International Journal of Innovative Computing, Information and Control
Volume **6**, Number **10**, October **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 4435–4448

IDENTIFYING HIGH-RATE FLOWS WITH USER-SPECIFIED ACCURACY

Yu Zhang[1,2] and Binxing Fang[1,2]

[1]Research Center of Computer Network and Information Security Technology
Harbin Institute of Technology
No. 92, West Da-Zhi Street, Harbin 150001, P. R. China
yuzhanghit@gmail.com

[2]Research Center of Information Intelligence and Information Security
Institute of Computing Technology
Chinese Academy of Sciences
No. 6, Kexueyuan South Road, Haidian District, Beijing 100190, P. R. China
bxfang@ict.ac.cn

Received May 2009; revised September 2009

Abstract. *Identifying high-rate flows is important for active queue management, traffic measurement and network security. Explicit measurement of high-rate flows is difficult because tracking the possible millions of flows needs correspondingly large high-speed memories. To reduce the measurement overhead, the deterministic* 1*-out-of-*k *sampling technique is adopted. Since the sampled packets are only a part of the whole traffic transmitted, it is critically important to identify high-rate flows correctly. However, there are no methods which are able to specify the identification accuracy. We develop two such methods. The first approach is based on Bayesian single sampling method (BSS) which is able to identify high-rate flows with user-specified false positive rate (*FPR*) and false negative rate (*FNR*). However, since BSS has to record every sampled flow during the measurement period, it is not efficient for memory. Therefore, the second novel approach based on Bayesian double sampling method (BDS) is proposed. BDS can remove the low-rate flows and identify the high-rate flows at the first sampling stage which can reduce the memory cost and identification time respectively. The experimental results show that both BSS and BDS can identify high-rate flows with user-specified* FPR *and* FNR*, moreover, BDS outperforms BSS in terms of less memory cost and identification time.*

Keywords: Traffic monitoring, High-rate flow, Identification, Bayes' theorem

1. Introduction.

The current Internet has no mechanism for controlling the throughput of each flow, which is however performed by end hosts using TCP. The UDP flows or malicious TCP flows which do not obey the TCP flow control mechanism will not reduce their packet-sending rates even when packet dropping is detected, therefore these flows are more likely to consume a large share of the link bandwidth [1]. A number of active queue management methods [2, 3, 4, 5] have been proposed to solve this problem and provide fairness in networks. The main idea is to identify high-rate flows and selectively drop their packets during times of congestion. A naive approach to identify high-rate flows is to keep per-flow counter for each arriving flow and identify flows of which the counter is bigger than a pre-specified threshold. However, it is difficult to directly identify high-rate flows in backbone links because tracking the possible millions of flows needs correspondingly large high-speed memories. To reduce the measurement overhead, the deterministic 1-out-of-k sampling technique is adopted which is widely used in today's operational networks, for instance, it has been implemented in Cisco routers. However, since the sampled packets