Measurement and Statistics of Application Business in Complex Internet

Lei Wang[1], Yang Li[2], Yipeng Li[3], Shuhang Wu[3], Shiji Song[1], and Yong Ren[3]

[1] Tsinghua University, Department of Automation,
100084 Beijing, China
leiwang03@mails.tsinghua.edu.cn
[2] Beijing Institute of Petrochemical Technology, School of Information Technology,
102617 Beijing, China
[3] Tsinghua University, Department of Electronic Engineerging
100084 Beijing, China

Abstract. Owing to independent topologies and autonomic routing mechanism, the logical networks formed by Internet application business behavior cause the significant influence on the physical networks. In this paper, the backbone traffic of TUNET (Tsinghua University Networks) is measured, further more, the two most important application business: HTTP and P2P are analyzed at IP-packet level. It is shown that uplink HTTP and P2P packets behavior presents spatio-temporal power-law characteristics with exponents 1.25 and 1.53 respectively. Downlink HTTP packets behavior also presents power-law characteristics, but has more little exponents γ = 0.82 which differs from traditional complex networks research result. Moreover, downlink P2P packets distribution presents an approximate power-law which means that flow equilibrium profits little from distributed peer-to peer mechanism actually.

Keywords: Internet, traffic, application business, peer-to-peer.

1 Introduction

Unlike the early description of the network's topology based on the random graph theory, the rapidly developing theory and methods of complex network have already been widely used in the research of Internet [1-4]. In 1999, the Faloutsos discovered the four kinds of power-law distribution characteristic involved in the topology of Internet AS(autonomous system) [5,6], which led to researches on the large number of topology models and growth mechanisms of Internet that have characteristics of power-law distribution, Inet, BRITE[7,8], etc. represented. The data resource of this kind of researches mainly comes from the measuring results of Internet topology based on Trace-route, provided by research institutions CAIDA and so on, such as Skitter, Oregon, etc[9]. On the other hand, after A.L.Barabási discovering the scale-free characteristic of complex network, a large number of information networks' topological characteristics based on the Internet application have been experimentally researched. A.L.Barabási, as a typical representative, used Robert to follow and catch

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 400 – 410, 2009.

the WWW (World-Wide Web) page links, and proved that the WWW which consists of pages and hypertext links have the characteristic of power-law distribution [10].

Aiming at measurement and improvement of Internet performance, the work above has greatly advanced people's comprehension of the Internet and the topological distribution of business in the Internet application layer. However, the default identity of network nodes in the statistical model leads to the research on application layer separated from that on the network layer. In the early work, we pointed out that the users` behaviors are a kind of logical application behavior which depended on the transmission capacity supplied by the Internet infrastructure, and meanwhile had independent topology and messaging rules [11,12]. The users` behaviors couple the users` nodes with the network nodes, which represent that the users` nodes launch business according to the logic of application layer, and after reaching the network nodes the business drives many changes of the network layer, such as the acceptance, routing, traffic, etc. This network structures which have coupled characteristic are already not able to be exactly described by the simple Internet power-law or WWW power-law.

With the number of internet users growing exponentially, the influence to the distribution of network traffic bring by the distribution of users` behaviors is increasing and can not be neglected any more. The latest statistics from ISC(Internet Society of China) show that the site of number one network traffic in China attracts 31% of the internet users to click it, and have the average click-through rate of 6.8/(day person). The large-scale gathering characteristics of the users` behaviors lead to the serious imbalance in network traffic, so that the load-balancing technology such as CDN (Content Delivery Network) has been proposed. We have already found that the virtual network behaviors had a malignant influence on the whole characteristics of internet, although people have gradually understood the power-law in the internet and the topological distribution of business in the Internet application layer, the distribution characteristics of users` behaviors which joined them have not been paid enough attention to. The scientific measurements and quantitative descriptions of the distribution characteristics of users` behaviors have not been seen in literature.

In order to understand the distribution characteristics of the behaviors of Internet application layer, this paper observes and analyzes the distribution characteristics of behaviors of Tsinghua University campus network, which is a local group of users, raises a topological constituting mechanism which uses bipartite graph model to couple users' nodes with network nodes, and observes the Traces traffic of the total export of Tsinghua University campus network. Through filtering and merging we select certain kinds of typical business which take a large proportion (such as HTTP, P2P), and simplify the users` behaviors to the packet level to form a conclusion of quantitative analysis. The research indicates that for a kind of directed logical network constituted by users` behaviors, The uplink and downlink HTTP packets and uplink P2P packets show a consistent characteristic of power-law distribution in both time and space. The difference from conclusion of the traditional complex network research is that the power index γ of the distribution of business behaviors are both less than 2. Downlink packets of the P2P business satisfy a kind of distribution similar to the power-law. In this paper we calculate and discuss the status when the γ changes, and point out the physical meaning involved.

2 Bipartite Graph Model

In the face of Internet, such a complex network system, we only focus on the whole measure of the users` behavior, so it can be abstracted to a structure with three network layers as shown in Figure 1. The core network takes charge of the main data transmission and exchange and the edge network is adjacent to the core network, which provides business and services. The access network is used to describe the physical topology of users` access nodes. The network business process is abstracted to: users produce business by the connecting network, and visit the edge network; The edge network answer the business request and drive the traffic distribution mainly of the core network, and also of other levels of networks.

In order to measure the distribution of users` behaviors, i.e. the interaction between the edge network and the connecting network, this paper bring the "bipartite graph" in graph theory to further describe the logical connection between the edge network and the connecting network. The logical connection is different from the physical connection in Figure 1, see in [12].

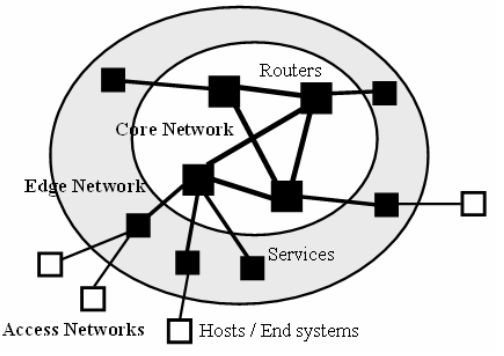

Fig. 1. Network architecture

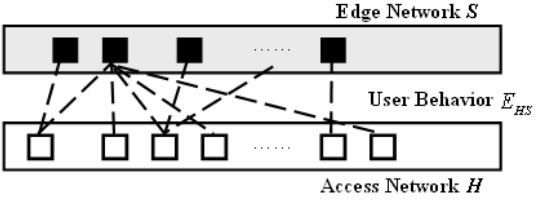

Fig. 2. Bipartite graph model of users' behavior

G is called bipartite graph, if there is a partition of the vertexes V of the undirected graph $G = <V,E>$, $V = V_1 \cup V_2$, $V_1 \cap V_2 = \Phi$, so that any for edge of G, its two ends are in V_1 and V_2 respectively.

Let S (services) to represent the set of nodes of the edge network, H (hosts) to represent the set of nodes of the connecting network, the users` behaviors in the

specified time period constitute the one-way set of logical connections E_{HS} between S and H. It is known through the network structure that $S \cap H = \Phi$, it does not matter to let $V = S \cap H$, so the one-way graph $G_{HS}=<V,E_{HS}>$ is a bipartite graph which describes the characteristic of behavior that the connecting network visits the edge network, the topology is shown in Figure 2.

By measuring uplink traffic E_{HS}, we can get the in-degree distribution of S, which can be seen as the distribution of the users behaviors to the network nodes. By the same token, let E_{SH} to represent the behaviors of answers of network to the visit of users, which constitutes the bipartite graph of downlink traffic. The out-degree distribution of S is whole reflected of the feedback behaviors of the application layer.

3 Application Business Packets Processing

As one of the largest campus networks in China, the Tsinghua University campus network owns a ten thousand million bandwidth backbone. To ensure the high enough sampling rate, the original velocity of flow got in the total export is about 250MBps, even if we analyze traffic of 1 minute, the quantity of date is about 15GB, so it will be difficult to measure and process the data in a long period. In fact, each IP datagram header contains a wealth of information, the filed of source address and destination address in the protocol tree can completely define the visiting behaviors of a group. This paper catches all packet headers of group in the experiment, and completes the data statistics of EHS by analyzing the source and destination address, which greatly reduces the data quantity.

For the two typical business which take the largest proportion of network traffic: HTTP and P2P (peer-to-peer), the original packet header files can not used directly for the statistics of users` behaviors, since there are a large number of different protocols, different sorts of business and useless “noise” packet headers, which should be filtered a second time. The experiment uses tools such as tcpdump and Ethereal to do multi-stage filter to the original traffic, and finally obtains the pure set of packet headers of HTTP and P2P.

Take the uplink bipartite graph as example, a visit behavior from connecting network H to the edge network S, is sliced up to several groups in the according IP layer. By classification and integration the set of packet headers obtained above, we can get the set of visit records from H to S, which is shown as E_{HS} in Figure 2. This paper decomposes users` behaviors to the scale of data packet so as to make a more basic analysis.

4 Performance Evaluation

In the experiment we used the uplink and downlink traces traffic in the Tsinghua University campus network. Data are collected in 7 different dates to ensure the statistical stability. The time period is about 1 hour, the number of groups is 10^6 orders, the data quantity of packet headers is more than 25GB, and the magnitude of the connecting network and the edge network are both 104 orders.

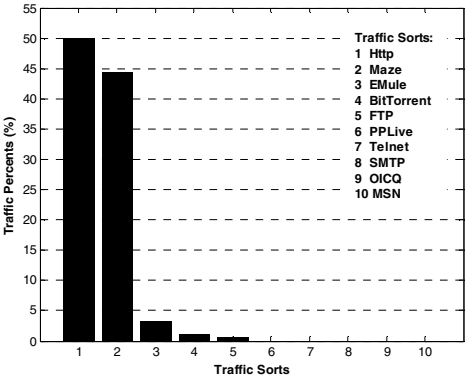

Fig. 3. Proportion of different Internet application business in backbone traffic

Figure 3 counts the uplink proportion distribution of several mainstream business (or application environment). The results indicate that the users` behaviors lead to the big difference between the proportions different business's traffic take. Maze, Emule, BiTTorrant and PPLive are all P2P application, they take the proportion close to HTTP, and the two take the vast majority of network resources. Actually, the experiment simplifies the connecting network H to the local range of Tsinghua University campus network, and the campus network is itself a huge Web Cache, inside which there are widely resource sharing, and the export traffic is reduced. So in actual situation the traffic of P2P business may take a greater proportion. As the mainstream businesses, HTTP and P2P accept the vast majority of users` behaviors, so we will mainly focus on these two businesses to do the statistics and discussions.

4.1 Statistics of HTTP Packets

This section discusses the behaviors of HTTP business between the connecting network and the edge network. This kind of behaviors in the application layer mostly result from the users clicking URL hypertext links to visit the web sites and download information, so it is a kind of two-way interactive behaviors. We first consider the uplink from the connecting network to the edge network. In the experiment, we filter the groups of HTTP businesses from the uplink trace flows and deal with them with relevant methods, and measure the in-degree distribution of the edge network nodes (web site) in different scales of time. It should be noted that the whole characteristic that the edge network accept the users` behaviors directly influence the basic network performance of the physical network such as load balance, quality of service and so on,[12] which is more significant relative to the distribution characteristics of the node degree of connecting network.

Because of the limitation of data quantity and processing ability, it is not able to do the traffic monitoring and analyzing in any large time scale at will. In order to confirm the statistical stability in time, we do the statistics of the grouping behaviors in two time scales. Figure 4 and Figure 5 shows the results of the uplink distribution of the grouping HTTP businesses, with the log-log coordinate system to be easily shown, and in which uplink traffic of 1 minute and 1 hour time intervals are

respectively used. In the figure, k represents the degree of the edge network nodes, $p(k)$ is the corresponding distribution function. The results indicate that the visits from HTTP groups to the edge network nods distribute as power law, and have consistent characteristics in different time intervals. Take the results of 1 hour observing time as example (Figure 5), the distribution of k satisfies

$$p(k) = C_l k^{-\gamma} \tag{1}$$

By fitting we can get $C_l = 0.69$, $\gamma_l = 1.25$, and the error is less than 0.1. In the case that the observing time is 1 minute, $\gamma_s = 1.23$, which is a little smaller. The statistics above reveals that the there exists power-law among the behaviors of grouping HTTP business, most of the edge nodes servers accept a relatively small part of business, while the minority of the nodes accept most of the user` behaviors, so the HTTP business have the apparent characteristics of cluster. Not like the scale-free properties pointed out by the complex network research, we confirm this phenomenon from the perspective of application-layer behaviors, and provide a direct explanation about the load imbalance of the actual network. Figure 6 can help us to understand the influence that scale-free HTTP business behaviors bring to the network more profoundly. After descending sorting the edge network nodes by degree, the top 10% node servers accept about 80% of the HTTP business, which leads to a local network peak, but a low efficiency. Therefore, to optimize the popular resource nodes has naturally become the main idea of the technology of Internet load balance at present.

In the statistics of the downlink distribution of HTTP grouping behaviors, the phenomenon of power-law is also discovered (Figure 7), and $\gamma = 0.82$. The resource nodes of the edge network will generally answer the users` uplink visiting behaviors, based on the analysis above, few popular resource nodes will produce the downlink answers which take a relatively large proportion (such as the file download, etc.), which leads to the distribution of scale-free behaviors, and aggravates the entirely load imbalance.

In the bipartite graph shown in Figure 2, N represents the maximal accessing number of the edge network nods in the specified period of time, which is the increasing function of the number of user M. For the sake of a convenient qualitative discussion, we uniformly use N to represent them. It is easy to see that the distribution of degree of the edge network nodes satisfies.

$$\sum_{k=1}^{N} P(k) = 1 \tag{2}$$

From formulas (1)(2) we can get the first-order and the second-order moment of k is

$$D_1 = \sum_{k-1}^{N-1} kP_d(k) = C\sum_{k-1}^{N-1} k^{1-\gamma} \tag{3}$$

$$D_2 = \sum_{k-1}^{N-1} k^2 P_d(k) = C\sum_{k-1}^{N-1} k^{2-\gamma} \tag{4}$$

$$C = \frac{1}{\sum_{k=1}^{N-1} k^{-\gamma}} \tag{5}$$

The mean and variance is $d_\mu = D_1$, $d_\sigma = D_2 - D_1^2$ respectively, and with the increase of γ, the D_1 and D_2 both decrease. When $\gamma \to \infty$, $d_\mu \to 1$, $d_\sigma \to 0$. So the increasing process of γ is a topological changing process that the heterogeneous characteristic of scale-free network gradually fades away, and tends to uniformity.

For the uplink business, $\gamma \in (1,2]$, when N is large enough, we have $D_1 = O(N^{2-\gamma})$, $D_2 = O(N^{3-\gamma})$, $d_\mu \ll N$. The mean value of the degree of the edge network node is significantly lower than the entire network level, while the mean and the variance both diverge, which is the obvious feature of the scale-free property. Compared with the traditional explanation about the flat network from complex network theory ($\gamma \in (2,3]$), the actual measured behaviors of application layer have stronger characteristic of cluster. Since the proportions of the traffic of uplink and downlink businesses are quite asymmetric, large number of downlink groups aggravates the heterogeneous characteristics of network, which results in the smaller power index $\gamma = 0.82$.

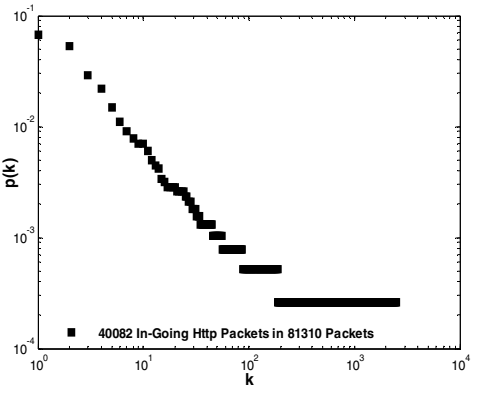

Fig. 4. Statistics of HTTP uplink packets. One minute uplink trace flow is measured.

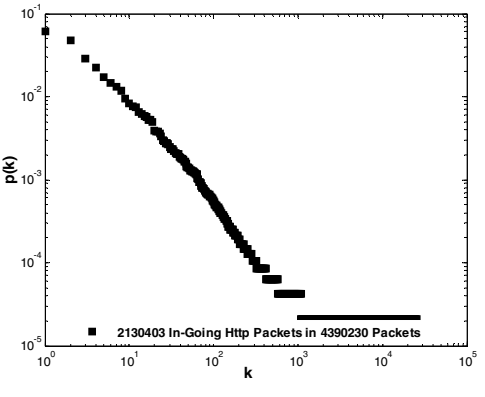

Fig. 5. Statistics of HTTP uplink packets. One hour uplink trace flow is measured.

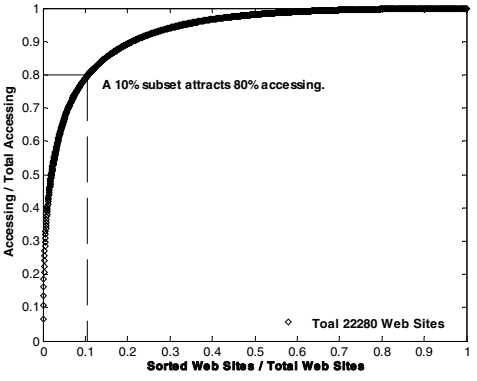

Fig. 6. Uplink HTTP packets collective behavior versus edge network nodes sorted by accessing times

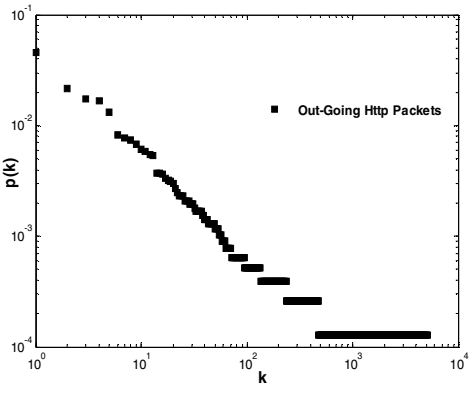

Fig. 7. Statistics of HTTP downlink packets. One hour downlink trace flow is measured.

4.2 Statistics of P2P Packets

With the development of broad band and the increasingly demands for the audio and video files, P2P businesses have gradually become the kind of business which use the most network resources. From the experiments, we can see that the uplink packets of P2P also entirely shows an obvious characteristic of power-law, with the power index $\gamma = 1.53$, a little bigger than that of HTTP grouping status, which indicates that the application layer behaviors diverges more for the assembly of P2P resources. Since the users' uplink behavior is a common reflect of information demands, the same with the characteristics of the HTTP uplink packets, a small number of popular resource nodes attracts the vast majority of grouping visits (Figure 9), represented by the assembly of users to the searching services nodes of P2P.

For the two mainstream businesses, HTTP and P2P, the downlink mechanisms of traffic and the resources using statuses are quite different, which is also reflected in the statistics of the downlink behavior of P2P business. As in Figure 10, it is a similar

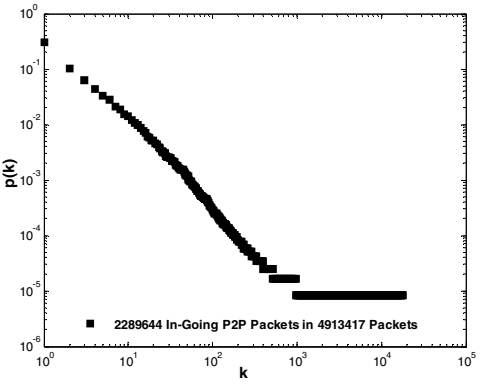

Fig. 8. Statistics of P2P uplink packets. One hour uplink trace flow is measured.

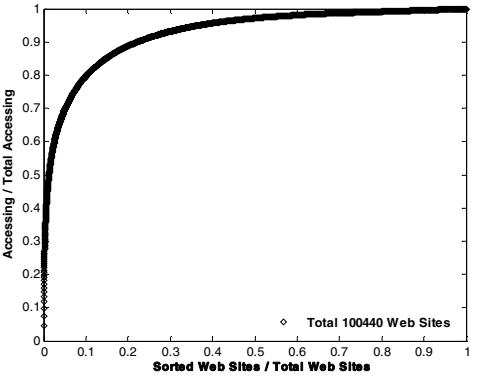

Fig. 9. Uplink P2P packets collective behavior versus edge network nodes sorted by accessing times.

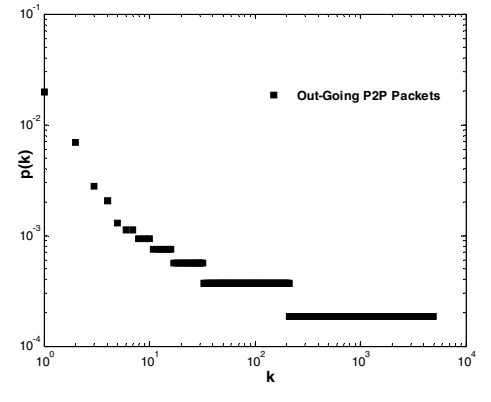

Fig. 10. Statistics of P2P downlink packets. One hour downlink trace flow is measured.

power-law distribution, the difference between nodes decreases, and the number of popular nodes is enlarged. The P2P business itself mainly shares long stream transmission. If we gather the statistics from the point of grouping, each node must have quite big degrees; and the mechanism of distributed peer-to-peer sharing resources, to a certain extent, curbs the flow imbalance and alleviates the power-law characteristic of business. However the grade difference of the network nods also exists obviously, for the edge network *S*, the node load is quite of imbalance, and the characteristic of assembling (diverging) still exists. The experiment indicates that the P2P mechanism improves the efficiency of resource sharing from the peer-to-peer perspective. However, it is a pity that it contributes very little to the flow balance in the physical network, and itself also occupies large part of network resources and redundantly costs the long stream, which will enlarge the influence of this kind of similar power-law characteristic to the traffic distribution, and this is a major conflict that can not be avoided.

5 Conclusion

In the former work, we use the simulation of abstract model to point out that the behaviors of application behaviors and the coupling of physical networks will influences the overall dynamics of Internet, with the development of network and the increasing of number of users, the influence from the behaviors of application layer to the distribution of the entire network traffic can not be ignored any more. This paper observes and researches the distribution characteristic behaviors of Tsinghua University campus network, such a local users group, based on the topology of bipartite graph which couples users nodes with the network nodes, in the scale of IP grouping we do the statistics and discussion about the two typical businesses, HTTP and P2P businesses, which take the most proportion of network traffic, and obtain the conclusions below: 1) The uplink distributions of HTTP business and the P2P business groups both show a consistent power-law distributed characteristic in time and space, the power index is respectively 1.25 and 1.53. A few popular resource nodes attract the vast majority of users` behaviors, which has a obvious characteristic of cluster. 2) The downlink packets of HTTP business show the power-law characteristic, and $\gamma = 0.82$, which is a little different from the traditional explanation about the flat network from the complex network theory. And the actual measured behaviors of application layer have a stronger characteristic of cluster. 3) Downlink packets of P2P business satisfy a distribution similar to the power-law, and the peer-to-peer mechanism contributes little to the traffic of physical network.

The general power-law characteristic shown in the behaviors of Internet application will heavily leads to the network load imbalance, the declining of the bandwidth utilization rate, and aggravates the rate of deterioration of the local network. The increasing P2P business occupies large quantities of network resources, however it doses not provide the obvious support for the physical network balance at the same time. This kind of actually measured characteristic of behaviors of application layer urges us to not only focus on the single layer of network or technical development in local range, but also entirely observe and research a series of theories and methods of analysis of network traffic, routing balance, topological design and so on.

Acknowledgment

This work is supported in part by the National Nature Science Foundation of China (NSFC) grants 60672142, 60772053, the National Basic Research Program of China (973 Program) grants 2007CB307100 / 2007CB307105, and by the China Postdoctoral Science Foundation grants 20080430400.

References

1. Fuks, H., Lawniczak, A.T.: Mathematics and computers in simulation 51, 101 (1999)
2. Li, Y., Liu, Y., Shan, X., Ren, Y., Jiao, J., Qiu, B.: Chin. Phys. 14, 2153 (2005)
3. Takayasu, M., Fukuda, K., Takayasu, H.: Physica A 274, 140 (1999)
4. Takayasu, M., Takayasu, H., Fukuda, K.: Physica A 277, 248 (2000)
5. Faloutsos, M., Faloutsos, P., Faloutsos, C.: ACM SIGCOMM Computer Communication Review 29, 251 (1999)
6. Siganos, G., Faloutsos, M., Faloutsos, P., Faloutsos, C.: IEEE/ACM Transactions on Networking 11, 514 (2003)
7. Medina, A., Lakhina, A., Matta, I., Byers, J.: Proc. of MASCOTS, Washington, 346 (2001)
8. Winick, J., Jamin, S.: Technical report CSE-TR-456-02, Department of EECS, Universtiy of Michigan (2002)
9. Chen, Q., Chang, H., Govindan, R., Jamin, S.: Proc. of IEEE INFOCOM, New York, 608 (2002)
10. Albert, R., Barabasi, A.-L.: Reviews of Modern Physics 74, 47 (2002)
11. Liu, F., Shan, X., Ren, Y.: Acta Phys. Sin. 53, 273 (2004)
12. Wang, L., Zhou, S., Yuan, J., Ren, Y., Shan, X.: Acta Phys. Sin. 56, 36 (2007)