International Journal of Grid and Distributed Computing
Vol. 9, No. 5 (2016), pp.303-312
http://dx.doi.org/10.14257/ijgdc.2016.9.5.26

Analyzing Spatiotemporal Characteristics of Education Network Traffic with Flexible Multiscale Entropy

Chen Yang, Renjie Zhou*, Jian Wan, Jilin Zhang and Yuyu Yin

School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China.
Key Laboratory of Complex Systems Modeling and Simulation, Ministry of Education, Hangzhou, China.
**rjzhou@ hdu.edu.cn*

Abstract

In this paper, we propose an analysis method of spatial and temporal characteristics of education network traffic based on flexible multiscale entropy (FMSE). As an improved method, flexible multiscale entropy has a significant improvement in stability and accuracy over multiscale entropy (MSE). We analyze network traffic in different time scales, space scales and traffic levels by building network traffic time-space analysis model and using flexible multiscale entropy as a method to quantify complexity of different network traffic subsequences. The results show that there are distinct characteristics in the complexity of different levels of network traffic. We also find that the existence of a large number of small network traffic flows has a significant influence on the complexity of network traffic.

Keywords*: education network, network traffic, flow, multiscale entropy, flexible multiscale entropy, spatiotemporal scale transformation*

1. Introduction

With the continuous development of the Internet, it has become an indispensable part of people's life and modern education. How to ensure safe, stable and efficient operation of education network has become an urgent problem needs to be solved. As the carrier of the data flow in network, the spatiotemporal characteristic of network traffic can help to understand the complex network structure and the dynamic characteristics of network, and it has important significance for designing, implementing and monitoring network. The self-similarity and long range dependence [1, 2], have been a hot research topic since they were discovered. As an important mathematical tool, the wavelet transform has been widely applied in the research of network traffic [3].

With the continuous development of machine learning, many machine learning methods have been applied into the analysis and research of network traffic [4-6]. However, there are few researches on the complexity of network traffic. As the network heterogeneity brings the difference of network behavior. Although the IP system in the user level unify all kinds of heterogeneous network and domain management, but the IP layer only shield the heterogeneity of network, and the difference in the network is still exist. According to the research of network complexity, we can reveal the intrinsic characteristics of network traffic, which can provide a strong theoretical basis for network traffic prediction and anomaly detection.

As a measure of system complexity, entropy often implies the level of whole system complexity. Entropy theory can help us to discover the inherent dynamics of network

* Renjie Zhou and Chen Yang are co-first authors who contribute to this work equally. Renjie Zhou is the corresponding author.

ISSN: 1738-9976 IJXX

traffic [7]. However, traditional entropy theory can only describe time series from a single scale, which cannot reveal the complexity of time series completely. Multiscale entropy combined entropy theory and multi scale idea for the first time, analyzing time series from different scales to discovery the correlation of time series in different scales. Multiscale entropy has been widely applied in different fields, e.g., Janne Riihijärvi use multiscale to study characteristics of networks and wireless communication [9]. Multiscale entropy has obvious advantages compared with information entropy and self-similar parameter H, and multiscale entropy also has a good effect on the prediction of network traffic. In this paper, we use flexible multiscale entropy as a characterization tool for time series complexity. As an improved algorithm of multiscale entropy, flexible multiscale entropy has a great improvement on the stability and accuracy of computation, and has less dependence on the length of time series.

The rest of the paper is organized as follows. In Section 2, we introduce flexible multiscale entropy and the time-space analysis model in detail. In Section 3, we apply flexible multiscale entropy and the analysis model to real network traffic data. In Section 4, we provide the result of traffic statistics and analysis. Finally, we conclude the paper in Section 5.

2. Methodology

2.1. Flexible Multiscale Entropy

Sample entropy (SampEn) is a measurement of time series complexity proposed by Richman [10], which can be used to quantify the complexity of time series. The calculation of time series complexity can provide a theoretical basis for the prediction and detection of time series. Costa found that sample entropy method applied to health and disease research often output contradictory results, and then proposed the multiscale entropy. And experiments were carried out with white noise and 1/f noise, the experimental results show that the multiscale entropy can accurately reflect the complexity of time series. However, the stability and accuracy of the multiscale entropy decreased significantly when the scale factor becomes larger. We proposed flexible multiscale entropy, which introduces the flexible coefficient f into computation，and the method of flexible accumulation is used in the calculation of sample entropy at the specific scale. By setting the flexible coefficient f to control the matching degree of template, the discrimination of different time series is improved. At the same time, the stability and accuracy of computation improved greatly compare to multiscale entropy. The calculation steps of flexible multiscale entropy are as follows:

Step1: A “coarse-graining” process is applied to a given time series $\{x(i)|1\le i\le N\}$. This process is based on composite multiscale entropy proposed in paper [11]. The transformed sequence is as follow:

$$\mathbf{y}_k^{(\tau)}=\{y_{k,1}^{(\tau)},y_{k,2}^{(\tau)},...,y_{k,P}^{(\tau)}\},P=\left\lfloor\frac{N-k+1}{\tau}\right\rfloor,1\le k\le\tau \tag{1}$$

For every scale factor τ , it will build τ new time series. The specific transformation formula is as follow:

$$y_{k,j}^{(\tau)}=\frac{1}{\tau}\sum_{i=(j-1)\tau+k}^{j\tau+k-1}x_i,1\le j\le P,1\le k\le\tau \tag{2}$$

Step2: For a given pattern length m, composing each new time series to m dimensional consecutive vector sequences, the result is as follow:

$$\mathbf{Y}_{k,m}^{(\tau)}=[y_{k,i}^{(\tau)},y_{k,i+1}^{(\tau)},...,y_{k,i+m-1}^{(\tau)}],1\le i\le P-m+1,1\le k\le\tau \tag{3}$$

Step3: We define vector distance $d[\mathbf{Y}_{k,m}^{(\tau)}(i), \mathbf{Y}_{k,m}^{(\tau)}(j)]$ as the maximum value of the difference between corresponding elements in the two vectors.

$$d[\mathbf{Y}_{k,m}^{(\tau)}(i), \mathbf{Y}_{k,m}^{(\tau)}(j)] = \max_{0 \le q \le m-1} | y_{k,i+q}^{(\tau)} - y_{k,j+q}^{(\tau)} | \tag{4}$$

Step4: For a given similarity criterion *r*, the ratio of "similar" vectors of which distance is smaller than *r*, denoted as $A_i^m(r)$.

$$A_i^m(r) = \frac{1}{P-m-1} num\{d[\mathbf{Y}_{k,m}^{(\tau)}(i), \mathbf{Y}_{k,m}^{(\tau)}(j)] < r\}, 1 \le i \le P-m, i \ne j \tag{5}$$

where *num* accumulates the number of similar vectors.

Step5: Define $A^m(r)$ as the average of $A_i^m(r)$.

$$A^m(r) = \frac{1}{P-m} \sum_{i=1}^{P-m} A_i^m(r) \tag{6}$$

Step6: In this step, calculate $A^{m+1}(f)$. Different from the accumulative function of sample entropy, we define a new accumulative function *s* here. *s* is a piecewise function that avoids the measurement of similar vectors changing suddenly between 0 and 1.

$$s = \begin{cases} 0 & d[\mathbf{Y}_{k,m+1}^{(\tau)}(i), \mathbf{Y}_{k,m+1}^{(\tau)}(j)] \ge f \\ 1 - \dfrac{d[\mathbf{Y}_{k,m+1}^{(\tau)}(i), \mathbf{Y}_{k,m+1}^{(\tau)}(j)]}{f} & d[\mathbf{Y}_{k,m+1}^{(\tau)}(i), \mathbf{Y}_{k,m+1}^{(\tau)}(j)] < f \end{cases} \tag{7}$$

where *f* is flexible similarity criterion.

The ratio of "similar" vectors for pattern length $m+1$ is as follow:

$$A_i^{m+1}(f) = \frac{1}{P-m-1} \sum_{j=1}^{P-m-1} s \tag{8}$$

Step7: Calculate $A^{m+1}(f)$ as follow:

$$A^{m+1}(f) = \frac{1}{P-m} \sum_{i=1}^{P-m} A_i^{m+1}(f) \tag{9}$$

Step8: Calculate sample entropy for coarse-grained time series $\mathbf{y}_k^{(\tau)}$

$$SampEn(\mathbf{y}_k^{(\tau)}, m, r, f) = -\ln[\frac{A^{m+1}(f)}{A^m(r)}] \tag{10}$$

Step9: Calculate flexible multiscale entropy.

$$FMSE(\mathbf{x}, \tau, m, r, f) = \frac{1}{\tau} \sum_{k=1}^{\tau} SampEn(\mathbf{y}_k^{(\tau)}, m, r, f) \tag{11}$$

Above is the calculation process of flexible multiscale entropy. In the following calculations, we appoint m=2, r=0.15， which means similarity criterion is 0.15**SD*. *SD* is standard deviation of time series $\{x(i) | 1 \le i \le N\}$. We appoint f=0.2, which means flexible similarity criterion is 0.2**SD*.

2.2. Network Traffic Analysis Model

Traditional methods usually consider only one single space-time scale, which are incapable of discovering some inherent characteristics of network traffic. In this paper, we introduce a new method to analyze the spatiotemporal characteristics of network traffic based on FMSE. First, we build the time-space analysis model as shown in Figure 1. Then, we select different space scales, time scales and network levels, cut the original traffic data sequence into different sub sequences and

calculate the complexity of each sequence with flexible mutiscale entropy. Finally, summarize and compare the results of each sub sequence, discovering the evolvement mechanism and the inherent characteristics of network traffic at different time scales, space scales and network levels. We will conduct analysis from two perspectives. One is macro level, and the other is pc level. We can first find a general law of network traffic from macro level, and then zoom in to pc level to test the validity of the law. Since different individuals have different behaviors, some particularities may exist in pc level traffic. We will do a further analysis to discover the causes of particularities and find a general law that is suitable for all dimensions of network traffic. Through analyzing network traffic from two perspectives, we can improve the accuracy and practicability of the law we discovered.

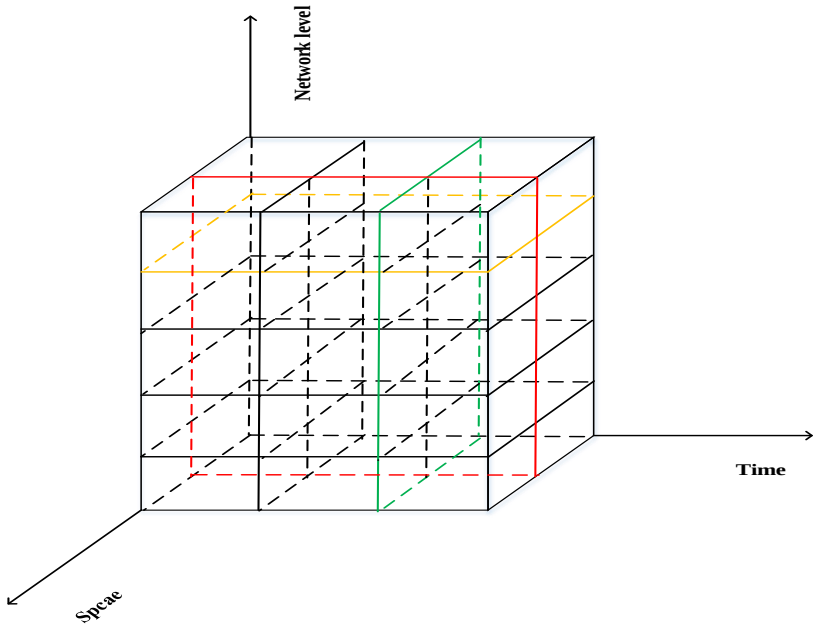


Figure 1. Time and Space Analysis Model of Network Traffic

3. Experiment Results

3.1. Experiment Setup

The data set used in our study was collected from the east China's backbone of the CERNET (China Education and Research Network) and was released by Jiangsu Key Laboratory of Computer Networking Technology on November 9, 2014. The backbone covers over 100 universities and high schools. Its bandwidth was updated to 10G from 2.5G in 2006 [12]. The duration of the data set is 1440 minutes and the size of the data set is 2240.06GB. The analysis method introduced in section 2 is implemented in Java language.

3.2. Macro Level Network Traffic Analysis

At first, analyze network traffic from macro level. We classify the original traffic data to different data sequences based on the network level, and then calculate the complexity of each sequence by the flexible multiscale entropy proposed in section 2. In the

calculation process, the pattern length m is 2, the similarity criterion is 0.15 and the flexible similarity is 0.2. The maximum scale factor is often set to 20 in the calculation. However, considering that the length of traffic data contains 86400 points (seconds), we set the maximum scale factor to 180 and the scale interval to 5. Since the length of time series is long enough, we can select different scale factors to analyze, so that it is easier to discover the inherent law of the time series.

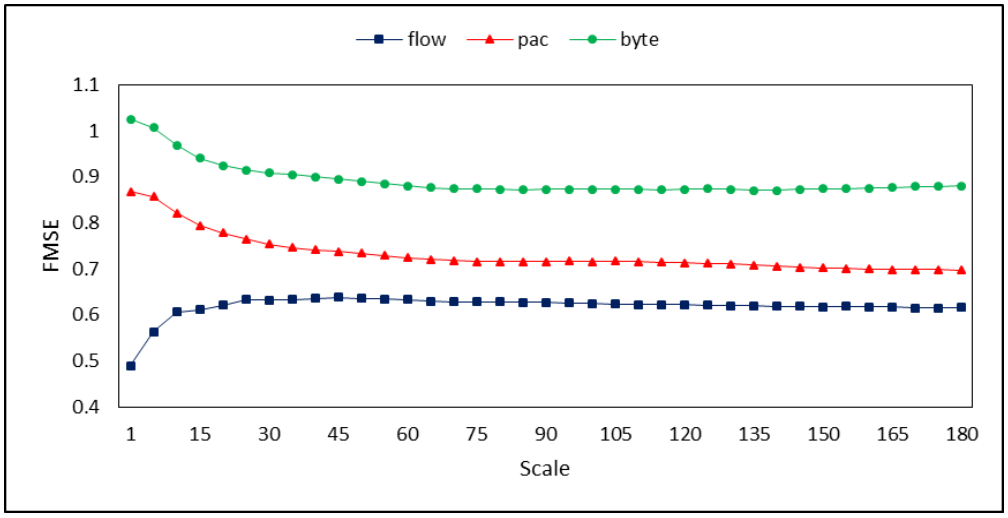


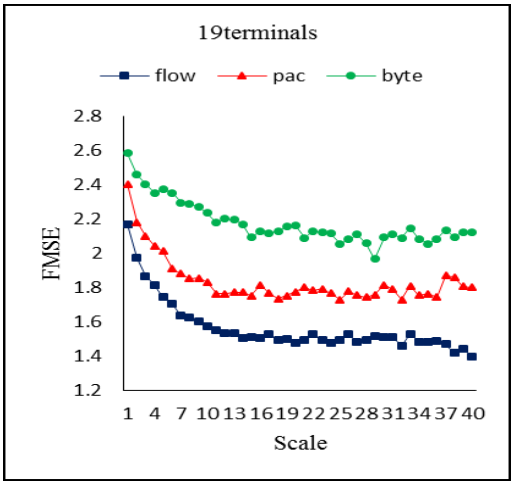
Figure 2. Entropy Values Curve Based on Network Level

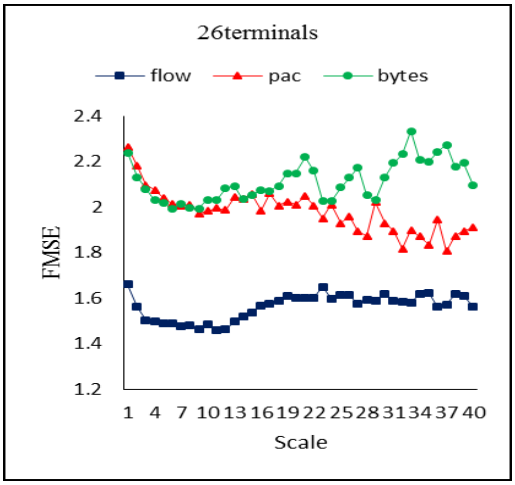
Figure 2 is the flexible multiscale entropy values of byte, packet, and flow series. In the figure, flow stands for traffic flow, which is based on the well-known five-tuple - the source IP address, destination IP address, source port, destination port and protocol fields. Pac stands for packet. The horizontal axis is the time scale in seconds; the vertical axis is the value of flexible multiscale entropy. The original data sequence is coarse-grained at scale factors from 1 to 180 and calculated with flexible multiscale entropy. From the figure, we can see the entropy value of traffic flow rises slowly as the scale factor increases, whereas the entropy value of packet and byte decreases slowly with the scale factor. Three entropy curves are relatively stable when the scale factor is more than 45. From whole scale, the entropy value is in a relatively stable range, which indicates that network traffic exhibits long-range dependence characteristic. The larger the scale factor is, the more information of the original data sequence is contained in the data sequence. Throughout all the scales, byte entropy is higher than packet and flow entropy, and the flow entropy is the lowest of the three. Furthermore, the relationship of complexity trend of the three is relatively stable. The relative relationship among byte, packet and traffic flow can be treated as an internal law of network traffic. And this law can be applied to analyze the behavior of network traffic, monitor and predict network traffic. The law will be further studied in the PC level traffic analysis.

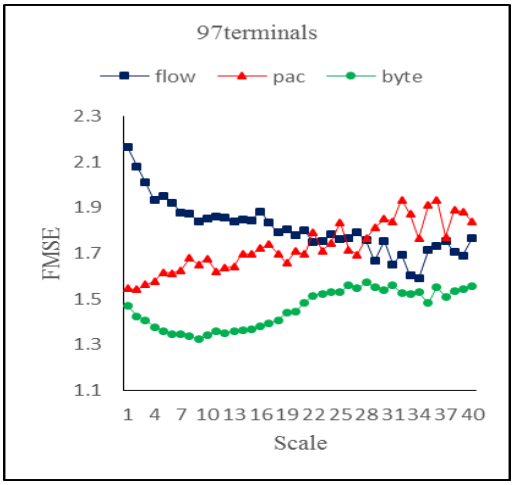
3.3. PC level Network Traffic Analysis

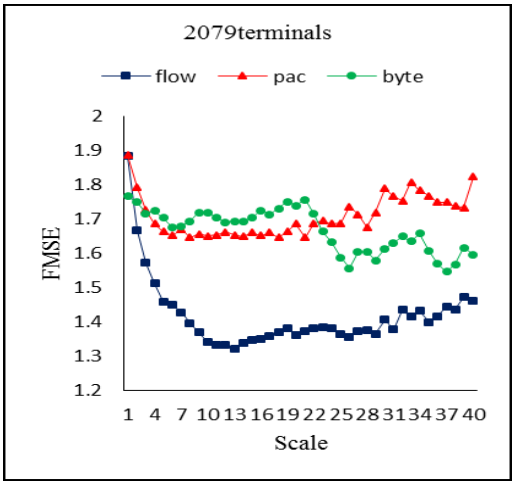
Based on the internal law of network traffic observed at macro level, we will investigate the terminal-level characteristics of network traffic in this section. In order to discover the law of common characteristics in network traffic, we analyze the traffic generated by different sets of terminals. In the analysis, we use one hour's traffic data with a size of 78GB. We select eight network segments from class A, B, C of the traffic data to analyze. The eight network segments contain 19, 26, 97, 2079, 44441, 50798, 61732 and 110349 terminals, respectively. These eight sets of data have a good continuity in an hour of time series, and have an extensive coverage ranging from 10 to 100

thousand terminals, and thus the results of the analysis are universal. By using analysis method in section 3.2, the traffic data of the eight groups were classified according to the network level, and calculate the flexible multiscale entropy value for each data sequence. By comparing the complexity of network traffic at different scales, finding the evolution mechanism of network traffic at different temporal and spatial scales.









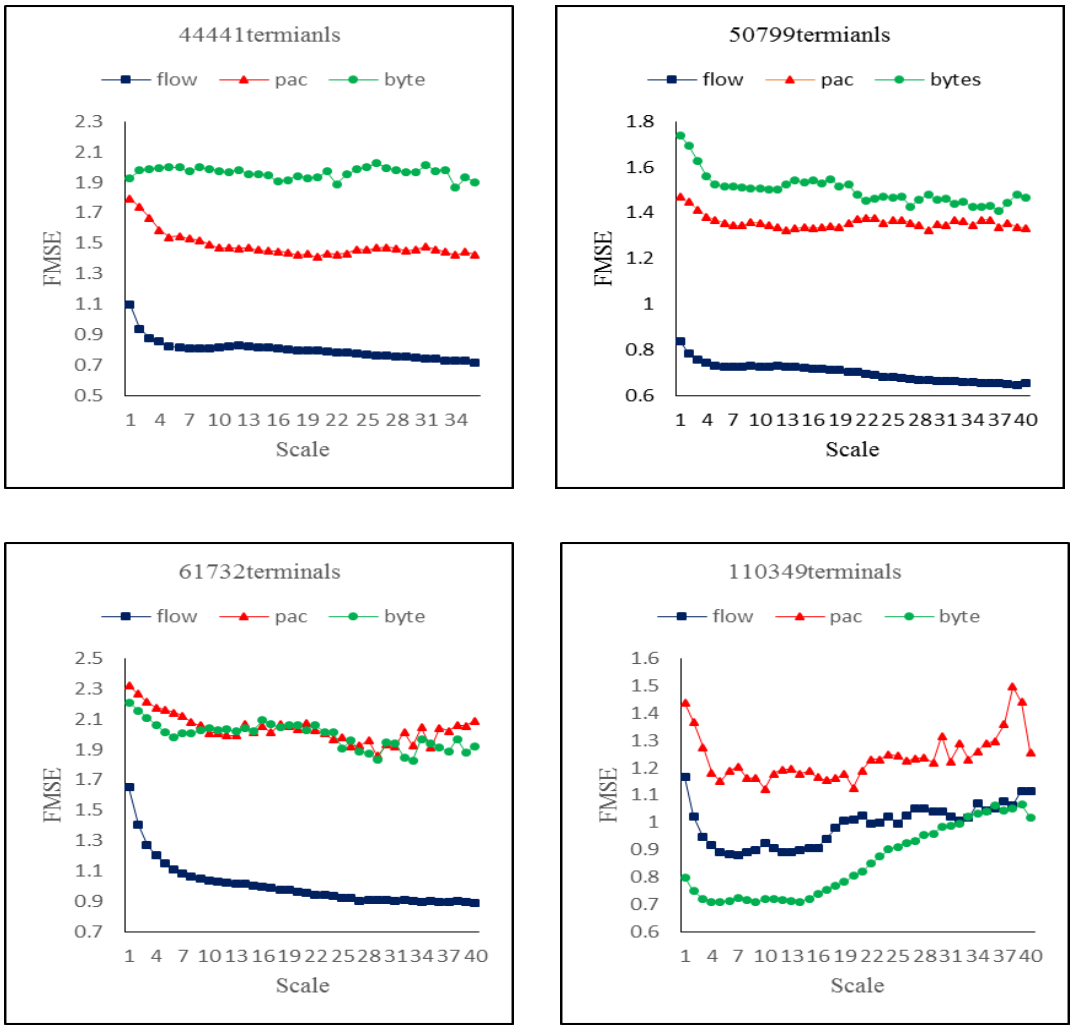


Figure 3. Entropy Value under Different Scales of Network

Figure 3 shows the entropy values based on network level from eight groups of traffic data at different scales. In the figure, flow stands for traffic flow, which is based on the well-known five-tuple - the source IP address, destination IP address, source port, destination port and protocol fields. Pac stands for packet. As can be seen from Figure 3, for different size of eight groups of traffic data, the entropy value curve of flow is always lower than the packet entropy curve throughout all the scales. This shows that flow sequence complexity is lower than the packet sequence complexity, and the relative relationship between the packet and flow does not change as traffic scale changes. At the same time, last section's experiment also shows that flow entropy is lower than packet entropy throughout all the scales. Whereas, in the eight figures, the complexity relationship between byte and packet is not obvious in the whole scale, it changes in different scales. Then, we focus on the complexity relationship between flow and byte sequence. From the figure, we can see the complexity of the flow is higher than the complexity of the bytes, except for 97 and 110349 terminal's entropy curves. And for the other six groups of data, the complexity relationship between flow and byte is relatively stable. Over the whole scale, the complexity of byte is higher than flow. In the rest of the paper, we focus on the particularity of the entropy curves of 97 and 11349 terminals. The above statistics is based on the number of bytes per second and the number of flow per second, however, the number of bytes include in each flow has a great difference. From the traffic data, we calculate the average bytes of a flow per second. By comparing the

complexity between flow sequence and the sequence of average bytes that one flow contained, discovering the common complexity relationship between flow and byte. We compare the complexity of two kinds of sequence measured by flexible multiscale entropy. As we compare the complexity of two sequence directly, we only calculate the flexible multiscale entropy at scale factor 1.

Table 1. Entropy Value at Different Scale of Network

Terminals	Flow-Byte Entropy	Flow Entropy	Entropy Difference
19	2.656	2.168	0.488
26	1.701	1.66	0.041
97	**1.29**	**2.167**	**-0.877**
2079	1.895	1.884	0.011
44441	1.85	1.098	0.752
50798	0.871	0.839	0.032
61732	2.203	1.65	0.553
110349	**0.858**	**1.166**	**-0.308**

In Table 1, Flow-Byte Entropy stands for the flexible multiscale entropy value of average bytes for one flow per second and Entropy Difference stands for the difference between the Flow-Byte Entropy and Flow Entropy. From Table 1, we can see that the entropy difference in the scale of 97 terminals and 110349 terminals is smaller than 0. Due to the complexity of flow-byte is lower than the complexity of flow, the complexity of flow is hence higher than the complexity of byte. This phenomenon may be due to many small size flows or short duration flows in the network traffic. In order to verify the results, we investigate the number of average bytes including in one flow per second.

Table 2. Relationship between Byte and Flow at Different Scale of Network

Terminals	Flow-Byte
110349	**2237.079949**
97	**5153.755829**
61732	6731.311003
19	7254.642316
26	9820.898529
2079	10005.38842
50798	10867.96524
44441	12068.47809
61	14648.05307

In Table 2, Flow-Byte stands for the number of bytes including in one flow per second. As show in the table 2, the bold line is the range of 110349 terminals and 97 terminals. We can see the values of flow-byte of these two scales are the smallest among all the groups. The traffic of these two groups includes a larger number of small flows than the others. The increase of small flows leads to the increased complexity of flow. The complexity between flow and byte is changed as the relationship between flow and byte changes. From above results, we can see complexity of flow is higher than the complexity of byte in most cases.

4. Conclusion and Future Work

In this paper, we adopted flexible multiscale entropy (FMSE). As an improved method of multiscale entropy, flexible multiscale entropy has an improved stability and accuracy of calculation, especially in large scale. We also proposed a method to analyze network traffic based on flexible multiscale entropy, which is used as a tool to describe the complexity of traffic sequence. By establishing the spatial and temporal characteristic network analysis model, describing complexity of network traffic clearly from different dimensions. Through the experiment carried on one day's network traffic, of which the size is more than 2TB. The result shows that there is a universal law among the different levels of network traffic. For all scales, the complexity of packet sequence is higher than that of flow sequence. In most cases, the complexity of byte sequence is higher than that of flow sequence. However, the complexity of the flow sequence will be higher than that of byte sequence when there are many small flows in network traffic.

In our future work, we will go further to analyze network traffic and build a network situational awareness model. We will use this model to analyze network behavior, detect anomaly among network and predict network traffic. Analyzing in real time is also an important part of our future work.

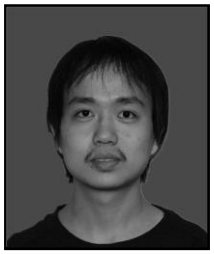
Acknowledgment

The authors are grateful to the anonymous reviewers for their valuable comments and to the editors for their work that improved this paper. This work was supported by NSF of Zhejiang under grant NO.LQ13F020017, LY14F020044, LY16F020018, and NSF of China under grant NO.61300211, 61572163, 61472112, and National Key Technology Research and Development Program of China under grant No.2014BAK14B04.

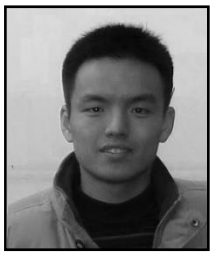
References

[1] W. E. Leland, M. S. Taqqu and W. Willinger, "On the self-similar nature of Ethernet traffic (extended version)", IEEE/ACM Transactions on Networking, vol. 2, no. 1, **(1994)**, pp. 1-15.
[2] A. Y. Privalov and A. T. Analysis, "Simulation of WAN Traffic by Self-Similar Traffic Model with OMNET", Proceedings of 10th International Wireless Communications and Mobile Computing Conference, Nicosia, Cyprus, **(2014)**.
[3] R. H. Riedi, M. S. Crouse and V. J. Ribeiro, "A multifractal wavelet model with application to network traffic", IEEE Transactions on Information Theory, vol. 45, no. 3, **(1998)**, pp. 992-1018.
[4] Y. Chen, B .Yang and Q. Meng, "Small-time scale network traffic prediction based on flexible neural tree", Applied Soft Computing, vol. 12, no. 1, **(2012)**, pp. 274–279.
[5] X. Liu, X. Fang and Z. Qin, "A Short-Term Forecasting Algorithm for Network Traffic Based on Chaos Theory and SVM", Journal of Network & Systems Management, vol. 19, no. 19, **(2011)**, pp. 427-447.
[6] Q. F. Yao, C. F. Li, H. L. Ma and S. Zhang, "Novel network traffic forecasting algorithm based on grey model and Markov chain", Journal of Zhejiang University, vol. 34, no. 4, **(2007)**, pp. 396-400.
[7] R. Xiang, J. Zhang and X. K. Xu, "Multiscale characterization of recurrence-based phase space networks constructed from time series", Chaos, vol. 22, no. 1, **(2012)**, pp. 127-131.
[8] M. Costa, "Multiscale entropy analysis of complex physiologic time series", Physical Review Letters, vol. 89, no. 6, **(2002)**, pp. 705-708.
[9] J. Riihijärvi, M. Wellens and P. Mahonen, "Measuring Complexity and Predictability in Networks with Multiscale Entropy Analysis", Proceedings of IEEE INFOCOM 2009, Rio de Janeiro, Brazil, **(2009)**.
[10] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy", American Journal of Physiology Heart & Circulatory Physiology, vol. 278, no. 6, **(2000)**, pp. H2039-H2049.
[11] S. D. Wu, C. W. Wu and S. G. Lin, "Time series analysis using composite multiscale entropy", Entropy, vol. 13, no. 3, **(2013)**, pp. 1069-1084.
[12] IPTRACE: http://iptas.edu.cn/src/system.php.

Authors



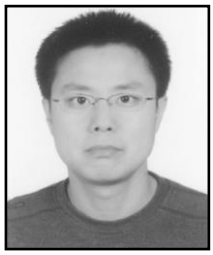
Chen Yang, he received the B.S degree in network engineering from North University Of China, Taiyuan, China, in 2011 and M.S degree from Hangzhou Dianzi University, Hangzhou, China, in 2016. His research interests include network security and complex system.



Renjie Zhou, he is an assistant professor in School of Computer Science and Technology at Hangzhou Dianzi University, Hangzhou, China. He received his Ph.D. degree from Harbin Engineering University, Harbin, China, in 2012. He was a visiting scholar in the Department of Electrical and Computer Engineering at the University of Massachusetts at Amherst. Currently. His research interests include analysis of online social networks, and network security.

JianWan, he is the Director of Grid and Service Computing Lab in Hangzhou Dianzi University, Hangzhou, China, and is the Dean of the School of Computer Science and Technology, Hangzhou Dianzi University. He received his Ph.D. degree from Zhejiang University, Hangzhou, China, in 1996. His research areas include parallel and distributed computing systems, virtualization, and grid computing. He is a member of the IEEE.

Jilin Zhang, he received the PhD degree in Computer Applied Technology from University of Science Technology Beijing, Beijing, China, in 2009. He is currently an associate professor at Hangzhou Dianzi University, China. His research interests include High Performance Computing and Cloud Computing.



Yuyu Yin, he received the Ph.D. degree in computer science from Zhejiang University, Zhejiang, China, in 2010. He is currently an Associate Professor with Hangzhou Dianzi University, Hangzhou, China. His research interests include service computing, cloud computing, and middleware techniques.