International Journal of Grid and Distributed Computing
Vol.6, No.6 (2013), pp.103-116
http://dx.doi.org/10.14257/ijgdc.2013.6.6.09

An Adaptive Trust Sampling Method for P2P Traffic Inspection

Hongwei Chen, Dongyang Yu, Chunzhi Wang and Shuping Wang
School of Computer Science, Hubei University of Technology, Wuhan, China
chw2001@sina.com

Abstract

This paper focuses on the sampling-based Deep Packet Inspection for the traffic of P2P file sharing systems, especially for BitTorrent, and proposes a logarithmic-based Adaptive Trust Sampling (ATS) strategy for P2P traffic identification. In the whole process of sampling identification for P2P traffic, the sampling ratio of the current node in a P2P network can automatically adjust and dynamically vary according to the estimator of P2P traffic ratio of historical cycles. The experimental results show that the Adaptive Trust Sampling strategy can adapt to the dynamic change of sample size, effectively reduce the total sample size, mitigate the consumptions of system resources to some extent, and achieve the purpose of P2P traffic sampling.

Keywords: *Peer-to-Peer (P2P), Traffic Inspection, Simple Random Sampling (SRS), Adaptive Trust Sampling (ATS), Deep Packet Inspection (DPI)*

1. Introduction

In more recent years, Peer-to-Peer (P2P) technology has come into vogue in the application fields of file sharing, Internet television, and music download, and ermerged many popular P2P applications such as BitTorrent, PPLive, eMule, FastTrack, eDonkey, PPStream and KuGoo. P2P traffic has become the main growth force of Internet traffic, and has brought immeasurable impact on the traditional Internet routing architecture [1]. Compared with traditional C/S network architecture, P2P application traffic have such characters as long connection time, high transmission rate, and huge bandwidth possession so on, resulting in a decline in normal network performance [2]. In the past few years, P2P applications consume a large amount of Internet network bandwidth [3]. P2P traffic identification can make Internet service providers (ISPs) better manage large network and improve the quality of network service [4].

The existing P2P traffic identification technology mainly includes: well-known port identification, Deep Flow Inspection (DFI) and Deep Packet Inspection (DPI), *etc.* [5]. The early P2P applications used well-known ports for peer-to-peer communication, and now most P2P applications take dynamic ports to hide its traffic through firewall or avoid traffic monitoring devices. DFI indentifies P2P traffic in terms of some flow characters, such as average length of packets, average arriving interval of flow, the duration of the flow, and the ratio of upstream flow to downstream flow and so on, to judge whether the traffic fits P2P traffic characters [6-11]. DPI is a kind of traffic detection and control technology based on application layer, and its test unit is a single integrated data packet. First, the DPI method predefines some specific signatures of P2P application as a feature library, and then scans the data packets to match signatures in the feature library to identify whether the packet is a P2P packet in the whole scanning process [12]. The core of DPI-based P2P traffic identification

ISSN: 2005-4262 IJGDC

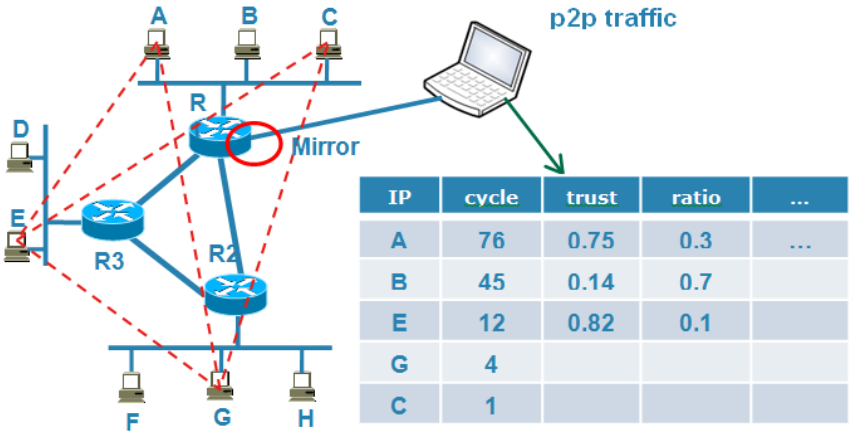
system is a multiple pattern string matching algorithm. At present, AC, Wu-Manber and SBOM are the most wildly used multiple pattern string matching algorithms. Because DPI methods scan the complete packet, it is difficult to scan all packets through ISP routers. At present, it is an effective means to recognize P2P traffic combined with sampling-based DPI methods [13-15].

On the basis of above AC matching algorithm, combined with the DPI sampling method and the trust strategy, an adaptive trust sampling (ATS) method is proposed in this paper. We set the trust value to the nodes of P2P networks, and increase sampling frequency to nodes with higher trust degree, while decreasing sampling frequency to nodes with lower trust degree. We can dynamically adjust some sampling characteristics to reduce the resource consumption of the system and assure the accuracy of sampling through this trust strategy. In this paper, Section 2 presents a system environment and proposes a framework for P2P traffic identification. Section 3 proposes an ATS strategy and algorithm. Then, we give experimental environment, sampling results and analyses in Section 4. Finally, Section 5 concludes the work.

2. System Environment and Proposed Framework

2.1. System Deployment Environment for P2P Traffic Inspection

Figure 1 is a schematic diagram of system deployment environment for P2P traffic inspection. There are three routers R1, R2 and R3. For example, if R1 wants to monitor P2P traffic status of its access nodes, R1 would deploy a mirror port and copy traffic through R1 to its mirror port. So, access nodes such as A, B and C would not be affected when R1 scans them. Then, we can connect the P2P traffic monitoring device to the mirror port. Theoretically, we can capture all packets through the router [15]. The internal of the P2P traffic monitoring device maintains a two-dimensional table for each access node of the router to allocate a record. Then, the relative parameters such as sampling cycle, trust degree value, sampling ratio, ratio estimator and so on are set up for each node. Due to the limitation of conditions, we only use one record to establish simulative experiment for a single node.



IP	cycle	trust	ratio	...
A	76	0.75	0.3	...
B	45	0.14	0.7	
E	12	0.82	0.1	
G	4			
C	1			

Figure 1. A Schematic Diagram of System Deployment Environment for P2P Traffic Inspection

2.2. An Adaptive Trust Sampling Framework for P2P Traffic Inspection

Figure 2 is an adaptive trust sampling framework diagram for P2P traffic inspection. When the system starts initialization, the strategy selection module selects strategy from the sample strategy set. According to selected strategy and initialized trust degree value, the sample ratio calculation module calculates the sample ratio of the first sampling cycle and delivers to the sampling packets module. Then, the sampling packets module extracts a portion of packets from the sniffer packets module according to sample ratio, and delivers to the DPI identification module. Next, the DPI identification module analyzes P2P packets in terms of the DPI feature lib and estimates the P2P traffic ratio of current sampling cycle. In the end, the trust calculation module calculates trust degree value of the next sampling cycle in terms of P2P traffic ratio of current sampling cycle, and delivers trust degree value of the next sampling cycle to the sample ratio calculation module. Hereto, the system enters the next sampling period.

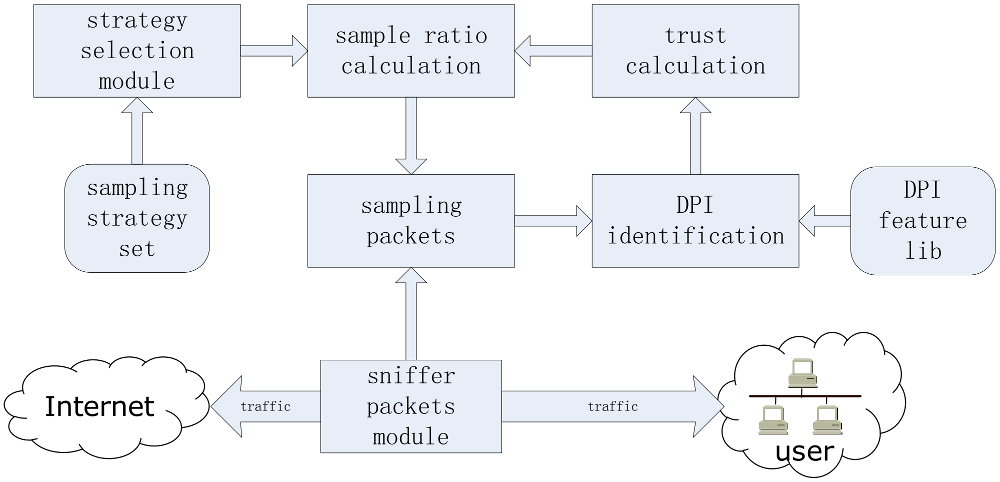


Figure 2. An Adaptive Trust Sampling Framework Diagram for P2P Traffic Inspection

3. Adaptive Trust Sampling Strategy and Algorithm

3.1. Sampling Strategies

At present, there are some mature sampling methods such as simple random sampling (SRS), stratified random sampling, multistage sampling and so on. We choose the SRS strategy and the linear systematic sampling strategy, and view each cycle as a whole sampling process.

(1) The Simple Random Sampling Strategy

The simple random sampling means each sample is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process. Supposed that the sample size is n, the number of population is N, the sampling ratio is $f = n / N$, and the theoretic or true proportion of P2P traffic is P. The n samples are formed a simple random sample while n samples is chosen. Estimator p is the approximate evaluation of population N proportion through sample observation, and p is an unbiased estimation of P. The mean square error $MES(p)$ of P is:

$$MSE(p) = V(p) = \frac{1-f}{n-1} p(1-p) \tag{1}$$

The confidence interval of estimator P under confidence coefficient $1-\alpha$ is:

$$p \pm (Z_{\alpha/2}\sqrt{MSE(p)} + 1/2n) \tag{2}$$

(2) The Linear Systematic Sampling Strategy

The main idea of linear systematic sampling is as follows: the random sequence of population N is labeled as 1,2…, N. Firstly, the sampling procedure randomly chooses a starting sample, then chooses the other samples according to the fixed interval k until n samples has been chosen.

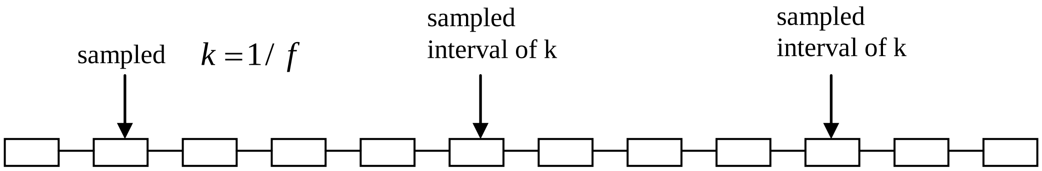


Figure 3. A Sketch Map of the Linear Systematic Sampling Strategy

p is an unbiased estimation of p. The mean square error $MES(p)$ of p has two formula modes. If the population N is a random sequence without any rules, then:

$$MSE(p) = V(p) = \frac{1-f}{n-1} p(1-p) \tag{3}$$

, else if the population N is a sequence with some rules, then:

$$MSE(p) = \frac{(N-1)}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2 \tag{4}$$

,where S^2 is population variance, and S_{wsy}^2 is variance of system group.

The population variance S^2 is:

$$S^2 = \frac{1}{N-1} \sum_{r=1}^{k} \sum_{j=1}^{n} (y_{rj} - \overline{Y})^2 \tag{5}$$

The variance of system group S_{wsy}^2 is:

$$S_{wsy}^2 = \frac{1}{k(n-1)} \sum_{r=1}^{k} \sum_{j=1}^{n} (y_{rj} - \overline{y}_r)^2 \tag{6}$$

,where $\overline{Y}$ is the theoretic overall average, and y_{rj} represents the j-th individual in the period of the r-th interval. The confidence interval of estimator p is:

$$p \pm (Z_{\alpha/2}\sqrt{MSE(p)} + 1/2n) \tag{7}$$

3.2. An Adaptive Trust Sampling Strategy

The main idea of P2P trust sampling is: the higher trust degree value of node, the least sampling ratio or sample size, vice versa. Usage of different sampling rate aiming at different situations not only can save system resources, but also enhance the sampling purpose. Based on the above idea, we propose an ATS strategy. Figure 4 is a sketch map of the ATS Cycles. In Figure 4, T, f and p respectively represent trust degree value, sampling ratio and estimator of ratio. Seen from Figure 4, the strategy presets trust degree value before a sample cycle, then starts sampling after sampling ration is determined, and obtains estimator of ratio and mean square error of estimator by sample observation. Before the start of the next sampling cycle, the strategy calculates trust degree value of the next sampling cycle in terms of estimator of the previous cycle, and determines the sampling ratio of the next sampling cycle, then starts sampling the next sampling cycle.

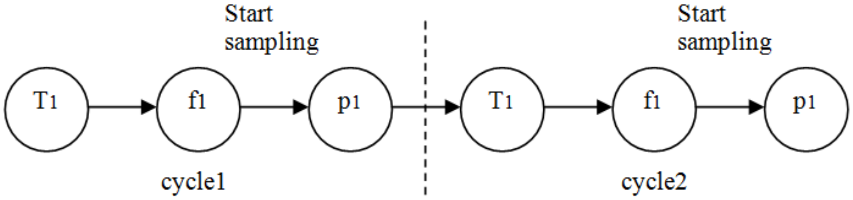


Figure 4. A Sketch Map of the Adaptive Trust Sampling Cycles

We can deduce the following functional relationships: formula 8 represents estimator of the previous sampling cycle p_{n-1} determines the trust degree value of the next sampling cycle T_n ; formula 9 represents sampling ratio f_n is determined by calculating the trust degree value of current sampling cycle T_n .

$$T_n = f(p_{n-1}) \tag{8}$$

$$f_n = g(T_n) \tag{9}$$

Formula 10 is the extension of formula 9:

$$f = g(p) = \frac{n}{N} = \frac{Z_{\alpha/2}^2}{(N-1)d^2 / p(1-p) + Z_{\alpha/2}^2} \tag{10}$$

We can extend formula 8 to achieve the ATS strategy purpose, the higher trust degree value of node, the least sampling ratio, vice versa.

$$T = f(p) = (N-1)d^2 / a\log_b^{(p+1)} \tag{11}$$

, where parameters a and b meet the condition of $a^2b = 1$.

$$f = g(T) = \frac{Z_{\alpha/2}^2}{T + Z_{\alpha/2}^2} \tag{12}$$

3.3. An Adaptive Trust Sampling Algorithm

Figure 5 is the flow chart of an adaptive sampling algorithm based on SRS method. Before starting sampling, the algorithm initializes trust degree value t, period counter N, and packet counter i. Then, the algorithm calculates sampling ratio f, and allocates dynamic sample pool space pos*(n). The algorithm uses reservoir sampling, and pre-allocates n positions in the sampling pool space. As the packets arrive, the corresponding mapping position represents the chosen sample of the packets. The reservoir sampling is realized based on Knuth sampling algorithm. Now, the traffic sampling begins. The counter i will add 1 when a packet flows, and the algorithm will judge whether counter i is equal to the position value of sampling pool. If equal, the packet will be chosen and delivered to the DPI module and the index of the sampling pool will plus 1, else the algorithm compares counter i and N for equality to judge whether the current sampling period is over. If i is less than N, the algorithm will continue monitoring the traffic, else the algorithm finished the current sampling cycle. Then, the algorithm calculates the estimator of ratio P in terms of the P2P traffic recognition results of the DPI module, resets the counter, and releases the sampling pool space. Then, the algorithm will process the next sampling period by calculating the trust degree value of the next sampling cycle according to estimator of ratio.

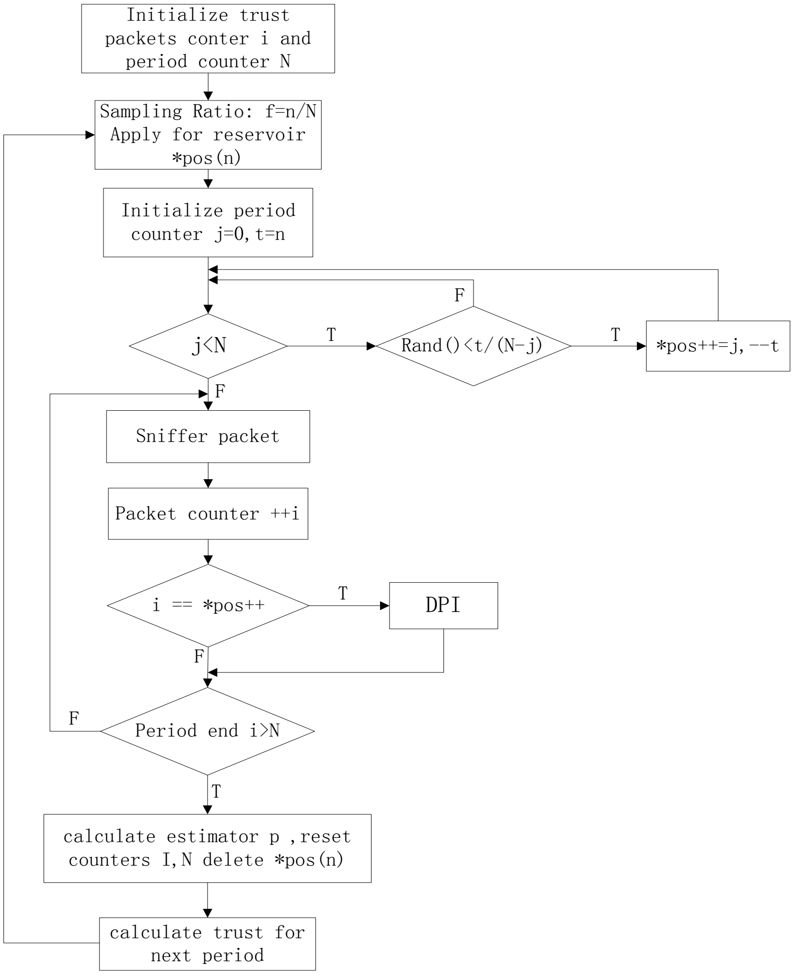

Initialize trust packets conter i and period counter N
Sampling Ratio: f=n/N Apply for reservoir *pos(n)
Initialize period counter j=0,t=n
j<N
T
Rand()<t/(N-j)
F
T
*pos++=j,--t
F
Sniffer packet
Packet counter ++i
i == *pos++
T
DPI
F
F
Period end i>N
T
calculate estimator p ,reset counters I,N delete *pos(n)
calculate trust for next period

Figure 5. The Flow Chart of an Adaptive Sampling Algorithm

4. Experiment Results and Analyses

4.1. Experimental Environment

We use a computer to simulate an ISP router, take P2P protocol BitTorrent as a test protocol, and utilize a BitTorrent-based P2P file sharing tool BitComet 0.59 to share 1G file in P2P network. Then, we use the test program developed by Visual C++ 6.0 to compare different P2P traffic sampling strategies such as SRS strategy and ATS strategy. The initialized parameters of the experiment are as follows. (1) SRS strategy: the absolute error threshold is d =0.05; the confidence coefficient is $1-\alpha$ =0.95; the population size is N =1000; the calculated sampling rate is f =0.286; the sample size is n =286. (2) ATS Strategy: the absolute error threshold is d =0.05; the confidence coefficient is $1-\alpha$ =0.95; the population size is N =1000; the initialized trust degree value is $T_1 = f(p_0 = 1.0)$; the calculated sampling rate is f_1=0.606; the sample size of the first sampling cycle is n_1 =606; the given lower limit of sample size in a cycle is n_m =30.

Table 1 is signatures of peer wire message protocol in BitTorrent. We use these signatures to establish DPI feature lib to recognize BitTorrent packets. Seen from the table, the starting position of signatures at BitTorrent packets is 0 or 3, and the length of signatures is lesser than 20 bytes. In order to reduce the system load and decrease the matching length, the sniffer packets module of the test program only extracts the first 40 bytes of all packets to deliver the DPI identification module.

Table 1. Signatures of Peer Wire Message Protocol in BitTorrent

Message Type	Protocol	Position	Feature Code	Length
handshake	tcp	0	13426974546f7272656e 747270726f746f636f6c	19
bitfield	tcp	0	000000e405	5
con data	tcp	0	0000000ea0	5
	tcp	0	0000000da1	5
	tcp	0	0000000da0	5
	tcp	0	0000000101	5
	tcp	0	0000000102	5
request	tcp	0	0000000d06	5
interested	tcp	0	0000000102	5
piece	tcp	0	0000400907	5
have	tcp	0	0000000504	5
d1:ad2	udp	3	64313a616432	6
d1:rd2	udp	3	64313a726432	6

4.2. Sampling Results and Analyses

The experiment captures about 333,000 packets. We set 1000 packets as a sampling cycle, so there are 333 sampling cycles in the experiment. Figure 6 and Figure 7 show that the true P2P ratio, the SRS ratio estimator of P2P ratio and the ATS ratio estimator of P2P ratio. In order to improve the discrimination of three fold lines, the ordinate values of the ATS ratio estimator of P2P ratio all plus 1 in Figure 6 and Figure 7, while the ordinate values of true P2P ratio plus 2 in Figure 6 and Figure 7. In the macroscopic view, there is not obvious variation among the true P2P ratio, the SRS ratio estimator of P2P ratio and the ATS ratio estimator of P2P ratio.

Figure 8 and Figure 9 show that the absolute error among true P2P ratio, the SRS ratio estimator of P2P ratio and the ATS ratio estimator of P2P ratio. Seen from figure 8 and figure 9, both the absolute errors of the SRS ratio estimator of P2P ratio and the ATS ratio estimator of P2P ratio are almost less than 0.05.

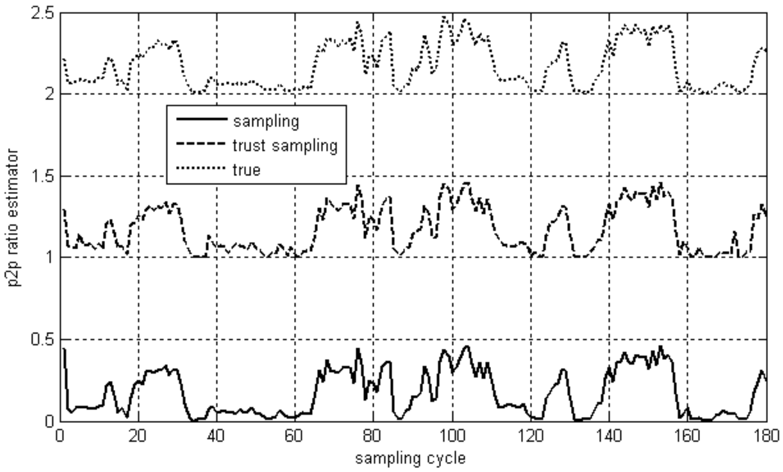


Figure 6. The Variation Diagram of P2P Ratio Estimator with Respect to Sampling Cycles

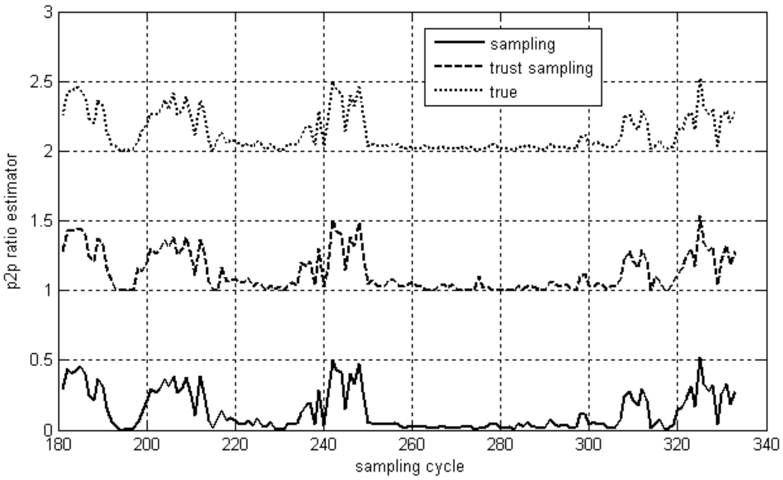


Figure 7. The Variation Diagram of P2P Ratio Estimator with Respect to Sampling Cycles (continue)

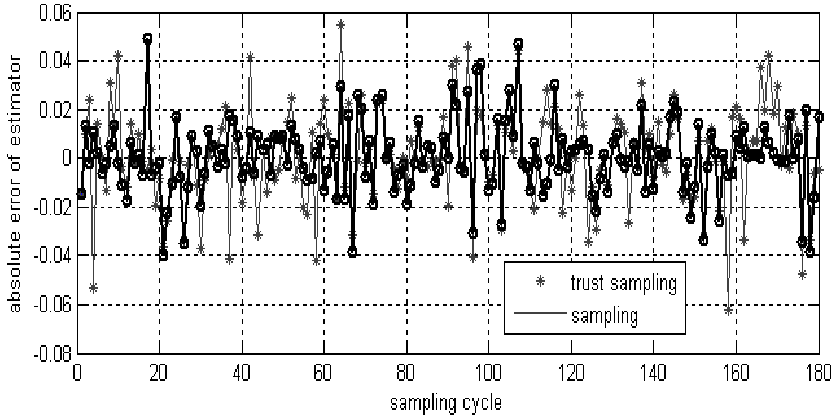


Figure 8. The Variation Diagram of Absolute Error of P2P Ratio Estimator with Respect to Sampling Cycles

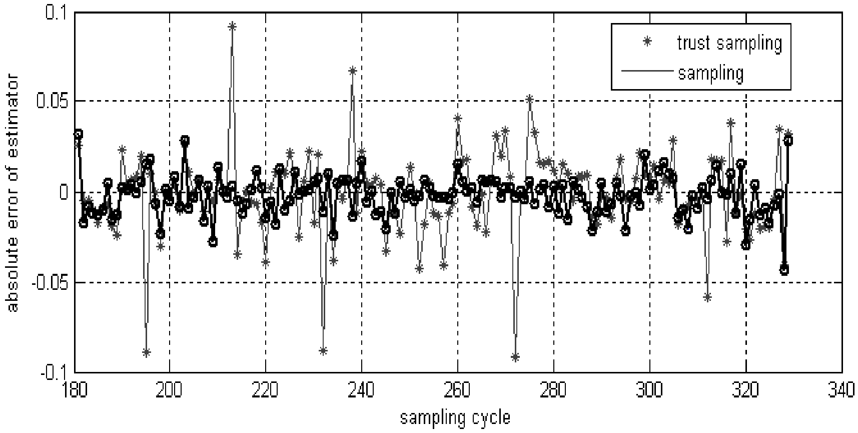


Figure 9.The Variation Diagram of Absolute Error of P2P Ratio Estimator with Respect to Sampling Cycles (continue)

Compared with SRS strategy, during the period of sampling process from the 181st cycle to the 333rd cycle, the absolute errors of P2P ratio estimator of ATS strategy appear obvious fluctuations. Through analysis and observation, the sample size is the key factor to lead to obvious fluctuations of the absolute errors of P2P ratio estimator of ATS strategy. The sample size of the ATS strategy in each sampling cycle ranged from 30 to 606. We can obtain function relationship between sample size n and absolute error d by calculating the mean values of absolute error d with respect to sample size n. Figure 10 is the variation diagram of absolute error of estimator with respect to sample size ranged from 30 to 666. Seen from Figure 10, the absolute errors of the ratio estimator decrease followed with the increasing of the sample size of each sampling cycle. As the sample size comes to the lower limiting value n=30, the average absolute error reaches the peak value 0.12. It is difficult to deal with the fluctuation of absolute error of estimator, because the sample size n is a dynamic value during the process of ATS, and the sample size of current period is determined by the trust degree value of the previous cycle.

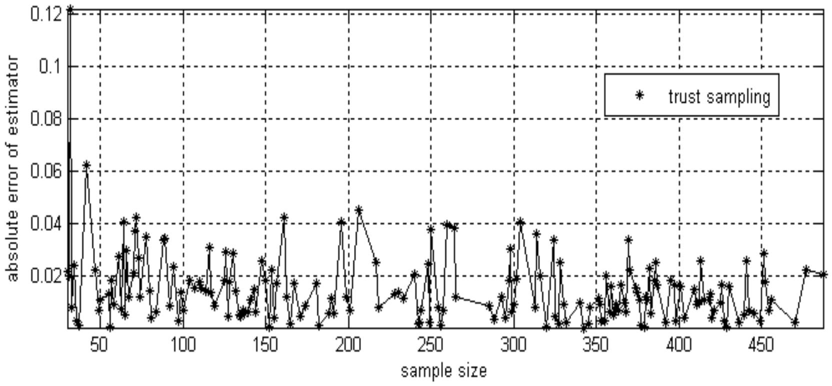


Figure 10. The Variation Diagram of Absolute Error of Estimator with Respect to Sample Size

Figure 11 is the variation diagram of absolute error of estimator with respect to P2P ratio estimator. The horizontal ordinate represents P2P ratio estimator ranged from 0.001 to 0.524, and the vertical coordinate represents the average absolute error of estimator corresponding to the SRS and the ATS. Seen from Figure 11, as the P2P ratio is lower than 0.225, the fluctuation of absolute error of ATS is obvious, and vice versa.

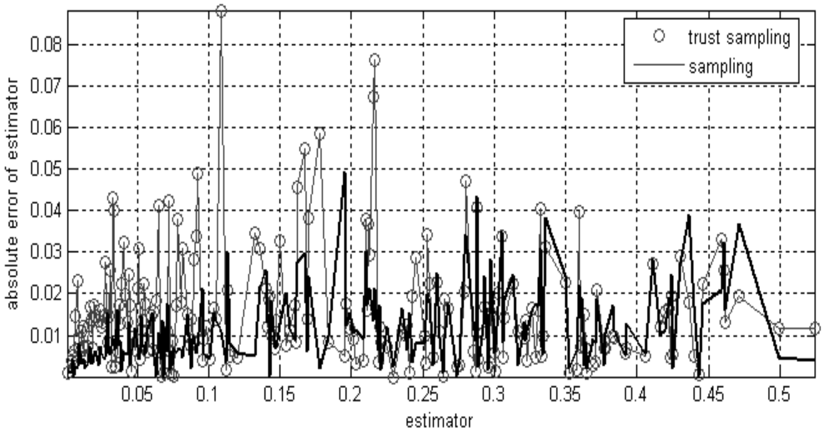


Figure 11. The Variation Diagram of Absolute Error of Estimator with Respect to P2P Ratio Estimator

Figure 12 is the diagram of accumulation of sample packets with respect to sampling cycles. In fact, the sample size of the SRS is constant (n=286), so the accumulative sample sizes of the SRS can be expressed a line in Figure 12. Compared with the SRS, the accumulative sample sizes of the ATS has obviously slow rising trend. The core of P2P traffic identification system based on DPI is a multiple pattern matching algorithm of a string. It is well known that the DPI algorithm has high real-time performance analysis requirements. We apply the AC multiple pattern matching algorithm to the test program and its time complexity is O (n), and n represents to the matching length of the string [16]. Although the data from Figure 12 can't reflect the time complexity of SRS strategy and ATS strategy, it can reflect that the sample packets processing time of the ATS-based DPI algorithm is less than that of

SRS-based DPI algorithm. Compared with the SRS-DPI algorithm, the ATS-based DPI algorithm can effectively increase the sampling efficiency.

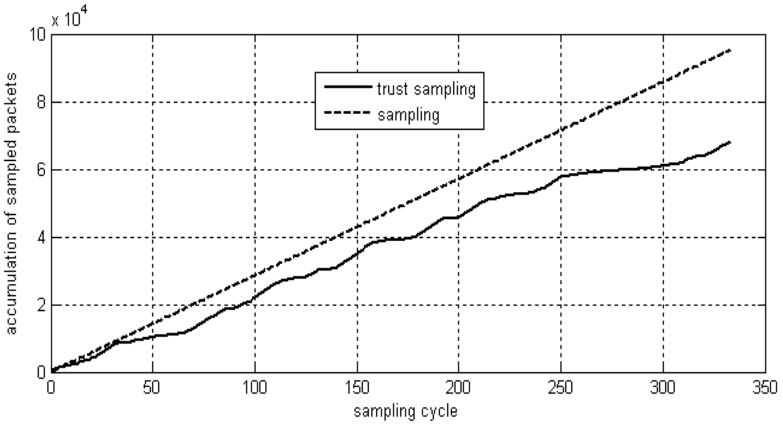


Figure 12. The Diagram of Accumulation of Sample Packets with Respect to Sampling Cycles

5. Conclusions

This paper briefly introduces the popular P2P traffic identification technology, and presents a system deployment environment and an ATS framework for P2P Traffic Inspection. Then, on the basis of some sample strategies such as the SRS strategy and the linear systematic sampling strategy and a logarithmic trust model, an ATS strategy and algorithm are presented. Finally, we build an experimental environment and sampling results show that, compared with the simple sampling method, on the premise of basic accuracy assurance, the ATS method can adapt to the dynamic change of sample size, effectively reduce the total sample size, mitigate the consumption of system resources to some extent, and achieve the purpose of P2P traffic sampling.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (No. 61170135, No. 61202287), the General Program for Natural Science Foundation of Hubei Province in China (No. 2013CFB020, No. 2010CDA011, No. 2011CDB075, No. 2012FFB00601), the Key Project for Scientific and Technological Research of Wuhan City in China (No. 201210421134).

References

[1] W. -m. Hong, "A Novel Method for P2P Traffic Identification", Procedia Engineering, vol. 23, (**2011**), pp. 204-209.

[2] J. Seibert, R. Torres, M. Mellia, M. M. Munafo, C. Nita-Rotaru and S. Rao, "The Internet-Wide Impact of P2P Traffic Localization on ISP Profitability", IEEE/ACM Transactions on Networking, vol. 20, no. 6, (**2012**), pp. 1910 -1923.

[3] J. Yang, L. Yuan, Y. He and L. -y. Chen, "Timely traffic identification on P2P streaming media", The Journal of China Universities of Posts and Telecommunications, vol. 19, no. 2, (**2012**), pp. 67-73.

[4] R. Keralapura, A. Nucci and C. -N. Chuah, "A novel self-learning architecture for p2p traffic classification in high speed networks", Computer Networks, vol. 54, no. 7, **(2010)**, pp. 1055-1068.

[5] J. V. Gomes, P. R. M. Inácio, M. Pereira, M. M. Freire and P. P. Monteiro, "Detection and Classification of Peer-to-Peer Traffic: A Survey", ACM Computing Surveys, vol. 45, no. 3, (**2013**), pp. 1-40.

[6] K. Xu, M. Zhang, M. Ye, D. M. Chiu and J. Wu, “Identify P2P traffic by inspecting data transfer behavior”, Computer Communications, vol. 33, no. 10, (**2010**), pp. 1141-1150.

[7] B. Xu, M. Chen and X. Wei, “Hidden Markov model-based P2P flow identification: a hidden Markov model-based P2P flow identification method”, IET Communications, vol. 6, no. 13, (**2012**), pp. 2091 - 2098.

[8] P. Bermolen, M. Mellia, M. Meo, D. Rossi and S. Valenti, “Abacus: Accurate behavioral classification of P2P-TV traffic”, Computer Networks, vol. 55, no. 6, (**2011**), pp. 1394-1411.

[9] M. Iliofotou, H. -c. Kim, M. Faloutsos, M. Mitzenmacher, P. Pappu and G. Varghese, “Graption: A graph-based P2P traffic classification framework for the internet backbone”, Computer Networks, vol. 55, no. 8, (**2011**), pp. 1909-1920.

[10] H. Chu, H. Yi and X. Zhang, “A new P2P traffic identification methodology based on flow statistics”, 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN), IEEE Press, (**2011**), pp. 277 - 281.

[11] G. Silvestre, S. Fernandes, C. Kamienski and D. Sadok, “Most Wanted Internet Applications: A Framework for P2P Identification”, 2010 Eighth Annual Communication Networks and Services Research Conference (CNSR), IEEE Press, (**2010**), pp. 341 - 347.

[12] N. Cascarano, L. Ciminiera and F. Risso, “Optimizing Deep Packet Inspection for High-Speed Traffic Analysis”, Journal of Network and Systems Management , vol. 19, no. 1, (**2011**), pp. 7-31.

[13] L. Chen, P. Lin, R. Cong and Y. Qiao, “Traffic classification method of P2P streaming application based on sampling technology”, 2010 2nd IEEE International Conference on Network Infrastructure and Digital Content, IEEE Press, (**2010**), pp. 184-189.

[14] W. Zhong, B. Raahemi and J. Liu, “Classifying peer-to-peer applications using imbalanced concept-adapting very fast decision tree on IP data stream” , Peer-to-Peer Networking and Applications, vol. 6, no. 3, (**2013**), pp. 233-246.

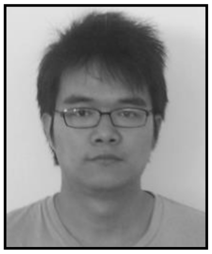
[15] B. Li, M. Ma and Z. Jin, “A VoIP Traffic Identification Scheme Based on Host and Flow Behavior Analysis”, Journal of Network and Systems Management, vol. 19, no. 1, (**2011**), pp. 111-129.

[16] D. Cantone, S. Faro and E. Giaquinta, “On the bit-parallel simulation of the nondeterministic Aho–Corasick and suffix automata for a set of patterns”, Journal of Discrete Algorithms, vol. 11, (**2012**), pp. 25-36.

Authors

Hongwei Chen

In 2006, he graduated from Nanjing University of Posts & Telecommunications and received PHD degree in China, majored in Communication and Information System. Now, he is an associate professor at School of Computer Science in Hubei University of Technology, Wuhan, China. From August of 2013, he is an academic visiting scholar at Temple University in USA. Currently, his major field of study is Peer-to-Peer Computing, Cloud Computing, Grid Computing, Network Security and Mobile Agent. He is a member of CCF, ACM and IEEE, and a paper reviewer for SCI journals such as Peer-to-Peer Networking and Applications, Knowledge-Based Systems, and International Journal of Network Management.



Dongyang Yu

He is from Hubei Province of China, master candidate of Hubei University of Technology, interested in Peer-to-Peer and Cloud Computing.

Chunzhi Wang

She is from Hubei province of China, PHD, Professor and dean at School of Computer Science, in Hubei University of Technology. She is interested in Peer-to-Peer, and network security. She is member of CCF, ACM and IEEE.

Shuping Wang

He is from Hubei Province of China, master candidate of Hubei University of Technology, interested in Peer-to-Peer, Game Theory and Information Security.