The Community Analysis of User Behaviors Network for Web Traffic

CAI Jun
Department of Electronic and Communication Engineering Sun Yat-Sen University, Guangzhou 510275, P. R. China; School of Electronic and Information Guang Dong Polytechnic Normal University, Guangzhou 510665, P. R.China
e-mail:gzhcaijun@gmail.com

YU Shun-Zheng and WANG Yu
Department of Electronic and Communication Engineering Sun Yat-Sen University, Guangzhou 510275, P. R. China
e-mail: syu@mail.sysu.edu.cn; hf98403@163.com

***Abstract*—Understanding the structure and dynamics of the user behavior networks for web traffic (To be convenient in next sections, we refer to replace it as UBNWT) that connect users with servers across the Internet is a key to modeling the network and designing future application. The Web-visited bipartite networks, called the user behavioral networks, display a natural bipartite structure: two kinds of nodes coexist with links only between nodes of different types. We obtained the result that the out-degree distribution of clients (the host initiating the connection), the in-degree distribution of servers (the host receiving the connection) and the strength distribution (the exchange bytes between clients and servers) are approximately power-law, whose exponential is between 1.7 and 3.4. The clustering coefficient of clients and servers is larger than that in randomized, degree preserving versions of the same graph, which indicate a modular structure of UBNWT. Finally, based on the algorithm of finding the community structure in bipartite network, we divided the clients into different communities, through manual examination of hosts in these communities, the typical normal (interest) and abnormal (DOS) communities were found. Interestingly, the loyalty of clients belonging to the same community in different time is higher than 80%. The structure analysis of UBNWT is very helpful for the network management, resource allocation, traffic engineering and security.**

Keywords- **complex networks; user behaviors; community; clustering coefficient; bipartite network**

I. INTRODUCTION

Although the Internet was first built as an infrastructure to support other research efforts, its overwhelming success has created a complex system that has become a scientific challenge in its own right. There are two types of topology that are the physical topology and the logical topology. The physical topology is how the computers and peripheral are connected and how the cable is run between them, in other words the way the network looks. The logical topology describes the way in which a network transmits information from one node to the next node. Previous much studies has already gone into determining the physical structure of the Internet at several levels of granularity, with the goal of developing an abstract representation of Internet topology in which nodes and edges represent either routers and their physical connections, or autonomous systems(ASes) and their peering relations. Although these studies have revealed much about the physical arrangement of the Internet, they told us relatively little about the virtual networks created by the users who now spend a significant portion of their daily lives online, carrying out a wide variety of activities in different media. A detailed understanding of the many facets of the Internet's logical topological structure is critical for evaluating the performance of networking protocols, for assessing the effectiveness of proposed techniques to protect the network from anomaly intrusions and attacks, or for developing improved designs for resource provisioning. Human behavior more than physical connectivity determines the structure of these networks.

Web traffic is the amount of data sent and received by clients to a web site (server), which composed a large portion of Internet traffic. In this paper, we define these client-to-server networks, whose topology is formed by mutual use of HTTP application rather than the physical structure of the network, as the user behaviors network for Web traffic (UBNWT). Understanding their properties is an essential basis for further work in modeling the structure and dynamics of Internet Web traffic and contributes to our overall understanding of complex and emergent systems.

The remainder of this paper was organized as follows. In Section Ⅱ, we review the related work about the applications of complex network in communication network. In Section Ⅲ, we describe the methodology in this paper, including constructing the behavioral network from Web traffic, statistical characteristics of complex network used in this paper and the algorithm for finding community structures. In sectionⅣ we analyze the experiment results. Finally, we present conclusions and future work in Section Ⅴ.

II. RELATED WORK

Three particular developments have contributed to the complex network theory: Watts and Strogatz's investigation of small-world networks[1], Barabasi and Albert's characterization of scale-free models[2], and Girvan and Newman's identification of the community structures

doi:10.4304/jsw.6.11.2217-2224

present in many networks[3]. Complex network structures, generally modeled as large graphs, have played an important role in recent real networks. A series of applications to real networks, including social networks [4,5], the Internet and the World Wide Web[6], Metabolic, protein, genetic networks[7], and brain networks[8], have attracted increasing attention.

There are many applications based on complex network theory to communication networks and it includes aspects such as measurement, analysis, modeling, and algorithms. Measurements are critical for most of the papers in this Special issue. The last few years have seen an increasing interest in designing and applying more accurate measurement methods for various complex networks of interest [9]. A main goal of analyzing complex networks is to extract interesting information and illuminating properties from them. More recently, there has been interest in analyzing graph structures that evolve over time, but the analysis of such dynamic graphs is still in its infancy [10]. Modeling is a main aspect of complex network research. Network models are used for a variety of purposes, including the generation of synthetic network structures for simulations, prediction of the evolutionary behavior of the network, and gaining an understanding of the key forces the structure and impact the evolution of the network [11]. Dealing with large-scale graphs poses many algorithmic issues. A typical example is the author's devise new and efficient Steiner-tree constructions for multicast communication in Ref [12].

In this paper, we present the first analysis of the structure characters and the community of UBNWT based on the theory of bipartite networks. The ultimate purpose about our study is to model the UBNWT.

III. METHODOLOGY

In this section we consider the methodology we applied to UBNWT. First we define the UBNWT. Then we explain the Statistical characteristic of Complex network that will be used in this paper. At last, we introduce the algorithm for finding community structures in bipartite networks.

A. Constructing the user behavior networks for web traffic (UBNWT)

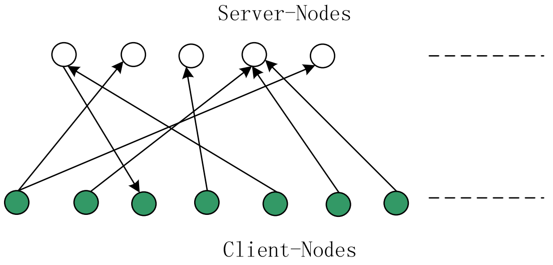


Figure 1. (Color online)The sketch map of the the bipartite network of user behaviors for web traffic (UBNWT). There are two types of nodes, i.e., client nodes and server nodes. Each link and its weight represent there being web traffic records between client nodes and server nodes and the accumulative bytes between them, respectively.

To perform the analysis presented in this paper we collected two weeks worth of flow records from a single site in a large campus environment connected by a private IP backbone and serving a total user population in excess of 16000 users. The flow records were collected from a boundary router using the Wireshark [13]. During the two week period we collected flow records corresponding to more than 400TByte of network traffic, then we removed weekend data from our data set and filtered out the web flow and ignored the network traffic among clients (that accounted for less than 0.02% of the total web traffic), stored traffic in 5 minute, one hour and one day intervals.

We can thus partition the set of all hosts into a subset $C=\{i_1,i_2,...,i_{Nc}\}$ of systems that act as clients and a subset $S=\{j_1,j_2,...,j_{Ns}\}$ of systems that act as servers. We constructed UBNWT in which the nodes represent individual hosts and edges represent the directed transmission of bytes between a pair of host sent client i to server j over the course of different time-intervals. Each weight w_{ij} represents the total amount of bytes sent from client i to server j over the course of different time-intervals, and w_{ji} represents the amount of bytes from server j back to client i .The sketch map of such bipartite network is shown in Figure 1.

One can transform a two-mode network into a one-mode network by considering, e.g. two clients linked if they co-communicate at least one server. Such a one-mode projection is shown in Ref [14]. Note that such a transformation is lossy, that is we no longer know which servers the clients communicate with. One can however assign weights to each edge corresponding to the number of shared connections in the two-mode networks. Since most network metrics are designed only for unweighted networks, and since omitting weights introduces a further loss of information, we will prefer to work with the full-bipartite network directly.

B. Statistical Characteristic of Complex Network

Graph theory is the natural framework for the exact mathematical treatment of complex network and, formally, a complex network can be represented as a graph. A graph $G=(V,E)$ consists of two sets $V(G)$ and $E(G)$. The degree of the complex network is the most simple and important unit. The degree distribution of the complex network $p(k)$ is defined as the probability that a node chosen uniformly at random has degree k or equivalently, as the fraction of nodes in the complex network having degree k . The same information is also sometimes presented in the form of a cumulative degree distribution function, the fraction of nodes with degree greater than or equal to k .

$$P_k=\sum_{k'=k}^{\infty}P(k') \qquad (1)$$

The average degree in the graph is defined as

$$<k>=\sum_{x\in V}\deg(x)/N \qquad (2)$$

Where N is the total number of nodes in the network. sIn the case of directed networks one needs to consider two distributions, $P(k^{in})$ and $P(k^{out})$.

For weighted networks, the definition of degree given above can be used, but a quantity called strength of i, s_i, defined as the sum of the weights of the corresponding edges, is more generally used:

$$s_i = \sum_{j \in N} w_{ij} \qquad (3)$$

It's cumulative degree distribution function is defined as

$$P_s = \sum_{k^{'}=k}^{\infty} P(s^{'}) \qquad (4)$$

Some networks, notably the Internet, the world wide web, and some social networks are found to have degree distributions that approximately follow a power law: $p(k) \sim k^{-\gamma}$, where γ is a constant. Such networks are called scale-free networks and have attracted particular attention for their structural and dynamical properties.

The clustering coefficient of a node i gives the probability that its neighbors are connected to each other. It is defined as

$$C_3(i) = \frac{2t_i}{k_i(k_i - 1)} \qquad (5)$$

Where t_i is the number of triangle observered, k_i is the number of neighbors of node i. However, in bipartite network, all nodes have $C_3 = 0$. To investigate the clustering properties of bipartite network, people usually project them into classical networks which are also called one-mode networks. However, the one-mode projection of a bipartite graph, generally loses some information of the original networks, brings an inflation of the number of edges and other drawbacks which are caused by the projection. Therefore, high clustering coefficients in projections may not viewed as significant properties: they are consequences of the bipartite nature of the underlying affiliation network. Some prior works have confirmed these[14]. So we used the define in Ref[15], the fraction of cycles with size four was used to define the clustering coefficient. The equation is shown as:

$$C_{4,mn}(i) = \frac{q_{imn}}{(k_m - \eta_{imn}) + (k_n - \eta_{imn}) + q_{imn}} \qquad (6)$$

Where m and n are the pair of neighbors of node i, and q_{imn} is the number of existent squares which include these three nodes. $\eta_{imn} = 1 + q_{imn} + \theta_{mn}$ with $\theta_{mn} = 1$ if neighbors m and n are connected with each other and 0 otherwise. Because of the definition of the bipartite networks, there is no link can exist between node m and n, neither among their neighbors. So $\eta_{imn} = 1 + q_{imn}$. k_m and k_n is respectively the degree of node m and n. Since the clustering coefficient $C_4(i)$ is easily obtained from eq.(6) just by suming the numerator and denominator separately over the neighbors of i.

The Algorithm for Finding Community Structures

In recent years, people have found that both of physical systems in nature and the engineered artifacts in human society can be modeled as complex networks, such as the internet, the World Wide Web, social networks, citation networks and etc. Although these systems come from very different domains, they all have the community structure in common[16], that is they have vertices in a group structure that vertices within the groups have higher density of edges while vertices among groups have lower density of edges. There are many successful methods for the identification of modules in unipartite network. However, a widely used one is the maximization of a modularity function. A ubiquitous function for unipartite networks is the Newman-Girvan's modularity. The rationale behind this modularity is that, in a modular network, links are not homogeneously distributed. Thus, a partition with high modularity is such that the density of links inside modules is significantly higher than the random expectation for such density. The modularity $M(P)$ of a partition P of a network into modules is

$$M(p) = \sum_{s=1}^{N_M} \left(\frac{l_s}{L} - \left(\frac{d_s}{(2L)} \right)^2 \right) \qquad (8)$$

Where N_M is the number of modules in a network, L is the number of links in the network, l_s is the number of links between nodes in modules s, and d_s is the sum of the degrees of the nodes in module s. Then l_s/L is the fraction of links inside module s, and $(d_s/2L)$ is an approximation to the fraction of links one would expect to have inside the module form chance alone. As the strategy for finding communities from given networks, modularity optimization is often employed. As for bipartite networks, there are two definitions of modularity: Guimera's bipartite modularity [17] and Barber's bipartite modularity[18]. But the weakness of Barber's bipartite modularity are: 1) the number of communities have to be searched in advance, and 2) the numbers of communities of both vertex types have to be equal[19]. Because the first weakness is fatal for practical community extraction since the search for the number of communities is commputationally expensive. The second weakness is also fatal for dividing real networks since the numbers of communities of both vertex types are often imbalanced. So we choosed the Guimera's bipartite modularity, whose basic theory is similar to Newman-Girvan's eq.(8), which is defined as the cumulative deviation from the random expectation of the number of the Y-vertex communities in which two vertices of type X are expected to be together:

$$M_B(p) = \sum_{s=1}^{N_M} \left(\frac{\sum_{i \neq j \in s} c_{ij}}{\sum_a m_a(m_a - 1)} - \frac{\sum_{i \neq j \in s} t_i t_j}{\left(\sum_a m_a \right)^2} \right) \qquad (9)$$

Where s is a X-vertex communities, N_M is the number of X-vertex communities, a is a Y-vertex community, m_a is the number of edges that are connected to the vertices in community a , c_{ij} is the actual number of Y-vertex communities in which vertices i and j are connected, and t_i and t_j are total number of Y-vertex communities to which vertices i and j are connected, respectively. We refined the results with a simulated annealing approach using the heat-bath algorithm [20].

IV. Experiments and Analysis

We now present the results of our analysis of the user behavior networks for web traffic gathered from 6 February 2008 to 20. In the course of two weeks, the flow collector received over 600 million flows involving almost 16000 hosts. Of these flows, 258 million (41.5%) were Web-related. We randomly selected 5 minutes, one hour and one day intervals traffic to analyze. Of the Web-related traffic, the number of behaving as clients or servers at different time-interval is as TABLE 1.

Table 1 The Number of clients and servers at different time-interval

	Clients	servers
Five minutes	1595	4457
One hour	3252	15605
One day	4888	66149

A. Degree distribution

We begin by considering the distribution of degree and strength for the nodes in UBNWT. Given a node N with i initial edges and j terminal edges, we define the degree as $d_N = i + j$ and the strength as

$$s_N = \sum_{k=1}^{i} w_{N,N_k} + \sum_{k=1}^{j} w_{N_k,N} \qquad (10)$$

Where $w_{a,b}$ denotes the weight of the edge between nodes a and b . In other words, the degree of a node in UBNWT reflects the total number of users with which it has exchanged data, and the strength reflects the total amount of bytes (bytes is a unit in kbytes) it has exchanged.

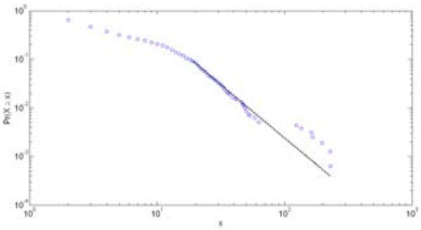
Because both the degree and strength distributions reflect the individual decisions made by a large population, it might seen plausible for their form to be roughly normal. This turns out to be far from the case, however, as shown in figure 2. All of the degree and strength distributions shown have extremely long tails. We are able to approximate both the degree and strength distributions with a power-law function $p(n) \sim n^{-r}$ over several orders of magnitude. Here we describe statistical techniques for making accurate parameter estimates for power-law data, based on maximum likelihood methods and the Kolmogorov-Smirnov statistic [21].

There are a variety of measures for quantifying the distance between two probability distributions, but for normal data the commonest is the Kolmogorov-Smirnov or KS statistic, which is simply the maximum distance between the CDFs of the data and the fitted model:

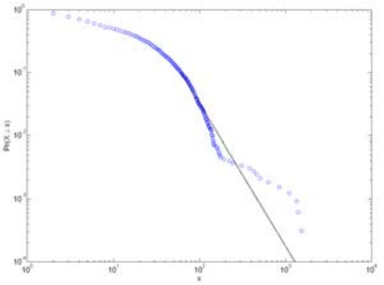
$$D = \max_{x \geq x_{\min}} |S(x) - P(x)| \qquad (11)$$

Here S(x) is the CDF of the data for the observations with value at least $x_{\min}$, $x_{\min}$ is the lower bound to x and $P(x)$ is the CDF for the Power-law model that best fits the data in region $x \geq x_{\min}$ .Our estimate $\hat{x}_{\min}$ is then the value of $x_{\min}$ that minimizes D .

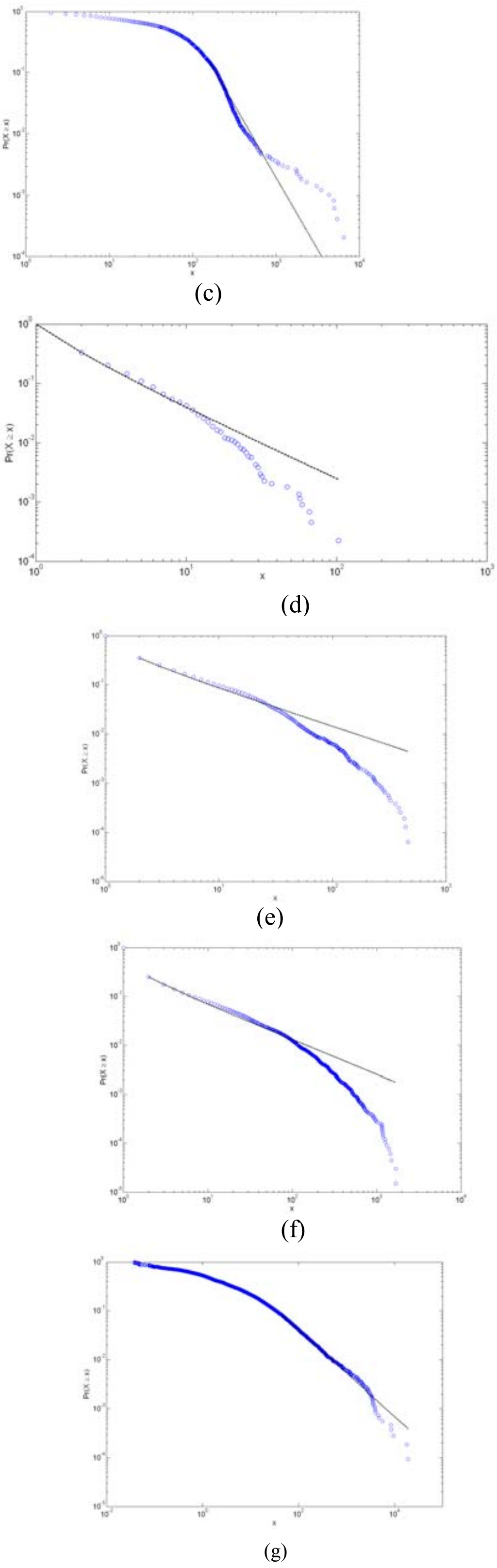
Let us examine the slopes of these power-law approximations, such as TABLE 2. When $1 < \gamma < 2$, as is the case for the both the degree of the servers(except 5 minutes interval) and strength, as for servers, the results show that there exist plenty of hub servers, that is, there exist hot spot website that are visited by the clients in the local area network; from the distribution of strength, we can conclude that traffic between the clients and servers in different time interval is different, there exist plenty of big traffic pairs. When $2 < \gamma < 3$, the edges are linearly dependent on the nodes, as is the case for the degree distribution of servers in 5 five minutes interval, there exist a certain amount hub servers, but it is not obvious comparing one hour and one day. When $3 < \gamma$, as is the case for the all clients in different time interval, as for the normal clients, the number of they visit website is homogeneous.



(a)



(b)



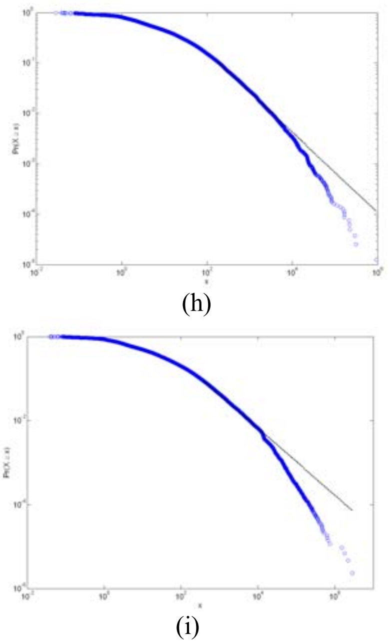
(c)

(d)

(e)

(f)

(g)



(h)

(i)

Figure 2. Probability distribution for degree of clients ((a), (b) and (c) separately denotes the 5 minutes, one hour and one day) and servers((d), (e) and (f) separately denotes the 5 minutes ,one hour and one day)and strength((g), (h) and (i) separately denotes the 5 minutes ,one hour and one day interval).

TABLE 2 . FITTING A POWER-LAW DISTRIBUTION

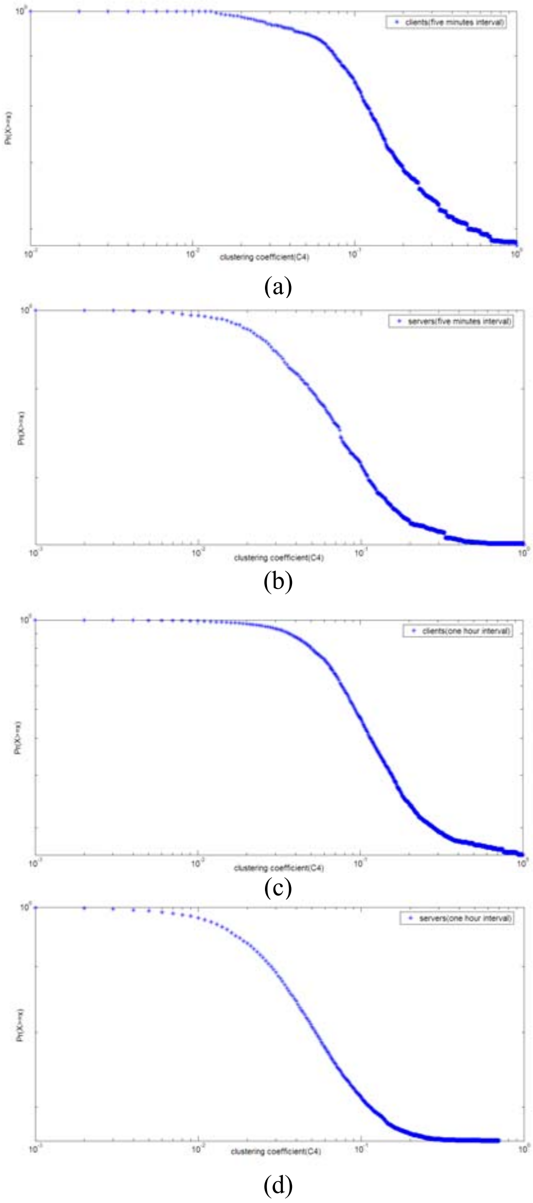
		Power exponent(γ)	D	Xmin
Five minute	clients	3.1700	0.0399	19
	servers	2.1800	0.0179	1
	Strengthen	1.8831	0.0120	44.7422
One hour	clients	3.2600	0.0611	60
	servers	1.7700	0.0362	2
	Strengthen	1.7964	0.0225	160.5170
One day	clients	3.3600	0.0568	160
	servers	1.7100	0.0330	2
	Strengthen	1.7949	0.0230	370.3130

B. *Clustering*

The basic cycle in bipartite networks is square. The cluster coefficient C_4 with squares is the quotient between the number of squares and total number of possible squares. Here we apply the eq.(6) to UBNWT. The results are displayed in TABLE 3, which is the average clustering coefficient for clients and servers at different time-intervals, which is significantly ($p < 10^{-12}$) larger than that in randomized, degree preserving versions of the same graph (C_4 =0.000263). The C_4 of clients in local area network is higher than that of servers, which show that clustering phenomena of clients is more obvious than that of the servers. From the above analysis of the clustering coefficient, we can see that in UBNWT, clients and servers are more likely to be correlated and overlapped than if they are randomly distributed.

Interestingly, it is observed that, in TABLE 3, the average clustering coefficients of clients becomes larger with the network size increasing, which explain that the interest of most clients is similar, with the number of clients increasing, the number of squares in network become large. But the average clustering coefficients of servers becomes smaller with the network size increasing, which suggests that some servers are visted by small fraction clients.

Figure 3 shows the distribution(complementary cumulative distribution function(CCDF))of the clustering coefficients of clients and servers in different time intervals on the double logarithm coordinates. They are approximately power-law distribution, which suggests that the large clustering coefficients of clients and servers are relatively rare.

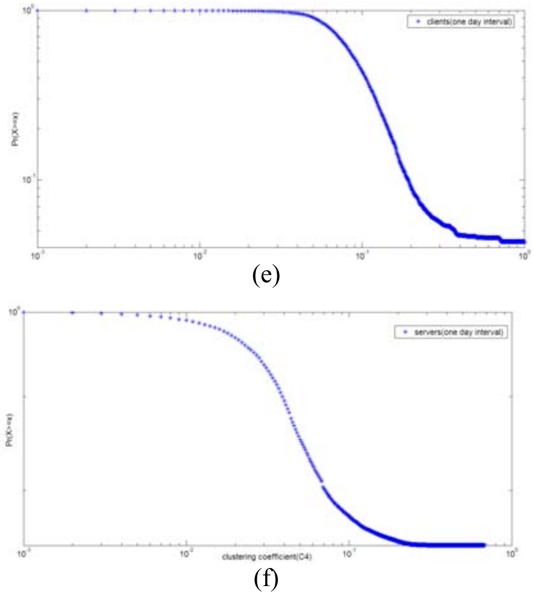


(a)

(b)

(c)

(d)



(e)

(f)

Figure 3. Distribution (CCDF plot) of the clustering coefficient (C_4) in different time interval ((a) and (b) separately denote C_4 of clients and servers in 5 minutes interval, (c) and (d) are corresponding to C_4 of one hour interval, (e) and (f) denote C_4 in one day interval).

TABLE 3. THE AVERAGE CLUSTERING COEFFICIENT AT DIFFERENT TIME-INTERVALS

Time-intervals	The average C_4	
	clients	servers
5 minute	0.0762	0.0250
One hour	0.0908	0.0203
One day	0.0995	0.0141

C. Community Structure Analysis

As we saw in the previous section, clients may be 'clustered', many of the clients co-visit in the same set of servers. We may be interested in identifying such communities because they may correspond to different interest communities or abnormal communities. For small networks, this question may be answered through a visual analysis, However, the node in the network exceed 100, it is difficulty to be clear visual. For example, figure 4 shows community feature of clients in five minutes interval (the number of clients and servers exceeds 6000), it is difficult to vision. Larger networks often need to be analyzed using a community-finding algorithm. We applied the algorithm for finding community structures in bipartite network described above to the clients of 5 minute-interval and one day interval web traffic. We sort the communities based on their size (number of clients inside). In 5 minutes interval, out of the 103 identified communities, 70 contain fewer than 10 hosts, with total of 613 clients falling into these small communities. On the other hand, the top 5 communities contain total 716 clients with an average size of more than 100 clients per communities. This indicates that clients with

HTTP application exhibit similar behaviors at very short interval. Manual examination of hosts in some communities from different five durations shows that they have some abnormal behaviors, such as a huge volume of outgoing traffic to a small number of destinations which resemble a DOS attack pattern, or brief communication with a large number of destinations, which resembles scanning traffic. At one day interval, the community character of clients is more obvious than that of 5 minutes interval, out of identified communities, 100 contain fewer than 50 hosts, with total of 706 clients falling into these small communities. On the other hand, the top 10 communities contain total 4320 clients with an average size of more than 400 clients per communities. On the other hand, more than 80% clients fall into communities larger than 300, and represent a routine usage of the majority of clients. The one day community analysis results show that most people have the similar interest, which can guide us integrate the information most people are interested in using the wide storage technology to satisfy the information most people need, then people will get information more economically and more quickly.

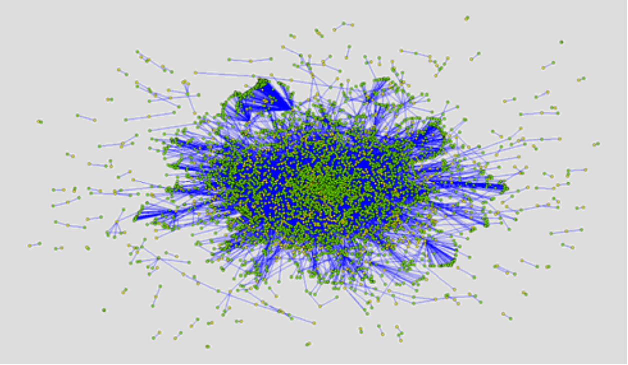


Figure 4. A spring layout of the network of clients and servers in five minutes interval drawn using the Pajek network analysis and visualization software[20], yellow solid circular is corresponding to servers, green one is corresponding to clients.

We evaluate loyalty of clients to communities in one day interval. This experiment tests the hypothesis that normal clients tend to fall into the same or a similar communities, despite of their varying behavior over time. To compare client behaviors for HTTP with the characteristics of their belonging the same or a similar communities, one day interval trace is randomly selected from the two-week collected data (weekend data is), we first apply community detecting algorithm described above to clients, then, tag each clients with ID of communities it belongs to. We call these communities the “control communities” for corresponding clients. The community change of clients is define as clients community change rate (CCCR):

$$CCCR = \frac{the\ number\ of\ changing\ community\ clients}{the\ total\ number\ of\ clients} \times 100\%$$

We then use remaining 10-day data to test loyalty of clients to communities. The results of these tests are shown in TABLE 3. It is observed that more than 80% clients have their control communities. This result verifies the hypothesis that a large number of clients exhibit steady behavior patterns over time. The above analysis results indicate that rapid client community changes would be an additional indication of anomalous network behavior. This holds true if a large number of clients are rapidly added to or removed from the community they do not belong to.

TABLE 3. THE CLIENTS COMMUNITY CHANGE RATE IN DIFFERENT DAY

date	6	7	8	11	12
CCCR(%)	16.3	17.9	19.1	19.6	17.8
date	14	15	18	19	20
CCCR(%)	19.2	18.3	19.6	18.3	17.9

V. SUMMARY AND THE FUTURE WORK

In this paper, we analyzed the structure characterizes of user behaviors for Web traffic, including the degree of clients and servers and strength between clients and servers, the clustering coefficient of clients and servers in different time-interval, the community structure analysis of clients. The pervasive presence of distributions with extremely long and heavy tails implies that client and server behavior rarely follows normal distributions, but is so diverse as to defy characterization with a mean value. High co-visit implies clustering-communities of clients that visit simultaneously the same server. We demonstrated that these clusters of clients may be discovered by using community finding algorithm in bipartite network. Finally, we evaluate loyalty of clients to communities in one day interval, the result show that more than 80% client fall into the same community.

We are continuing the presented work by moving from the presented structure characterization of user behaviors for web traffic to finer grained per-host characterization for different traffic, such as P2P, FTP and DNS. Our ongoing work aims to provide models that accurately capture client community behavior. And our ultimate goal is to be able to apply such models to network management, resource allocation and security.

ACKNOWLEDGEMENTS

We would like to thank all the members in our research group, for the valuable discussions about the ideas presented in this paper. Funding from Project supported by the National High Technology Research and from Development Program of China (Grant No.2007AA01Z449) and the Key Program of NSFC-Guangdong Joint Funds (Grant No.U0735002).

REFERENCES

[1] D.J. Watts and S.H. Strogatz, “Collective dynamics of small-world networks,” Nature 1998 Jun 4: 393(6684): 440-2

[2] Albert-László Barabási, Réka Albert, “Emergence of Scaling in Random Networks,” Science 15 October 1999 vol.286. no. 5439, pp.509-512

[3] M. Girvan and M.E.J. Newman, "Community structure in social and biological networks," PNAS June 11, 2002 vol.99 no. 12 7812-7826

[4] Willinger W, Doyle J.Robustness, "the Internet: Design and evolution, " 2002. http://netlab.caltech.edu/Internet/.

[5] S.Boccaletti, V.Latora, Y.Moreno, M.Chavez and D.-U.Hwang, "Complex networks: Structure and dynamics," Physics Reports, 2006, 424(4,5): 175-308.

[6] S.Wasserman, K.Faust, "Social Networks Analysis," Cambridge University Press, Cambrigdes, 1994.

[7] R. Pastor-Satorras, A. Vespignani, "Evolution and Structure of the Internet: A Statistical Physics Approach," Cambridge University press, Cambridge, 2004

[8] L.H.Harwell, J.J.Hopfield, S.Leibler, A.W.Murray, "From molecular to modular cell biology," Nature, Vol.402, C47-C52, 1999.

[9] M.Fraiwan, G. Manimaran, "Scheduling algorithms for conducting conflict-free measurements in overlay networks," Computer Networks 52 (2008)2819-2830.

[10] A.Scherrer, P.Borgnat, E.Fleury, J.-L.Guillaume, C.Robardet, "Description and simulation of dynamic mobility networks," Computer Networks: The International Journal of Computer and Telecommunications Networking. Volume 52, Issue 15 (October 2008) 2842-2858

[11] Andrej Vilhar and Roman Novak, "Policy relationship annotations of predefined AS-level topologies," Computer Networks Volume 52, Issue 15, 23 October 2008, Pages 2859-2871

[12] Knut-Helge Vik, Pål Halvorsen and Carsten Griwodz, "Evaluating Steiner-tree heuristics and diameter variations for application layer multicast," Computer Networks Volume 52, Issue 15,23 October 2008, 2872-2893

[13] http://www.wireshark.org/

[14] Matthieu Latapy , Clemence Magnien , Nathalie Del Vecchio, "Basic Notions for the Analysis of Large Affiliation Networks / Bipartite Graphs," arXiv:cond-mat/0611631v1

[15] Peng Zhang, Jinliang Wang, Xiaojia Li, Menghui Li, Zengru Di, Ying Fan, "Clustering coefficient and community structure of bipartite networks," Physica A 387(2008) 6869-6875.

[16] M. E. J. Newman, "Modularity and community structure in networks," PNAS 103 (23) 2006 8577-8582.

[17] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral, "Module identification in bipartite and directed networks," PHYSICAL REVIEW E, 036102(2007).

[18] M. J. Barber, "Modularity and community detection in bipartite networks," Physical Review E, 76(066102): 1-9, 2007.

[19] Tsuyoshi Murata, "Modularities for bipartite networks," Proceeding of the 20th ACM conference on Hypertext and hypermedia, pages: 245-250 (2009).

[20] Roger Guimerà and Luís A. Nunes Amaral, "Functional cartography of complex metabolic networks," Nature 433, 895-900 (24 February 2005)

[21] A. Clauset, C.R. Shalizi, and M.E.J. Newman, "Power-law distributions in empirical data," arXiv:0706.1062

[22] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

CAI Jun received the B.S degree from Hunan normal university, Changsha, China, and M.S degrees from Jinan University, Guangzhou, China.

He is presently a PHD candidate with the communication information system at department of electrical and communication engineering, Sun Yat-Sen University. He is also an instructor at Guangdong Polytechnic Normal University.