On Traveling Diameter of an Instance of Complex Networks – Internet

Ye Xu[*], Zhuo Wang, and Wen-bo Zhang

College of Information Science and Engineering, Shenyang Ligong University,
Shenyang 110168, China
{xuy.mail,zhuow,wenbozhang}@gmail.com

Abstract. As an instance of complex networks, Internet has been a hot topic for both complex networks and traditional networks research fields. Internet Traveling Diameter (ITD) is an important property defined in this paper representing the dynamic flow of Internet performance, and was mainly discussed. Short-term forecast model of ITD was firstly studied, then after it was proved that the short-term one was not good enough for long term forecast due to the complexity of Internet, the long-term model was studied. Both short-term and long-term model were given their mathematical descriptions at last.

Keywords: Internet traveling diameter; Logistic Model; GA; chaos; correlation dimension.

1 Introduction

With great improvements in the Internet research fields recently, more and more emerging research approaches, either from new research viewpoints or by a crossover study with other subjects, are applied to Internet related studies.

Studies in reference [1-4] made use of many new research approaches on Internet from complex networks point of view. In these approaches, Internet was regarded as an example of complex network due to its large scale and complicated variations, and definitions such as power law distribution [1,3,4], spectrum density [5], scale-free [2] and so on were utilized to depict qualitatively or compute quantitatively properties of Internet.

Referring to the idea of these research approaches, together with considering the fact that the hops of Internet datagram traveling from one node (router) to another are closely related to Internet performance, a definition of Internet traveling diameter, short for Internet diameter, would be discussed. Besides, the computing approaches of Internet diameter in both short-term type and long-term type would be discussed detailedly in this paper.

1.1 Definition of Internet Diameter

Assume that one datagram's transferring from one node (router) to a direct link node counts for 1 hop in Internet, and the datagram transferred from one source node to a destination

[*] Corresponding author. Ye XU, ph.D in computer application technology, associate professor, his current research interests include complex networks modeling, adaptive signal processing and pattern recognition.

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 90–99, 2009.

counts for *J* hops, where *J* is a number greater than or equal to 1. A definition of Internet traveling diameter is brought forward to represent the average hops, or statistical hops, of millions of datagram transferred from any source to any end at any time in Internet.

Definition 1. Assume J_i represents the hops of No. *i* datagram in Internet, the size of total datagram sample is *N*, and the frequency of No. *i* datagram in the sample is F_i, its probability is p_i, then Internet Diameter is

$$D = \frac{1}{N}\sum_{i=1}^{N} J_i F_i = \sum_{i=1}^{N} J_i p_i \tag{1}$$

1.2 Datagram Samples

As is well known, the validity of statistical results are entirely dependent on the size of statistical samples, the larger the sample is, the more accurate results would be.

Sample in this paper comes from three CAIDA[1] monitor nodes -- riseling node in SanDiego, CA, US, k-peer node in Amsterdam, NorthHolland, NL and apan-jp node in Tokyo, Kanto, JP[2]. Sampling time is from July 1999 to Jun. 2004[3]. The sample comprises more than 75,000,000 items in all and is of large scale. What's more, several other CAIDA monitors are also referred to sometime for a better accuracy.

There are datagram that could not reach the destination and are discarded by routers in Internet, and these datagram were depicted as "incomplete" or "*I*" in our sample, and those that could reach the target were depicted as "complete" or "*C*". Table 1 gives the detail.

Table 1. Sample details

Monitor node	Size of *C*	Size of *I*	*C*%	Total
riseling	16448329	13669728	**54.6%**	30118057
k-peer	9479028	9555955	**49.8%**	19034983
apan-jp	15172408	10423192	**59.3%**	25595600
Total				74748640

Since the unreachable datagram is simply discarded by routers in Internet, we ignore them and focus on the reachable ones in the sample. We could see from table 1 that the reachable ones account for larger percentage of the whole sample, and could still be regarded as a sample with great size.

1.3 A Quick Look at Hops in the Sample

Basically there are two techniques for us to have a quick look at properties of Internet Diameter. The first method was called "space-dimension analysis", by which differences

[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 distributed throughout the whole world, measuring and monitoring the variations of Internet.

[2] The reason selecting these monitors is that the nodes are separately located in three different continents on earth. This way of selection might provide a more general view of Internet throughout the whole world.

[3] Data items of one day out of one month are drawn out to build up the sample in this paper.

between different monitor nodes from different continents was illustrated in figure 1(a). And the other method is "time-dimension analysis", by which features of data items from only one node but at different time were illustrated in Fig. 1(b).

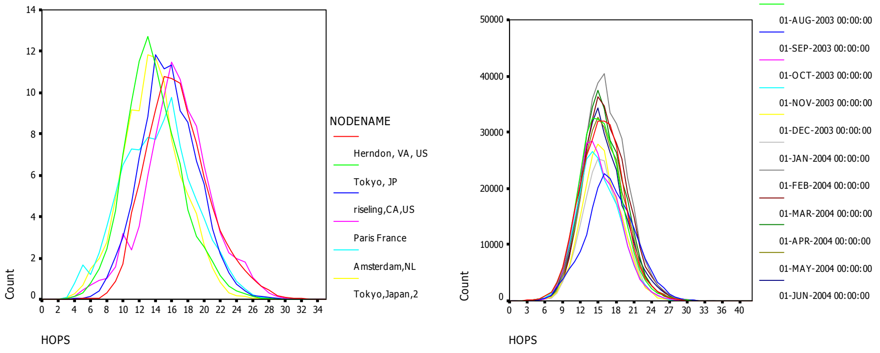
Another three monitors, a-root node in Herndon, VA, US, cdg-rssac node in Paris, France and nrt node in Tokyo, Kanto, JP were added to current sample for space-dimension analysis. And for "time-dimension analysis", a twelve months dataset (from 2003 to 2004) from riseling was selected, this selection does not lose generality because samples from all six nodes represent great consistency in figure 1(a).

We can see From fig. 1 that summits of six curves lie between 10 and 18 hops, which indicates that, although the data items in the sample were drawn out from different monitors at different time, they presents highly similar characters.

From figure 1(a) and (b), we see that all hops lie in an interval of [2, 32]. Taking errors into account, we enlarge the interval from [2, 32] to [1, 36], by which we could ensure that the entire sample was included.

Then for Traveling Diameter, according to equation (1), D is an average hops, so

$$D \in [1,36] \tag{2}$$



(a) Spatial-dimensional analysis

(b) Temporal-dimensional analysis

Fig. 1. A quick look at Internet Diameter

1.4 A Quick Look at Internet Diameter

We then begin to look at properties of Internet Diameter (D) by constructing a Cartesian coordinate system with D as Y axis and time as X axis, which is illustrated in Fig. 2.

It's obvious that all three curves have consistent variations in Fig. 2. With t growing longer, D is decreasing gradually. Though there is much difference among $D_{riseling}$, D_{kpeer} and D_{apanjp} at the beginning, the difference gets narrowed rapidly. Three Ds reach at a very small difference interval when t is around 60.

From Fig. 2, we can conclude that $D_{riseling}$, D_{kpeer} and D_{apanjp} represent great consistency when t is getting longer. Then, for simplicity of calculation, we average the three Ds and illustrate it in Fig. 3.

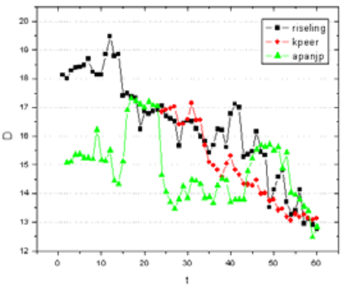


Fig. 2. A quick look at Internet Diameter (*D*). There are three monitor nodes selected in the graph, riseling, k-peer, apan-jp node, and the time is from 1999.07 to 2004.06. Y axis is value of *D* calculated by equation (1). X axis is time with a resolution of month, there are totally sixty months from 1999.07 to 2004.06.

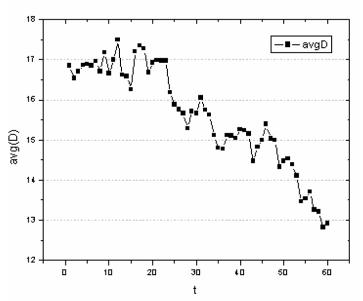


Fig. 3. Average *D* from 1999.07 to 2004.06.

AvgD in Fig. 3 is the basic element for short-term and long-term computing approaches.

In short-term part, variations of avgD in Fig. 3 is much similar to a Logistic Model, so a Logistic Model is selected as a fundamental framework model, but additional adjustment models are necessary [6].

2 Short-Term Model of Traveling Diameter

2.1 Improving Logistic Model

After applying avgD sample to Logistic Model [7,8,11], we get a nonlinear differential equation:

$$\frac{dD}{dt} = rD(1 - \frac{D}{k}) \tag{3}$$

where $D(\geq 0)$ means avgD at t, $t(>0)$ is time(month), $k(>0)$ means upper bound of avgD and $r(>0)$ means growth rate. Solving equation (3), we get an equation

$$D = \frac{k}{1 + \frac{k}{D_0 - 1} e^{-rt}} = \frac{k}{1 + me^{-rt}} \tag{4}$$

where $D_0 = D(t=0), m = \frac{k}{D_0 - 1}$ ·

(1) Transform one: from increasing function to decreasing one

Standard Logistic Function is an increasing function, but avgD is decreasing. Transform is needed.

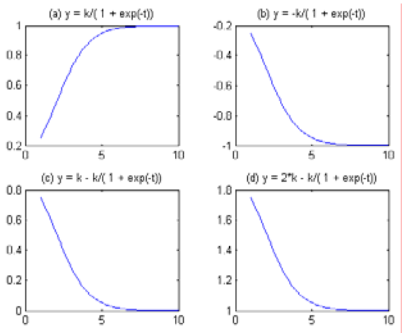


Fig. 4. Transform standard Logistic function into decreasing one through four steps. Step 1, standard Logistic function, *y*=*k*/(1+*e*^-*t*), *k* is 1. Step 2, add a minus sign to the function, then *y*=-*k*/(1+*e*^-*t*). Step 3, *y*=*y*+*k*, then *y*=*k*-*k*/(1+*e*^-*t*). And step 4, add another *k* to function because avgD is greater than 1, then *y*=2*k*-*k*/(1+*e*^-*t*).

Take a standard Logistic Function, as in equation (5), for an example,

$$y = \frac{k}{1+e^{-t}} . \quad \text{where } k = 1 \tag{5}$$

It is transformed into a decreasing one through four steps illustrated in Fig. 4(a)~(d).

In Fig. 4(d), the transition amount along Y axis is 2**k*, but this is not necessary. It could be any real number greater than *k* since avgD belong to an interval of [1, 36] and its lower limit is 1 (according to Fig. 4(c)). Then the improved Logistic Function is

$$D = d - \frac{k}{1+me^{-rt}} \tag{6}$$

(2) Transform two: additional adjustment models

There is a kind of quasi-periodic vibrations in the variations of avgD in figure 3, and according to reference [6], this kind of vibration could be simulated by adjustment models composed of sine and cosine functions. In reference [6], only sine function is referred to. However, we import both sine and cosine functions for a better efficiency in this paper.

After two transforms, we get a composite Logistic Function as:

$$D=d-\frac{k}{1+me^{-(v+p_1e^{-g_1t}\sin[\pi(\frac{t}{h1}+u1)]+p_2e^{-g_2t}\cos[\pi(\frac{t}{h2}+u2)])t}} \tag{7}$$

where $D(\geq 0)$ is avgD at t, d is adjustment parameter, p_1,p_2 is amplitude of vibration of avgD, h_1,h_2 is half period of vibration in month, u_1,u_2 is the initial phase and g_1,g_2 are parameters of modulus decay.

2.2 Curve Fitting

Float-point Genetic Algorithm [6][9][10] listed in table 2 is used for curve fitting of equation (7).

Table 2. An implementation of GA

procedures	information	equations and algorithms
(1) Definition of genes in GA.	Randomly initializing a gene group comprising 100 genes.	$x=(d,k,m,v,p_1,p_2,$ $g_1,g_2,h_1,h_2,u_1,u_2)$
(2) Definition of evaluation function	Assume *D(t)* is the value of avgD out of the composite model at *t*, and *D**(*t*) means real value at *t* from sample. Then evaluation function *f*(*x*) and its score function in GA are set up to be:	$f(x)=\sum_{i=1}^{60}\lvert D(t_i)-D^*(t_i)\rvert$ $score(x)=1/f(x)$
(3) Selection	Genes were sorted by scores from high to low in the gene group, and the first *m**N genes, *m* is a random number (0<*m*<1), were selected for the next round calculation by GA. Then we duplicate the selected genes, and deleted the last *m* genes to keep the group size remaining the same.	
(4) Crossover	Randomly select two genes, *xi*(*vi*…) 、 *xj*(*vj*…) out of the group to perform the crossover	$v_i'=v_i(1-\alpha)+\beta v_j$ $v_j'=v_j(1-\alpha)+\beta v_i$
(5) Mutation	Randomly select two genes, *xi*(*vi*…) 、 *xj*(*vj*…) out of the group to perform the crossover.	$v_i=v_i(1+\alpha)\ \ if\ \ \gamma\geq 0.5$ $v_i=v_i(1-\alpha)\ \ if\ \ \gamma<0.5$
(6) Termination conditions	Basically there are two termination conditions in GA in this paper. The first condition is when score of the best gene in the group is greater than a threshold *s*, *s* is set to be 0.1 in the algorithm. The other condition is when the number of calculation rounds in GA gets greater than a threshold *n*, and *n* is set to be 50000.	

After the curve fitting of GA, the result, i.e., the short-term model of equation (7) is:

$$D=16.9495-\frac{6.8262}{1+58.2418\times e^{-(A+B)t}} \tag{8}$$

where $t>0$, $t\in N$, A and B are:

$$A = 0.007844 + 0.001218\ e^{-0.001191\ t} \times \sin[\ \pi(\frac{t}{0.000866} + 0.014911\ )]$$

$$B = 0.082215\ e^{-0.004712\ t} \times \cos[\ \pi(\frac{t}{0.001} + 0.003242\ )]$$

2.3 Problem of Long-Term Computing with the Short-Term Model

Experiments indicate that the accuracy degree of the short-term model is getting worse with time getting longer, i.e., the short-term model is not suitable for long-term computation. So study on long-term computing approaches is to be introduced.

3 Long-Term Model of Traveling Diameter

3.1 Long-Term Forecast Principle in Chaos System

According to reference [12], it's very difficult to perform long-term forecast or computing a Chaos system due to the initial sensitivity. Mr. Lorenz E. N. had proved during his study on weather forecast that it is impossible to perform a long-term forecast in a Chaos system [17].

However, if "odd attractor, OA" exists in the Chaos system, long-term forecast of the system during a limited space and a limited period of time, is computable [12,13]. Though the forecast time is limited into a period of time, it's still much longer than that in short-term part, so we call the model under such conditions as long-term forecast model.

Since OA only exists in Chaos system, what we ought to do now is to prove whether variations system of Internet Diameter is a Chaos or not.

3.2 Chaos Proof

To prove whether a system is a Chaos, a new notion "correlation dimension, D_2" has to be introduced. If D_2 could be obtained out of the target system, the system is confirmed to be a Chaos, and OA is proved to exist [12].

Table 3. Algorithm for *D*2

Algorithm: Correlation dimension solving algorithm
Input: list /*Time sequence sample of Internet Diameter (1999.07~2004.06)*/

```
    /* Initialization */
    tao = 3; length = list.length;
    /* calculate $D_2$ with an increaing m */
loop when $m$=4,8,10,12 … until $D_2$ is convergent
    /*Constructing a coordinate system of length-(m-1)*tao dimensions */
    vecgroup = zeros($m$, length-($m$-1)*tao);
    vecgroup = getValue(list);
    /* Calculate distances between vector $i$ and vector $j$ in the coordinate system */
    rij = calcRIJ(vecgroup(:,$i$), vecgroup (:,$j$));
    /* Get a Matrix r*/
    r = [maxRij:(maxRij-minRij)/15:minRij]
    /* Calculating correlation integral, cr */
    cr = calcCR($r$, $rij$);
    /* plot */
    plot(log($r$), log($cr$));
    /* get $D2$, $D2$ is the slope of curve in plot. If $D2$ exists, the curve in the plot would
be similar to a straight line and the slope of curves would increase till a limited bound
with the increasing $m$. */
    calcD2();
/* end of loop while $m$=4,8,10,12,14,16,18 */
end loop
```

For calculating D_2, a special sample called "time sequence" would have to be drawn out first from the sample ranging from 1999.07 to 2004.06.

With the time sequence sample, we then construct an *m*-dimensional coordinate system, where *m* is an integer and is usually less than 100. After this, different *m* in an increasing order would be input into the algorithm [12,14~16] to construct different *m*-dimensional system. The output of the algorithm D_2 is the slope of curve in the output plot, and it would increase very fast when m increases. If D_2 increases to an infinite number such as tan(*pi*/2) finally, the system is proved to not be a Chaos. On the contrary, if D_2 increases till a limited upper bound, the system is confirmed to be a Chaos with OA. The algorithm is listed in table 3 [12][14][15][16].

Reference [12] suggests that tao should lie between (4, 8), but since the size of the time sequence sample of Internet Diameter is rather small (only sixty months), so we set tao to be 3 in the algorithm. Result of D_2 solution is illustrated in Fig. 5.

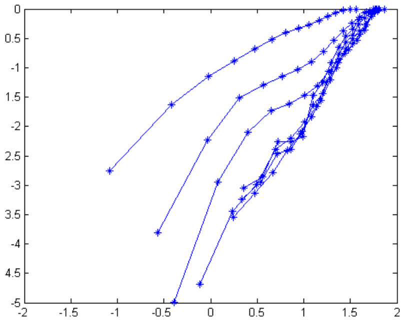


Fig. 5. *D*2 solutions diagram with lnC2(*r*,*m*) as Y axis and ln(*r*) as X axis. Algorithm parameters are: τ=3, *m*=2,5,8,11,13,15,16. Seven curves represent seven *m* from left to right, when *m*=2,5,8, 11,13,15,16.

From Fig. 5, there is good convergence of curves when *m*=11,13,15,16 -- the last four curves from left to right. The reason is that, firstly, the four curves are much similar to straight lines. Secondly, the slopes of the lines remain stable with *m* getting larger.

So D_2 in the variations system of Internet Diameter exists and is:

$$D_2(m_c = 16, \tau = 3) = 2.2444 \tag{9}$$

3.3 Chaos Forecast Model for Long Time System

The existence of D_2 proves that the system is a Chaos with OA, and could provide further useful information for setting up long-term models.

Study in reference [12] indicates that only a model with at least $\lceil D_2 \rceil$ dimensions, i.e., a function set with $\lceil D_2 \rceil$ functions, could perform a long-term forecast with relatively acceptable accuracies. According to equation (9), $\lceil D_2 \rceil$ is three, then the model should be:

$$\begin{cases} x = f_1(x, y, z, x^{(n)}, y^{(n)}, z^{(n)}) \\ y = f_2(x, y, z, x^{(n)}, y^{(n)}, z^{(n)}) \\ z = f_3(x, y, z, x^{(n)}, y^{(n)}, z^{(n)}) \end{cases} \tag{10}$$

where *x*~*z* represents important physical variables in target system, and $x^{(n)}$~$z^{(n)}$ means *n* orders of differential coefficient of the corresponding variable.

Equation (10) is only a framework of a Chaos model. It's still very difficult to confirm the parameters of the model because Chaos is a system with great complexity and the researches on it are still not very clear nowadays, as well as the time range of the time sequence sample is rather small (only from 1999 to 2004). And this would be our future work.

4 Conclusions

In short-term part, a model based on a Logistic Model and additional adjustment models is brought forward. Parameters of the model were finally confirmed through GA experiments. In long-term part, correlation dimension D_2 of Internet Diameter is calculated out of an algorithm. With D_2, a long-term model with three differential coefficient functions is brought forward. However, parameters of the model are still uncertain due to a short time range of sequence sample and complexity of the target system. And this would be our next work.

References

1. Floyd, S., Paxson, V.: Difficulties in simulating the Internet. IEEE/ACM Trans. on Networking 9(4), 392–403 (2001)
2. Watts, D., Strogatz, S.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
3. Aiello, W., Chung, F., Lu, L.Y.: A random graph model for massive graphs. In: Proc. of the ACM STOC 2000, pp. 171–180. ACM Press, Portland (2000)
4. Zhang, Y., Zhang, H.-L., Fang, B.-X.: A Survey on Internet Topology Modeling. Journal of Software 15(8), 1221–1226 (2004)
5. Farkas, I.J., Derényi, I., Barabási, A., Vicsek, T.: Spectra of 'real-world' graphs: Beyond the semicircle law. Physical Review E 64(2), 1–12 (2001)
6. Yin, C.Q., Yin, H.: Artificial Intelligence and Expert System, pp. 291–295. China Waterpower Publishing House (2002)
7. Yang, Z.J., Xu, Z.R.: Forecast of the Population Growth in the Country of HEILONGJIANG by the Forecast Method of Dynamic Logistic. Journal of Agriculture University of HEILONGJIANG 9(2), 23–28 (1997)
8. Wu, S.L.: Forecast of development of China Numerical Library by Logistic Model. Journal of Information 4 (2004)
9. Wang, J.M., Xu, Z.L.: New crossover operator in float-point genetic algorithms. Control Theory And Applications 19(6) (December 2002)
10. Rudolph, G.: Covergence properties of canonical genetic algorithms. IEEE Trans. on Neural Networks 5(1), 96–101 (1994)

11. Zhang, H.: Two New Population Growth Equation. Journal of Bimathematics 10(4), 78–82 (1995)
12. Huang, R.S.: Chaos and its application, vol. 1.1, pp. 191–239. WU HAN University Press (2000)
13. Zhang, Q.C., Wang, H.L., Zhu, Z.W.: Split and Chaos, vol. 1.1, pp. 256–281. TIAN JIN University Press (2005)
14. Kenneth, J.: Falconer. Fractal Geometry – Mathematical Foundations and Applications, vol. 8, pp. 41–75. Northeastern University Press (1991)
15. Xu, P.: Analysis of attractor of dam observations. Journal of Dam Observations 10 (1992)
16. Kang, Z., Huang, J.-W., Li, Y., Kang, L.-S.: A Multi-Scale Mixed Algorithm for Data Mining of Complex System. Journal of Software 14(7), 1229–1237 (2003)
17. Lorenz, E.N.: Deterministic nonperiodic flow. J. Atoms, Sci. 20 (1963)