第 53 卷 第 11 期 2004 年 11 月
1000-3290/2004/53(11)/3695-06

物 理 学 报
ACTA PHYSICA SINICA

Vol.53,No.11,November,2004

具有幂率度分布的因特网平均最短路径长度估计*

李 † 山秀明 任 勇

(清华大学电子工程系,北京 100084)

(2004 年 5 月 20 日收到;2004 年 6 月 25 日收到修改稿)

针对具有幂律度分布的复杂网络的平均最短路径长度进行了研究.给出了一个计算因特网平均最短路径长度 $< l >$ 的公式.提出因特网的整体构造实质是以最小代价换取最大收益,从此出发通过对因特网这类复杂网络平均最短路径长度影响因素分析,推断出网络最短路径长度分布 $P(l)$ 的基本性质,进而构造了一个只含参数 α 的 $P(l)$ 的简洁形式,直接打通了拓扑具有幂律度分布的因特网的度分布 $P(k)$ 与网络最短路径长度分布 $P(l)$ 之间的关系.然后说明了导出的 $< l >$ 公式的物理意义,即参数 α 代表网络节点连接方式和网络的总边数对平均最短路径长度 $< l >$ 的影响.此公式意味着只要掌握幂律 $P(k) \sim k^{-\alpha}$ 中 α 值,就可以直接计算相应网络的平均最短路径长度 $< l >$.通过对已知数据的计算,验证了公式的有效性.阐明了参数 α 对网络设计的重要性和对网络功能的影响.

关键词:复杂网络,幂律,度分布,平均最短路径长度

PACC:0540J,8980H

1. 引 言

因特网的功能就是使人们更快、更好、更便捷地进行通讯,其速度的快慢是因特网功能的最重要体现之一.因特网速度由网络协议和网络拓扑决定,其中网络拓扑为信息传递速度的基础.但因特网是当前人工制造的最复杂网络结构,并且规模在不停地扩张,结构在不停地变化,如何有效估计这样的复杂网络系统中最短路径问题由于其理论和现实意义而引人注目,但也由于其解决上的难度而进展缓慢.

对于因特网这样复杂的网络,人们甚至无法确切掌握其内部结构的全部信息,加之网络规模的巨大,传统图论提供的最短路径计算方法已经失效.而正在兴起的复杂网络(complex networks)理论给相关研究提供了另外的思路.

复杂网络是基于网络统计特性,研究网络的诸如平均最短路径等结构元素的方法.起端为 1959 年 Erdös 和 Rényi 研究随机图论(random graphs)开始,后由 Barabási 等人进行了发展[1—5].由于因特网研究者获取的多是网络结构的统计信息,因此应用统计信息研究因特网具有现实可行性.

网络度(每个网络节点与其他节点连接的边数)分布是复杂网络理论中对网络特征的重要描述,现实中包括因特网拓扑在内很多复杂网络,如 WWW 网[5]、生物细胞网[6]、演员合作网[7]、人类性接触网[8],在拓扑上属于标度自由(scale-free)网络,度分布(幂律 distribution of degree)符合重拖尾分布,即 $P(k) \sim k^{-\alpha}$.这些复杂网络的另外一些重要特性还有小世界(small-world)特性和高度聚集(high clustering)特性,这使这些复杂网络的平均最短距离在网络规模增长很快时增长却不大,比如 WWW 网在网络规模为 325729 时平均最短路径仅为 11.2,这种特性按照随机图论表述为 $< l > \sim \ln(N)$,$< l >$ 为任意节点间的平均最短路径长度,N 为网络节点数.所谓复杂网络的路径长度是指网络中从任意一个节点跳到任意其他节点经过的路径中包含的边的数目,那么最短路径就是连接任意两点的最短路径中边的数目.而平均最短路径长度就是网络中任意节点间连接的平均最小边数.很多研究都在研究复杂网络的平均最短路径长度,Albert, Jeong and Barabási[5] 给出公式 $0.35 + 2.06\log(N)$ 用来计算 WWW 网的平均最短路径长度 $< l >$,还有 Lu 等人[9]也进行了相应工作.主要思路都是以网络规模(N)为基础,针对不同的复杂网络给出计算公式.但这样的估计由于只包含了网络节点数 N,很难解释

* 国家自然科学基金(批准号:90204004 和 90304005)资助的课题.

† 通讯联系人. E-mail:liy@mail.ee.tsinghua.edu.cn

网络拓扑与平均最短路径之间的关系,因为即便是同样的网络规模(N),在以不同方式连接时,$< l >$会有很大不同.另外,对同一属性的网络,比如两个WWW网,在符合$P(k) \sim k^{-\alpha}$分布时,指数α往往不同,这是由于其拓扑结构上的细节分布差异造成的,而只包含一个N参数的公式无法涵盖不同α对此造成的影响.

针对以上问题,本文首先分析了在具有幂律特征的复杂网络中影响$< l >$大小的主要因素,并以网络的度分布的指数α和网络规模N为主要参数,针对因特网给出了计算最短平均路径长度$< l >$的公式(此公式比只含N参数的公式具有更强的物理意义),通过对真实因特网的自治域(autonomous system,AS)水平和路由(router)水平的$< l >$进行计算,结果证明公式有效.

2. 复杂网络最短路径长度影响因素分析

复杂网络的平均路径长度$< l >$由很多因素影响,包括网络规模N,网络的总边数E和各节点间的连接方式等等.由*Erdös*和*Rényi*创建的随机图理论[1],给出

$$< l > \sim \frac{\ln(N)}{\ln(< k >)},$$

k为每个节点的度,$< k >$为网络的度平均.这个关系说明,网络中平均最短路径长度与网络规模N的对数成正比,与网络度平均的对数成反比.实际上,网络的总边数和节点的连接方式都与网络图的度k的分布直接相关,即与$P(k)$有关:网络的总边数为

$$\sum k/2 = \sum P(k) \cdot k/2,$$

即每个节点的度的和的1/2,而节点的连接方式可以由度的分布方式代表,即$P(k)$的形式代表.因此$< l >$可以认为由$P(k)$和N决定,将$< l >$看作$P(k)$和N的函数,即$< l > = f(P(k), N)$.

3. 路径长度分布$P(l)$性质分析与构造,及平均最短路径长度$< l >$的导出

因特网这样的网络的每个局部,在构造中蕴涵了人们的设计准则,去掉细节上的差异,它最根本的目标就是要以最小的代价换取最大的收益,具体而言就是以最经济的网络连接(总边数E)实现最快的信息传递速度(最小平均路径长度$< l >$).但这是一对矛盾目标,小$< l >$往往意味大E.如果人们只追求效率,要求以最短距离构造节点的连接,那么最终可以看到的是一个全连接网络,各节点之间只需要一跳就可到达,但这时,网络的成本会显得不可接受,每连接一个新节点都需要把它和已经存在的所有点进行连接,其网络付出的代价太大.这里将这种希望以小代价换取大效益的问题表述为以下的多目标优化问题:

$$\min(< k >) = \min\Big(\sum_{k=1}^{k_{\max}} P(k) \cdot k\Big),$$

$$\min(< l >) = \min\Big(\sum_{l=1}^{l_{\max}} P(l) \cdot l\Big),$$

$$k = 1,2,\cdots, k_{\max}, \quad l = 1,2,\cdots, l_{\max}, \qquad (1)$$

其中$< k >$为平均度分布,之所以将$< k >$代替E作为优化目标函数,是因为

$$E = \sum k/2 = \sum P(k) \cdot k/2 = < k > \cdot N/2,$$

$< k >$代表每个节点连接边的情况,与网络规模无关,代表了网络的平均成本,更具代表性.$k_{\max}$为可能最大度,$P(k)$为度分布,$< l >$为平均路径长,$l_{\max}$为最大路径长度($\leqslant N-1$),$P(l)$为路径长度分布函数.这两个目标函数并非相互独立,如上文所述,通过$< l > = f(P(k), N)$相关联.

幂律$P(k) = k^{-\alpha}$的度分布促使网络整体代价趋向最小,那么什么样的$P(l)$能够促使$< l >$显现小世界特性是这里需要思考的问题.

因为平均最短距离$< l >$为$P(l)$的函数,所以计算$< l >$的关键就是找到$P(l)$的形式.下面将通过分析推断$P(l)$的性质.

首先,为达到平均路径最小,路径长短的概率分布$P(l)$,从整体上也需要路径短的比重大,路径长的比重小,即$P(l)$是l的递减函数,才能使得$< l >$趋向尽量小,显现小世界的特性.因此我们推断$P(l)$是l的递减函数.

此外,因为因特网的度分布$P(k) \sim k^{-\alpha}$(一般地$2 < \alpha < 3$),又有

$$< l > = \sum_{l=1}^{l_{\max}} P(l) \cdot l = f(P(k), N),$$

因此$P(l)$一定与指数α有关,$P(l)$一定可以表述成α的函数形式.再有,应用随机图论的公式

$$< l > \sim \frac{\ln(N)}{\ln(< k >)},$$

将$< l >$和$< k >$代入其中,得

$$\sum_{l=1}^{l_{\max}} P(l)\cdot l \sim \frac{\ln(N)}{\ln\left(\sum_{k=1}^{k_{\max}} P(k)\cdot k\right)}. \qquad (2)$$

(2)式说明 $P(l)$与 $P(k)$为反比关系,其具体涵义为:当 $P_{\alpha 1}(k) > P_{\alpha 2}(k)$,那么 $P_{\alpha 1}(l) < P_{\alpha 2}(l)$,后面的内容还会详细解释这一点.

从以上分析我们推断 $P(l)$具有以下几个性质:

$P(l)$ 为 l 的减函数形式;

$P(l)$ 可以表示成参数 α 的函数形式;

$P(l)$ 与 $P(k)$为反比关系.

下面的工作就是找到一个 $P(l)$ 的形式,满足以上三个性质.这里由以上分析结果,并结合实际情况调整,针对因特网构造了(3)式的表达形式来计算 $P(l)$,在性质上满足以上三个要求.

$$P(l) = \left(1-\frac{1}{\alpha}\right)^{1.7}\cdot l^{-\frac{3\ln(5.55-\alpha)}{5\ln\alpha}-1}, \qquad (3)$$

其中 α 为幂律度分布 $k^{-\alpha}$ 的指数,l 为自然数序列 1,2,…

下面分析构造的 $P(l)$的性质及其物理意义:

首先,按照(3)式的定义,$P(l)$ 为 l 的减函数,从图 1 可以看到这个性质,这个性质使 $< l >$ 可以保持尽量最小,在网络规模增长很大时,$< l >$ 保持一个缓慢的增长速度.

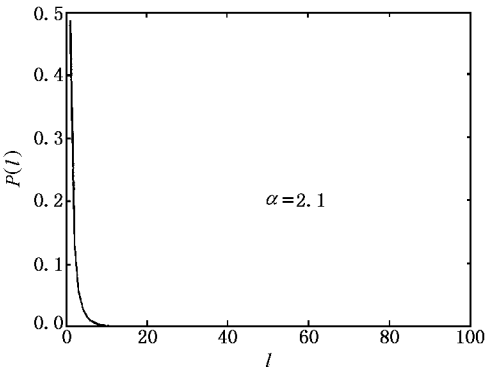


图1　$P(l)$为 l 的减函数　——为 $P(l) = (1-1/\alpha)^{1.7} \times l^{-3\ln(5.55-\alpha)/(5\ln\alpha)-1}$

其次,$P(l)$ 为参数 α 的函数,并且只含 α 一个参数.

最后,$P(l)$ 与 $P(k)$为反比关系,如图 2 所示:对于 $P(l)_{\alpha=2.4} < P(l)_{\alpha=2.9}$,相应 $P(k)_{\alpha=2.4} > P(k)_{\alpha=2.9}$.

由(3)式给出的 $P(l)$表达式反应了网络度分布 $P(k)$和路径长度分布 $P(l)$之间的联系,参数 α 构成了 $P(k)$和 $P(l)$之间的纽带.实际上,(3)式中指数 α 包含了节点的连接方式和总边数对平均最短路径长度的影响.因此(3)式给出的 $P(l)$在理论上包含了影响平均最短路径长度的因素,所以逻辑上合理.下一步即将计算 $< l >$:

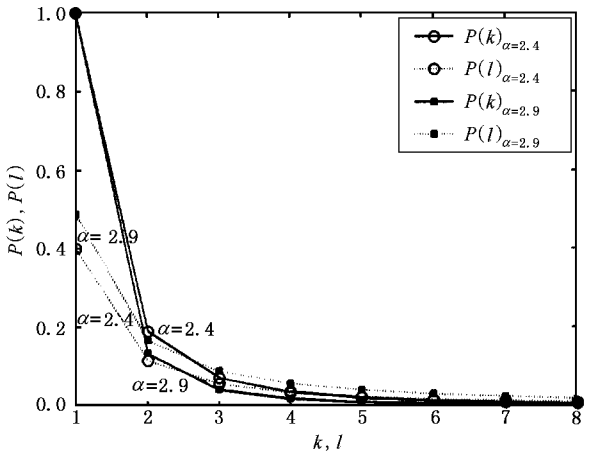


图2　$P(l)$ 与 $P(k)$为反比关系

$$\begin{aligned} < l > &= \sum_{l=1}^{l_{\max}} P(l)\cdot l \\ &= \sum_{l=1}^{l_{\max}} \left(1-\frac{1}{\alpha}\right)^{1.7}\cdot l^{-\frac{3\ln(5.55-\alpha)}{5\ln\alpha}-1}\cdot l \\ &= \left(1-\frac{1}{\alpha}\right)^{1.7}\cdot \sum_{l=1}^{l_{\max}} l^{-\frac{3\ln(5.55-\alpha)}{5\ln\alpha}} \\ &= \left(1-\frac{1}{\alpha}\right)^{1.7}\cdot \sum_{l=1}^{l_{\max}} l^{-\frac{0.6\ln(5.55-\alpha)}{\ln\alpha}}, \end{aligned} \qquad (4)$$

其中 $l_{\max}$为可能的最大路径长度,它一定小于等于 $N-1$,因此将(4)式近似为

$$\begin{aligned} < l > &= \left(1-\frac{1}{\alpha}\right)^{1.7}\cdot \sum_{l=1}^{l_{\max}} l^{-\frac{0.6\ln(5.55-\alpha)}{\ln\alpha}} \\ &\approx \left(1-\frac{1}{\alpha}\right)^{1.7}\cdot \sum_{l=1}^{N-1} l^{-\frac{0.6\ln(5.55-\alpha)}{\ln\alpha}}, \end{aligned} \qquad (5)$$

其中 $< l >$ 由 N 和 α 两个参数确定,代表了影响最短平均路径的三个因素:网络规模、网络的总边数和节点间的连接方式,具有直观的物理意义.

4. 数值验证

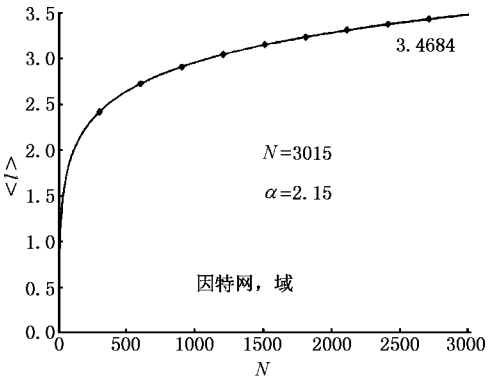
因特网的拓扑研究可以从不同水平进行,路由层次和自治域层次.在路由层次,一个节点是一个路由;在自治域层次,一个节点是一个域.文献[10—12]给出因特网在不同层次和规模的数据,这

里应用(5)式来计算$\langle l\rangle$,如图 3 和表 1 所示.

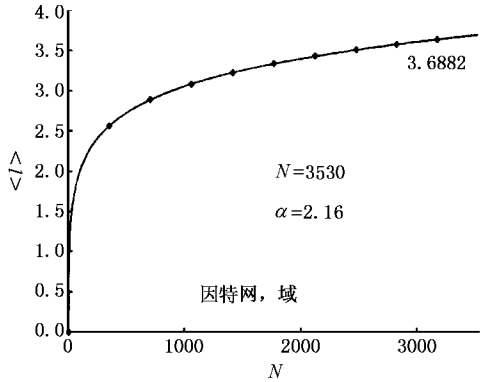
表 1 中误差用

$$\frac{|\ l_{\text{真实}} - l_{\text{预测}}\ |}{l_{\text{真实}}} \times 100$$

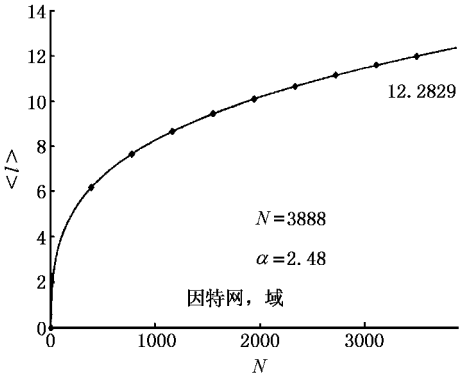
来计算.计算所得的预测值 $l_{\text{预测}}$ 和真实值 $l_{\text{真实}}$ 很接近.图 4 是真实值与预测值的比较,及误差示意图.



(a) No. 1



(b) No. 2

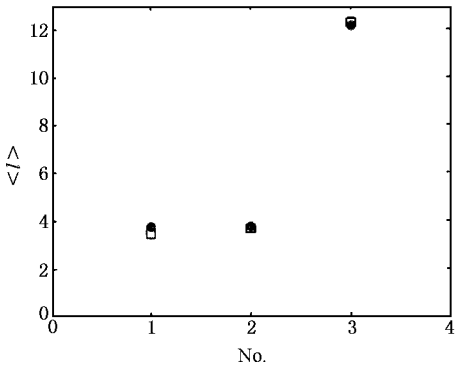


(c) No. 3

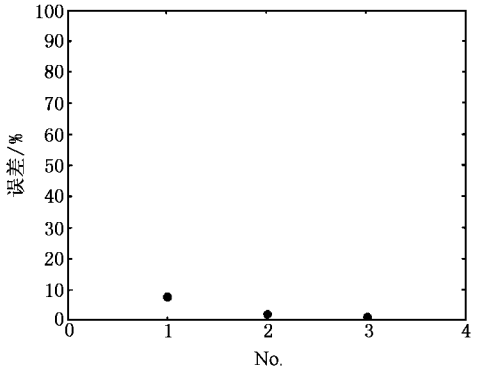
图 3 应用(5)式计算 $< l >$ 值

表 1 数值验证

网络	N	α	$l_{\text{真实}}$	$l_{\text{预测}}$	误差/%	No.
因特网,域	3015	2.15	3.76	3.468	7.7544	1
因特网,路由	3530	2.16	3.77	3.688	2.1701	2
因特网,路由	3888	2.48	12.15	12.283	1.094	3



(a)真实值(●)与预测值(□)的比较



(b)误差图

图 4 计算结果

计算结果比较鼓舞人心,最大误差为 7.7544%,最小误差为 1.094%,所有误差均小于 10%. 因此(5)式给出的对因特网拓扑最短平均路径的估计比较准确.

5. 进一步分析

(5)式含有参数 α 和 N,即可以写为 $< l > = g(\alpha, N)$.通过对(5)式的分析,在同样的网络规模 N 下,随着 α 的增长,$< l >$ 也会增长.图 5 表明当 α 只有一个微小变化(从 2.2 变为 2.21),$< l >$ 就有一个显著的改变.这说明 $< l >$ 对参数 α 非常敏感.实际上,参数 α 控制着平均最短路径.这提醒我们可以通过调整因特网拓扑的度分布,即通过调整 α 来有效控制网络的平均最短路径.这对于因特网拓

扑设计有启示意义.

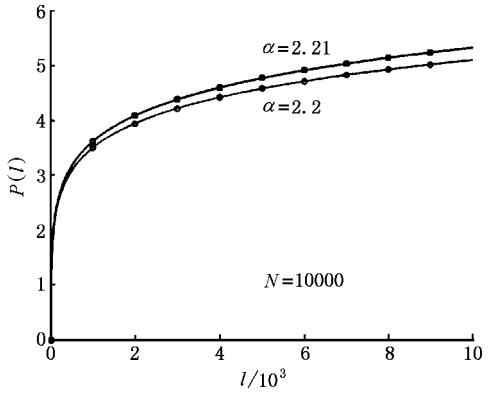


图 5　α 的微小变化能够带来 $<l>$ 的明显变化

6. 结论与展望

本文从因特网希望以最小代价换取最大收益的构造实质出发,通过对影响因特网这类复杂网络平均最短路径长度因素的分析,推断出网络最短路径长度分布 $P(l)$ 的基本性质,进而构造了一个只含参数 α 的 $P(l)$ 的形式,直接打通了拓扑具有幂律度分布的因特网的度分布 $P(k)$ 与网络最短路径长度分布 $P(l)$ 之间的关系,说明了幂律度分布中的参数 α 代表着网络节点连接方式和网络的总边数,直接影响平均最短路径长度.这意味着只要从统计上掌握幂律 $P(k)\sim k^{-\alpha}$ 中 α 的值,就可以直接计算相应网络的平均最短路径长度 $<l>$.通过对已知数据的计算,验证了公式的有效性.进而阐明了参数 α 对网络设计的重要性,α 的微小变化对网络性能带来明显影响.因此,未来的工作就是要探明各种设计因素对 α 的影响,进而可以更好地控制网络性能,实现网络功能.另外,本文的分析和结论显示对具有幂律度分布因特网的 $P(l)$ 可以用 α 的函数形式表示,因此作为推论其他具有重拖尾度分布的复杂网络的 $P(l)$,也一定具有相似的形式,只不过需要在表达形式中调整相应的参数,这也是下一步需要进行的工作之一.

[1] Erdös P and Rényi A 1959 *Publ. Math.* **6** 290

[2] Yuan J *et al* 2001 *Acta Phys. Sin.* **50** 1221 (in Chinese) [袁　坚等 2001 **物理学报 50** 1221]

[3] Liu F *et al* 2004 *Acta Phys. Sin.* **53** 373 (in Chinese) [刘　峰等 2004 **物理学报 53** 373]

[4] Yuan J *et al* 2000 *Acta Phys. Sin.* **49** 398 (in Chinese) [袁　坚等 2000 **物理学报 49** 398]

[5] Albert R, Jeong H and Barabási A L 1999 *Nature* **401** 130

[6] Fox J J and Hill C C 2001 *Chaos* **12** 809

[7] Barabási A L and Albert R 1999 *Science* **286** 509

[8] Albert R and Barabási A L 2002 *Rev. Mod. Phys.* **74** 47

[9] Fan C and Lu L Y 2001 http://euclid.ucsd.edu/llu/diameter

[10] Faloutsos M, Faloutsos P and Faloutsos C 1999 *Proc. Comput. Commun. Rev.* **29** 251

[11] Govindan R and Tangmunarunkit H 2000 *Proc. IEEE Infocom.* **3** 1371

[12] Siganos G, Falortsos M, Faloutsos P and Faloutsos C 2003 *IEEE/ACM Transactions on Networking* **11** 514

Average path length of Internet with power law degree distribution*

Li Ying† Shan Xiu-Ming Ren Yong

(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

(Received 20 May 2004; revised manuscript received 25 June 2004)

Abstract

The average path length of Internet with power law degree distribution is studied in this paper. The main work is to give a formula to compute the average path length $< l >$. The essential of Internet or many other complex networks is to realize the least cost (the smallest total number of edges) and the best benefit (the short average path length), which can be expressed as two optimization factors to minimize the average degree $< k >$ and the average path length $< l >$. By analyzing the main factor affecting $< l >$, we find the main property of path length distribution $P(l)$, and then give a formula of $P(l)$ only including one parameter α. This work shows that $P(l)$ and a direct relation with $P(k)$, and α represent the effective link between nodes and the total number of networks on $< l >$. This formula makes it possible that if we get the α of $P(k) \sim k^{-\alpha}$, we can calculate $< l >$ easily. The simulation shows the formula is quite exact. Our method is successful. In the end, we discuss that the parameter α is very important for Internet design.

Keywords: complex network, power law, degree distribution, average path length

PACC: 0540J, 8980H

* Project supported by the National Natural Science Foundation of China (Grant Nos. 90204004 and 90304005).

† Corresponding author. E-mail: liy@mail.ee.tsinghua.edu.cn