第34卷 第21期 计 算 机 工 程 2008年11月
Vol.34 No.21 Computer Engineering November 2008
·网络与通信· 文章编号：1000—3428(2008)21—0092—03 文献标识码：A 中图分类号：TP393

Internet的复杂网络统计规律研究与分析

关 沫[1,3]，李 波[2]，赵 海[3]
(1. 沈阳工业大学信息科学与工程学院，沈阳 110023；2. 中国人民解放军65021部队，沈阳 110162；
3. 东北大学复杂网络研究中心，沈阳 110004)

摘 要：运用CAIDA Skitter项目提供的Internet监测数据和统计手段对Internet的平均最短路径、节点度分布所具有的统计规律进行研究与分析。统计结果显示Internet的拓扑结构是不均匀的，其内部集聚程度较高，网络中任意2个节点间的最短距离很小，整个网络的平均最短路径约为14跳，且Internet节点的度分布服从幂律分布，证实了Internet不是一种随机网络，而是一种同时具有“小世界”和“无尺度”网络特征的复杂网络。

关键词：复杂网络；平均最短路径；幂律；度分布；无尺度

Research and Analysis of Complex Networks Statistical Regularity for Internet

GUAN Mo[1,3], LI Bo[2], ZHAO Hai[3]
(1. School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110023;
2. 65021 Troop of PLA, Shenyang 110162;
3. Complex Networks Research Center, Northeastern University, Shenyang 110004)

【Abstract】 Using the Internet monitoring data offered by CAIDA Skitter project and the method of statistics, the statistical regularity of average shortest path and node degree distribution of the Internet are studied and analyzed. Statistical result reveals the topology of the Internet is not homogeneous. It is a highly clustered network, which means the shortest distance between two random nodes is short. The average shortest path of the Internet is about 14 hops. The node degree distribution is power-law distribution. The statistical results also show that the Internet is not random network, but complex network with the characteristics of “small world” and “scale-free”.

【Key words】 complex networks; average shortest path; power-law; degree distribution; scale-free

1 概述

Internet是迄今为止最复杂的人工系统，表面看来杂乱无章，但是在其复杂的表象下却隐藏着大范围模式的秩序与结构[1-2]。这里的大范围模式是指Internet中自发涌现出来的秩序以及产生这种秩序的机制。

本文利用国际 Internet 数据分析联合组织(Cooperative Association for Internet Data Analysis, CAIDA)提供的Internet监测数据，试图从亿万IP活动中找出Internet宏观有序度，并借助统计手段来揭示其所蕴藏的某些宏观统计规律。介绍了研究背景、数据来源以及几个典型的网络特征量和网络分类。通过对CAIDA Skitter项目提供的5 500余万条有效测试样本数据的研究，分析了 Internet 平均最短路径以及节点入度和出度分布所存在的统计规律。

2 研究背景及数据来源

CAIDA是一个对全球范围Internet结构及数据进行研究的国际合作机构，主要研究内容包括：Internet网络的产生、发展和演化趋势，Internet网络行为、动力特征以及Internet宏观拓扑结构的变化规律等。CAIDA在世界范围内的参与者共有30余家，分布于北美洲、欧洲、大洋洲、亚洲等众多国家的研究院所、军事机构及高等院校中。目前亚洲仅有3个监测节点：2个在日本东京，另外1个在中国东北大学复杂网络研究中心[3]。

CAIDA的Skitter项目主要是对分布于美洲、欧洲、大洋洲、亚洲的源主机进行监测。通过跟踪某个源地址到其他多个目的地址的前向IP地址的方式来获取Internet的拓扑结构；通过收集ICMP协议的TTL值绘制Internet节点间的关系；通过探测到达IPv4 地址空间的多个目的IP地址路径来实现网络连通性的可视化[4]。

本文选取 Skitter 项目中具有代表性的 k-peer(位于荷兰阿姆斯特丹)和 i-root(位于瑞典斯德哥尔摩)2 个节点的监测数据进行研究。数据经预处理后，从数据库中初步选取的样本数量为55 132 205条有效样本，时间跨度从2000年9月到2004年7月。对于 Skitter 项目数据，本文使用 SPSS Inc.公司的SPSS(Statistical Package for the Social Science)统计分析软件，并采用数理统计的方法揭示 Internet 随机现象背后所蕴藏的宏观规律性。

3 网络分类及其特征比较

3.1 网络特征量

以下介绍几个典型的网络特征量。

基金项目：辽宁省博士科研启动基金资助项目(20071024)
作者简介：关 沫(1976－)，女，讲师、博士，主研方向：复杂网络，嵌入式计算；李 波，工程师、硕士；赵 海，教授、博士生导师
收稿日期：2008-05-20 **E-mail：**guanmosy@163.com

度：设N是个网络，$V(N)$是所有网络节点的集合，$E(N)$是所有边的集合。节点v的度$d_v(v \in V)$是与此节点v连接的边的数量。

入度：对于有向网络 N，以节点 v 为终点的边的数目，称为节点 v 的入度。

出度：对于有向网络 N，以节点 v 为起点的边的数目，称为节点 v 的出度。

最短路径：在网络 N 中，任意 2 个节点 i 和 j 之间的所有通路中，连通这 2 个节点的最少边数，称为节点 i 和 j 之间的最短路径。网络中所有节点对的最短路径的平均值，称为网络的平均最短路径，是网络的全局特征。

集聚程度：指网络集团化的程度，即考察连接在一起的集团节点各自的近邻之中有多少是共同的近邻。网络中所有节点集聚程度的平均值称为网络的平均集聚程度，是网络的局部特征。

3.2 网络分类

大量真实网络既不是规则网络，也不是随机网络，而是具有与两者皆不同的统计特征的网络，这样的网络被称为复杂网络。

1998 年开展的一个描绘万维网的项目揭示了令人惊诧的事实：万维网是由少数高连接性的页面串联起来的，80%以上页面连接数不到 4 个，然而只占节点总数不到万分之一的极少数节点却有高达上百万乃至几十亿个链结[5]，这恰与“Internet是随机网络”的预期相背离。由于其缺乏一个描述问题的特征尺度，研究者把这种包含重要集散节点的网络称为“无尺度网络”。

按照网络的各种特征量对网络进行分类，可分为规则网络、随机网络、小世界网络和无尺度网络 4 种，其特征比较如表 1 所示。

表 1　4 种网络的特征比较

网络种类	平均集聚程度	平均最短路径	节点的度分布
规则网络	大	大	δ 分布
随机网络	小	小	泊松分布
小世界网络	大	小	类似泊松分布
无尺度网络	大	小	幂律分布

对这 4 种网络的说明如下：

规则网络：是一个规则的有迹可循的晶格点阵，网络中各节点的度相同。

随机网络：由 n 个顶点构成的图中，可以存在 C_n^2 条边。对于 C_n^2 中任何一个可能连接，以概率 p 进行连接而构成的网络称为随机网络。

小世界网络：网络节点数变化时，任意两节点间的平均最短路径变化相对缓慢，随节点数的增加呈对数增长，且网络具有较高的集聚程度的现象称为小世界效应。具有小世界效应的网络，称为小世界网络。

无尺度网络：网络中的节点度服从幂律(power-law)分布，即某个特定度的节点数目与这个特定的度之间的关系可以用一个形如$Y \propto X^C$的幂函数近似表示。其中，对于变量X和Y，存在一个常数C，使Y与X的C次幂成比例。将具有这种特征的网络称为无尺度网络。

4 Internet 的统计规律分析

4.1 Internet 的平均最短路径

平均最短路径是指网络中所有节点之间的最短路径的平均值，由于网络路由过程中要按最短路径原则进行选路，因此以网络中的路由跳距离(hop)代替网络平均最短路径。

(1)将 Skitter 项目的 k-peer 和 i-root 两个节点的 5 500 余万条有效样本进行综合统计，全部样本数据的 Internet 平均最短路径分布统计结果如图 1 所示。

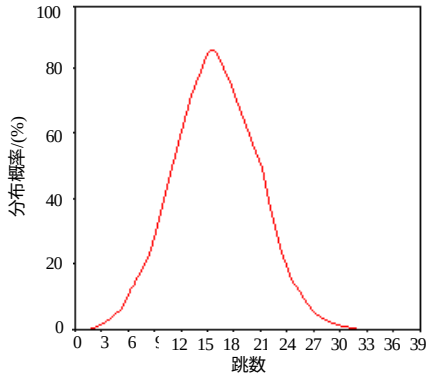


图 1　全部样本数据的 Internet 平均最短路径分布

从图 1 中可以看出其分布特征近似符合正态分布，偏态峰值很小，均值D_{means}=14.21，方差$Std.$=4.654。

(2)分别选取 i-root 和 k-peer 两个节点 2004 年 4 月 23 日的数据参与统计，其 Internet 平均最短路径分布的统计结果如图 2 所示。

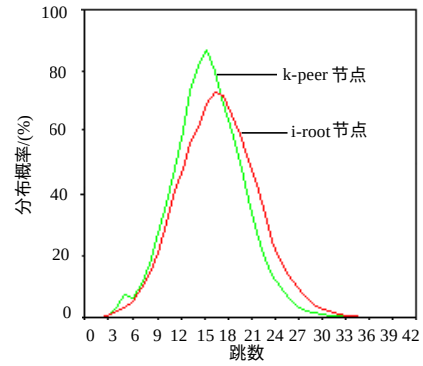


图 2　不同监测点 Internet 平均最短路径分布

从图 2 中可以看出 i-root 节点和 k-peer 节点的数据局部分布规律有所差异：k-peer 在跳数为 4 的附近有小的突起，表示该监测点数据中集聚状态较 i-root 监测点要好，但整体统计规律无太大变化，都近似符合正态分布。

通过以上对大样本容量的 Internet 平均最短路径的数据分析可以看出，高冗余数据表现出较强的统计规律性，即 Internet 的网络拓扑结构是不均匀的，其平均最短路径曲线峰度值集中收敛在[12, 18]之间，平均最短路径约为 14 跳左右，呈现出“小世界”网络特征，即网络的集聚程度相对高于随机网络，整个网络的平均最短路径较小，网络中存在捷径，这与规则网络和随机网络都不同。

4.2 Internet 节点的度分布

节点的度表示一个节点与其他节点的连接程度。如果一个节点直接连接的其他节点数越多，其度值就越高，表明其连通性越好。

随机网络理论有一项预测：尽管连接是随机的，但由此形成的网络却是高度民主的，网络中节点度的分布符合泊松分布。

Internet 中节点的连接由于受多种因素的影响而带有很大的随意性。这意味着 Internet 中的每个节点与其他节点都

有等可能的连接。

按照上述随机网络理论的预测，可以推测 Internet 的整体拓扑结构应该是均匀的，即 Internet 应该是一个随机网络，其节点的度分布应该服从泊松分布。但近年来的研究发现却否定了这个推测，即 Internet 的节点的度分布不服从泊松分布，而是呈“幂律”分布。

以下对 Internet 节点的度分布分别从节点的入度和出度两方面所具有的统计规律进行分析，从统计的角度证明 Internet 节点的度分布符合“幂律”分布，并给出其幂指数。

4.2.1 入度分布统计

选取 2004 年 4 月 23 日的数据对 Internet 节点的入度分布进行统计。对数据样本中节点的入度数据进行频数统计，其统计结果如图 3 所示。

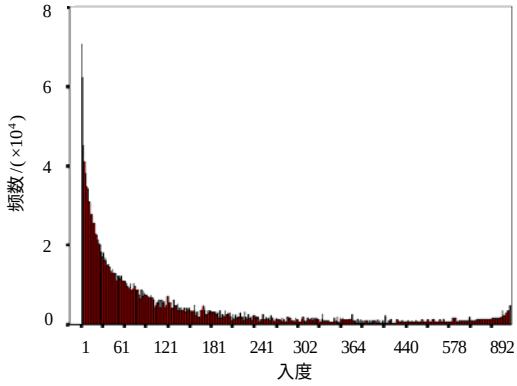


图 3 Internet 节点入度和频数的统计关系

从图 3 中可以看出 Internet 节点入度的分布曲线近似于幂律分布。

将节点入度的统计频数进行曲线拟合，拟合时选用多种模型：LIN(线性函数)，LOG(对数函数)，INV(逆函数)，QUA(二次函数)，CUB(三次函数)，COM(复合函数)，POW(幂函数)，S(S 形曲线)，GRO(生长函数)。Internet 节点入度的曲线拟合数据见表 2。

表 2 Internet 节点的入度曲线拟合表

Mth	Rsq	d.f.	F	Sigf	*b*0	*b*1	*b*2	*b*3
LIN	0.08	58	5.07	0.02	7 429.29	-6.458 8		
LOG	0.70	58	141.03	0.00	42 489.1	-6 885.5		
INV	0.88	58	208.73	0.00	3 050.42	71 296.9		
QUA	0.22	57	8.34	0.01	12 124.4	-26.380	0.009	
CUB	0.39	56	12.18	0.00	18 379.7	-70.115	0.065	-2.E-05
COM	0.11	58	7.65	0.08	2 695.60	0.999		
POW	0.91	58	111.60	0.00	13 7662	-2.887 4		
S	0.20	58	15.09	0.00	7.413 9	4.345 6		
GRO	0.11	58	7.65	0.08	7.899 4	-0.009		

在表 2 的各模型检验参数中，Rsq(决定系数R^2)是表示样本观测值聚集在样本回归线周围的紧密程度的指标，其数值越接近 1 说明其拟合程度越好。从表 2 的拟合结果看，Internet 节点入度分布拟合优度最高的是POW分布，其Rsq值在所有模型中最大，即Rsq=0.918。

POW 模型的拟合公式为

$$Y=b_0X^{b1}$$

其中，$b1$ 为其幂指数。

从以上数据分析可知，Internet 节点的入度分布符合幂律分布，其幂指数为-2.887 4。

4.2.2 出度分布统计

选取 2004 年 4 月 23 日的数据对 Internet 节点的出度分布进行统计。对数据样本中节点的出度数据进行频数统计，其统计结果如图 4 所示。

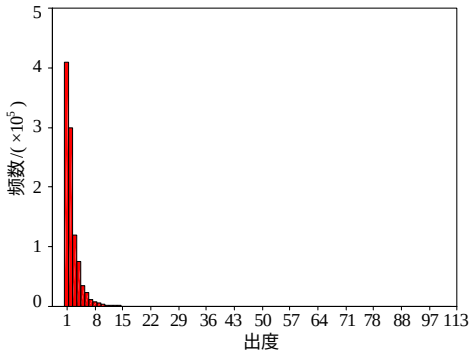


图 4 Internet 节点出度和频数的统计关系

从图 4 中可以看出 Internet 节点出度的分布曲线也近似于幂律分布。

将节点出度的统计频数进行曲线拟合，同样在拟合时也选用多种模型：LIN, LOG, INV, QUA, CUB, COM, POW, S, GRO。Internet 节点出度的曲线拟合数据见表 3。

表 3 Internet 节点的出度曲线拟合表

Mth	Rsq	d.f.	F	Sigf	*b*0	*b*1	*b*2	*b*3
LIN	0.156	50	1.85	0.20	688.3	-676.4		
LOG	0.626	50	16.77	0.02	23 496	-5 621		
INV	0.907	50	432.41	0.00	-14	61 332		
QUA	0.371	49	2.66	0.12	13 229	-305 4	13.886	
CUB	0.601	48	4.02	0.05	20 661	-900 0	104.62	-0.34
COM	0.590	50	14.39	0.04	105.8	0.944		
POW	0.969	50	311.20	0.00	22 824	-3.040		
S	0.747	50	29.57	0.00	1.793	23.323		
GRO	0.590	50	14.39	0.04	6.990	-0.057		

从表 3 的拟合结果来看，节点出度分布拟合优度最高的也是 POW 分布，其 Rsq 值为 0.969。因此，Internet 节点的出度分布也符合幂律分布，其幂指数为-3.040。

通过对Internet节点入度和出度数据的统计分析，证明了Internet中节点的连接度分布呈“幂律”分布：对于节点的入度分布，其幂指数为-2.887 4；对于节点的出度分布，其幂指数为-3.040。这与实际值(-2.9±0.1)很接近[6]，也进一步证实了Internet不是一个随机网络，而是一个具有“无尺度”特征的复杂网络。

幂律分布与泊松分布的显著差别在于：在泊松分布中，由于快速的指数衰减，使得一个具有很大连通度的节点是几乎不可能出现的；在幂率分布中，出现高度连通的节点则是完全可能的。这种情况可以解释为在网络中少数节点有较高的“度”(称为集散节点)，而大多数节点的“度”较低。“度”较高的节点与其他节点的连接较多，通过它找到待查资源的概率就比较大。而对于新的节点在接入 Internet 时，更倾向于加入到与自己最近、连接度较高的节点中。随着时间的推进，在这种“富者愈富，穷者愈穷”的“马太效应”过程中，这些节点就会拥有比其他节点更多的连接数目，此即无尺度网络特征的宏观体现。这种现象在现实的 Internet 中就有很充分的体现，如万维网上一些像 Google, Yahoo，百度等知名网站的连接数要远多于其他一般网站的连接数。

5 结束语

本文利用 CAIDA Skitter 项目提供的大样本容量的 Internet 监测数据，采用统计手段对 Internet 所具有的统计规律性进行研究与分析。通过对 Internet 平均最短路径和 Internet 节点的度分布两方面的数据分析，证实了 Internet 是一种复杂网络。

（下转第 97 页）