Useful Techniques of Power-law and Spectra in Modeling Internet Topology Structure

Ye XU[1], Hai ZHAO[2] and Zhuo Wang[1]

[1]College of Information Science and Engineering, Shenyang Ligong University, Shenyang 110168, China
[2]School of Information Science and Engineering, Northeastern University, Shenyang 110004, China

Summary

Modeling of Internet topology has become the focus of Internet-related research fields recently, and was also the topic of this paper. First, the topology measuring results were collected and processed by tools of IP Alias resolution and sampling bias handling. Then, frequency-degree power-law, degree-rank power-law and so on were performed on these topology data to find power-law properties of the Internet topology. With the power-law achievements, an Internet topology model was constructed based on BA model after two steps of improvements. The first improvement is to optimize parameters of BA model through Genetic Algorithm and SLS in order to make the model complying with frequency-degree power-law analysis result; and the second one is to modulate the improved model again according to the degree-rank power-law analysis results. Generating algorithm for the topology model was finally given.

Key words:

Genetic Algorithm; Power-law distribution; SLS (Signless Laplacian spectra); Topology modeling;

1. Introduction

In recent years, many emerging useful researches techniques such as the power-law distribution[1][2] and spectra analysis[33] are getting to be key approaches in Internet-related researches, especially the Internet topology modeling research[1][2][4].

In power-law part, frequency-degree power-law was firstly used by Faloutsos to character the topology of both AS-level and router-level Internet in 1999, thereafter, degree-rank power-law, eigenvalue-rank power-law and CCDF(d)-degree power-law were brought forward[1][3].

In spectra part, the signless Laplacian spectra (SLS)[33] and normalized Laplacian spectra (NLS)[33] are research techniques from Graph Theory but recently have been considered to be quite useful tools in topology modeling researches. So, power-law together with spectra analysis would be mainly used in studies of Internet topology modeling, based on a giant set of measured samples of router-level Internet topology in this paper.

A. The measured samples of Internet topology

1) Measuring methods

Static methods based on the BGP route table and the dynamic methods based on the active probing are the main ways to measure the router-level Internet topology[16]. And the static methods are gradually replaced by the dynamic ones due to their lack of the redundant routers measures [16].

The dynamic methods, at present, are mainly divided into three categories[19]: (1) single-monitor-measuring by recording all routers in the route path, such as the Internet Mapping Project (IMP) in Bell Lab.[20], and the Mercator[21] projects; (2) active measuring based on the Public Traceroute Server (PTrS), such as the ISP topology measuring project by Boston University[22]. (3) multi-monitor-measuring or measuring-from-multiple-vantage-points by self-developed software engines, such as the CAIDA projects[17][18], and the Active Measuring Project by Harbin Institute of Technology[19].

In the upper three methods, the PTrS (method No.2) is quite limited due to the following reasons[19]. Firstly, PTrS are quite unevenly distributed in Internet and not all ISPs render services of PTrS. Studies in [19] indicated that only one of nine ISPs providing PTrS, so PTrS method is not as reliable as the others. Secondly, it's rather hard to transfer or gain the control of PTrS from the ISPs due to security considerations, which directly resulted in the inefficiency of measuring Internet topology.

The first method is similar to the third one (e.g., CAIDA), they are all based on traceroute or the traceroute-like programs[17][18], but the first method is inferior to the third one since it‘s totally upon single-monitor-measuring tools. CAIDA, however, could implement multi-monitor-measuring and consequently yield better measuring results[17][18]. The Active Measuring Project by Harbin Institute of Technology (HIT) also used multi-monitor-measuring tools, but it had fewer monitors in its project than CAIDA has, what's more, the HIT project was mainly focused on the Internet topology in China part2][19], on the contrary, CAIDA project measured the world-wide Internet. So CAIDA[1] measuring methods were selected for studies in this paper.

2) Problems of the measuring results

[1] CAIDA, the Cooperative Association for Internet Data Analysis, is a worldwide research center on Internet-related research fields. CAIDA has more than thirty monitor nodes which are distributed throughout the whole world, measuring and monitoring the variations of Internet. Three of the monitors are located in Asia.

Manuscript received June 5, 2008.
Manuscript revised June 20, 2008.

The measuring results from CAIDA monitors are complete but in coarse granularity. There are two main problems in it: IP Alias problem and the sampling bias problem due to single-monitor- measuring[6][19].

3) Problems of IP Alias

[Def 1] IP Alias[23][24]: Different ports of one Internet router are assigned with different IP addresses, and they are mistaken for different routers in the active measuring. And this problem is known as IP Alias.

IP Alias Resolution[25] is a way to distinguish the IP addresses and solve the problem of IP Alias. However, the researches on IP Alias Resolution is still in progress, and only a few methods or tools are provided at present and they still could not solve the whole problem, only to some extent[23][24]. Among these tools, three of them are comparatively practicable, and they are iffinder tool[26] from CAIDA, Mercator[27] and Rocketfuel tool[28] from Boston University. Rocketfuel tools distinguished aliased IP addresses by some complicated algorithm such as recognizing the TTL segment of the ip datagram. And some researches[28] found Rocketfuel tool could find Alias IP addresses three times more than the other present tools. So it was selected as IP Alias Resolution tool in this paper.

4) Problems of Sampling Bias

Some recent researches[6][19] found that the measuring results were usually different from real network topology and tended to show stronger power-law (frequency-degree power-law) properties than what the real network actually has when only one monitor or less monitors was used during the active measuring by the traceroute-like tools.

Sampling bias is directly associated with the number of measuring monitors[6][19]. Though it's still hard right now to find perfect approaches solving the sampling bias problems, we still found an easy and effective way to solve, in some extent, the problem of sampling bias. That is to use as many monitors as possible when measuring a target network[6][19]. And this is how we handle the measuring results of Internet topology from CAIDA monitors in this paper.

5) The router-level Internet measuring samples after IP Alias Resolution and Sampling Bias handling

The rough measuring results in this paper are the router-level Internet topology data measured at 30th, Jan. 2006[2] from as many as twenty-one CAIDA monitors[3]. And after the IP Alias resolution, we get twenty-one set of measuring samples.

Then we move on sampling bias handling process. Firstly, we gather them together (the twenty-one monitor measuring results) to form a complete testing sample in order to reduce the impact of sampling bias to an extreme extent. And this best copy of sample is undoubtedly regarded as our key sample in experiments of the paper.

However, we still made several other inferior or incomplete testing samples for comparison reasons, and they are sample(1) comprising data from only one monitors (arin monitor), and sample(2) from two monitors (arin, b-root), till sample(20) from as many as twenty monitors.

Now we eventually had twenty-one set of measuring samples including the key testing sample.

B. Mathematical description of power-law distribution

Power-law distribution is mathematically described as $y = cx^{-r}$, where *x*, *y* are random variables, and *c*, *r* are constants greater than 0. Perform logarithm on it, we then get $\ln y = c'\ln x$. It's easy to see that there is a linear relationship between $\ln y$ and $\ln x$, i.e., there would be a straight line if we plot the relationship between them in a dual-logarithmic coordinates. And this linear relationship, or the straight line in dual-logarithm plot, would be regarded as a primary judgment identifying whether power-law distribution is suited or not.

Three important power-law distributions mostly used in Internet topology researches are listed in table I[3][4], and the parameters are listed in table II.

TABLE I
THE BASIC EQUATIONS OF POWER-LAW DISTRIBUTIONS

Power-law distributions	Mathematical models
frequency-degree	$p_v \propto d_v^R$
degree-rank	$d_v \propto r_v^R$
CCDF(d)-degree	$D_d \propto d^D$

TABLE II
DEFINITIONS OF THE PARAMETERS AND SYMBOLS

Variable	Definition
G	Undirected graph
N	Number of the nodes in a graph
E	Number of the links in a graph
d_v	Degree of node v
d	Average degree of a graph, $\overline{d} = 2E/N$
p_v	Frequency of node whose degree is v
D_d	CCDF(complementary cumulative distribution function)
r_v	Order of node v
λ	eigenvalues of $N*N$ Matrix A: $X{:}X \in R^N \setminus\{0\}$ and $AX=\lambda X$

[2] The reason why measuring topology data at 30th, Jan. 2006 is that there are as many as twenty-one monitors providing effective measuring data that day. For other days round that period of time, the fact is, there would be fewer effective monitors.

[3] The twenty-one monitors are arin, b-root, cam, cdg-rssac, champagne, d-root, e-root, h-root, i-root, iad, ihug, k-root, lhr, m-root, mwest, neu1, nrt, riesling, sjc, uoregon and yto. And all monitors are distributed into different continents for better measuring Internet throughout the whole world.

C. Mathematical description of spectra

1) Spectra

A non-directed graph G could be denoted by it symmetrical adjacency matrix ***A***. If there is a link between node *i* and node *j* in G, then $\boldsymbol{A}_{ij}=\boldsymbol{A}_{ji}=1$, otherwise $\boldsymbol{A}_{ij}=\boldsymbol{A}_{ji}=0$. Eigenvector of G are the eigen values of ***A***, and they are denoted as $\lambda_1, \lambda_2 \cdots \lambda_n$. Researches in Graph Theory show that eigenvector of a graph are closely related to the structural properties of the graph topology. So studies on a graph's eigenvector are useful in topology research. Spectra of a graph G is denoted by a set of the eigen values and their tuples[2], as is equation (1).

$$Spec(G) = \begin{pmatrix} \lambda_1 & \dots & \lambda_n \\ m_1 & \dots & m_n \end{pmatrix} \qquad (1)$$

where *m* is the tuple of the corresponding eigen value. Spectral density $\rho(\lambda)$, is the eigen value density of **A**, and it could be denoted as[2][5][35]:

$$\rho(\lambda) = \frac{1}{N}\sum_{i=1}^{n}\delta(\lambda - \lambda_i) \qquad (2)$$

where λ_i is the *i*th eigen value of adjacency matrix ***A***, *N* is the number of the eigenvector.

2) Signless Laplacian spectra (SLS)

An SLS matrix |***L***| of a graph G is defined to |***L***|=***D***+***A***, where matrix ***D*** is a diagonal matrix representing G's degree, and ***A*** is G's adjacency matrix[2][5]. SLS is the eigenvector of |***L***|. Some researches in graph theory indicated that SLS is the best spectra in distinguishing different graphs[5]. So SLS would be mainly used in analysis of properties of Internet topology structure.

2. Power-law Analysis

A. Frequency-degree power-law

Calculate the frequency and degree from one-monitor sample, two-monitor sample, five-monitor sample and twenty-one-monitor sample (the key sample) and make the illustration in Fig.1. The power-law curve fitting results were also illustrated in Fig.1.

There is clear power-law relationship between variable frequency and degree since the curve fitting result - straight line shown from Fig.1. Besides, the curve fitting results (the straight line) are close to the sample, and all four fitting ACCs (Absolute value of the correlation coefficient) are greater than 0.95, meaning that the curve fitting results are acceptable.

Though the results in four sub-graphs show clear power-law relations, their power-exponents |R|, however, are different. We list them in table III.

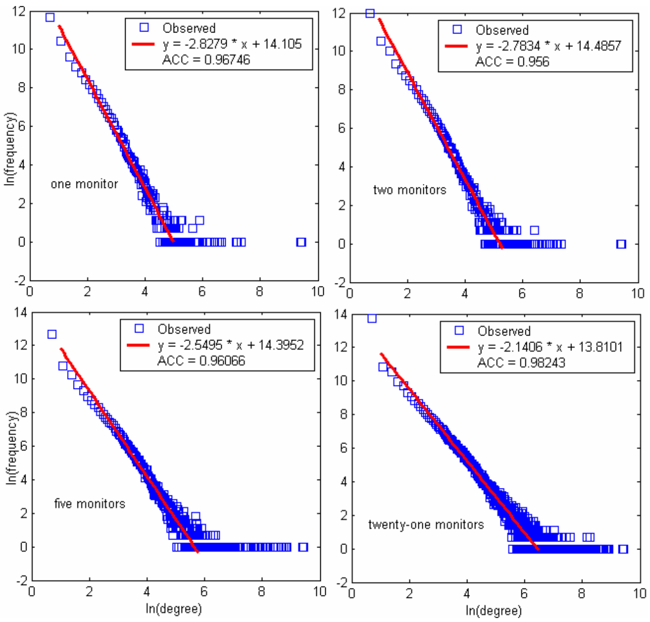

Fig. 1 The frequency-degree power-law analysis on the router-level Internet topology and the curve fitting results.

TABLE III
POWER EXPONENTS OF THE FREQUENCY-DEGREE POWER-LAW ANALYSIS

Number of monitors	ACC	|*R*|
1	0.9675	**2.8279**
2	0.9560	**2.7834**
5	0.9601	**2.5495**
21	0.9824	**2.1406**

|R| is decreasing with increasing monitors. Considering the fact that a greater |R| means a stronger power-law relationship, we find that the power-law relationship of Internet topology is getting weaker with increasing monitors. This conclusion, however, is not so much correct because the sampling bias problem in measuring topology might tend to produce extra stronger power-law relations than what the real network actually has. Then, the reason of decreasing |R| with increasing monitors is easy to figure out now. And what was found here on the router-level Internet in Fig.1 is quite similar to the research outcomes in [5], proving the correctness of our experiments.

In Fig.1(the 4th sub-graph), the power-law property might be least influenced by the sampling bias since the number of monitors reaches most. Obvious power-law relations still exist under such conditions, indicating that there is definite power-law property in Internet topology.

From table III, the frequency-degree power exponent of the router-level Internet topology is 2.1406 (out of the key sample in the paper), quite close to the power-exponent 2.2 of AS-level Internet topology in reference [6][7][8]. As we know, AS-level Internet topology is a coarse granularity of router-level Internet topology, the two research outcomes are expected to be similar to each other. And the analogs, in return, testify the accuracy of the frequency-degree power-law research results in this paper.

B. Degree-rank power-law

In degree-rank power-law analysis, we first sort the degree in descending order, then perform the logarithm operation on the degree and its order (rank) to form dual-logarithmic coordinates. The power-law analysis is illustrated in Fig.2.

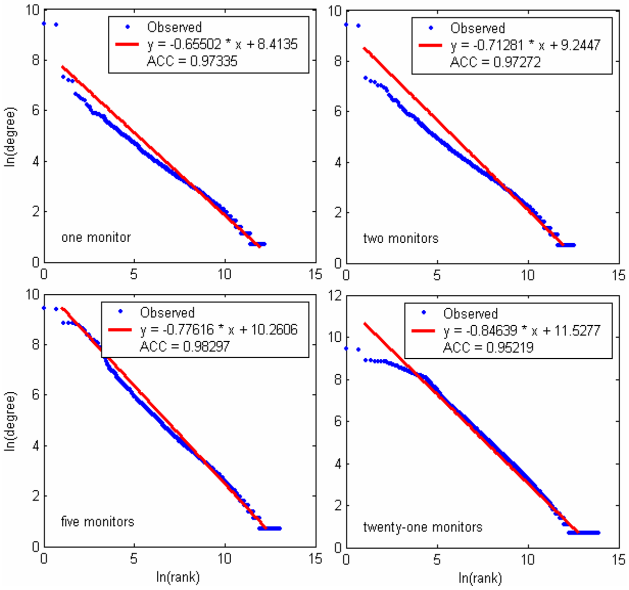

Fig. 2. The degree-rank power law analysis and curve fitting results.

It's obvious that there is power-law relation in Fig.2. And the fitting ACCs are greater than 0.97, meaning the fitting result is good. The power-exponent (|R|) is listed in table IV.

From table IV, |R| is increasing with increasing monitors. To better explain this phenomenon, we make reference to the research results of [2] that the power-exponent |R| would increase or decrease exactly with increasing or decreasing $\mathrm{Num}_{ld}/\mathrm{Num}_{sld}$[2] in degree-rank power-law analysis. What was found in table IV is quite the same, proving that the results of the degree-rank analysis in this paper are correct.

TABLE IV
POWER EXPONENT OF THE DEGREE-RANK POWER-LAW ANALYSIS

Number of monitors	ACC	|*R*|	$\mathrm{Num}_{ld}/\mathrm{Num}_{sld}$
1	0.9734	**0.6550**	**3.3921**
2	0.9727	**0.7128**	**4.2578**
5	0.9830	**0.7762**	**6.7064**
21	0.9941	**0.8464**	**17.4633**

Note: Num_{ld} is the number of nodes with the least degree, and Num_{sld} is the number of nodes with the second least degree in the Internet topology graph.

After further studies on Fig.2, we find that there are bad curving fitting parts when ln(rank) is less than around 3 in all four sub-graphs, and especially in sub-graph 4. Since sub-graph 4 is out of the key sample of the paper, we would perform further degree-rank power-law studies on it.

In sub-graph 4, we find that there might be another kind of power-law relationship when ln(rank) is less than around 3, then the curve fitting is performed and the result is illustrated in Fig.3.

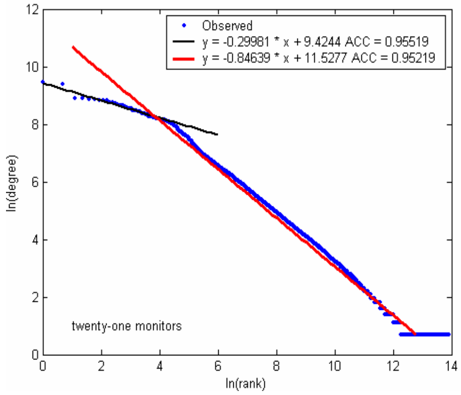

Fig. 3. Analysis of two phases of degree-rank power-law properties and the curve fitting results.

The cross position of two straight lines in Fig.3 is 3.6 on axis x. Besides the power-law relationship when ln(rank) is greater than 3.6 as we discussed above, the straight line when ln(rank) less than 3.6 also indicates that another power-law property is suited since the fitting ACC of this part is greater than 0.95. Thus, two phases of degree-rank power-law relations are found in Internet topology graph, and power exponents are 0.29981 and 0.84639, respectively.

The power exponents could be used to quantitatively depict the power-law properties of Internet topology and would be used in Internet topology modeling later.

C. CCDF(d)-degree power-law

There are several mathematical models to calculate CCDF, and they are listed in table V.

TABLE V
FOUR COMPLEMENTARY CUMULATIVE DISTRIBUTION FUNCTIONS (CCDFS)

Function name	PDF	CCDF
Power law	$f(x) = Cx^{\alpha}(C > 0, \alpha < -1)$	$F'(x) = -\frac{C}{\alpha+1}x^{\alpha+1}$
Power law(2)	$f(x) = Cx^{\alpha} + D(C > 0, \alpha < -1)$	$F'(x) = -\frac{C}{\alpha+1}x^{\alpha+1} + Dx$
Weibull(2-parameter)	$f(x) = \frac{c}{b}(x/b)^{c-1}e^{-(x/b)^{c}}$	$F'(x) = e^{-(x/b)^{c}}$

Apply different CCDFs on the samples, and the results are listed in Fig.4 and table VI.

TABLE VI
CURVE FITTING RESULTS OF CCDFS

Number of monitors	Function style	**SSSR$_1$**
	Power law	**12455.6927**
1	Power law(2)	**219431.0825**
	Weibull(2-parameter)	**11594.8785**
	Power law	**24215.0629**
2	Power law(2)	**303397.4291**
	Weibull(2-parameter)	**20133.3965**
	Power law	**114594.8493**
5	Power law(2)	**503785.6687**
	Weibull(2-parameter)	**59191.7273**
	Power law	**485010.9747**
21	Power law(2)	**1160172.4009**
	Weibull(2-parameter)	**221809.1604**

Note: SSSR is standard square sum of residual, and it equals to sqrt(SSR).

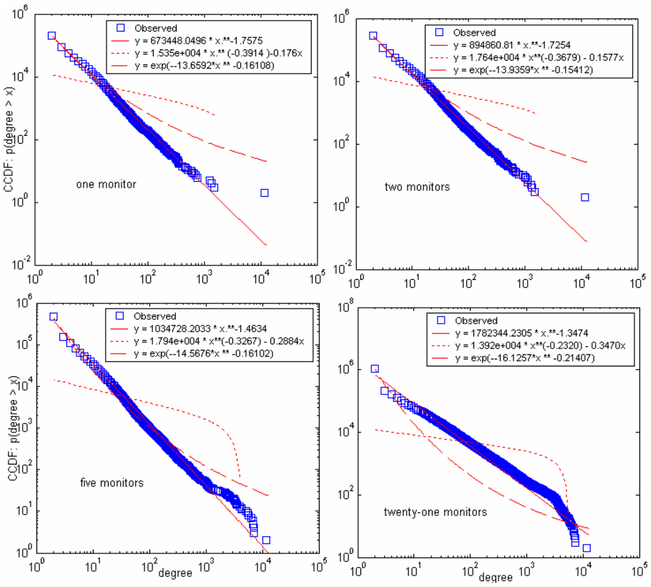

Fig. 4 The CCDF(d)-degree power-law relationships (in log-log plot) and curve fitting results..

There are three fitting curves in Fig.4 representing the CCDFs of power-law (line), power-law2 (dotted line) and Weibull(2-parameter) distributions (the long dotted line), respectively. It's obvious that the four experiment results illustrated in all four sub-graphs present great uniformity, indicating the effectiveness of the research approach.

It's hard to directly distinguish the best fitting curve because of the log-log coordinate. Then the SSSR(standard square sum of residual) from table VI is introduced to make estimations.

First, SSSR of the CCDF of power-law(2) is greater than the other two CCDFs, so power-law(2) is the worst in three. For the other two CCDFs, we see in table VI that SSSR of power-law in all four sub-graphs is greater than that of Weibull(2-parameter), thus Weibull(2-parameter) is better than power-law in fitting the Internet topology samples.

Then, the CCDF(d)-degree power-law distribution might not be the best way to quantitatively character the Internet topology when compared with Weibull(2-parameter) distribution. And this research result is completely identical to the studies in [9][10][11].

D. Power-law analysis conclusions

Firstly, frequency-degree power-law relationship is clearly found in the router-level Internet topology, with a power exponent of 2.1406.

Secondly, two phases of power-law relationships are found in degree-rank power-law studies, the first phase locates in an interval where ln(rank) is less than 3.6, and its power-exponent is 0.29981. The second part lies in where ln(rank) is greater than 3.6, and its power-exponent is 0.84639.

Finally, the CCDF(d)-degree power-law distribution is proved not to be the best approach to character router-level Internet topology and thus would not be used for modeling Internet in this paper.

3. MODELING OF THE INTERNET TOPOLOGY STRUCTURE

A. Introductions of the way to construct Internet model by results of power-law analyses

With the power-law analyses outcomes, it's a natural idea to set up a topology model complying with their power exponents.

1) Step 1: according to the frequency-degree power-law exponent

The power exponent of frequency-degree power-law is |R|=2.1406. Some researches[4][14] indicated that, a network having frequency-degree power-law properties is a kind of scale-free networks, and the traditional model - BA model[29] is regarded as one of the best choices to generate such scale-free networks. With this, we use BA model as a base to form the Internet topology model.

A short description of the algorithm of BA model is: generate $m_0(m_0 > 1)$ nodes, and link them randomly; repeat the following step: for network G(t-1), add one new node with n links to G(t-1) and form a new network G(t). The n links could be connected between the new added node and any selected node i in current network if node i's $\Pi_i = k_i / \sum_j k_j$ is greater than a given threshold, where i, j are nodes existed in G(t-1) and k_i, k_j are degree value of corresponding nodes.

Networks generated by the upper algorithm conform to a frequency-degree power-law distribution of $p(k) \sim k^{-\alpha}$, where the power exponent α is irrelevant to m_0 and n.

Researches in [4], [14] showed that the power exponent of the network generated by BA model is usually 3, which is different from 2.1406 in this paper. So improvement of BA model is necessary.

Researches on how to improve the power exponent of BA model are still scarce at present. Reference [15] gave an algorithm but is too complicated to fit for the improvement requirement in this paper for using limit calculations. Reference [7] gave another way of improvement during its studies in AS level Internet topology. And this approach is briefly depicted as: according to the probability model of linking nodes (as mentioned in the upper BA algorithm description):

$$\Pi_i = k_i / \sum_j k_j \qquad (3)$$

where k_i, k_j are degree value of node i and j. If it's changed to:

$$\Pi_i = k_i^{1+\varepsilon} / \sum_j k_j^{1+\varepsilon} \qquad (4)$$

Then the power exponent of BA model would be modulated to be around 2.2 when parameter ε is set in an interval [0.1, 0.3][7]. Since value 2.2 is close to value 2.1406 in this paper, this method seemed to be effective for our requirement and would be adopted in this paper.

As to the optimization of the specific value of ε, a composite approach of Genetic Algorithm (GA)[30][31] and spectra[2][5][35] (SLS) would be used in the paper, and will be detailed described later.

2) Step 2: according to the degree-rank power-law exponents

Studies on AS-level Internet topology in [32] indicated that nodes in a network would not definitely conform to a power-law distribution with only one power exponent, especially the CCDF(d)-degree power-law and degree-rank power-law distribution. Likewise, the outcome of degree-rank power-law analysis is divided into two parts with two different power exponents (according to Fig.2 and Fig.3). So the model designed in this paper should be modulated according to this property so as to generate a network with two phases of degree-rank power-law distributions.

B. Implementation of Step 1: improve BA model by GA & SLS

1) Choice of the improvement approaches

There is a choice that we might find or optimize parameter ε through thorough searches with a certain increment step in a given interval. And if the network generated by the improved BA model with a certain value of ε could produce power exponent close to 2.1406, ε is optimized. Or else, continue the algorithm by move up to another value by an increment.

This method, however, is of low efficiency. Genetic Algorithm (GA)[30][31] is a kind of approach similar to but better than this method. GA also tries to find and optimize parameter ε in a certain interval, but differs in that, GA generate many random ε values in a certain interval and automatically find better ε out of all by operations such as cross, mutation and selection, etc. Experiments indicate that GA is much more efficient than the thorough searches approach.

GA, however, is still in low efficiency because GA could not evaluate the quality of a randomly selected ε till the power exponent of the generated network is calculated. This calculation of power exponent, however, is rather slow due to the process of the statistical operation and curve fitting (just as the power-exponent gained out of Fig.1, 2 and 3).

To solve this problem, SLS (signless Laplacian spectra)[2][5][35] is introduced into GA as an evaluation tool of parameter ε. The reason is, firstly, SLS is proved to be capable of quantitatively charactering a network topology; secondly, calculation of SLS is completely in matrix form and could be easily implemented by computer programs.

2) SLS of the Internet topology

Apply SLS on four 3000-node samples originated from the key testing sample (the twenty-one-monitor Internet topology) in this paper, and the outcome is illustrated in Fig. 5.

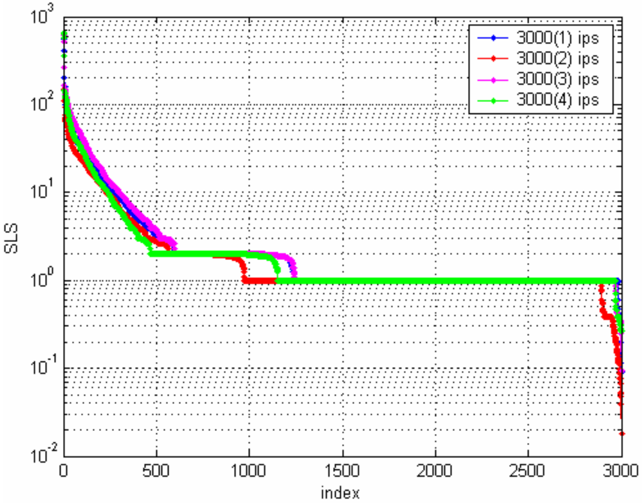

Fig. 5 SLS analysis results of four 3000-node Internet topology, axis y is in logarithm scale, and axis x is the SLS eigenvector sorted in descending order.

From Fig.5, all four curves show high similarities although the four samples are completely random and different from each other, which could be regarded as a proof that SLS is efficient in charactering Internet topology in this paper.

Besides, there are two evident horizontal lines when SLS equals to 1(10^0) and 2, which means that there are the most nodes in the Internet topology graph when SLS equals to 1, and the second most nodes when SLS=2. All four curves conform to these same properties. Again, correctness of SLS for samples in the paper is proved.

Finally, the four 3000-node samples are selected in a totally random way from the key testing sample, and the analogy of four and more samples (experiments show that more samples still conform to each other) proves that Internet topology, just like other scale-free network, is of a property of self-similarity in topology[4][14]. So, a conclusion could be inferred that the 3000-node topology is capable of being used as a representative of the whole Internet topology, i.e., the SLS of 3000-node topology could be regarded as a good and effective example of the whole Internet topology in the paper.

3) Evaluations of the differences between topologies by SLS

The SLS eigenvector is a sequence of values representing the topology characteristics of target graph. Common approaches are of little values in evaluation of such sequence. So an algorithm of cross-correlation[34] from communication theory is introduced here.

Cross-correlation algorithm is capable of distinguishing and identifying the differences between sequences in an absolutely quantitative way[25], i.e., it helps to determine how much two topologies are alike. It's mathematically defined as:

$$r_{xy}(n) = \frac{1}{N}\sum_{k=0}^{N-n-1} x(k)y(n+k) \qquad (5)$$

where *x*, *y* represents eigenvectors of the two topologies respectively, *k* is the order of the sequence, *N* is the sequence length, r_{xy} is outcome of cross-correlation calculation.

Cross-correlation would result in a maximum outcome only when the two SLS eigenvectors are totally identical, if there are some differences between them, the outcome would decrease[25]. And the larger the difference is, the smaller the gained result would be, which means more differences between topologies.

Equation (5) generally involves *n* rounds of calculations and *n* outcomes would be gained. Shift of SLS eigenvector sequence is auto-operated before each round of calculation in order to ensure that all possible conditions would be included in the calculation results. And the max value (i.e., the best outcome) out of the *n* outcomes would be selected as the final evaluation value of cross-correlation[34].

4) Final implementations of Step 1 by GA & SLS

Take cross-correlation algorithm as the evaluation function of GA, implementation of improvement of BA model by optimizing its parameter ε could be finally performed through GA. And the algorithm is described as follows: repeat the following steps till the termination conditions are met.

i) Gene code: We define a gene code ***x*** as a vector comprising primary parameters to be optimized. Of course, parameter ε is the only one to be optimized in this paper. So,

$$x = (\varepsilon) \qquad (6)$$

ii) Random initialization of gene group: Assuming the size of the gene group is *N* (*N* is set to be 100 in this paper), we randomly initialize a gene group having *N* genes, i.e., 100 copies of randomly selected parameter ε.

iii) Evaluation function: The choice of ε should minimize the difference between the generated network and real Internet, i.e., the cross-correlation outcome should be maximized. So the evaluation function should be:

$$f(x) = |\, r(x_\varepsilon, y)\,| \qquad (7)$$

where $r()$ is the cross-correlation operation, x_ε and y are SLS eigenvectors of generated network and the Internet topology (an eigenvector randomly selected from the analysis results of the four 3000-node Internet topology in Fig.5), respectively. The evaluation function is expected to score the genes. Superior genes have higher scores (value).

iv) Selection: Genes were sorted in descending order by their corresponding scores in the gene group, so all good genes were list in front. The first *m***N* genes, *m* is a random number (0<*m*<1), were directly selected for the next round of calculation by GA. Thereafter, we duplicate these *m***N* genes, and together with the genes that were not selected, i.e., *N*(1-*m*) genes, we get the gene group with size of 2**m***N* + *N*(1-*m*) = *N*+*mN*.

In order to keep the size of group remaining unchanged, we remove the last (worst) *m***N* genes and then the size of group gets back to *N*. This group is ready for next round of calculation in GA.

v) Crossover: Crossover operation is:

$$\begin{aligned} \varepsilon_i' &= \varepsilon_i(1-\alpha) + \beta\varepsilon_j \\ \varepsilon_j' &= \varepsilon_j(1-\alpha) + \beta\varepsilon_i \end{aligned} \qquad (8)$$

where α, β are random numbers, and $0 < \alpha < 1, 0 < \beta < 1$.

vi) Mutation: Mutation operation is:

$$\begin{aligned} \varepsilon_i &= \varepsilon_i(1+\alpha) \ \ if \ \ \gamma \geq 0.5 \\ \varepsilon_i &= \varepsilon_i(1-\alpha) \ \ if \ \ \gamma < 0.5 \end{aligned} \qquad (9)$$

where α, γ are random numbers, and $0 < \alpha < 1, 0 < \gamma < 1$.

Unlike crossover operations, not all genes have to be mutated. We set up a threshold of 0.3 in the algorithm, which means only 30% genes would be performed by mutation.

vii) Termination conditions: Basically there are two termination conditions in GA.

Firstly, GA would be terminated right after the best gene is found when evaluation function (Equation 7) result in the highest score or a maximum value. As mentioned above, maximized outcome from cross-correlation only occurs when the two SLS eigenvectors are totally identical. And in this paper, it's quite obvious that $r(y,y)$ (as mentioned, y is the SLS eigenvector of real Internet topology from Fig.5) is the maximum we are looking for, which means the generated network is completely equivalent in topology to real Internet.

This maximum value, however, is hard to achieve, since it's hard to generate a network exactly same as real Internet. We then set up a threshold as $0.95 \cdot r(y,y)$ to replace $r(y,y)$. A best optimized parameter ε is regarded to be found and GA will stop running if the evaluation result out of Equation 7 is great than this threshold.

The second termination condition is when GA have repeated for more than 1000 times before finding the best gene (parameter ε). If so, terminate the algorithm. This is done to ensure ending GA in an appropriated way, or else GA might run a very long time.

Parameter ε was finally optimized to be 0.10812 by GA in this paper.

C. Implementation of Step 2: the second improvement with two phases of degree-rank power exponents

After improvement of BA model by frequency-degree power exponent (step 1), we come to step 2 –another improvement by two phases of degree-rank power exponents.

The outcome of degree-rank power-law analysis is divided into two parts with two different power exponents (according to Fig.2 and Fig.3). And the nodes where ln(rank)<3.6 complied with a power exponent of 0.29981. So, the generated network should be modulated again to conform to this property, i.e., the degree-rank power-law |R| of the generated network should equal to 0.29981 when ln(rank) is less than 3.6, and equal to 0.84639 when others.

This improvement could be implemented as a periodical modulation model in the Internet model constructed in this paper, and its algorithm is directly depicted as part of the generating algorithm of the final Internet model in table VII.

D. Generating algorithm of the final Internet model

The final algorithm is listed in table VII.

E. The incomplete part of the model

The network generated by this final model only comprises nodes with degree greater than one, due to its primary inheritance from BA model. Internet topology, however, has a large amount of nodes whose degree is one. So the research on Internet topology modeling comprising those nodes would be our next work.

TABLE VII
THE GENERATING ALGORITHM OF THE CONSTRUCTED INTERNET MODEL

step	contents
(1)	Input number *N*. *N* is the number of the nodes in the to-be-generated network; /* *N* should be input by users */
(2)	Loop steps (3)(4)(5) and (6) until a network with *N* nodes is generated ;
(3)	/* Growth by the frequency-degree power-law properties */ Add a new node to the current network, and it would be linked to the randomly selected *m* nodes in the present network according to the linking probability function (shown in Equation (4) with parameter ε optimized to be 0.10812), and *m* is less than or equal to the total number of the nodes in the network. If the outcome out of the linking probability function is greater than a threshold t0=0.6, then a link between node *i* and the new added node will be added to the network. Or else, the link would not be added to the network. /* Threshold t0=0.6 is set by the program, and it helps avoid constructing a network with too many or too few links */
(4)	Define a threshold t1=10%, if the increment percentage of the new added nodes is greater than t1, then go to step (5) for degree-rank power-law modulation operation; or else go back to step (2).
(5)	/* Degree-rank power-law modulation */ Sort the nodes of the present network in descending order, for each node lying in an interval where ln(rank) is less than 3.6, calculate its degree by the degree-rank power-law distribution with the power-exponent of |R|=0.29981. If node *i*'s calculated degree is less than its present degree, then add links by algorithm step (3). Loop the operation till the degree equals to the calculated degree. If node *i*'s calculated degree is greater than its present degree, delete links. Randomly select node *j*, if the linking probability between *i* and *j* out of Equation (4) is greater than t0=0.6 and there is a link between node *i* and *j*, then delete it. Loop the operation till node *i*'s degree equals to the calculated degree.
(6)	Go back to step (2);

4. CONCLUSIONS

Frequency-degree power-law, degree-rank power-law and CCDF(d)-degree power-law distributions on the router-level Internet topology measuring samples were studied in this paper. The frequency-degree power-law relation is obvious and the power-exponent is found to be 2.1406. While for the degree-rank power-law, two phases of power-law relationships were found with power-exponents of 0.29981 and 0.84639, respectively. However, the CCDF(d)-degree power-law relationships were not clearly found in the research.

With the power-law relations and power-exponents found in the experiments, we began to construct an Internet topology model by two steps. Step 1, we improved traditional BA model through optimizing its parameter ε by GA and SLS eigenvector. Step 2, modulation by two-phase degree-rank power-law distributions was introduced to promote accuracy of the model. Generating algorithm of the constructed model was finally given in this paper.

REFERENCES

[1] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology[J]. ACM SIGCOMM ComputerCommunication Review, 1999,29(4):251-262.

[2] Jiang Y, Fang B.X., Hu M.Z. An Example of Analyzing the Characteristics of a Large Scale ISP Topology Measured from Multiple Vantage Points[J]. Journal of Software, 2005,16(5):846-856.

[3] Siganos G, Faloutsos M, Faloutsos P, Faloutsos C. Power laws and the AS-level Internet topology[J]. IEEE/ACM Trans. on Networking, 2003,11(4):514-524.

[4] Wang X.F., Li X., Chen G.R., Complex networks theory and its application[M]. Beijing:QsingHua Press, 2006,49-70.

[5] Dam E, Haemers WH. Which graphs are determined by their spectrum? [J]. Linear Algebra and its Applications, 2003,373:241-272.

[6] Lakhina A, Byers JW, Crovella M, Xie P. Sampling biases in IP topology measurements[C]. In: Proc. of the IEEE INFOCOM 2003,Vol 1. San Francisco: IEEE, 2003. 332~341.

[7] Sagy B, Mira G, Avishai W. An incremental super-linear preferential Internet topology model[C]. Proc. 5the Annual Passive and Active Measurement Workshop, LNCS 3015, 2004,53-62.

[8] Sagy B, Mira G. Avishai W. A geographic directed preferential Internet topology mode[C]. Arxiv:CS,2005,NI/0502061.

[9] Cao L.B., Dai R.W., The intelligent Information System—Internet[M]. Beijing: Science Press, 2001,121-130.

[10] Broido A, Claffy KC. Internet topology: Connectivity of IP graphs[C]. In: Fahmy S, Park K, eds. Scalability and Traffic Control in IP Networks (Proc. of the SPIE ITCom Vol. #4526). Washington: SPIE Press, 2001. 172-187.

[11] Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with rocketfuel[J]. ACM SIGCOMM Computer Communication Review, 2002,32(4):133-145.

[12] Waxman BM. Routing of multipoint connections[J]. IEEE Journal on Selected Areas in Communications, 1988,6(9):1617~1622.

[13] Zhang W.B. Research on the Life Characteristic and Evolution of Internet macroscopic Topology[D]. Shenyang: Northeastern University, 2005,6-23,49-67.

[14] Barabási AL, Albert R. Emergence of scaling in random networks[J]. Science, 1999,286(5439):509~512.

[15] P.L. Krapivsky, S. Redner and F. Leyvraz, Connectivity of Growing Random Networks[J], Phys. Rev. Lett., 85(2000), 4629-4632.

[16] Huffaker B, Plummer D, Moore D, et al.Topology discovery by active probing[EB/OL]. http://www.caida.org/outreach/papers/2002/SkitterOverview/. Jan. 2002.

[17] Skitter, CAIDA. http://www.caida.org/tools/measurement/skitter/

[18] Mapnet: Macroscopic Internet Visualization and Measurement, CAIDA. http://www.caida.org/tools/visualization/mapnet/

[19] Jiang Yu, Fang Binxing, Hu Mingzeng. Mapping Router-level Internet Topology from Multiple Vantage Points[J]. Telecommunications Science,2004(9):12-17.

[20] Cheswick B, Burch H, Branigan S. Mapping and visualizing the Internet[C]. In: Proc of the 2000 USENIX Ann Technical Conf, San Diego, California, USA, June 2000.

[21] Govindan R, Tangmunarunkit H. Heuristics for Internet map discovery[C]. In:Proc of IEEE INFOCOM 2000.

[22] Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with rocketfuel[J]. ACM SIGCOMM Computer Communication Review, 2002,32(4):133-145.

[23] R. Teixeira, K. Marzullo, S. Savage, and G. Voelker, In search of path diversity in ISP networks[C]. Proceedings of the USENIX/ACM Internet Measurement Conference, (Miami, FL, USA), October 2003.

[24] S. Bilir, K. Sarac, and T. Korkmaz, End to end intersection characteristics of Internet paths and trees[C]. IEEE International Conference on Network Protocols (ICNP), (Boston, MA, USA), November 2005.

[25] Huffaker B, Plummer D, Moore D, et al.Topology discovery by active probing[EB/OL]. http://www.caida.org/outreach/papers/2002/SkitterOverview/. Jan. 2002.

[26] iffinder, CAIDA. http://www.caida.org/tools/iffinder/.

[27] Govindan R, Tangmunarunkit H. Heuristics for Internet map discovery[C]. In:Proc of IEEE INFOCOM 2000.

[28] Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with rocketfuel[J]. ACM SIGCOMM Computer Communication Review, 2002,32(4):133-145.

[29] Ebel H, Mielsch L I, Bornholdt S. Scale-free topology of e-mail networks[J]. Phys. Rev E, 2002, 66, 036103-1-035103-4.

[30] WANG Jianming, XU Zhenlin.New crossover operator in float_point genetic algorithms[J]. CONTROL THEORY AND APPLICATION,2002.12 19(6).

[31] Rudolph G. Covergence properties of canonical genetic algorithms[J]. IEEE Trans.on Neural Networks, 1994, 5(1):96-101.

[32] Jared Winick, Sugih Jamin. Inet-3.0: Internet topology generator. Technical Report, CSE-TR-456-02, Ann Arbor: University of Michigan, 2002.

[33] Goh KI, Kahang B, Kim D. Spectra and eigenvectors of scale-free networks[J]. Physical Review E, 2001,64(5):1-5.

[34] Rorabaugh. COMPLETE DIGITAL SIGNAL PROCESSING[M]. US:McGraw-Hill, 2005.

[35] Farkas IJ, Derényi I, Barabási A, Vicsek T. Spectra of ‘real-world’ graphs: Beyond the semicircle law[J]. Physical Review E, 2001,64(2):1-12.

Contact Author: XU Ye
Contact Email: xuy.mail@gmail.com
Institutions / companies: Shenyang Ligong University, China
Contact Address: College of Information Science and Engineering, Shenyang Ligong University, Shenyang city, Liaoning Province, China. 110168

Xu Ye, got his ph.D degree in Computer application technology in 2006 from Noreastern University, China. And his research interests now include complex network modeling and information processing technology.
Email: xuy.mail@gmail.com