链路预测的网络演化模型评价方法

王文强，张千明

(电子科技大学互联网科学中心 成都 611731)

【摘要】在网络演化研究领域，以前工作中对于网络演化机制之间的比较并没有公平、统一的标准。该文基于链路预测理论，采用极大似然估计思想建立了一套用于评价网络演化模型的体系。在基于自治系统的数据实验中，比较了GLP和Tang两个演化模型，结果显示GLP优于Tang，而且得到的最优参数也与其提出者给出的均不相同。实验结果表明基于一定规模为真实网络使用新参数生成的网络更加接近真实网络，并且本文的评价框架可以为模型参数的选取提供建议。

关 键 词 评价方法； 链路预测； 极大似然估计； 网络演化模型

中图分类号 TP391 **文献标识码** A doi:10.3969/j.issn.1001-0548.2011.02.003

New Method of Assessing Network Evolving Models Based on Link Prediction

WANG Wen-qiang and ZHANG Qian-ming

(Web Science Center, University of Electronic Science and Technology of China Chengdu 611731)

Abstract As the previous evaluation methods of evolving models can not be credible because of its inconformity, in this paper, we propose a new method by applying the theory of link prediction and maximum likelihood estimation. Based on the Internet autonomous system networks, we find that GLP is better than Tang which is not agreed with previous results. Moreover, the optimal parameters of these two evolving models are different from the original ones. The experimental results support our optimal parameters with which the corresponding models can generate more real networks.

Key words evaluation; link prediction; maximum likelihood estimation; network evolving model

近年的研究显示一些社会网络如科学家合作网，以及一些技术网络如WWW、Internet自主系统(internet autonomous system)等都具有十分吸引人的统计特性[1]，如小世界效应[2]，无标度特性[3]，而对这些网络的研究，在众多领域都具有重大意义。文献[3]将真实系统通过自组织形成无标度网络归于节点增加(growth of nodes)和偏好连接(preferential attachment)机制两个方面，并提出了一个依赖于该两种机制的简单模型(BA)[3]。该模型要求网络在演化过程中的每个时间步增加一个新节点，并由新节点产生若干条连边连接于现存节点，而现存节点被选中的概率正比于它们的度。通过计算机模拟和理论推导，证明该模型可以促使网络的度分布呈现幂率特性，从而说明统计特性呈现的原因之一是网络的演化。因此，这些社会网络奇妙统计特性的涌现很大程度上取决于网络内在的演化机理，而研究网络演化、剖析网络特性呈现的内在机理并构建演化模型，则是复杂网络研究的重要方向之一。

除了小世界效应和无标度性等特征以外，不同的网络还具有自己独特的统计性质。Internet AS作为一种典型的技术网络得到各界广泛关注。一系列的研究揭示它还具有富人俱乐部效应[4]、分层结构[5]、分形[6]、环结构[7]及异配性[8]等特征。因此对于像Internet AS这样的特定网络，需要构建特定的网络演化模型对其统计特性与演化机理进行描述解释。

近年来，相关领域的专家学者针对不同的网络提出了许多模型。以Internet AS网络为例，其演化模型就有BRITE[9]、GLP[10]、IG[11]、PFP[12]、DP[13]及Tang[14]等十几种。这些模型除了能够刻画Internet AS网络的无标度特性以外，还可以刻画其他特性，例如：Tang模型可以很好地描述Internet中的叶子节点的比例[14]；基于PFP模拟模型生成的网络具有真实网络所具有的富人俱乐部特性[12]。这些模型在提出时都被证明其比以往的模型更优越，但是在评价时

收稿日期：2011－01－28
基金项目：国家自然科学基金(11075031)
作者简介：王文强(1989－)，男，主要从事信息物理和计算机科学方面的研究.

仅仅选择或者构造了一些统计指标[13]，当其模拟生成的网络对应的统计指标相比于其他模型更加接近真实网络，则说明该模型更胜一筹[10-14]。然而这样的评价方式存在明显的缺陷：1) 不能说明这些模型能够很好地刻画没有被选中的统计特征；2) 不同的模型使用的统计指标不同，使得不同的模型之间的优劣难于比较；3) 不同的演化模型与真实网络比较时使用的数据集不同，有可能造成在某种数据集上表现良好，但是在另一种数据集上的表现会差强人意[13]。文献[13]为了解决数据集的不同对演化模型的说服力造成影响的问题，构造了动态和静态两类统计指标，但难于取得学术界的广泛认可。因此，需要一种构建在统一数据集上的、公平公正的评价方法，以对不同的网络演化模型进行评估。

本文尝试使用链路预测及其相关思想构建一个评判框架，对不同的网络演化模型进行评价。链路预测指的是基于节点的属性和被观测边的信息，估计两个节点之间存在连边的可能性[15]。网络演化的过程是节点、边的增删过程，演化模型中边的增加机制实则对边的存在进行概率描述。换句话说，一种网络演化机制本质上对应一种链路预测算法[15]，这就是链路预测与网络演化模型的内在联系。使用链路预测的相关理论构建网络演化模型评价体系的一种直观的方法是使用链路预测算法的评价方法——AUC、Precision等——直接评价不同的演化模型，即在给定真实演化网络中每条边建立的时间信息的情况下，可以把某时间点之后出现的连边作为测试集[15]，在此以前的连边作为训练集，然后把各种演化模型的规则应用到当前已知的网络结构上，比较新生成的连边和测试集中的连边，就可以评估相应演化模型的准确程度。但是这种方法存在一个巨大挑战，即它受制于数据：需要知道真实网络中每条边建立的时间信息。但这种数据本身是十分稀少的，而且现存的数据由于统计方式的局限性会出现很多边的建立时间相同的问题，为评价带来极大的困难，因此必须另辟蹊径。本文试图提出普适的评价框架，利用极大似然思想避开数据的局限性，对不同的网络演化模型进行评估。

1　方法框架

假设某网络在t时刻的边集为E_t，经过演化在t+1时刻变为E_{t+1}，则集合$E_{\text{new}}=E_{t+1}-E_t$为网络演化生成的新边。这些新边有3种类型：新加入节点(新节点)之间的边、新节点和现存节点之间的边以及现存节点与现存节点之间的边。为了方便计算每条边的似然，忽略节点的增长，即假设演化过程中新加入节点以孤立节点的形式存在，则演化的过程就是向网络中添加边的过程，并且假设所有边的加入是无序的，通过多次独立重复演化实验就可以从统计的角度考察比较演化机制的优劣。根据t时刻的网络拓扑结构以及网络演化机制，可以计算出E_{new}中的每一条边的似然P_e。如果将真实网络看成是其网络系综中的一个网络，则每条边似然之积 $\prod_{e\in E_{\text{new}}} p_e$ 便，是该网络在其系综中呈现的似然。若网络模型中的演化机制能够更好地描述真实网络的演化过程，那么它应该具有更大的似然，并且对于同一种网络演化机制取不同的参数得到的似然也会不同，最大似然对应的参数应该有它的统计意义。下面给出这种框架的详细描述。

假设某种网络演化模型采用的演化机制为：

$$\prod(k_i)=\begin{cases} f(k_i;a) & k_i\neq 0 \\ C, & k_i=0 \end{cases} \tag{1}$$

式中，k_i表示节点i的度；a为参数；$k_i=0$代表该节点为新节点；$\prod(k_i)$ 表示在演化过程中节点i被选为新加入边的一个端点的概率大小。由于在大多数的网络演化模型中两个端点的选择是相互独立的，则在演化过程中一条新边(i,j)的出现概率大小就为：

$$p_{(i,j)}=\prod(k_i)\times\prod(k_j)$$

为了方便不同模型之间的比较，该似然需要经过归一化处理，即：

$$p^n{}_{(i,j)}=\frac{p_{(i,j)}}{\sum_{(a,b)\in E^N} p_{(a,b)}} \tag{2}$$

式中，E^N是所有不存在的边的集合(the set of non-existed links)。根据式(2)，给定参数a，将E_{new}中每一条边的似然相乘便可得到在该参数下的网络似然。取不同的参数a，即可得到不同的网络似然。

2　模　型

本文关注的是Internet AS网络演化模型的评估。选用了两个面向Internet AS的演化模型，且都传承自BA模型。

第一个为GLP(generalized linear preferential)模型[10]，它将BA的演化机制进行推广，其采用的网络演化机制为：

$$\prod(k_i)=\frac{k_i-\beta}{\sum_j(k_j-\beta)} \qquad \beta\in(-\infty,1) \tag{3}$$

该模型开始于包含m_0个顶点的初始网络，它由m_0–1条边连接起来，每个时间步执行下列步骤：1) 以概率$p\in[0,1]$添加m条新边，m≤m_0，每条边的每一个端点i均按照式(3)选取；2) 以概率$(1-p)$添加一个新节点和该新节点的m条新边，每一条边的另一个端点按照式(3)在现存节点中选取。这样对于GLP模型，具体计算演化生成新边的似然的方法为：对于两个节点都是现存节点的新边(i, j)，其似然为$\prod(k_i)\times\prod(k_j)$；对于一个节点是现存节点、另一个节点是新节点的新边，它的似然是$\prod(k_i)\times C$；对于两个都是新节点的新边，其似然计算方法为$C\times C$。常数C可以取1，因为可以把新节点的添加看成是一种必然事件。

第二个为Tang模型[14]，该模型的演化机制为：

$$\prod(k_i)=\frac{k_i^{1+\varepsilon}}{\sum_j k_j^{1+\varepsilon}}\quad \varepsilon\in R \tag{4}$$

该模型要求初始网络有m_0个节点，每一个时间步增加一个新节点和m条边$m\leqslant m_0$，其中一条边连接新节点和现存节点，剩下的m–1条边连接现存节点。m–1条边中，每条边的一个端点随机选取，另一个端点按照式(4)选取。该模型的似然计算的具体方案是：1) 对于新节点与新节点之间以及新节点与现存节点之间的边，计算方法同GLP；2) 对于老节点由于该模型要求其中一个端点在现有节点中随机选取，另一个端点按照式(4)中的超线性连接机制选取，但是只凭借数据无法判断哪一个顶点是随机选邓的，哪一个顶点是按照模型提供的演化机制选取的，因此取两种情况的几何平均值，即对于一条边(i, j)，其似然计算方法为：

$$\sqrt{\prod(k_i)\times\frac{1}{n}\times\prod(k_j)\times\frac{1}{n}}$$

式中，n为现存节点的数量(相当于t时刻网络的节点数目)。

除此之外，还考察了BA的线性连接机制和ER式的完全随机的机制，以对比特定网络演化模型是否优于BA或者是完全随机的情况。对于BA的线性连接机制，本文将其进行推广，以便可以计算新节点和新节点以及现存节点之间的连边的似然[13]，它的似然计算方法与Tang的类似，定义为：

$$\prod(k_i)=\frac{k_i}{\sum_j k_j} \tag{5}$$

设n为现存节点的数量，完全随机机制的似然计算方法为：对于现存节点和现存节点之间的连边，其似然为$(1/n)^2$；现存节点与新节点之间的连边的似然为$(1/n)\times C$；而新节点之间的连边的似然为$C\times C$。值得注意的是，BA线性连接的似然相当于Tang模型ε=0的情况，完全随机机制相当于Tang模型的$\varepsilon=-1$的情况。

3 数 据

本文使用的是Routeviews Project[16]收集的2006年6月和2006年12月的AS数据，并采取如下处理方法：由于Internet AS网络中边的重连现象以及边和节点的消失现象出现的频率不大，假设t时刻的网络拓扑是t+1时刻网络的子集，将2006年6月的Internet AS网络中的边并到2006年12月的边集中，然后挑出12月份相对于6月份的新边集，从而得到新边的数量为9 723。数据的详细情况如表1所示。

表1 数据统计情况

时间	节点数	边数
2006年6月	22 960	49 545
2006年12月	24 403	52 826
2006年12月(处理)	25 103	59 268

4 实验结果

4.1 评价结果

按照框架中的设计，实验中分别计算了GLP、Tang、BA、ER这4个模型在不同参数下的网络似然，得到的结果如图1所示。4种演化机制所得到的最大似然等数值如表2所示。

由于BA的线性连接机制和ER式的完全随机机制不含有参数，因此它们的图像是直线。从计算结果中可以清楚地看到，似然最低的为ER，BA的线性连接机制明显好于完全随机机制，说明这种偏好连接机制要比完全随机机制好得多，符合之前的工作。GLP的结果明显好于Tang，Tang的似然只在最大值附近才比BA的线性连接机制好少许，说明Internet AS的专用模型比BA和完全随机机制的模型好。除此之外，最大似然对应的参数并不与模型的理论推导相符[10,14]。导致这种结果的原因可能是：1) 所有演化模型都强调网络从一个规模很小、结构极简单的初始网络开始演化，但是在使用框架进行评判时，使用了一个富含很多信息的网络作为起始状态，而模型的原配参数只适用于前者；2) 在演化过程中，驱动网络演化的机制会发生变化[19]。

为了说明评价框架是合理的，并揭示使用的评价方法得到的参数的统计意义，设计了下面的实验：

使用评价方法得到的参数和模型提供的原配参数，以2006年6月的网络拓扑结构为基础，操作演化模型演化到2006年12月的规模，考察新增网络部分的统计量，将之与真实网络进行对比。为了操作演化模型，首先需要确定模型中的各种参量。

表2　4种演化机制的极大似然

模型名称	最大似然
ER	$4.17\times10^{-132\,356}$
BA	$2.26\times10^{-124\,449}$
GLP	$3.54\times10^{-120\,497}$
Tang	$9.77\times10^{-124\,442}$

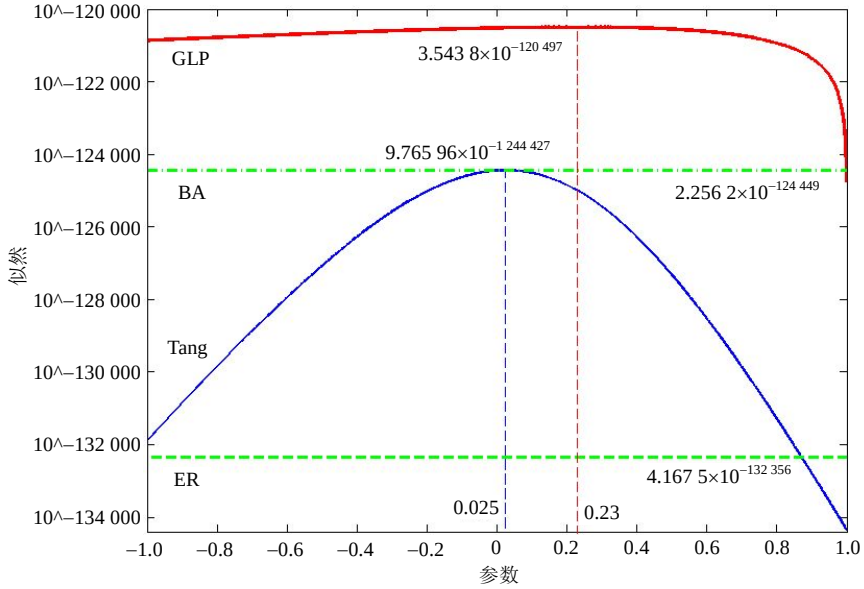

图1　4种演化机制的似然

4.2　GLP模型的参数确定

4.2.1　m的确定

文献[17]指出，自制系统中86.82%的新生节点在刚刚出现时的度为1，12.67%的新生节点的度为2，其他情况只占0.53%的。因此文章仅考虑新生节点度为1和2的情况。而m可以看成是新节点的初始度，因此m取平均新生节点的平均度为：

$$m=0.13\times2+0.87\times1\approx1.13$$

4.2.2　其他参数的确定

文献[10]使用平均场方法经过一系列推导得到:

$$\begin{cases}p=\dfrac{N_{\mathrm{E}}-mN_{\mathrm{V}}}{N_{\mathrm{E}}}\\[2ex]\dfrac{2m-\beta(1-p)}{(1+p)m}=-\alpha\end{cases}\tag{6}$$

式中，N_{E}为网络中边的数量；N_{V}为网络中节点的数量；α 为度的complementary distribution ($p(k\geqslant k_i)$)幂指数。通过线性拟合的方式得到的$\alpha=-1.1431$。由于需要操作模型使AS从2006年6月的规模演化到处理过的2006年12月的规模，所以取处理后的2006年12月的数据为：

$$\begin{cases}N_{\mathrm{V}}=25\,103\\N_{\mathrm{E}}=59\,268\end{cases}$$

将它们连同m代入式(6)得到：

$$\begin{cases}p=0.521\,4\\\beta=0.616\,0\end{cases}$$

4.3　Tang模型的参数确定

Tang模型[14]并没有得出ε的解析解。但是文献[14]发现ε=0.2时网络的度分布最接近真实值，因此可以认为其最优参数为ε=0.2。

4.4　演化结果

试验选取了新节点的平均度、连接密度、叶子节点在新节点中的比例这几种常见的统计指标，每种参数(详见表3)的模型分别演化了100次，并使用Box and whisker图[18]展示结果，如图2～4所示。图2～4都由4部分组成，从左至右依次为GLP(本文的参数)、GLP(原配参数)、Tang(本文的参数)、Tang(原配参数)。其中虚线代表Internet AS的真实值，真实的新节点平均度为1.596 36，新节点的连接密度为$1.917\,08\times10^{-5}$，新节点中度为1的比例为0.516 099。

表3　最优参数

GLP(原配)	GLP	Tang(原配)	Tang
0.616 0	0.23	0.2	0.025

图中显示，对于网络的新增部分，使用极大似然参数生成的演化网络更加贴近真实网络，并且两

种模型同类间(极大似然对应的参数和模型原配参数)GLP模型相比于Tang模型更加接近真实值。值得一提的一个有趣结果是，文献[13]指出Tang模型可以很好地描述网络的叶子节点，然而实验结果显示其对叶子节点描述的真实度却不如GLP，其中隐藏的原因尚待进一步的探讨。以上说明，提出的评价框架很好地显示了演化模型的优劣，并且最大似然对应的参数是有其统计意义的。更重要的是，对于网络的新增部分，使用本方法得到的最优参数进行演化可以得到更贴近真实的网络。

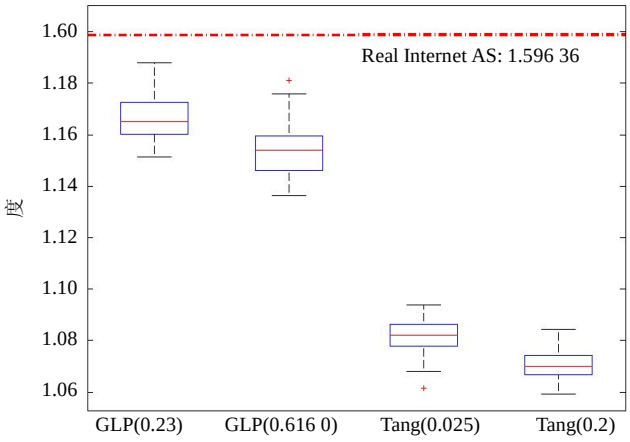

图2 新节点的平均度

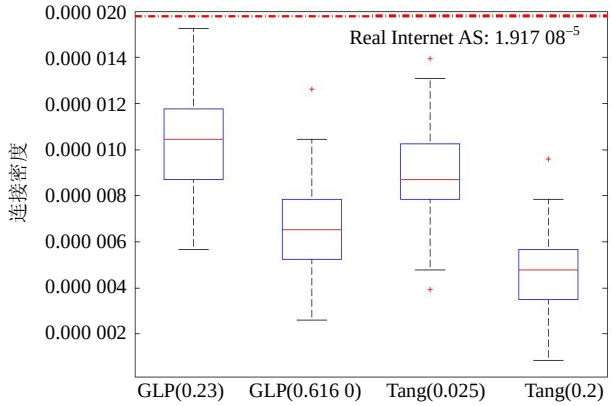

图3 新节点的连接密度

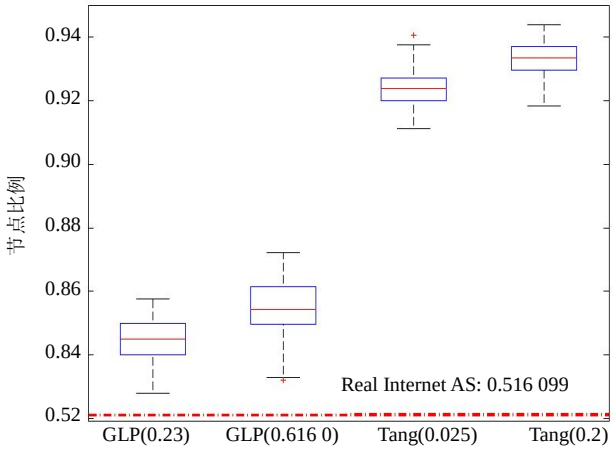

图4 新节点中叶子节点的比例

5 结论与讨论

本文针对网络演化机制传统比较、评价方法的局限，提出了一种基于链路预测理论、利用极大似然估计思想建立起来的相对公正、并且独立于数据集的评价方法。通过多次重复实验，发现在各自最优参数的作用下，GLP模型相比于Tang模型能够更好地描述Internet自制系统的演化过程，且二者均好

于BA和完全随机演化机制。然而，本文得出的两个模型的最优参数与其各自的原配最优参数是不同的，实验结果表明使用本文的最优参数模拟生成的网络更贴近真实网络的统计特征，究其原因可能在于：基于网络演化是一个从无到有的过程的假设，而且驱动网络演化的机制并不是一成不变的。现今模型的建立对这两个方面的关注尚不足，希望本文的工作能够对网络“从少到多”演化过程的研究起到抛砖引玉的作用。

然而本文方法仍然存在缺陷。一方面，为了方便计算似然，将网络的中新增部分的边看成是一种静态过程，该假设有可能会对计算的结果造成影响；另一方面，一些演化机理十分复杂的演化模型，它们的似然计算将会变得十分复杂，对于这些模型的评价，可能需要使用更加高级的数学技巧和计算方法。作为一个起步工作，本文试探性地提出这一个评价体系，希望为之后提出的网络演化模型能够更好地刻画真实网络有所帮助。

感谢电子科技大学互联网科学中心的周涛教授为本文的研究和成文提供的宝贵建议！

参 考 文 献

[1] DOROGOVTSEV S N, MENDES J F F, Evolution of networks[J]. Advances in Physics, 2002, 51(6): 1079 - 1187

[2] WATTS D J, STROGATZ S H. Collective dynamics of ‘small-world’ networks[J]. Nature, 1998, 393: 440-442.

[3] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286: 509-512.

[4] ZHOU S, MONDRAGÓn R J. The rich-club phenomenon in the Internet topology[J]. IEEE Commun Lett, 2004, 8(3): 180-182.

[5] RAVASZ E, BARABáSi A L. Hierarchical organization in complex networks[J]. Phys Rev E, 2003, 67: 026112.

[6] CALDARELLI G, MARCHETTI R, PIETRONERO L. The fractal properties of Internet[J]. Europhys Lett, 2000, 52: 386.

[7] BIANCONI G, CALDARELLI G, CAPOCCI A. Loops structure of the internet at the autonomous system level[J]. Phys Rev E, 2005, 71: 066116.

[8] MAHADEVAN P, KRIOUKOV D, FOMENLOV M, et al. Lessons from three views of the internet topology[J]. 2005, arXiv: cs. NI/0508033.

[9] MEDINA A, LAKHINA A, MATTA I, et al. BRITE: an approach to universal topology generation[C]//Proceedings of MASCOTS. Washington: IEEE Computer Society, 2001: 346-353.

[10] BU T, TOWSLEY D. On distinguishing between Internet power law topology generators[C]//Proceedings of INFOCOM. New York: IEEE Press, 2002, 2: 638-647.

[11] ZHOU Shi, MONDRAGON R J. Towards modeling the internet topology – the interactive growth model[C]//18th International Teletraffic Congress. Berlin: IEEE Press, 2003, 0303029.

[12] ZHOU Shi, MONDRAGON R J. Accurately modeling the internet topology[J]. Phys Rev E, 2004, 70: 066108.

[13] PARK S T, PENNOCK D M, GILES C L. Comparing static and dynamics measurements and models of the Internet’s topology[C]//Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies. Hong Kong: IEEE Press. 2004, 3: 1616-1627.

[14] SAGY B, MIRA G, AVISHAI W. An incremental super-linear preferential Internet topology model[J]. LNCS. 2004, 3015: 53-62.

[15] 吕林媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 35(5): 651-661.
LÜ Lin-yuan. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 35(5): 651-661.

[16] Advanced Network Technology Center, University of Oregon. University of Oregon route views project[EB/OL]. [2011-01-26]. http://www.routeviews.org.

[17] The origin of power law in internet topologies revisited[C]//Proceedings of INFOCOM. New York: IEEE Press, 2002.

[18] TUKEY J W. Exploratory data analysis[M]. MA: Addison-Wesley, Reading: 1977.

[19] ZHANG Guo-qing, ZHANG Guo-qiang, YANG Qing-feng, et al. Evolution of the Internet and its cores[J]. New Journal of Physics, 2008, 10: 123027.

编 辑 蒋 晓