Correlation dimension based nonlinear analysis of network traffics with different application protocols*

Wang Jun-Song(王俊松)[a)b)†], Yuan Jing(袁 静)[b)], Li Qiang(李 强)[b)], and Yuan Rui-Xi(袁睿翕)[b)]

[a)] *School of Biomedical Engineering, Tianjin Medical University, Tianjin* 300070, *China*
[b)] *Department of Automation, Tsinghua University, Beijing* 100084, *China*

(Received 29 May 2010; revised manuscript received 5 December 2010)

This paper uses a correlation dimension based nonlinear analysis approach to analyse the dynamics of network traffics with three different application protocols—HTTP, FTP and SMTP. First, the phase space is reconstructed and the embedding parameters are obtained by the mutual information method. Secondly, the correlation dimensions of three different traffics are calculated and the results of analysis have demonstrated that the dynamics of the three different application protocol traffics is different from each other in nature, i.e. HTTP and FTP traffics are chaotic, furthermore, the former is more complex than the later; on the other hand, SMTP traffic is stochastic. It is shown that correlation dimension approach is an efficient method to understand and to characterize the nonlinear dynamics of HTTP, FTP and SMTP protocol network traffics. This analysis provided insight into and a more accurate understanding of nonlinear dynamics of internet traffics which have a complex mixture of chaotic and stochastic components.

Keywords: application protocol, network traffic, correlation dimension, chaos

PACS: 05.45.–a, 47.52.+j **DOI:** 10.1088/1674-1056/20/5/050506

1. Introduction

Most complex phenomena observed in nature arise from the nonlinear nature of the underlying dynamics. The development of nonlinear dynamics theory has brought new methodologies to analyse nonlinear behaviours in complex system; hence techniques from nonlinear dynamics are increasingly applied in diverse fields such as biology, economics and network science.[1−4] Computer networks are highly complicated systems, both in the temporal and spatial behaviour of their traffics.[5,6] Recently, it was found that the Internet traffic could exhibit complex nonlinear dynamical behaviours.[7−12] Nowadays, the application of nonlinear theory in the analysis of network traffic is a very active research field.

Chaos analysis is a powerful non-linear method which enables the extraction of characteristic quantities.[1,4,9] It offers tools which bridge the gap between experimentally observed irregular behaviour and chaos theory and have been used in many related fields such as power system and biomedical engineering as well as applications to internet traffic.[1,4,9] One of the most widely used methods for quantifying the nature of such data is the calculation of the correlation dimension D_2 by using the Grassberger–Proccacia (GP) algorithm.[13,14] The results of the analysis are often useful for indicating whether the nature of complexity in the data is chaotic, stochastic or a mixture of the two. Nowadays, the correlation dimension has been widely used as a powerful tool for interpreting irregular signals in electrical, mechanical and other engineering domains.[15,16]

Correlation dimension based non-linear analysis for practical applications; especially for internet traffic was reported in previous research work.[4] Nevertheless, to our best knowledge, characteristics analysis of network traffics with different application protocols is still an open problem. Here, we introduce a method for the analysis of different internet application protocol traffics. One of the main objectives is to explore the nonlinear behaviours of three typical traffics with different application protocols—Hypertext (HTTP), file transfer (FTP) and simple mail transfer protocol (SMTP), respectively.

In this paper, we introduce a correlation dimension approach for the dynamic analysis of different internet application protocol traffics, i.e. HTTP, FTP and SMTP. First, a mutual information method is employed to reconstruct the phase space of network traffic. Secondly, correlation dimension based chaos analysis is conducted and the analysis results have

*Project supported in part by the National High Technology Research and Development Program of China (Grant No. 2007AA01Z480).

†Corresponding author. E-mail: wjsong2004@126.com

© 2011 Chinese Physical Society and IOP Publishing Ltd

shown that HTTP traffic and FTP traffics are chaos and SMTP traffic is stochastic, which indicates that the dynamics of three typical traffics with different application protocols is different in nature.

2. Data acquisition

In our study, network traffics for analysis are acquired by the Netscope, which is a system developed by our research group and used to monitor network traffic, works at the gateway of the Local Area Network (LAN) and captures all the in and out traffics and packets, which is stored at a database for further study. The scheme of the Netscope system is shown as Fig. 1.

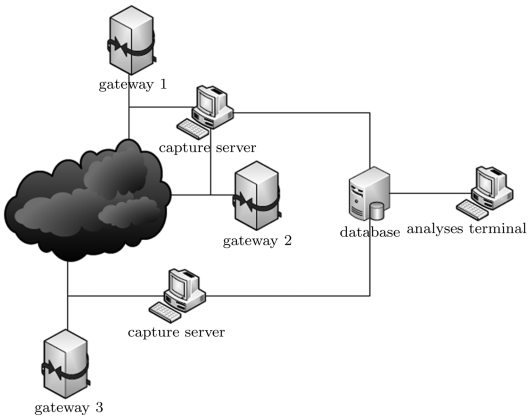
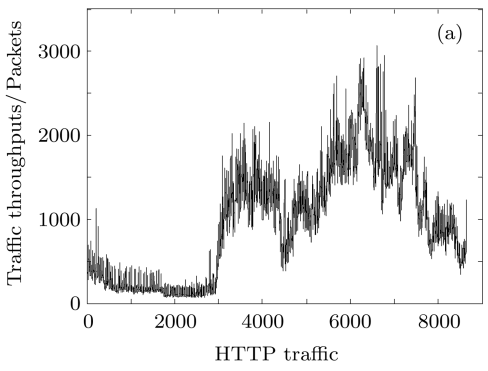


Fig. 1. The schematic diagram of data acquisition system. —Netscope.

The Netscope system is located at the main building in Tsinghua University, and captures the input traffic of the Tsinghua University LAN. The recording has been performed with a sampling time up to 10 seconds. The recorded traffic data correspond approximately to 1 day (8640 records) of measurements. Measured traffic throughputs of HTTP, FTP and SMTP are respectively shown in Fig. 2.



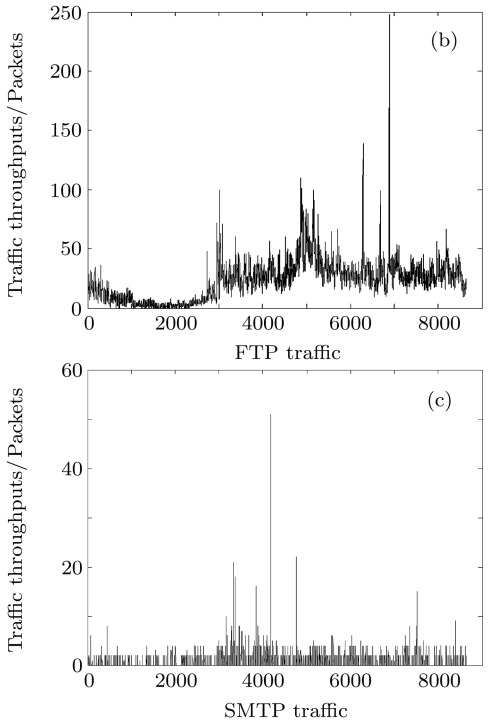


Fig. 2. Network traffics with different application protocol. (a) HTTP, (b) FTP, (c) SMTP.

3. Phase space reconstruction of network traffic based on mutual information method

The embedding procedure is the first step of the phase space reconstruction of a dynamical system from the observation of a single variable. The most common phase space reconstruction technique is the method of delays (MOD) proposed by Takens *et al.*[17,18]

Giving a time series $\{x_i, i = 1, 2, \ldots n\}$, we build a new m-dimensional phase-space in the following way:

$$\begin{aligned} \boldsymbol{Y}_1 &= [x_1, x_{1+\tau}, \ldots, x_{1+(m-1)\tau}], \\ \boldsymbol{Y}_2 &= [x_2, x_{2+\tau}, \ldots, x_{2+(m-1)\tau}], \\ &\ldots \\ \boldsymbol{Y}_N &= [x_N, x_{N+\tau}, \ldots, x_{N+(m-1)\tau}], \end{aligned} \tag{1}$$

where $N > n - (m + 1)\tau$, m is the embedding dimension of the reconstructed phase space and τ is the lag time measured in unit of sampling interval. Takens *et al.* proved that such systems converge to the real dimension if $m > 2D + 1$, where D is the fractal dimension (FD) of the real system. In our study, only embedding time delay τ is wanted to estimate for the computation of correlation dimension, so mutual information method is employed to determine the embedding time delay.[19]

Let us consider two random vectors $(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$, denoted by $p_{(\boldsymbol{Y}_i,\boldsymbol{Y}_j)}$, $p_{(\boldsymbol{Y}_i)}$ and $p_{(\boldsymbol{Y}_j)}$, the joint and marginal densities of the random vectors. The mutual information can be interpreted as the H-information obtained from the Shannon entropy. The Shannon entropy of a density p with respect to μ, is defined by

$$H(p) = -\int p\log(p)\,\mathrm{d}\mu. \tag{2}$$

In the discrete case, $H(p)$ always exists, is positive and can be interpreted as an uncertainty or an information measure.

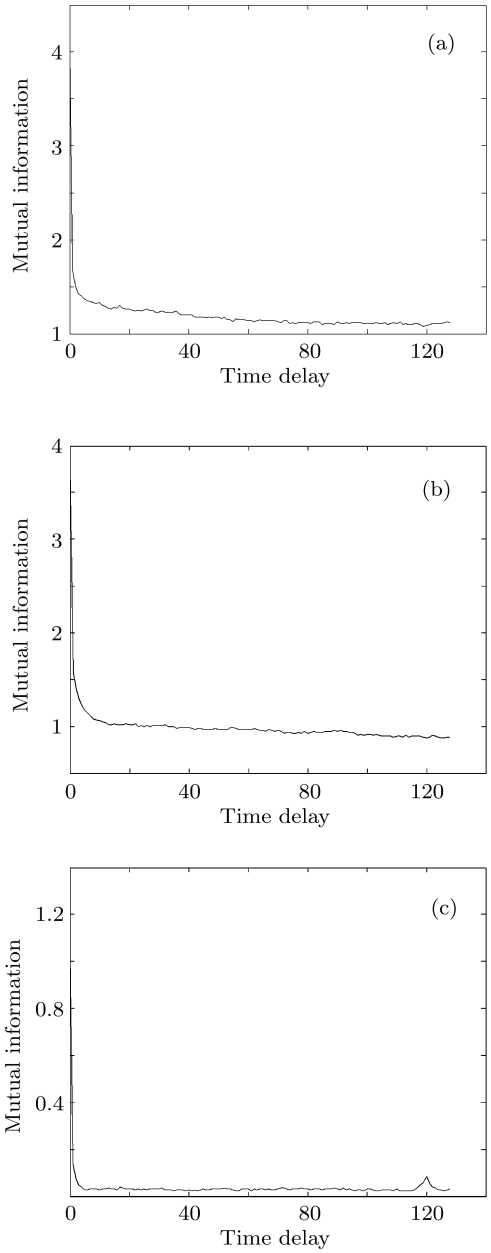


Fig. 3. The embedding time delay of network traffics: (a) the HTTP traffic (time delay $\tau = 9$), (b) the FTP traffic (time delay $\tau = 13$), (c) the SMTP traffic (time delay $\tau = 6$).

With respect to the Shannon entropy, the mutual information between $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$ can be easily rewritten as

$$\begin{aligned} I(\boldsymbol{Y}_i;\boldsymbol{Y}_j) &= H(p_{\boldsymbol{Y}_i}) - E_{p_{\boldsymbol{Y}_j}}[H(p_{\boldsymbol{Y}_i|\boldsymbol{Y}_j=\boldsymbol{y}_j})] \\ &= H(p_{\boldsymbol{Y}_j}) - E_{p_{\boldsymbol{Y}_i}}[H(p_{\boldsymbol{Y}_j|\boldsymbol{Y}_i=\boldsymbol{y}_i})]. \end{aligned} \tag{3}$$

Hence, the mutual information can be interpreted as the reduction in the uncertainty of $\boldsymbol{Y}_i$ due to the knowledge of $\boldsymbol{Y}_j$. Rewriting the expectation in Eq. (3), we obtain

$$E_{p_{\boldsymbol{Y}_j}}[H(p_{\boldsymbol{Y}_i|\boldsymbol{Y}_j=\boldsymbol{y}_j})] = H(p_{(\boldsymbol{Y}_i,\boldsymbol{Y}_j)}) - H(p_{\boldsymbol{Y}_j}). \tag{4}$$

Therefore the mutual information between $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$ is

$$I(\boldsymbol{Y}_i;\boldsymbol{Y}_j) = H(p_{\boldsymbol{Y}_i}) + H(p_{\boldsymbol{Y}_j}) - H(p_{(\boldsymbol{Y}_i,\boldsymbol{Y}_j)}). \tag{5}$$

In practice, the first local least value of the mutual information defined as Eq. (3) can be estimated as the embedding time delay for phase space reconstruction of the time series. Phase space reconstruction results of network traffics with three different application protocols are illustrated as Fig. 3, i.e. the embedding time delay for HTTP traffic is $\tau = 9$, $\tau = 13$ for FTP traffic and $\tau = 6$ for SMTP traffic.

4. Correlation dimension analysis of network traffics with different application protocols

The correlation dimension of non-linear time series is of an estimated value of the number of freedom degrees, which represents the complex degrees of the time series. To calculate the correlation dimension, a correlation integral $C_m(r)$ is defined as the probability with which a pair of points chosen randomly with respect to the natural measure is separated by a distance less than r in the reconstructed phase space of m dimensions. The correlation integral $C_m(r)$ is then approximated by[15,20,21]

$$C_m(r) = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\Theta(||\boldsymbol{Y}_i - \boldsymbol{Y}_j|| - r), \tag{6}$$

where Θ is the Heaviside function

$$\Theta(x) = \begin{cases} 0, & \text{if } x \le 0, \\ 1, & \text{if } x > 0, \end{cases} \tag{7}$$

and r is the distance parameter, $||\boldsymbol{Y}_i - \boldsymbol{Y}_j||$ is the distance between two reconstructed vectors.

When $N \to \infty$, the scaling property is derived from the inherent characteristics of the correlation integral as

$$C_m(r) \propto r^{D_2}. \tag{8}$$

So the correlation dimension D_2 of the reconstructed time series is then defined asymptotically as

$$D_2 = \lim_{r \to 0} \frac{\ln C_m(r)}{\ln r}. \tag{9}$$

The slope of $\ln C_m(r)$ versus $\ln r$ converges to D_2, the slope is generally computed by using the method of least squares fit of a straight line over a certain range of r called the scaling region.

In the case of finite data the limit cannot be formed. In practice, one usually locates a linear scaling region from the plot of $\ln C_m(r)$ vs. $\ln r$ and estimates the slope of the curve over the linear region. This slope is then considered as an estimation of the correlation dimension D_2 of the reconstructed time series. If D_2, plotted as a function of m, reaches a plateau for a range of large enough values of m, the value is then taken to be an estimation of the true correlation dimension D_2 for the system and which further shows that the analysed signal is chaos, otherwise, it is stochastic.

We outline the procedure of calculating the correlation dimension from the time series as follows. First, the correlation integral $C_m(r)$ is calculated for different values of r (radius of the sphere) for each embedding dimension 1 to m. From the logarithmic plot of these values, the scaling region to obtain the slopes can be determined. Secondly, a linear regression is computed with the points contained in the scaling region for each embedding dimension, the resulting slope is the dimension D_2.

The collected traffics were analysed by the correlation dimension (D_2) method. The relationships between $\ln C_m(r)$ and $\ln r$ and the slope of $\ln C_m(r)$ vs. $\ln r$ for HTTP, FTP and SMTP traffics is presented in Figs. 4(a), 5(a) and 6(a) respectively. So the function of D_2 and m is plotted as Figs. 4(b), 5(b) and 6(b) respectively and the estimated correlation dimension of HTTP traffic is $D_2 = 4.28$ and FTP traffic $D_2 = 1.90$ respectively, this fact indicates that the two typical traffics are chaos, furthermore, the former is more complex than the later, on the other hand, SMTP traffic is stochastic, since D_2 does not reach a plateau for a range of large enough values of m.

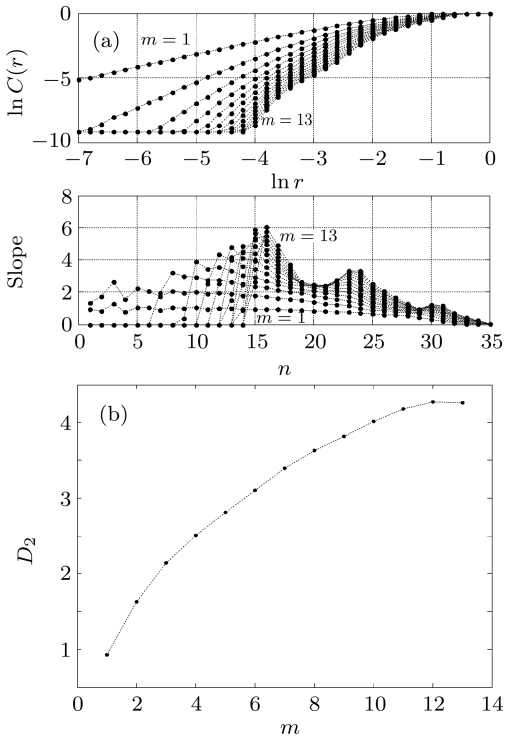


Fig. 4. Correlation dimension analysis results of HTTP traffic. (a) Correlation dimension results of HTTP traffic (embedding dimension $m = 1$ to 13 from left to right), (b) relationship between correlation dimension D_2 and embedding dimension m.

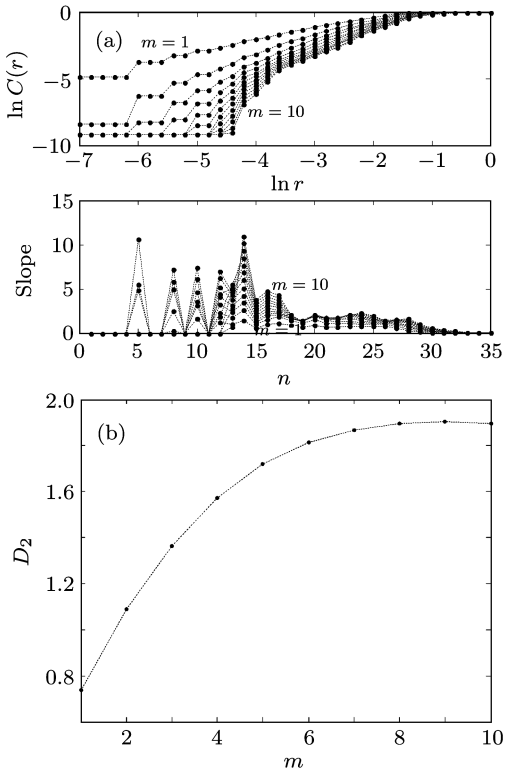


Fig. 5. Correlation dimension analysis results of FTP traffic. (a) Correlation dimension results of FTP traffic (embedding dimension $m = 1$ to 10 from left to right), (b) relationship between correlation dimension D_2 and embedding dimension m.

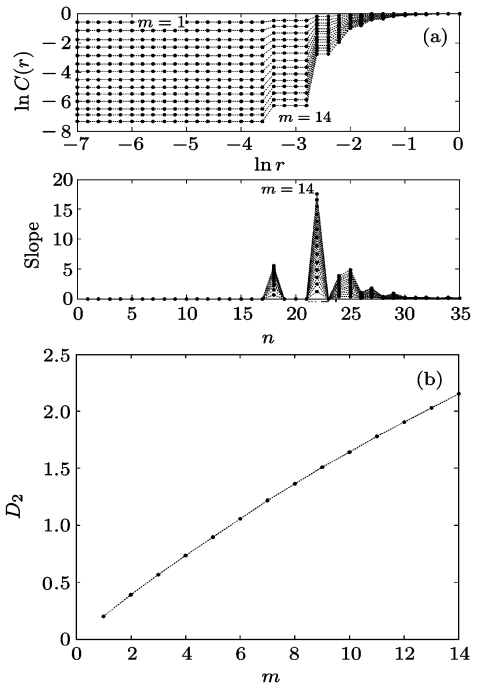


Fig. 6. Correlation dimension analysis results of SMTP traffic. (a) Correlation dimension results of SMTP traffic (embedding dimension $m = 1$ to 14 from left to right), (b) relationship between correlation dimension D_2 and embedding dimension m.

5. Conclusion

We have analysed the dynamical behaviours of three typical application protocol traffics (HTTP, FTP and SMTP) respectively by using the correlation dimensional method. The results indicate that the dynamics of the three protocol traffics is different from each other. While HTTP and FTP traffics are low dimension chaos in nature, SMTP traffic, on the other hand, is stochastic in nature. The differences may arise from the users' corresponding behaviours in the respective applications. For HTTP and FTP, the users are involved more interactively, hence creating more complexity and long tailed distributions; for SMTP, the mail servers are responsible for messages delivery, thus more uniformly distribute the loads. Our analysis provides a new insight and a more accurate understanding of the characteristics of three kinds of typical Internet traffics, which are valuable for developing efficient schedulers to achieve quality of service, and therefore to improve the capability of networks in general.

References

[1] Nitza T, Thomas J H, Charles L W and Joseph P Z 2001 *Phys. Lett.* A **279** 94

[2] Lan K C and John H 2006 *Computer Networks* **50** 46

[3] Masugi M and Takuma T 2007 *Physica* D **225** 119

[4] Feng H F, Shu Y T and Oliver W W Y 2009 *Nonlinear Analysis: Real World Application* **10** 1021

[5] Takayasua M, Takayasu H and Fukud K 2000 *Physica* A **277** 248

[6] Bigdeli N, Aeri M, Choobkar S and Jannesari F 2007 *Physica* D **233** 138

[7] Palmieri F and Fiore U 2009 *Computer Network* **53** 761

[8] Masugi M 2006 *IEEE Trans. on Circuits and Systems — I: Regular Papers* **53** 2318

[9] Ding D W, Zhu J and Luo X S 2008 *Chin. Phys.* B **17** 105

[10] Liu Y L, Zhu J and Luo X S 2009 *Chin. Phys.* B **18** 3772

[11] Li Y, Cao H D, Shan X M, Ren Y and Yuan J 2009 *Chin. Phys.* B **18** 1721

[12] Sun H L, Jin Y H, Cui Y D and Cheng S D 2009 *Chin. Phys.* B **18** 4760

[13] Grassberger P and Proccacia I 1983 *Physica* D **9** 189

[14] Grassberger P and Proccacia I 1983 *Phys. Rev. Lett.* **50** 346

[15] Harikrishnan K P, Misra R, Ambika G and Kembhavi A K 2006 *Physica* D **215** 137

[16] Alberto R N, Maria E and Montesino O 2005 *Mechanical Systems and Signal Processing* **19** 939

[17] Takens F, Rand D and Young L S 1981 *Dynamical Systems and Turbulence* (Berlin: Springer)

[18] Takens F 1981 *Lecture Notes in Math.* **898** 361

[19] Fraser M and Swinney H L 1986 *Phys. Rev.* A **33** 1134

[20] Peng H W, Li R Q, Chen S Z and Li C B 2008 *Chin. Phys.* B **17** 637

[21] Albona M, Muench J and Schwartz C 1988 *Phys. Rev.* A **38** 3107