A Time-series Decomposed Model of Network Traffic[1]

Cheng Guang
Computer Dept. Southeast Univsity
Nanjing, Jiangsu , China (210096)
8625-83794000-213
gcheng@njnet.edu.cn

Gong Jian
Computer Dept. Southeast Univsity
Nanjing, Jiangsu , China (210096)
8625-83792150
jgong@njnet.edu.cn

Ding Wei
Computer Dept. Southeast Univsity
Nanjing, Jiangsu , China (210096)
8625-83792313
wding@njnet.edu.cn

Abstract:

Traffic behavior in a large-scale network can be viewed as a complicated non-linear system, so it is very difficult to describe the long-term network traffic behavior in a large-scale network. In this paper, according to the non-linear character of network traffic, the time series of network traffic is decomposed into trend component, period component, mutation component and random component by different mathematical tools. So the complicated traffic can be modeled with these four simpler sub-series tools. In order to check the decomposed model, the long-term traffic behavior of the CERNET backbone network is analyzed by means of the decomposed network traffic. The results are compared with the ones of ARIMA model. According to the autocorrelation function value and predicting error function, the compounded model can get higher error precision to describe the long-term traffic behavior.

Categories and Subject Descriptors

H.4.3 Communications Applications

General Terms

Measurement

Keywords

Non-linear, Traffic Behavior, Compounded Model, Network Measurement.

1 Introduction

The traffic behavior and its tendency have bothered network managers for quite a long time, and is still a not fully understood problem for network management and planning. The underlying problem is that it is difficult to describe and model traffic of large-scale network. In order to research network behavior, firstly it is necessary to analysis the measured traffic and to find its statistical laws, such as the work done by Thompson in 1997[1]. Secondly, according to the statistical rules found, some traffic models are built, such as the establishment of self-similar model for Ethernet network traffic in 1994 [2]. If the time-scale of network traffic is considered, the network traffic behavior will be different in different time-scale. Paxson and Floyd [3] showed that the traffic behavior of millisecond-time scale isn't self-similar by the influence of network protocol. Due to the influence of environment, the traffic behavior whose time-scale is larger than ten minutes isn't also self-similar and is a non-linear time-series. Only the traffic behavior in second-time scale is self-similar. In the paper, the traffic behavior model whose time-scale is larger than ten minutes, is concerned.

It is very important to research traffic time-series model in order to describe traffic behavior. The traditional long time-scale traffic model can only model smoothing process and some special non-smoothing process. AR [4] (Auto Regressive) model, MA (Moving Average) model, and ARMA [5] (Auto Regressive Moving Average) model can deal with smoothing process. ARIMA [5] (Auto Regressive Integrated Moving Average) model and ARIMA seasonal model [6] can describe the uniform non-smoothing process. A large-scale network itself is a complex non-linear system, and is influenced by many environment factors, which is similar with water-volume time series that can be decomposed into mutation item, periodic item, trend item, random item [7]. So network traffic also can be considered to be the combination of periodic item, trend item, random item, and mutation item, which is very difficult to describe these traffic characters with a traditional traffic time-series model.

According to these characters of traffic behavior, in the paper the traffic time-series is decomposed into four simple sub-components: mutation component, trend component, period component, and random component. Firstly, based on the fact that median is the robust estimation of mean, the mutation component of long-term traffic is removed. Secondly, the trend component is separated from the rest traffic components by the GM(1,1) model of gray system theory. Thirdly, the period component is separated from the rest traffic components by the period wave theory. Last, the rest components are modeled on the AR(P) model of time-series theory. So the long-term traffic model can be obtained by combining the theory models of three sub components: trend component, period component, and random component. Finally, in the paper some CERNET traffic data are modeled with the compounded model

[1] This research was partially supported by 863 program of China under grant No. 2001AA112060, NNSFC under grant No. 90104031, and the young staff promotion program of Southeast University under grant No. 3009001112.

and the traditional ARIMA seasonal model respectively, and the analyzed results of two kinds of models are compared and analyzed.

2 Traffic Compounded Model

Network user behavior is influenced by environment, so network traffic behavior includes both rule and abnormity. In addition, a large-scale network itself is a non-linear system, so the non-linear long time-scale traffic can behave the characters of mutation, trend, period, and randomness. According to the traffic characters, the network traffic long time-scale time series X(t) can be separated into trend component A(t), period component P(t), mutation component B(t), and random component R(t). So the long-scale time series can be described as

$$X(t)=B(t)+A(t)+P(t)+R(t) \quad (1)$$

In the equation (1), X(t) is the rule traffic time-series. B(t) is effected by exterior environment mutation factors, and A(t) reflects the long term changed trend of network usage or environment factors, and P(t) reflects the periodic movement of traffic phenomena. B(t), A(t), P(t) show the determinate factors of traffic time series change. The random component R(t) can be decomposed into the smoothing random time series component S(t) and simple random component N(t) continuously.

$$R(t)=S(t)+N(t) \quad (2)$$

In the five components, the mutation component and pure random component belong to zero memory components. A(t), P(t), S(t) are the memory components that describe the long term trend, period, and smooth process of network traffic behavior. If the three component models can be modeled as a(t), p(t), and s(t) respectively, then the traffic model x(t) of X(t) can be modeled as

$$x(t)=a(t)+p(t)+s(t) \quad (3)$$

Therefore, according to equation (1) and (2), the network traffic can be divided into five sub-components with different mathematics tools respectively. Then the trend component, period component, and smoothing random component are modeled separately. The raw traffic time-series model can be obtained by the equation (3).

2.1 Decomposing Mutation Item

The basic idea of decomposed mutation item is to produce a smoothing estimate of curve firstly, then we can obtain a error time-series that is subtracted the smoothing curve from measurements of network traffic. If the error point is larger than the appointed threshold, then that point is considered a mutation point. The method bases on the fact that the median is a robust estimation of mean. The algorithm that the mutation component is removed from traffic time series X(t) is as following.

Step 1. A new time series X’(t) is constructed by traffic time series X(t).

$$X'(t) = middle(X(t-2), X(t-1), X(t), X(t+1), X(t+2)) \quad (4)$$

Where middle() is a function that obtains a median from data in bracket, and t∈[2, n-2].

Step 2. X’’(t) is constructed from X’(t).

$$X''(t) = middle(X'(t-1), X'(t), X'(t+1)) \quad (5)$$

Where t∈[3, n-3].

Step 3. X’’’(t) is constructed with X’’(t).

$$X'''(t) = X''(t - 1)/4 + X''(t)/2 + X''(t + 1)/4 \quad (6)$$

Where t∈[4, n-4].

Step 4. If |X(t) – X’’’(t)| > k， then X(t) is replaced by the linear inner inserted value. Where t∈[4, n-4], k is a predefined value.

Every measuring point in the [4, n-4] aggregate is computed with the fourth step repeatedly, till the mutation item B(t) is separated from the traffic time-series X(t).

2.2 Trend Item Decomposed Model

GM(1,1) is used to separate A(t) from the complex traffic time series X1(t) that includes trend item, period item, and random item. The algorithm is described as following.

Step 1. Accumulated equation is constructed. Traffic series X1(t) is expressed as equation (7).

$$X1^{(0)} = \{X1_0^{(0)}, X1_1^{(0)}, ..., X1_i^{(0)}, ..., X1_n^{(0)}\} \quad (7)$$

Where X1(0) is equal to X1(t) that doesn’t contain the mutation component, and X1i(0) means traffic bandwidth on ith time, and i∈[0, n]. The equation (7) is accumulated in turn, and X1(1) is obtained.

$$X1^{(1)} = \{X1_0^{(1)}, X1_1^{(1)}, \cdots, X1_i^{(1)}, \cdots, X1_n^{(1)}\} \quad (8)$$

Where $X1_i^{(1)} = \sum_{t=0}^{i} X1_t^{(0)} = X1_{i-1}^{(1)} + X1_i^{(0)}$, i∈[1, n], X10(1)=X10(0), X1i(1) is the network traffic throughput from time 0 to time i. Because the distribution of the series X1(1) can be simulated by exponential function, so the smoothing discrete coefficient can be expressed with differential equation. The one order differential coefficient function is expressed by function (9),

$$\frac{dX1^{(1)}}{dt} + aX1^{(1)} = b \quad (9)$$

and its result is computed by equation (10),

$$X1_t^{(1)} = (X1_0^{(1)} - \frac{b}{a})e^{-at} + \frac{b}{a} \quad (10)$$

where a and b are the parameters that must be estimated.

Step 2. The parameters a and b are estimated. The evaluation can obtain with the method of least squares,

$$Y = XB \quad (11)$$

where

$$X = \begin{bmatrix} -\frac{1}{2}\left|X1_0^{(1)} + X1_1^{(1)}\right| & 1 \\ -\frac{1}{2}\left|X1_1^{(1)}(2) + X1_2^{(1)}\right| & 1 \\ \cdots & \cdots \\ -\frac{1}{2}\left|X1_{n-1}^{(1)} + X1_n^{(1)}\right| & 1 \end{bmatrix}$$

$$Y = \begin{bmatrix} X1_1^{(0)} \\ X1_2^{(0)} \\ \cdots \\ X1_n^{(0)} \end{bmatrix} \quad B = \begin{bmatrix} a \\ b \end{bmatrix},$$

$$B = (X'X)^{-1}X'Y.$$

Step 3. If the parameters a and b are estimated, then the A(t) can be extracted from X(t). The model a(t) of A(t) is expressed in equation (8).

$$a(t) = X1_t^{(1)} - X1_{t-1}^{(1)} \text{ t}\in[1, \text{n}] \qquad (12)$$

According to the equation (10) and (12), the traffic model can be obtained by equation (13).

$$\text{a(t)} = (e^{-a} - 1)(X1_0^{(1)} - \frac{b}{a})e^{-a(t-1)} \text{ t}\in[1, \text{n}] \quad (13)$$

Let X2(t) = X1(t) – a(t) = P(t) + R(t) be the rest time-series that B(t) and A(t) be separated from X(t), so X2(t) is a new time series, whose axes is A(t). The advantage of this new time-series is that it emphasizes the effect of P(t).

2.3 Period Item Decomposed Model

X2(t) is considered as the superposition of some different period waves. So firstly some obvious periods are separated from X2(t) in turn, then these periods are accumulated into P(t). Its algorithm is described as following.

Step 1. Lists all possible period in X2(t). Before the period is analyzed, the number of periods contained in the time series is unknown, so each period must be tested separately.

$$K = \left\{\frac{n}{2}\right\} = \begin{cases} \frac{n}{2}, \ n \text{ is even} \\ \frac{n+1}{2}, \ n \text{ is odd} \end{cases} \qquad (14)$$

Where n is the length of X2(t), and K is the maximum possible period number.

Step 2. Computes the square sum of deviating mean, which include the square sum of deviating mean of both inner team (equation 15) and between teams (equation 16).

$$Q_2^2 = \sum_{j=1}^{k}\sum_{i=1}^{m}(y_{ij} - \bar{x}_j)^2, \ \bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} y_{ij} \qquad (15)$$

$$Q_3^2 = \sum_{j=1}^{k} m(\bar{x}_j - \bar{x})^2 \quad \bar{x} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_i \qquad (16)$$

where k is the chosen period length, m is the number of a team, yij is the sequence value X2(t) whose freedom degree is f2 = n – k, f3 = n - 1.

Step 3. Compute the variance ratio between different test periods.

$$F = \frac{Q_3^2 / f_3}{Q_2^2 / f_2} \qquad (17)$$

Step 4. Verifies the variance. A confidence limit α is chosen, e.g. α is equal to 0.05. Then the F distribution table [7] is checked to get Fα. If F> Fα, then the test period exists, otherwise the test period does not exist, and skip the step five.

Step 5. Tests k from 2 to K, where K = ⌊n/2⌋ and n is the length of X2(t), until the time series does not have obvious period.

2.4 Random Item Decomposed Model

Let X3(t) = X2(t) – p(t) = R(t) = S(t) + N(t), and a smoothing random item S(t) is expected to be extracted from X3(t).

X(t) =

$$x(t) = \beta_{p,1}x(t-1) + \beta_{p,2}x(t-2) + \ldots$$

$$+ \beta_{p,p}x(t-p) \qquad (18)$$

where βp,j (j=1,2,…,P) is the auto-regression coefficient, and P is the order number. The algorithm is described as following.

Step 1. Computes the model coefficient.

$$\begin{cases} \beta_{1,1} = \gamma_1 \\ \beta_{k,k} = \dfrac{\gamma_k - \sum_{j=1}^{k-1}\beta_{k-1,k}\gamma_{k-j}}{1 - \sum_{j=1}^{k-1}\beta_{k-1,j}\gamma_j} (k = 2,3,\ldots) \\ \beta_{k,j} = \beta_{k-1,j} - \beta_{k,k}\beta_{k-1,k-j} (j = 1,2,\ldots,k-1) \end{cases} \qquad (19)$$

whereβi,j is the auto-regression coefficient, and γk is the k order auto-correlation coefficient of X3(t),

$$\gamma_k = \sum_{t=1}^{n-k} X3_t X3_{t+k} \Big/ \sum_{t=1}^{n} X3_t^2 .$$

Step 2. Computes the order number of model that can be ensured according to AIC rule.

$$AIC = \min\left\{ n \ln \frac{\sum (X3(t) - \overline{X3})^2}{n - P - 1} \right\} \qquad (20)$$

where $\overline{X3}$ is the mean of X3(t), n is the length of X3(t).

Step 3. Computes the smooth time-series model s(t). Firstly p values of data in the preceding s(0) are defined, then s(t) is evaluated from X3(t) reversely in order to obtain the data s(-1), s(-2), … , s(-p) before the measuring points, so as to get equation (21).

$$s(t) = \beta_{p,1} s(t-1) + \beta_{p,2} s(t-2) + \ldots$$
$$+ \beta_{p,p} s(t-p) \qquad t\in[0, n] \qquad (21)$$

Random item of the measuring points can be estimated by means of equation (21).

3 Analysis of Network Traffic

In order to verify the compound model, three different time-scale network traffics are analyzed and modeled. The first group of 11 days data (CERNET01, in figure 1), whose timescale is one hour, comes from one backbone router of CERNET in 2001. The second group of 121 days data (CERNET02, in figure 2), whose time-scale is one day, also comes from CERNET. The third group of data (NSFNET, in figure 3), whose timescale is one week, comes from one national backbone route of NSFNET from Aug. 1, 1988 to Jun. 30, 1993 [8].

Figure1 cernet01 time-series

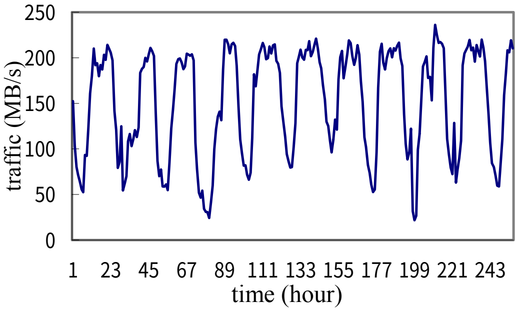


Figure 2 cernet02 time-series

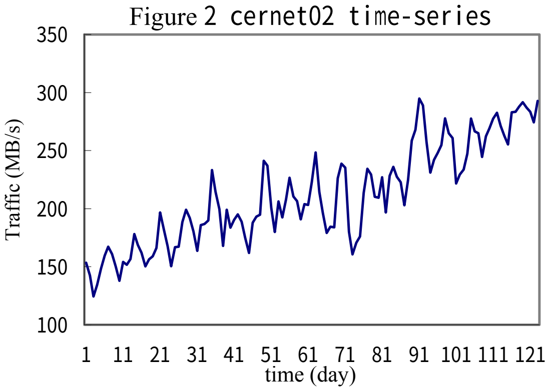
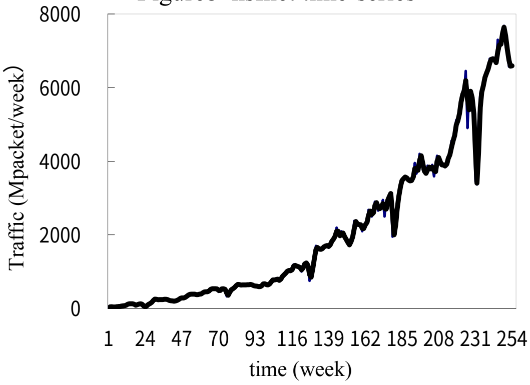


Figure3 nsfnet time-series



3.1 Model Parameters

Three kind traffic trace are modeled by the compound model algorithm, and these model parameters are listed in table 1.

Table1: Model Parameters of three Trace

trace			cernet1	cernet2	nsfnet
Model Parameters	A(t)	A	-0.0009	-0.0052	-0.018
		B	121.109	149.592	108.96
	P(t)	period	24	7	26
		A	-0.0056	0.0003	-0.406
		B	152.802	50.804	84.47
	S(t)	AR(p)	0.6863	0.6842	0.702
			0.0352	0.1016	0.008
				-0.0090	0.039
				-0.1706	-0.144
				0.0639	
				0.0711	
				-0.0011	

3.2 Model Analysis

The auto-correlation function ACF(i) can reflect the traffic long term behavior. Because the trend component dominates the whole traffic behavior, so the period behavior is hidden in Figure 2. Figure 4 shows the ACF(i) of traffic time-series that removes B(t) and A(t) components. Because the period component dominates the rest traffic behavior, so in Figure 4 the 7 days period of time-series is obvious.

Figure 5 shows the auto-correlation function curve including the trend item and period item, and the mutation item is removed. In fig5, the ACF shows the lag character. When lag is increasing, ACF tends to 0, the behavior is fit of AP(p) model. So the ACF shows that the decomposed model is very rational.

Fig4: CERNET period and random

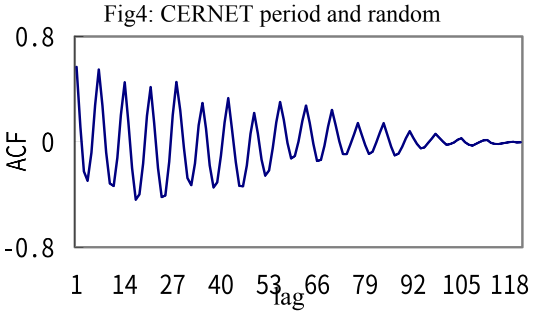
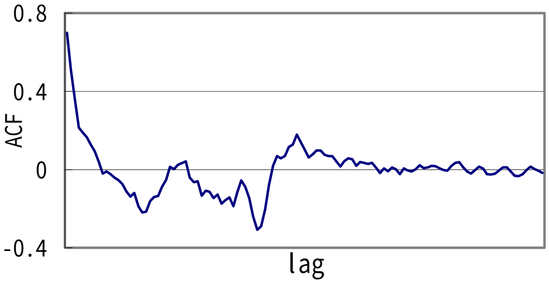


Fig5 CERNET traffic random item



4 Comparison with ARIMA Seasonal Model

CERNET01 and NSFNET traces are compared with their forecasting result error, which is defined as equation (22). CERNET02 trace is compared with the simulated result SSN, which is defined as equation (23).

$$error = \sqrt{\frac{\sum_{i=n+1}^{n+1+r}(X_i - \hat{X}_i)^2}{r}} \qquad (22)$$

Where n is the time-series length of the model, r is forecasting length. n of CERNET01 trace is 240, and r is 24. In NSFNET trace, n is equal to 253, and r is 52.

$$SSD(m) = \frac{1}{n-1}\sum_{i=1}^{n}(ACF_m(i) - ACF_s(i))^2 \qquad (23)$$

Where SSD(m) is the auto-correlation sample variance of model m, and ACFm(i) is the ith order auto-correlation function of model m, ACFs(i) is the ith order auto-correlation function of measuring sample series. The statistical metrics reflects the auto-correlation of the model, the less the value, the better the effect.

The ARIMA seasonal models of three kind traces are as following.

The first forecasting model of CERNET01 trace is ARIMA(2, 0, 2) ×(0, 1, 0) 24, and its parameters are $(\beta_1, \beta_2, \theta_1, \beta_2)$ = (0.1652, -0.676, 0.8705, -1294).

The second forecasting model of NSFNET t race is ARIMA(2, 2, 1)×(2, 2, 0) 52, and its parameters are $(\beta_1, \beta_2, \theta_1, \beta_3, \beta_4)$ = (-0.176822, -0.000685, 0.993894, -0.273511, 0.653531).

The simulated model of CERNET02 trace is ARIMA(7, 0, 0)×(0, 1, 0) 7, its parameters are $(\beta_1, \beta_2, \ldots, \beta_7)$ = (0.6606, 0.1631, -0.0805, -0.1232, -0.0085, 0.1721, -0.2153).

Figure 6 is the ACF of both CERNET02 decomposed simulating model and CERNET02 trace, and Figure 7 is the comparison between the ARIMA seasonal model and the CERNET02 trace. Figure 6 and Figure 7 show that the precision of the compounded model is better than the ARIMA seasonal model.

figure 6 the ACF both trace and compounded model

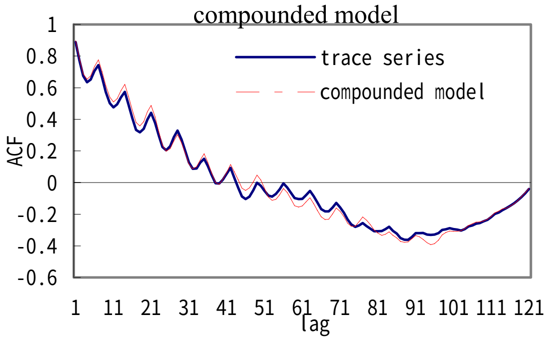


figure 7 the ACF both ARIMA seasonal model and trace

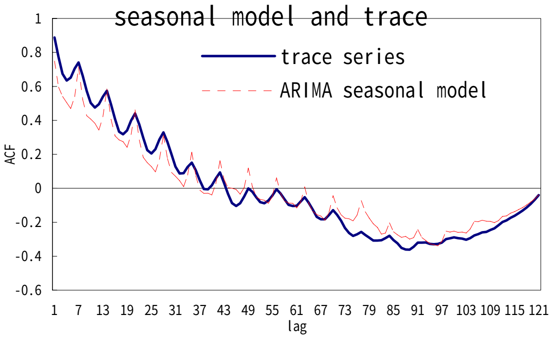


Table2 is the SSD of the decomposed models of CERNET02 trace, and table 3 is the forecasting error statistics of CERNET 01 trace and NSFNET trace. From table 2 and table 3, we can know that the decomposed model suggested in the paper is very effective.

Table2 SSD of two models

Model	SSD
Decomposed model	0.000984
ARIMA seasonal model	0.005471

Table 3 error of two models

model	cernet01 error	nsfnet error
Compounded model	6.81	209.31
ARIMA seasonal model	10.26	421.92

5 Conclusion

In the paper, a decomposed model of long time-scale network traffic in a large-scale network is suggested and verified with two groups of CERNET traffic traces and one group of NSFNET trace. The analysis shows that CERNET user behavior has the periodicity of hour, day, and week. The experiment result proved that the prediction precision of the decomposed model is better than the ARIMA seasonal model's.

The decomposed model has three main advantages. Firstly, the decomposed model uses multi-types sub-models to describe traffic behavior, and has more parameters, so it can describe traffic behavior more accurately and perfectly than the traditional ARIMA model. Secondly, the decomposed model is composed of four sub-models that can describe different aspects of the traffic character. Finally, according to the measured traffic behavior, one or multi sub-model can replaced by others sub-model. For example, in order to research self-similar traffic behavior, the random item sub-model can be replaced by FARIMA model or FGN model.

6 Reference

[1] Kevin Thompson, Gregory J. Miller, and Rick Wilder. Wide-Area Internet Traffic Patterns and Characteristics[J].IEEE Network, November / December 1997, 5(6): 10-23.

[2] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic [J]. IEEE/ACM Transaction on Networking, Feb. 1994, 2(1): 1-15.

[3] V. Paxson, S. Flod. Wide-area traffic: The failure of poisson modeling. IEEE/ACM Transactions on Networking [J], June 1995, 3(3):226-244.

[4] Rich Wolski. Forecasting Network Performance to Support Dynamic Scheduling Using the Network Weather Service [DB/OL]. UCSD Technical Report, TR-CS96-494(1996). http://citeseer.nj.nec.com/wolski98dynamically.html.

[5] S. Basu and A. Mukherjee. Time series models for internet traffic [J]. Proc. IEEE INFOCOM'96, San Francisco, CA., March, 1996, 2: 611-620.

[6] N. Groschwitz, G. Polyzos. A Time Series Model of Long-term Traffic on the NSFnet Backbone [j]. In Proceedings of the IEEE International Conference on Communications(ICC'94), New Orleans, LA, May 1994

[7] Jing Guangyan, Random Analysis of Hydrology and Water Resource, Press of Chinese Science and Technology, in Beijing, 1992. 5, pp:406-436.

[8] K. Claffy, G. C. Polyzos, and H. W. Braun. Traffic Characteristics of The T1 Nsfnet Backbone [J]. proceedings IEEE INFOCOM'93, San Francisco, California, March 28-April 1, 1993: 885-892.