Analyzing Traffic Characteristics of Application Using Chi-Square Statistic

Liang Chen
College of Computer Science
Southeast University
Nanjing, China
lchen@njnet.edu.cn

Jian Gong
College of Computer Science
Southeast University
Nanjing, China
jgong@njnet.edu.cn

***Abstract*—The knowledge of traffic behavior characteristics of every application has vital effect on the accuracy and granularity of application identification. Based on analyzing the Chi-Square Statistics, a novel method named ABSA is proposed to analyze the traffic behavior characteristics of applications. The ABSA method does not focus on any certain applications; in contrast, it aims at providing a quantitative measurement for describing the behavior differences among applications. So that the traffic behavior characteristics and their significances of any applications can be determined. The theoretical analysis and experiments results also indicate that, the ABSA method can keep the sequence of behavior characteristic significance unchanged in packet sampling environment, which is often used by NetFlow and many other flow information collecting systems, to simplify the process of characteristic re-selection when sampling rate changes.**

Keywords-Internet traffic; Traffic identification; Behavior characteristic; Chi-square test; Packet sampling

I. INTRODUCTION

Accurate Internet application identification can provide much useful information to QoS implementation, network monitoring, traffic billing and intrusion detection. Due to the inaccuracy of port-based application identification [1, 2], the high complexity and privacy issue of payload-based identification, traffic behavior based identification now becomes a new direction [3-6]. However, the accuracy, granularity and generality of the identification methods presented these years are not so satisfying. One vital reason is that such studies stay on applying existing classification methods to application identification domain, but not deeply analyzing which metrics are the traffic behavior characteristics of a certain application. Application traffic behavior characteristics is some traffic metrics of that application, which have distribution differences from other applications in real network environment, including metrics about flow behavior in time dimension, metrics about host behavior in space dimension, and topological metrics between hosts.

For these shortcomings, this paper presents a novel method named ABSA (Application Behavior Significance Assessment) to analyze the behavior characteristics of applications and assess their significances, using Chi-Square Statistic and Test. The ABSA method does not focus on any certain applications; in contrast, it aims at providing a quantitative measurement for describing the behavior differences among applications. So that the traffic behavior characteristics and their significances of any applications can be determined. The theoretical analysis and experimental results also indicate that, the ABSA method can keep the significance sequence of behavior characteristics unchanged after packet sampling, to simplify the process of characteristic re-selection when sampling rate changes. This method can easily cooperate with popular flow information collecting systems.

The remainder of this paper is organized as follows. Alongside the chi-square statistic and test, section II presents the problems and solutions when applying chi-square statistic to behavior characteristic analysis, and proposes ABSA method. Section III analyzes some traffic behaviors of eDonkey application and that of the overall network traffic, to show the correctness of ABSA. Section IV details the influence of packet sampling on the behavior difference and chi-square statistic. Section V concludes this paper and provides directions for future work.

II. TRAFFIC CHARACTERISTIC ANALYSIS

The nature of behavior characteristic determination is to judge whether the distributions of a behavior metric in two opposite sample spaces are identical or not. If the distributions are identical, then the two applications which are indicated by the sample spaces have that same behavior. Any identification method can not use this metric to distinguish between the two applications. This metric can not be used as a characteristic of that application. Otherwise, that behavior of the two applications is different, the metric can be a characteristic, and the obvious degree of the distribution difference determines the significance of that characteristic. The determination of application behavior characteristic should not lie on any identification method, and be independent of the proportion of the samples made up by that class.

A. Chi-Square Statistic

Chi-square Test [12] can judge whether the two sample spaces arising from a same distribution. Consider two sets of samples, whose value distributions are shown in the following table:

This work is supported by the National Basic Research Program of China (973 Plan) under Grant No. 2009CB320505; the State Scientific and Technological Support Plan Project under Grant No. 2008BAH37B04

Sample Space	Distribution of value			
	Interval 1	***Interval 2***	**…**	***Interval v***
Sample 1	n_{11}	n_{12}	…	n_{1v}
Sample 2	n_{21}	n_{22}	…	n_{2v}

Thus, we can raise the hypothesis:

H_0: distribution of Sample 1 is identical to Sample 2.

And the appropriate chi-square statistic:

$$\chi^2 = n\left(\sum_{C=1}^{v}\sum_{R=1}^{2}\frac{n_{RC}^2}{n_R n_C} - 1\right) \quad (1)$$

While n_{RC} is the value at *R*th row and *C*th column in the above table, n_R is the sum of values in the *R*th row, n_C is the sum of values in the *C*th column. n is the total number of samples in two sets. If n is large enough, the asymptotic distribution of the χ^2 is χ^2-distribution with v-*1* degrees of freedom. Once we have the samples of a metric, we could calculate its χ^2 statistic value using (1), set the confidence level α, and consult the χ^2-distribution table to achieve the determination of whether accepting the hypothesis H_0.

B. Interval Division

χ^2-test described above is the ideal situation in statistics theory. However, the different classification of the values in the above table will cause the value of χ^2 statistic changed, and may further cause different judgment. The different ways of value classification include:

1. The same number of intervals (the same degree of freedom), different interval divisions.
2. Different numbers of intervals.

In order to minimize the impact of different interval divisions on the judgment result, we take two refinements, as follows.

To issue 1, the degree of freedom is a constant, suppose its value is v. The interval division of χ^2-statistic has to further meet the following two constraints [13].

Constraint 1. Suppose T_{RC} is the theoretical frequency of n_{RC}, $T_{RC} = n_C n_R / n$. Then to every value of R and C, T_{RC} should not be less than 1.

Constraint 2. To the issues only having two sample spaces, the size of the set $\{T_{RC} \mid 1 \le T_{RC} < 5, R = 1,2, C = 1,2,\cdots, v\}$ should be no more than $2(v+1)/5$.

According to the above constraints, we can get our heuristic algorithm for interval division, called Divide_Interval. Suppose the minimum and maximum values of sample variables are *min* and *max*.

Algorithm Divide_Interval (v)
  Generate random numbers $a_1, ..., a_v$ in (*min*, *max*);
  $a_0 = min$; $a_{v+1} = max$;
  flag = TRUE;
  while (*flag*)
    flag = FALSE; *count* = 0;
    for C = 1 to v+1 and R = 1 to 2
      Count n_{RC}, n_R and n_C in every interval (a_{C-1}, a_C];
      $T_{RC} = n_R n_C / n$;
    for C = 1 to v+1 and R = 1 to 2
      if (T_{RC}<1)
        Merge (a_{C-1}, a_C] and T_{RC} with the adjacent interval;
        flag = TRUE; break;
      else
      if (1≤ T_{RC}<5)
        ++*count*;
      if (*count* ≥ 2(v+1)/5)
        Select smallest T_{RC};
        Merge (a_{C-1}, a_C] and T_{RC} with the adjacent interval;
        flag = TRUE; break;
      if (*flag*)
        Selete the largest T_{RC};
        Generate random number a in (a_{C-1}, a_C);
        Splite (a_{C-1}, a_C] into (a_{C-1}, a] and (a, a_C];
        Remark a_1, …, a_v;

After applying Divide_Interval, the ultimate division will not only meet the requirement of random and the constraints of theoretical frequency, but also trend to be consistent with the metric distribution. That is, in the range containing dense values, the intervals will be divided into small ones; while in the range containing sparse values, the span of intervals is relative larger. Therefore, the χ^2 statistic could represent the distribution differences more stably and precisely. As shown in Figure 1, after using Divide_Interval, the results of χ^2 statistic are more stable nearby the mean value, the variance of χ^2 is only 3.62% of that when not using Divide_Interval (The experimental data is introduced in Section III, metric *pkts*).

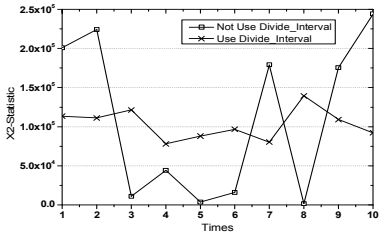

Figure 1. Improvement of χ^2-Statistic after using Divide_Interval

To issue 2, we use multiple degrees of freedom, and adopt the judgment result by voting.

C. Assessing the Significance of Characteristic

The significance of behavior characteristic is the obvious degree of distribution difference between two sample spaces. Reference [12] indicates that, if the distribution difference is more obvious, the χ^2 statistic will be greater. However, the result of χ^2 statistic also depends on its degree of freedom.

Some researches suggest that the ratio of χ^2 statistic to its corresponding degree of freedom can be used as a more precise measurement. However, the growth of χ^2 statistic is often linear lower than the growth of degree of freedom, as the line χ^2-*stat* and χ^2 */ freedom* shown in Figure 2 (χ^2-*stat* is a χ^2 value, χ^2*/freedom* is the ratio of χ^2-*stat* to its degree of freedom. Experimental data can be found in [12]). On the other hand, once the confidence level α is set, the critical quantile $\chi^2_{\alpha,v}$ can be used to average the statistic, whose effect is far

better than using degree of freedom (see line χ^2*/quantile* in Figure 2). This is because the critical quantile not only increases with degree of freedom, but also contains more information about the variable distribution under that degree of freedom.

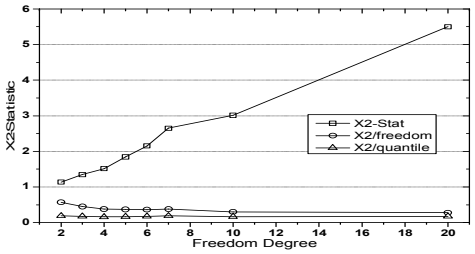

Figure 2. Relationship between χ2-Statistic and degree of freedom

Based on the above analysis, we can define the significance of application behavior characteristic as:

$$BS = \frac{1}{m}\sum_{i=1}^{m}\frac{\chi_{v_i}^2}{\chi_{\alpha,v_i}^2} \tag{2}$$

m is the number of degree of freedom we use. In order to further eliminate the impact of interval division on χ^2 statistic, the value of $\chi_{v_i}^2$ in (2) is the average value of multiple χ^2 statistic, every single χ^2 calculation is based on the algorithm Divide_Interval described above.

Synthesizing the all analysis about χ^2 statistic, we can arrive at our algorithm for assessing the significance of application behavior characteristic, which we call Application Behavior Significance Assessment or ABSA for short. Suppose the voting function is:

$$Vote(\chi^2, v, \alpha) = \begin{cases} -1 & \chi^2 < \chi_\alpha^2(v) \\ 1 & \text{Otherwise} \end{cases}$$

Then the ABSA algorithm could be briefly described as

Algorithm ABSA()
  for i = 1 to m
    for j = 1 to t
      Divide_Interval (v_i);
      Compute χ_{ij}^2 according to Equation (1);
      χ_i^2 += χ_{ij}^2;
    BS =χ_i^2 / (t * $quantile_{vi}$)
    result += **Vote** (χ_i^2 / t, v_i, α);
  BS /= m;

If *result* < 0, the metric can not serve as a characteristic of the application. Otherwise, the metric can be a characteristic, and the value of BS is the corresponding significance; the larger it is, the more significant the characteristic is.

III. EXPERIMENTAL RESULTS

In order to show the reasonableness and correctness of ABSA, this section takes eDonkey application [13] as an example, uses ABSA to analyze some of its behaviors and their distribution differences from the overall network traffic.

The experimental data is collected at 15:00 to 16:00 on August 20, 2008, lasts for one hour. The site monitored is the border of JSERNET to CERNET backbone network [14]. Although the behavior metrics are derived only using packet header information, the eDonkey traffic is derived using a content-based analysis (use the eDonkey pattern of l7-filter [15]), for the purpose of accuracy.

Let m=5, v=5, 10, 20, 30, 40, α=0.05, the BS values are in Table I.

TABLE I. BS VALUE OF SOME FLOW METRICS

Metric	BS	Description
TCPflags	2.1E+5	cumulative OR of TCP flags
pkt_size	9.0E+4	average bytes in IP packet
bytes	6.5E+4	total number of bytes observed
pkts	5.3E+4	total number of packets observed
duration	1.1E+4	end_time - start_time
Bps	3.1E+3	bytes per second, average byte rate
pps	2.9E+3	packets per second, average packet rate
pkt_size_ratio	1.7E+3	ratio of average packet size between two directions (≥1)
bytes_ratio	1.6E+3	ratio of number of bytes between two directions (≥1)
pkts_ratio	1.6E+3	ratio of number of packets between two directions (≥1)
head_size	5.1E+2	average bytes of packet header (IP header + TCP/UDP header)

The values of *result* in ABSA indicate that all behaviors in Table I have some distribution differences between eDonkey and the overall IP traffic; however, the significance varies greatly. To understand this result clearly and show whether ABSA can measure the significance accurately, we choose three behavior metrics (*bytes*, *Bps* and *head_size*, the BS value of each metric is in a different order of magnitude), further analyze their distribution differences. The PDFs of *bytes*, *Bps* and *head_size* are shown in Figure 3~5.

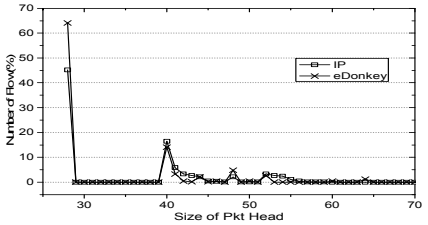

Figure 3. PDF of average packet head size of flow

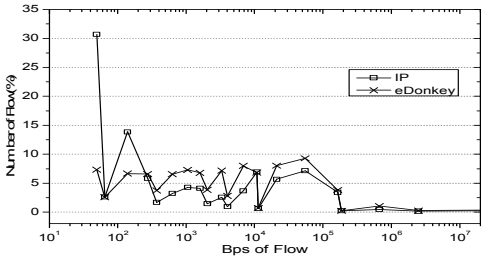

Figure 4. PDF of average Bps of flow

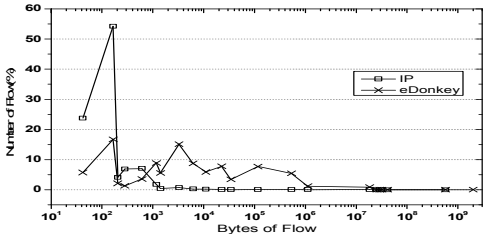

Figure 5. PDF of average Bytes of flow

Noticing the scaling of the vertical axis, the distribution difference becomes gradually obvious from Figure 3 to Figure 5, which is consistent with the BS values in Table I. The above analysis can further verify the relationship between BS value and the obvious degree of distribution difference: if the distribution difference of a certain metric is more obvious, then its BS value calculated by ABSA will be greater.

Therefore, the following conclusions can be drawn:

- ABSA can judge whether a metric can be a behavior characteristic of an application accurately.
- For those behavior characteristics already chosen by ABSA, the corresponding BS values can reflect their relative significances correctly and reasonably.

IV. IMPACT OF PACKET SAMPLING

In order to reduce the resource consumption, many high-end routers form flow statistics from a sampled substream of packets [16, 17]. If we can uncover the influence of such packet sampling on the results of χ^2 statistic and ABSA, it is possible to directly link ABSA method to the output of the popular flow statistics information collecting system, such as NetFlow, to improve the practicality of ABSA.

The impacts of packet sampling on the behavior distribution include two aspects: 1. the impacts of packet sampling on the flow behavior in the time dimension; 2. the impacts of flow sampling caused by packet sampling on the host behavior in the space dimension.

The former situation is typically like the distribution of *flow length* (the total number of packets in a flow). For its distribution differences in packet sampling environment, several theorems are as follows.

Theorem 1. If the original distributions of *flow length* have no difference, after packet sampling at any rate, the distributions still have no difference.

Proof: Suppose that the original distribution of flow length is $P_b(Len) = p_i$, sampling rate is p, then the distribution of *flow length* after sampling is

$$P_a(Len=l)=\sum_{i=l}^{\infty} P_b(Len=i)C_i^l p^l (1-p)^{i-l}$$

Since there is no difference in the original distributions, to every value of i, the value of $P_b(Len=i)$ is the same in the two distributions of applications. Meanwhile, when i and l is settled, C_i^l, p^l and $(1-p)^{i-l}$ are all constants. So, to every value of l, the values of $P_a(Len=l)$ are the same for the two applications, which means the distributions are still the same after packet sampling. #

Theorem 1 shows that, if ABSA indicates that *flow length* can not be used as a characteristic when no sampling used, then after packet sampling at any rate, the values of χ^2 statistic and BS are still close to 0, and ABSA can still indicate that it should not be a characteristic.

Theorem 2. If the original distributions of *flow length* have differences, then the obvious degree of distribution difference will become small after packet sampling, and the values of χ^2 statistic and BS also become small.

Proof: After sampling at rate p, the probability of a single flow containing i packets being selected is

$$P_{flow}(Len=i)=1-(1-p)^i \tag{3}$$

According to (3), flows containing different numbers of packets have different probabilities of being selected. But if just consider the flows containing i packets, the probability of them being selected is a constant. Therefore,

(a) If the original probability $P_b(Len=i)$ have difference at the point of $Len=i$, then the samples with larger original probability will be taken away more proportion of its total number of flow after packet sampling.

(b) If $P_b(Len=i)$ are the same in the two sample spaces, then they will be taken away the same proportion of flows after packet sampling.

From (a) and (b), we see that the distribution difference will become small after packet sampling, and also the values of χ^2 statistic will. #

Inference 1. If the original distributions of *flow length* have differences, then the lower the sampling rate is, the smaller the distribution difference after sampling is.

Proof: Suppose p_1 and p_2 are two sampling rates, $1>p_1>p_2>0$. Then $\exists\ p=p_2/p_1$, $0<p<1$, and $p_2=p_1\times p$.

The result of sampling at p_2 is equivalent to the final result after taking the following two steps: (a). Sampling at p_1 to the original data; (b). Sampling at p to the result of (a). According to Theorem 2, $BS_{p2} = BS_{p1\times p} < BS_{p1} < BS_{original}$. #

Inference 2. There is a sampling rate threshold p_0, when the actual sampling rate $p<p_0$, the distributions after sampling have no difference statistically.

Proof: According to (3), when $p=0$, the flow sampling rate $P_{flow}=0$. In this situation, distributions have no difference, $\chi^2=0$.

While according to Inference 1, when $p\to 0^+$, the distribution difference $\to 0^+$, $\chi^2\to 0$. Therefore, it is known from the definition of limit existence, to $\forall\ \delta>0$, $\exists\ p_0>0$, when $0<p<p_0$, there must be $(\chi^2 - 0) < \delta$. Let $\delta=\chi_\alpha^2$, then when $p<p_0$, $\chi^2<\chi_\alpha^2$, there is no difference between the distributions. #

The analysis of other time-dimensional behavior metrics is similar to the above analysis of *flow length*. However, it is necessary to pay attention to the metric *average packet size* and those metrics about ratio, such as *pkt_ratio*. Claffy [7]

shows that the distribution of *packet size* does not change after packet sampling. However, this research only focuses on the distribution of overall packets on the network, rather than the average packet size per flow. It is known that if the times of probability experiment are not enough, the experimental result may be deviated. So, as the number of packets in each flow is much smaller, it has a certain chance that the packet with certain size being selected, and may cause the value of average packet size changed. Similarly, although sampling is independent of packet direction, the distributions of the metrics about ratio may also change, and the distribution differences will be smaller than the original ones.

To those space-dimensional metrics, if taking the traffic of an application between two hosts as a general flow, and the flow sampling rate as the corresponding packet sampling rate, the analysis of their distributions is similar to the corresponding metrics in the time dimension. The conclusions of Theorem 1 and Theorem 2 are still correct.

Theorem 2 and its inferences show that, once the actual sampling rate is lower than a threshold, the distributions will become almost the same, even if there is a great gap between the original distributions. Therefore, this metric can no longer provide any useful information to the identification of that application, and can not be used as a characteristic of that application any more. This threshold varies with different applications and metrics. However, in any sampling rate situations, the sequence of the distribution difference degree can be guaranteed not to change, see Theorem 3.

Theorem 3. In any sampling rate situations, the obvious degree sequence of distribution difference is the same, and the value sequence of χ^2 statistic and BS is unchanged.

Before proving Theorem 3, we first put forward the following lemma.

Lemma 1. To any metrics, $\exists\ \varepsilon>0$, when the sampling rate $p\in(1-\varepsilon,\ 1]$, the overall distribution difference after packet sampling is no more than one unit smaller to the original distribution difference.

Proof: Since $p\rightarrow 1^-$, Δ(distribution difference)$\rightarrow 0^+$.

According to the definition of limit existence, to $\forall\delta>0$, $\exists\varepsilon>0$, when $0<1-p<\varepsilon$, there must be (Δ(distribution difference) $-0)<\delta$. Let $\delta=1/$(*number of samples*), Lemma 1 is proved. #

Thus, Theorem 3 can be proved.

Proof: Suppose M_1 and M_2 are two metrics, and the distribution difference degree of M_1 is greater than M_2 in original situation, $BS(M_1)>BS(M_2)$. Sampling rate is p, $0<p\leq 1$.

According to Lemma 1, $\exists\ \varepsilon_1>0$, when sampling rate is in $(1-\varepsilon_1,\ 1]$, the distribution difference of M_1 after sampling becomes smaller with no more than one unit; and $\exists\ \varepsilon_2>0$, when sampling rate is in $(1-\varepsilon_2,\ 1]$, the distribution difference of M_2 after sampling becomes smaller with no more than one unit. Let $\varepsilon=\min(\varepsilon_1,\ \varepsilon_2)$.

If $p\in(1-\varepsilon,\ 1]$, then according to Lemma 1, the distribution difference of M_1 after packet sampling is still greater than or equal to that of M_2, $BS(M_1)\geq BS(M_2)$, the theorem is proved.

Otherwise, $p\leq 1-\varepsilon$, assume that after sampling, the sequence of distribution difference degree changes, $BS(M_1)<BS(M_2)$. We decompose $p=p_{11}\times p_{12}$ ($p<p_{11}<1$, $p<p_{12}<1$), the result of sampling at p is equivalent to the result of sampling consequently at p_{11} and p_{12}, and there is exactly only one change in the BS sequence between the two samplings, supposing it is p_{11}.

Similarly, if $p_{11}\in(1-\varepsilon,\ 1]$, the sequence of BS should not change after sampling at p_{11}. The assumption is not correct.

If $p_{11}\leq 1-\varepsilon$, then decompose p_{11} again, and let the sampling rates which cause the sequence of BS value changed be p_{i1}.Since $p<p_{11}<p_{21}< \ldots <p_{i1}<\ldots<1$, there is $p_{i1}\rightarrow 1^-$. Then there must be a $p_{n1}\in(1-\varepsilon,\ 1]$, which makes $BS(M_1)>BS(M_2)$. It is also in conflict with the assumption. So the obvious degree sequence of distribution difference does not change after packet sampling. #

Theorem 3 shows that the sequence of distribution difference degree is independent to sampling rate. And it also guarantees that the sequence of characteristic selection is unique. Therefore, when sampling rate changes, it is not necessary to completely re-examine the current distribution difference of every metric, and re-select the characteristics. We just need to expand or curtail the current set of characteristic according to the sequence of BS value, which greatly simplifies the process of characteristic re-selection when sampling rate changes.

TABLE II. BS VALUE OF SOME METRICS AFTER PACKET SAMPLING

Metric	BS value		
	p=0.1	***p=0.01***	***p=0.001***
TCPflags	4.3E+4	9.3E+3	1.8E+3
pkt_size	1.2E+4	1.9E+3	2.3E+2
bytes	8.7E+3	1.3E+3	1.8E+2
pkts	6.3E+3	6.3E+2	1.3E+2
duration	7.9E+2	5.4E+2	44.8
Bps	7.7E+2	3.5E+2	38.5
pps	7.7E+2	3.2E+2	31.4
pkt_size_ratio	7.6E+2	2.4E+2	30.2
bytes_ratio	7.6E+2	2.1E+2	26.1
pkts_ratio	4.4E+2	14.3	12.7
head_size	1.5E+2	5.0	1.1

Table II shows the BS value of those metrics in Table I after packet sampling at different rate. Comparing the corresponding values in Table I and Table II, we can verify the correctness of the above theorems and inferences, as well as the analysis of the changes of distribution difference and BS value in packet sampling environment. Thus, the following conclusions can be drawn:

1. If the original distributions of a metric have no difference between applications, then there will be still no difference

after packet sampling, the ABSA method can still judge it should not be a characteristic.

2. If the original distributions of a metric have difference between applications, then the difference degree and the value of its BS will become small after packet sampling. And the lower the sampling rate is, the smaller its BS value is. Further more, there is a threshold of sampling rate, when the actual sampling rate is lower than that threshold, the distributions after sampling have no difference, the metric can not be used as a behavior characteristic of that application any more. The value of threshold varies with different applications and metrics.

3. In any sampling rate situation, the sequence of distribution difference degree can be guaranteed not changed. The sequence of characteristic selection is independent to sampling rate.

V. CONCLUSION

The current behavior-based application identification methods are not so satisfying in the aspects of overall accuracy and granularity, and lacks generality. The main reason is the lack of the knowledge on the behavior characteristic of every application. According to this situation, this paper proposes a method named ABSA to judge whether a behavior metric can be used as a behavior characteristic of an application, and further assess its significance. The ABSA method is based on chi-square test, and uses a heuristic algorithm for interval division, to ensure the stability of the result of χ^2 statistic. The method gets judgment result by voting among multiple degrees of freedom, to eliminate the impact of degree of freedom on χ^2 statistic. It also uses the critical quantile of every degree of freedom to average χ^2 statistic, which can guarantee the weights of χ^2 statistic under each degree of freedom are almost the same, in order to assess the significance of behavior characteristic precisely. The applicability of ABSA in packet sampling environment is also studied detailedly.

The theoretical analysis and experiments results show that, if the distributions of a metric between applications have difference, then the ABSA method can judge it being a behavior characteristic of that application correctly, and assess the characteristic significance according to its distribution difference reasonably. In packet sampling environment, the actual distribution difference degree becomes smaller than the original one, but ABSA can still guarantee the sequence of distribution difference degree not changed, which ensures the uniqueness of the characteristic selection sequence when sampling rate changes, and simplifies the process of characteristic re-selection.

REFERENCES

[1] Andrew W. Moore, Konstantina Papagiannaki. “Toward the Accurate Identification of Network Applications”. In: Proc. of *PAM 2005*. Boston, USA, 2005, pp. 41-54.

[2] Myung-Sup Kim, Young J. Won, James Won-Ki Hong. “Application-Level Traffic Monitoring and an Analysis on IP Networks”. *ETRI Journal*, 2005, vol. 27, no. 11, pp. 22-42.

[3] A. McGregor, M. Hall, P. Lorier, J. Brunskill. “Flow Clustering Using Machine Learning Techniques”. In: Proc. of *PAM 2004*. Antibes Juan-les-Pins, France, 2004, pp. 205-214.

[4] Thomas Karagiannis, Konstantina Papagiannaki, Michalis Faloutsos. “BLINC: Multilevel Traffic Classification in the Dark”. In: Proc. of *ACM SIGCOMM 2005*. Philadelphia, USA, 2005, pp. 229-240.

[5] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield. “Class-of-service mapping for QoS: A Statistical Signature-based Approach to IP Traffic Classification”. In: Proc. of *ACM SIGCOMM IMC 2004*. Taormina, Italy, 2004, pp. 135-148.

[6] Andrew W. Moore, Denis Zuev. “Internet Traffic Classification Using Bayesian Analysis Techniques”. In Proc. of *ACM SIGMETRICS 2005*. Banff, Canada, 2005, pp. 50-60.

[7] K. C. Claffy. “Internet traffic characterization”. San Diego: University of California, 1994.

[8] J. Pitkow. “Summary of WWW characterizations”. *World Wide Web*, 1999, vol. 2, no. 2, pp. 3~13.

[9] C. Dewes, A. Wichmann and A. Feldmann. “An Analysis of Internet Chat Systems”. In: Proc. of *ACM SIGCOMM IMC’03*. Miami Beach, Florida, USA, 2003, pp. 51-64.

[10] Louis Plissonneau, Jean-Laurent Costeux, Patrick Brown. “Analysis of Peer-to-Peer Traffic on ADSL”. In: Proc. of *PAM’05*. Boston, USA, 2005, pp. 69--82.

[11] Fabian Schneider, Sachin Agarwal, Tansu Alpcan, Anja Feldmann. “The New Web: Characterizing AJAX Traffic”. In Proc. of *PAM 2008*. Cleveland, USA, 2008, pp. 31-40.

[12] Cao Zhenhua, Zhao Ping, Hu Yueqing. The Theory of Probability and Mathematical Statistic. Nanjing: Southeast University Press. 2003.

[13] The eMule/eDonkey protocol. http://www.cs.huji.ac.il/labs/danss/p2p/resources/emule.pdf. 2005.

[14] CERNET. http://www.edu.cn/cernet_fu_wu_1325/index.shtml. 1994.

[15] quadong, sommere. SourceForge.net: Linux layer 7 packet classifier. http://sourceforge.net/projects/l7-filter. 2009.

[16] Cisco IOS NetFlow Introduction. http://www.cisco.com/warp/public/732/Tech/NetFlow. 2006.

[17] Huawei Technologies Co., Ltd. NetStream Technology White Paper. http://www.huawei.com/cn/products/datacomm/pdf/view.do?f=269. 2007.