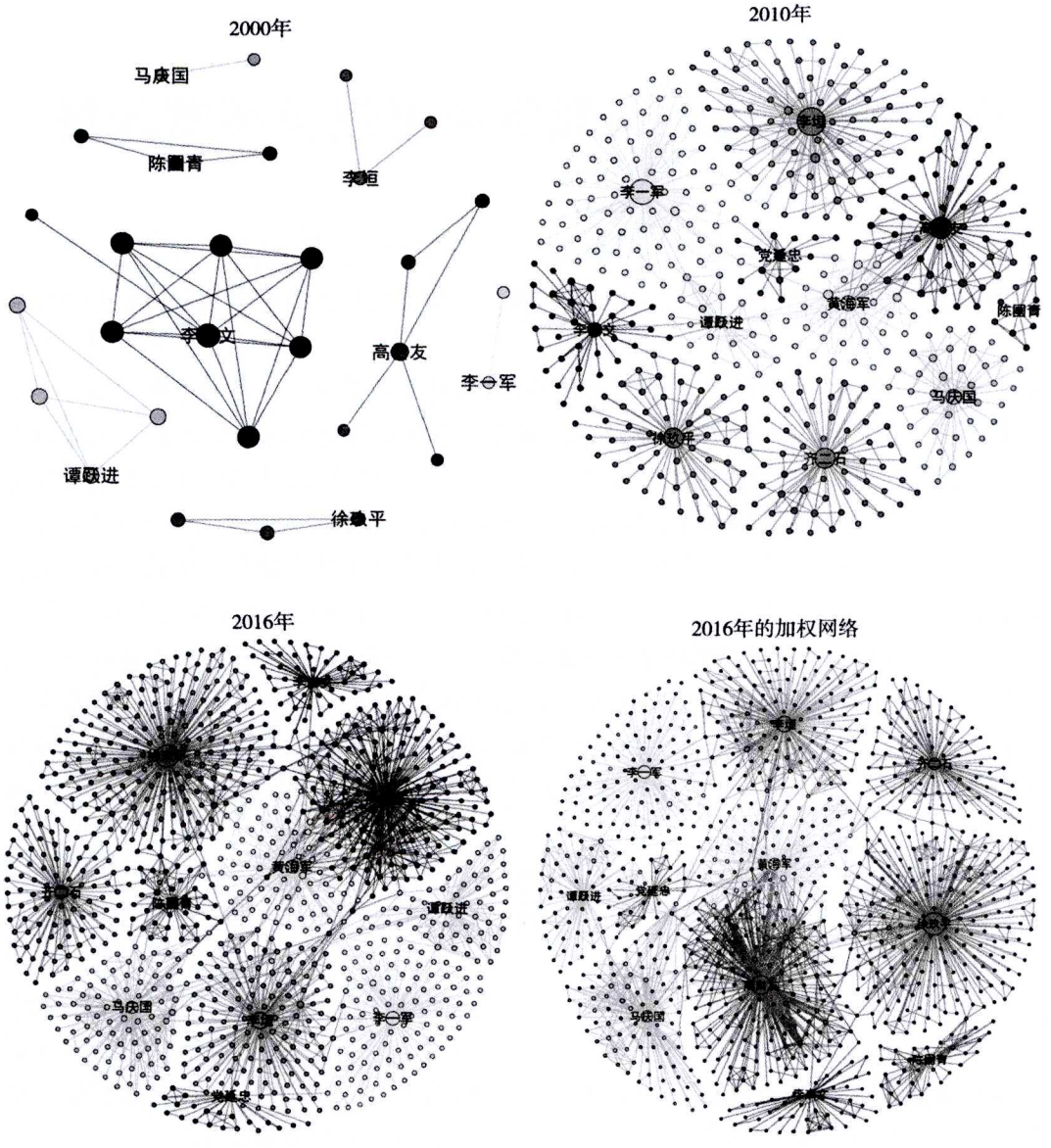

图 6　2000 年、2010 年和 2016 年管理科学与工程领域部分作者合作网络

Fig. 6 The author's cooperation networks of Management Science and Engineering of China in the year 2000, 2010 and 2016 respectively