文章编号: 1000_0887(2006) 11_1292_05

复杂网络幂律度分布和层次聚集函数标度指数的一种新的估计方法*

杨 波[1], 段文奇[2], 陈 忠[1]

(1. 上海交通大学 安泰经济管理学院, 上海, 200030;
2. 浙江师范大学 工商学院, 浙江 金华 321004)

(刘曾荣推荐)

摘要: 提出一种估计复杂网络幂律度分布和层次聚集函数标度指数的新方法, 并给出求解这些指数的数值算法• 该方法可以克服目前网络研究中采用的图形线性拟合估计方法存在的偏差和不准确等不足之处• 此外, 通过对 10 个 CNN 网络进行 KS 检验统计量的比较, 证实该方法比图形方法具有更好的拟合效果•

关 键 词: 参数估计; 复杂网络; 幂律; 度分布; 层次模块性
中图分类号: N94 文献标识码: A

引 言

近年来学术界兴起了复杂网络研究[1]• 在大量实证结果当中, 一个有趣的现象是很多现实中的自然网络和社会网络都具有两个一般特征: 幂律度分布和层次模块性• 前者是指, 一个被随机选中的结点有 k 条边(即度为 k) 的概率服从 $p(k) \sim k^{-\gamma}$[1], 其中 γ 为度指数• 具有幂律度分布的一些网络包括万维网[2], 因特网[3], 性接触网络[4], 引文网络[5] 等• 层次模块性在一些实际网络中也普遍存在, 如演员网络[6], 语言网络[7], 万维网[8], 新陈代谢网络[9]等• 层次模块性用标度律来刻画就是 $C(k) \sim k^{-\alpha}$[9~ 11], 其中 α 为层次指数, $C(k)$ 是度为 k 的结点的平均聚集系数, 即指度为 k 的结点其邻居结点之间存在边的概率• 此特征反映在网络拓扑结构上就是, 在具有层次模块性的网络中, 很多内部关联密集的小规模结点组之间松散关联从而形成更大规模的拓扑模块•

虽然实际网络的上述两个一般特征引起很多关注, 但目前为止, 相关研究还没有提供一个对于标度指数 γ 和 α 的准确度量方法• 很多研究者对目前复杂网络文献中采用的统计分析持怀疑态度, 提出了合理的批评[12]• 在目前的文献中, 普遍采用简单的图形方法来估计幂律函数形式的标度指数, 例如对原始数据在双对数图上直接进行线性拟合[1], 或者对对数化分间隔后的数据进行线性拟合[2]• 然而, 仅需一个简单的试验[13]就可说明, 基于双对数坐标线性拟

* 收稿日期: 2005_09_30; 修订日期: 2006_07_06
基金项目: 国家自然科学基金(重大) 研究项目(70431002); 国家自然科学基金资助项目(70401019)
作者简介: 杨波(1979—), 女, 江西人, 博士(联系人. E_mail: brendayang7920@ yahoo. com. cn);
段文奇(1976—), 男, 湖南人, 博士(E_mail: wenqiduan@ 126. com)•

合的图形方法来拟合幂律分布是有偏, 不准确的• 此外, 由于目前研究者在网络分析中运用不同的图形拟合方法, 因此得到的标度指数估计值之间不一致, 不具有可比性• 而且, 网络模型的标度指数估计值不准确, 将给建立在这些网络模型之上的进一步研究, 如无标度网络上各种动力学过程的研究[14,15], 带来很大的影响•

本文旨在提出一种新方法估计复杂网络的幂律度分布和层次聚集函数(即上所指层次模块性)的标度指数, 试图克服图形估计方法的不足•

1 方　　法

估计幂律度分布指数　给定一个由 N 个结点组成的网络, 假定其度分布服从幂律分布, 则可表达为 $p(k;\gamma)=k^{-\gamma}/\zeta(\gamma)$[13], 其中 γ 是待估计的未知常数参量, $\zeta(\gamma)$ 为黎曼 zeta 函数 $\sum_{k=1}^{\infty}k^{-\gamma}$, $p(k)$ 满足 $\sum_k p(k)=1$• 对于这个给定的网络, 可得到每个结点的邻接点数目的观察值, 共有 N 个, 分别为 $k_1, k_2, \ldots, k_N$, 且是相互独立的• 则似然函数由下列乘积形式给出

$$L(k_1,k_2,\ldots,k_N\mid\gamma)=L=\prod_{i=1}^{N}p(k_i;\gamma)=\prod_{i=1}^{N}\frac{k_i^{-\gamma}}{\zeta(\gamma)}, \tag{1}$$

式(1)两边取对数, 得到对数形式的似然函数为

$$\Lambda=\ln L=\sum_{i=1}^{N}(-\gamma\ln(k_i)-\ln(\zeta(\gamma)))=-\gamma\sum_{i=1}^{N}\ln(k_i)-N\ln(\zeta(\gamma)), \tag{2}$$

对参数 γ 进行最大似然估计(MLE), 即在式(2)中对 γ 求导后求解0点, 所得到的0点(最大值)即为度指数 γ 的估计值• 此过程可得到方程(3)

$$\frac{\zeta'(\gamma)}{\zeta(\gamma)}=-\frac{1}{N}\sum_{i=1}^{N}\ln(k_i), \tag{3}$$

其中 $\zeta'(\gamma)$ 为黎曼 zeta 函数的导数• 方程(3)的解即为度指数 γ 的最大似然估计值•

下面根据方程(3)给出度指数 γ 估计的数值求解过程• 为方便读者的运用, 过程描述中在必要的情况下提供了计算中用到的 Matlab 代码或函数•

(ⅰ) 预处理: 给定网络的邻接矩阵, 通过调用 Matlab 函数 sum 可获得每个结点的度数 $k_1, k_2, \ldots, k_N$ (这里得到的数据是未分间隔的原始数据)•

(ⅱ) 估计: 调用 Matlab 函数 fsolve 解方程(3), 解的初始估计值一般设定在 1.5 至 3 之间(此范围的设定得到大量实证研究的支持[1])• 在给定的初始解估计值下, 如果得到的解 γ 小于0或者 γ 的求解过程中止后未得到解, 则需重新设定解的初始估计值, 并重复(ⅱ)•

估计层次指数　层次模块性的表达式为 $C(k)=C_0k^{-\alpha}$, 其中参数 α 为层次指数• 参数 α 和系数 C_0 是需要估计的量• 根据定义[10], $C(k)$ 是网络中所有度为 k 的结点的平均聚集系数, 可写为

$$C(k)=\frac{\sum_{j=1}^{N_k}C_j(k)}{N_k}=\frac{\sum_{j=1}^{N_k}\frac{C_j(k)}{N}}{N_k/N}=\frac{\sum_{j=1}^{N_k}\frac{C_j(k)}{N}}{S(k)}, \tag{4}$$

其中 $C_j(k)$ 是度为 k 的结点 j 的聚集系数• N_k 为网络中度为 k 的结点数目• $S(k)=N_k/N$ 是由样本数据直接得到的经验度分布• 注意到 $C(k)=C_0k^{-\alpha}$, 由(4)可得

$$\sum_k(C_0k^{-\alpha}S(k))=\frac{1}{N}\sum_k\left(\sum_{j=1}^{N_k}C_j(k)\right)=\frac{1}{N}\sum_{i=1}^{N}C_i, \tag{5}$$

其中, C_i 为结点 i 的局部聚集系数• 式(5) 中右边项恰为整个网络的聚集系数 C[1]• 另一方面, 根据 $C(k)$ 的定义可得(6) 式成立•

$$\sum_k C(k) = \sum_k C_0 k^{-\alpha} = \sum_k \left(\sum_{j=1}^{N_k} \frac{C_j(k)}{N_k} \right), \tag{6}$$

联立方程组(5) 和(6) 可知, 在给定实际网络数据从而结点局部集聚系数已知的情况下, 该方程组仅有两个变量 α 和 C_0•

类似地也可以得到层次指数 α 估计的数值求解过程

(ⅰ) 预处理: 给定网络的邻接矩阵, 计算经验度分布 $S(k)$; 每个结点的局部集聚系数 $C_i = 2E_i/(k_i(k_i-1))$ $(i = 1, 2, \cdots, N)$[16]; C_i 的平均值 C 以及网络中度为 k 的 N_k 个结点的局部聚集系数的平均值 $\sum_{j=1}^{N_k} \frac{C_j(k)}{N_k}$•

(ⅱ) 估计: 联立方程组(5) 和(6) 求解 α 和 C_0, 调用的 Matlab 函数为 fsolve, 根据 Dorogovtsev 等[17] 和 Ravasz 等[10] 的研究发现, 这里解的初始估计值 $[\alpha, C_0]$ 一般设定为 $[1, 1]$• 在给定的初始解估计值下, 如果得到的解 $[\alpha, C_0]$ 小于 0 或者 $[\alpha, C_0]$ 的求解过程中止后未得到解, 则需重新设定解的初始估计值, 并重复(ⅱ)•

2 应用举例

下文将采用 CNN 模型(Connecting Nearest_Neighbor model)[18] 来检验本文提出的方法的估计效果• 这 10 个 CNN 网络是由重复执行下述规则生成的: 1) 以概率 1- μ 引入一个新结点, 它与随机选中的结点 j 之间建立一条边; 2) 以概率 μ 随机选中一条潜在边, 将其转换为实际边• 这样得到的网络具有幂律度分布, 度指数 γ 是 μ 的函数• 而且也具有层次模块性的特征, 其层次指数 α 约为 0.6• 运用本文提出的方法, 可获得 γ 和 α 的估计值, 如表 1 所示• 结果证实了 CNN 模型能够复制出幂律度分布和层次模块性的特征•

在目前的复杂网络研究中, 有关标度指数的估计几乎没有考虑对带估计指数值的幂律函数假设的拟合优度检验• 因此, 获得的估计结果以统计研究者的观点来看并不严格• 虽然拟合优度检验也并非本文想阐述的重点, 但为了证实所估计结果具有更高可信度, 这里以幂律度分布为例, 通过运用 Kolmogorov_Smirnov(KS) 检验方法, 对本文新方法(记为 typeⅠ) 和图形方法(记为 typeⅡ) 的估计效果进行比较• 表 2 给出了对 10 个 CNN 网络采用两种不同方法得到的 KS 检验统计量的比较• 结果表明, 运用本文提出的方法估计度指数 γ 的幂律度分布的拟合优度要高于图形估计方法下得到的幂律度分布的拟合优度•

表 1 新方法下 10 个 CNN 网络的 γ 和 α 的估计值

	γ	α		γ	α
CNN01	1.803 6	0.779 9	CNN06	1.698 2	0.798 1
CNN02	1.768 2	0.790 0	CNN07	1.672 2	0.733 6
CNN03	1.764 3	0.728 2	CNN08	1.664 7	0.760 4
CNN04	1.732 7	0.746 2	CNN09	1.646 0	0.758 4
CNN05	1.719 5	0.758 7	CNN10	1.616 5	0.747 1

表 2 对 10 个 CNN 网络采用两种不同方法得到的 KS 检验统计量的比较

	TypeⅠ	TypeⅡ		TypeⅠ	TypeⅡ
CNN01	0.151 8	0.226 9	CNN06	0.166 9	0.288 9
CNN02	0.163 6	0.253 9	CNN07	0.163 6	0.297 9
CNN03	0.145 9	0.237 9	CNN08	0.178 0	0.315 9
CNN04	0.169 8	0.275 9	CNN09	0.184 8	0.331 9
CNN05	0.162 8	0.274 9	CNN10	0.193 0	0.354 9

3 结　语

本文提出的估计标度指数 γ 和 α 的方法对于复杂网络研究具有重要价值• 该方法能够有效克服图形估计方法的不足,适用于估计实际网络的两个标度指数• 此外,对于实证网络数据分析而言,它提供了一个具有鲁棒性的标准数值计算方法,便于比较不同的实际网络之间在幂律度分布和层次模块性特征上存在的差异,有助于对实际网络进行分类• 再者,对于网络模型的理论建构而言,评价模型的重要一点就是看理论模型和现实网络的符合程度,例如评估理论模型生成的网络在幂律度分布和层次模块性方面与现实网络符合的程度• 然而目前大多数的网络模型都并非静态的[16],因此需要提出方法能够动态的评估这些理论模型• 本文给出的方法有助于达到这一目的•

[参　考　文　献]

[1] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, **45**(2): 167—256.

[2] Albert R, Jeong H, Barabási A_L. Diameter of the World Wide Web[J]. Nature, 1999, **401**(6749): 130—131.

[3] Faloutsos M, Faloutsos P, Faloutsos C. On power_law relationships of the Internet topology[J]. ACM SIGCOMM Computer Communications Review, 1999, **29**(4): 251—262.

[4] Liljeros F, Edling C R, Amaral L A N, et al. The web of human sexual contacts[J]. Nature, 2001, **411**(6840): 907—908.

[5] Redner S. How popular is your paper? an empirical study of the citation distribution[J]. Eur Phys J B, 1998, **4**(2): 131—134.

[6] Albert R, Barabási A_L. Topology of evolving networks: local events and universality[J]. Phys Rev Lett, 2000, **85**(24): 5234—5237.

[7] Sigman M, Cecchi G. Global organization of the Wordnet lexicon[J]. Proc Nat Acad Sci USA, 2002, **99**(3): 1742—1747.

[8] Eckmann J_P, Moses E. Curvature of co_links uncovers hidden thematic layers in the World Wide Web[J]. Proc Nat Acad Sci USA, 2002, **99**(9): 5825—5829.

[9] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, **297**(5586): 1551—1555.

[10] Ravasz E, Barabási A_L. Hierarchical organization in complex networks[J]. Phys Rev E, 2003, **67**(2): 026112.

[11] Vázquez A, Dobrin R, Sergi D, et al. The topological relationship between the large_scale attributes and local interaction patterns of complex networks[J]. Proc Nat Acad Sci USA, 2004, **101**(52): 17940—17945.

[12] Alderson D, Doyle J C, Li L, et al. Towards a theory of scale_free graphs: definition, properties, and implications[J]. Internet Math, 2005, **2**(4): 431—523.

[13] Goldstein M L, Morris S A, Yen G G. Problems with fitting to the power- law distribution[J]. Eur Phys J B, 2004, **41**(2): 255—258.

[14] Zhou T, Wang B H. Catastrophes in scale_free networks[J]. Chinese Phys Lett, 2005, **22**(5): 1072—1075.

[15] Duan W Q, Chen Z, Liu Z R. Phase transition dynamics of collective decision in scale_free networks

[J]. Chinese Phys Lett, 2005, **22**(5): 2137—2139.

[16] Albert R, Barabási A_L. Statistical mechanics of complex networks[J]. Rev Modern Phys, 2002, **74**(1): 47—97.

[17] Dorogovtsev S N, Goltsev A V, Mendes J F F. Pseudofractal scale_free web[J]. Phys Rev E, 2002, **65**(6): 066122.

[18] Vazquez A. Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations[J]. Phys Rev E, 2003, **67**(5): 056104.

A New Method to Estimate Scaling Exponents of Power_Law Degree Distribution and Hierarchical Clustering Function for Complex Networks

YANG Bo[1], DUAN Wen_qi[2], CHEN Zhong[1]

(1. Antai College of Economics & Management, Shanghai Jiaotong University, Shanghai 200052, P. R. China;
2. School of Business Administration, Zhejiang Normal University, Jinhua, Zhejiang 321004, P. R. China)

Abstract: A new method and corresponding numerical procedure were introduced to estimate scaling exponents of power_law degree distribution and hierarchical clustering function for complex networks. This method could overcome the biased and inaccurate faults of graphical linear fitting methods commonly used in current network research. Furthermore, it has been verified to have higher goodness_of_fit than graphical methods by comparing the KS test statistics for 10 CNN networks.

Key words: parameter estimation; complex networks; power_law; degree distribution; hierarchical modularity