논문 2008-45TC-6-6

수동적 인터넷 측정을 위한 샘플링 기법 비교: 사례 연구를 통한 검증

(Comparison of Sampling Techniques for Passive Internet Measurement: An Inspection using An Empirical Study)

김 정 현*, 원 유 집**, 안 수 한***

(Junghyun Kim, Youjip Won, and Soohan Ahn)

요 약

인터넷이 일상생활에서 중요한 위치를 차지함에 따라 인터넷에서 발생되는 트래픽의 특성을 밝히는 것은 매우 중요한 연구 과제로 주목을 받고 있다. 그러나 인터넷 트래픽은 대용량이므로 쉽게 다룰 수 없다. 이러한 문제는 인터넷 트래픽 측정 연구에 가장 큰 장애다. 많은 연구자들은 다양한 샘플링 기법을 통해 트래픽을 다룰 수 있는 양으로 샘플링하여 분석하고 있다. 본 연구에서는 기존의 인터넷 측정 연구에서 사용된 샘플링 기법을 비교 분석하고, 가장 효과적인 샘플링 방안을 제시하고자 한다. 연구에 비교 사용된 샘플링 기법은 규칙적 샘플링, 단순 랜덤 샘플링, 층화 샘플링이며, 샘플링 단위는 1/10, 1/100, 1/1000을 사용하였다. 분석한 항목은 트래픽 크기 분석, 엔트로피 분석, 패킷 크기 분석이다. 단순 랜덤 샘플링은 무난한 결과를 보였고, (간격을 패킷 개수로 설정한) 규칙적 샘플링은 대상과 샘플링 강도에 상관없이 고른 결과를 보였다. 한편, 간격을 시간으로 설정한 규칙적 샘플링은 매우 좋지 않을 결과를 나타내었다. 전송층 프로토콜을 기준으로 층화 샘플링 수행할 경우 더욱 좋은 결과를 얻을 수 있었다. 연구 결과를 통해 샘플링 기법이 시간에 따른 트래픽의 흐름을 얼마나 잘 유지하는가가 샘플링 성능을 좌우함을 알 수 있었다. 또한 엔트로피 분석은 샘플링에 강하고, 이상 트래픽 탐지에 매우 적절함이 확인되었다. 그러나 병목 현상에 의한 트래픽 크기 감소는 잘못된 엔트로피 분석 결과를 유발할 수 있음을 발견하였다. 마지막으로, 패킷 크기 분포는 패킷 샘플링 방식이나 강도에 영향을 받지 않음을 발견하였다.

Abstract

Today, the Internet is a part of our life. For that reason, we regard revealing characteristics of Internet traffic as an important research theme. However, Internet traffic cannot be easily manipulated because it usually occupy huge capacity. This problem is a serious obstacle to analyze Internet traffic. Many researchers use various sampling techniques to reduce capacity of Internet traffic. In this paper, we compare several famous sampling techniques, and propose efficient sampling scheme. We chose some sampling techniques such as Systematic Sampling, Simple Random Sampling and Stratified Sampling with some sampling intensities such as 1/10, 1/100 and 1/1000. Our observation focused on Traffic Volume, Entropy Analysis and Packet Size Analysis. Both the simple random sampling and the count-based systematic sampling is proper to general case. On the other hand, time-based systematic sampling exhibits relatively bad results. The stratified sampling on Transport Layer Protocols, e.g.. TCP, UDP and so on, shows superior results. Our analysis results suggest that efficient sampling techniques satisfactorily maintain variation of traffic stream according to time change. The entropy analysis endures various sampling techniques well and fits detecting anomalous traffic. We found that a traffic volume diminishment caused by bottleneck could induce wrong results on the entropy analysis. We discovered that Packet Size Distribution perfectly tolerate any packet sampling techniques and intensities.

Keywords : Internet Measurement, Sampling Technique, Entropy Analysis, Anomaly Detection

---

* 학생회원, ** 정회원, 한양대학교 전자컴퓨터통신공학과
(Dept. of Electronics and Computer Engineering, Hanyang University)
*** 비회원, 서울시립대학교 통계학과
(Dept. of Statistics, University of Seoul)
※ 본 연구는 한국학술진흥재단 문제해결형인력양성지원사업 (KRF-2006-511-D00370)과 과학기술부/한국과학재단 우수연구센터 육성사업 (R11-2000-073-00000)의 지원으로 수행되었음.
접수일자: 2008년4월3일, 수정완료일: 2008년6월19일

Ⅰ. 서 론

1990년대에 WWW (World Wide Web)의 등장으로 일반 사용자에게도 인터넷의 사용이 급격히 늘어나게 되었다. 2000년대에는 인터넷을 통한 쇼핑, 교육, 전화, 방송 등의 새로운 서비스가 보급되면서 인터넷은 생활에 없어서는 안 될 필수요소가 되었다. 그러나 인터넷의 보급과 사용의 확대에 비해, 인터넷에서 발생되는 트래픽에 대한 연구가 많이 부족한 실정이다[1]. 세계적으로 인터넷 측정 (Interment Measurement) 연구가 본격적으로 시작된지 10여년에 불과하다. 새로운 인터넷 서비스의 등장은 새로운 형태의 트래픽을 생성하게 되지만, 아직까지 심도있는 트래픽 특성 연구는 일부에 불과하다[1]. 많은 네트워크 구성이 트래픽의 특성을 고려하지 않고 있으며 다양한 문제점이 나타나고 있다. 특히 악의적인 목적으로 만들어진 트래픽들이나 의도적이지는 않지만 네트워크 구성요소에게 부정적인 영향을 주는 이상 트래픽 (Anomalous Traffic)도 매우 다양하게 나타나고 있다[1~4]. 그러나 소수의 공격도구나 웜바이러스 (Worm Virus)[5~7]만이 세부적으로 분석된 상황이다. 예를 들어, 2000년에는 DoS (Denial of Service) 공격에 의해 Yahoo, eBay, E*trade가 피해를 입었고, 2001년에는 유사한 DoS 공격으로 인해 MS의 네임서버 (Name Server)가 다운되었다[2]. DoS 공격에 대한 많은 대응책 연구로 인해, 현재는 일정 수준이하의 DoS 공격에는 충분히 대응할 수 있는 상황이다[3]. 또다른 예로, 2003년에는 슬래머웜 (Slammer Worm)[5, 7]이 전 세계적으로 큰 피해를 주었다. 특히 슬래머웜에 의해 국내의 인터넷이 마비되는 사건이 비슷한 시기에 발생하기도 했다. 슬러머웜에 대한 심도있는 분석과 대응책을 통해, 현재는 이와 관련한 문제가 거의 발생하지 않게 되었다. 아직도 다양한 웜바이러스와 그 변종들이 만들어지고 있으며, 대부분 상세하게 분석되지 않은 상황이다. 다양한 웜바이러스들이나 DoS 공격 도구들이 발생시키는 이상 트래픽이 심도있게 분석되고, 미리 대응할 수 있다면 인터넷은 더욱 안정적으로 운영될 수 있을 것이다.

인터넷 측정 연구의 중요성은 명확하지만 인터넷 측정을 어렵게 하는 문제는 아직까지 완전히 해결되고 있지 못하다. 인터넷 측정 방식은 크게 능동적 방식 (Active Method)과 수동적 방식 (Passive Method)으로 나눌 수 있으며, 수동적 방식에서 명확히 해결해야할 트래픽 용량문제가 있다[1, 8]. 트래픽을 임의로 생성하여 어떤 측정 결과를 얻는 능동적 방식(Active Method)은 네트워크의 성능을 측정할 때 유용하다. 다행히 능동적 방식에서는 특별히 감당하기 어려운 문제는 발생하지 않는다. 반면에, 트래픽을 캡쳐하여 트래픽 자체의 특성을 연구하는 수동적 방식에서는, 캡쳐된 트래픽의 용량을 감당하기 쉽지 않다는 것이 큰 문제다. 해결책으로, 감당할 수 있을 정도로 트래픽을 샘플링하고 처리하여 연구결과를 얻고 있다. 일반적으로 많이 사용되는 샘플링 기법은 단순 랜덤 샘플링 (Simple Random Sampling)[8, 11], 규칙적 샘플링 (Systematic Sampling) 등[8, 12, 18]이 있다. 이외에도 다양한 방식이 존재하지만, 실제 인터넷 측정 연구에서 사용되는 것들은 위와 같은 단순한 방식들이다. 그리고 샘플링 강도는 1/10, 1/100, 1/1000 등[8~12, 18]이 주로 사용되고 있다. 인터넷 측정 연구에 샘플링이 많이 사용되고 있음에도 샘플링에 관한 검증 연구는 흔하게 볼 수 없는 실정이다. 최근 몇년간 샘플링 기법을 검증한 연구를 살펴보면, 2004년 N. Duffield가 수동적 방식을 위한 샘플링 기법의 전반적 사항을 소개하였다[8]. 샘플링에 대한 전반적 소개와 분석은 충분하지만, 실증적 내용이 부족하다. 2006년에는 이상 탐지 (Anomaly Detection)와 샘플링에 관한 몇가지 연구가 발표되었다. D. Brauckhoff는 랜덤 샘플링 (Random Sampling)을 기본으로 샘플링 강도 (Sampling Intensity)가 이상 탐지에 어떤 영향을 미치는지 분석하였으며[11], J. Mai는 기존에 소개된 트래픽 샘플링 방식이 이상 탐지에 어떤 영향을 주는지 분석하였다[12]. 두 연구는 주로 샘플링이 이상 탐지에 어떤 영향을 주는지 분석하였으며, 플로우 (Flow)와 패킷의 수 (Packet Count)에 중점을 두었다. 아직 부족한 부분은 패킷 크기(Packet Size)와 관련된 분석이 부족하다는 점이다.

본 연구에서는 이상 탐지를 위한 짧은 시간 단위의 인터넷 측정 연구에서 샘플링이 어떠한 영향을 미치는지 살펴본다. 연구에 사용된 방식은, 적용과 구현이 쉬워 많이 사용되는 단순 랜덤 샘플링, 규칙적 샘플링, 층화 샘플링을 선택하였다. 샘플링 강도는 1/10, 1/100, 1/1000를 선택하였다. 연구 대상 데이터는 우리나라의 백본망의 트래픽을 캡쳐한 것이다. 캡쳐된 데이터 중에서 UDP 기반의 DoS 공격이 발견된 구간을 위주로 샘플링의 영향을 주로 분석한다. 세부 분석 항목은 "트래픽 크기 분석", "엔트로피 분석", "패킷 크기 분석"이다. 트래픽 크기 분석에서는 시간 기반 (Time-based) 샘플링 방식들은 트래픽 분석에 부적합한 것을 검증하였다.

연구 과정에서, 층화 샘플링 방식이 프로토콜 단위의 모니터링에 유리한 것을 발견하였다. 엔트로피 분석에서는, 기존 연구결과와 같이 샘플링 강도가 엔트로피에 적은 영향을 주는 것을 확인하였고, 이상 트래픽으로 인한 잘못된 엔트로피 분석의 가능성을 실증적 데이터를 통해 알 수 있었다. 패킷 크기 분석에서는 패킷 크기 분포가 샘플링 방식과 강도에 상관없이 잘 표현됨을 알 수 있었다. 이는 다른 연구에서는 아직까지 확인되지 않는 결과다.

본 논문의 Ⅱ장에서는 인터넷 측정 방식과 트래픽 샘플링 방식에 대해 간단히 소개를 한다. Ⅲ장에서는 트래픽 크기 분석과 엔트로피 분석을 통해 샘플링 방식과 강도에 따른 트래픽의 변화를 살펴본다. 특히 패킷 크기 분석으로 통해서는 패킷 크기 분포는 샘플링의 영향을 거의 받지 않음을 보인다.

Ⅱ. 인터넷 측정과 트래픽 샘플링

본 장에서는 인터넷 측정 (Internet Measurement)과 트래픽 샘플링 (Traffic Sampling) 기법을 소개하고 각각의 방식의 장단점을 간단히 비교한다.

1. 인터넷 트래픽 측정의 필요성

인터넷 트래픽 측정에 대해서 많은 관심이 집중되고 있다. 이러한 이유는 상업적 (Commercial), 사회적 (Social), 기술적 (Technical)인 이유가 있다[1].

가. 상업적 이유 (Commercial Reasons)

인터넷에서 제품을 판매하거나 정보를 제공하는 경우, 인터넷 측정으로부터 매우 유용한 정보를 얻어 대응할 수 있다. 예를 들면, 인터넷 트래픽으로부터 사용자 통계를 얻어 활용할 수 있다. "얼마나 많은 사용자가 어떤 서버에 어떤 서비스를 받기 위해서 접속하는지?", "어떤 사용자가 어떤 경로 (초고속 통신망, 전화망 등)로 인터넷 서비스에 접속하는지?", "무선 인터넷 사용자의 인터넷 사용 패턴은 어떠한지?" 등을 인터넷 트래픽 측정으로부터 알 수 있다.

효과적인 인터넷 상거래를 위해서는 네트워크 성능 특성 (Network's Performance Properties)에 대한 이해도 매우 중요하다. 예를 들어, "기업 사이트 (Vendor's Site)에서 웹페이지를 다운로드 받는 시간이 얼마인가?", "얼마나 자주 네트워크 문제 (Network Problems)가 발생하여 정상적인 데이터 전송에 지장을 주는가?"와 같은 질문들은 모든 인터넷 기반 기업에게 매우 중요한 정보다.

나. 사회적 이유 (Social Reasons)

정부나 연구자, 기업은 인터넷 사용의 사회적 영향에 대한 정보를 원한다. 즉, 다양한 사이트와 프로토콜의 동작이 발생시키는 트래픽의 통계는 사회적 이슈에 대한 중요한 판단을 가능하게 한다. 예를 들어, 유명한 웹사이트는 엄청난 양의 트래픽을 발생시킨다. 이는 해당 사이트가 사회적인 영향을 크게 미치고 있음을 판단하는 기준이 될 수 있다. 또한 해당 사이트의 주제와 내용은 현재 사회적 이슈가 되고 있음을 예측할 수 있는 근거자료가 된다.

다. 기술적 이유 (Technical Reasons)

네트워크 구성요소와 프로토콜 설계는 반드시 인터넷 워크로드 (Internet Workload)에 따라서 달라져야 한다. 예를 들면, 라우터 설계는 네트워크 트래픽의 통계적 특성과 패킷 크기 분포 (Packet Size Distribution)에 크게 영향을 받는다. 또한 웹페이지 (Web Pages)의 통계적 특성은 웹서버 (Web Servers)와 웹브라우저 (Web Browsers)의 성능과 설계에 영향을 준다. 즉, 인터넷 측정 결과는 향후 개발될 네트워크 구성요소와 관련 응용프로그램들의 설계에 큰 영향을 주게 된다.

2. 인터넷 측정 방식

인터넷 측정의 방식에는 크게 능동적 방식 (Active Method)과 수동적 방식 (Passive Method)이 있다[1, 8].

가. 능동적 방식 (Active Method)

인터넷 측정을 위해서 패킷 (Packet)을 사용한다. 두 호스트 (Host) 간에 패킷을 발생시켜 네트워크의 성능 측정을 하는 방식이 대표적인 예가 될 수 있다. 이러한 능동적 방식의 인터넷 측정을 위해서 사용되는 일반적인 도구에는 Ping[14]과 Traceroute[15]가 있다. Ping을 이용할 경우, 호스트 간의 경로의 RTT (Round Trip Time) 측정을 통해서, 패킷 전달 지연시간 측정할 수 있다. 그러나 세부적인 전달 경로는 매번 달라질 수 있다. Traceroute를 이용하면 어떤 호스트에서 또다른 호스트까지 패킷이 어떤 네트워크 경로로 전달되는지를 검사할 수 있다. 능동적 방식을 통해서 연구된 네트워크 이상 현상으로, PFL (Persistent Forwarding Loop)[13]이 있다. 이는 라우터의 버그로 인해서 발생되

는 문제로써, Traceroute를 통해서 분석되었다.

나. 수동적 방식 (Passive Method)

인터넷 사용자들이 발생시킨 트래픽을 캡처하여, 트래픽 자체의 특성을 분석하는 방식이다. 즉, 대량의 트래픽을 저장소에 저장하고, 트래픽 다양한 특성을 연구한다. 트래픽을 저장하는 방식에는 패킷 저장 방식과 플로우 (Flow) 저장 방식이 있다. 패킷 저장 방식은 인터넷에서 발생된 패킷 헤더 (Packet Header)를 저장하거나, 패킷 헤더와 패킷 데이터 (Payload)의 일부를 저장하다. 플로우 저장 방식에서는 발생한 패킷의 헤더를 분석하여, protocol, source IP address, source port, destination IP address, destination port의 5개의 속성이 동일한 패킷들을 하나의 플로우로 저장한다. 패킷 저장 방식에 비해 플로우 저장 방식에서 필요한 용량이 작기 때문에 플로우만을 대상으로 하는 연구에서는 매우 유용하다. 하지만, 패킷과 플로우 모두를 대상으로 하는 연구에서는 패킷 저장 방식이 유리하다. 패킷 저장 방식으로 저장된 트래픽으로부터 플로우 정보를 손실 없이 얻을 수 있기 때문이다.

대부분의 인터넷 측정 연구들이 수동적 방식을 이용하고 있지만, 대용량의 트래픽을 관리하는 문제 때문에 연구에 어려움을 겪고 있다. 이 때문에 다양한 방식의 트래픽 샘플링 기법이 사용되고 있는 것이다.

3. 인터넷 측정의 어려움

(특히 수동적 방식의) 인터넷 측정을 쉽게 않게 만드는 여러 원인이 있다[1, 8, 16]. 크게 나누어, 개인 사생활 (Individual Privacy), 트래픽 용량 (Traffic Volume), 데이터 신뢰성 (Data Reliability) 문제로 나눌 수 있다.

가. 개인 사생활 (Individual Privacy)

인터넷 측정을 위해서 캡쳐되거나 모니터링 되는 트래픽은 개인의 사생활 정보를 포함하고 있다. 저장된 트래픽을 추적할 경우에는 어떤 한 개인이 언제, 어떤 사이트에서, 어떤 정보를 검색하고, 어떤 내용의 E-Mail을 주고 받는지 알 수 있다. 트래픽의 캡쳐와 분석은 마치 전화 감청과 같은 효과를 가진다. 따라서 네트워크 관리의 책임이 있는 기관에서는 트래픽을 캡쳐되는 것과 캡쳐된 데이터가 유출되는 것을 매우 꺼리고 있다. 이러한 상황으로 인해, 인터넷 측정 분야의 연구자들은 연구 대상 데이터를 얻는데 어려움을 겪고 있다. 트래픽 캡쳐로 인한 개인 사생활 정보 유출 문제를 해결하기 위해 트래픽을 구성하는 패킷의 source 및 destination IP address를 암호화하거나 다른 값으로 치환하는 방법을 사용하고, 패킷의 데이터 (Payload)는 제거한다. 이러한 과정을 트레이스 민감성 완화 (Trace Desensitization)라고 한다.

나. 트래픽 용량 (Traffic Volumes)

캡쳐되는 트래픽은 대량의 패킷으로 구성된다. 따라서 대용량의 트래픽을 다루는 것은 쉽지 않다. 캡쳐된 트래픽 데이터를 저장하기 위한 저장소 (Repository)가 대용량이어야 하고 (적어도 수 테라바이트단위의 RAID 저장 장치), 캡쳐 시스템부터 저장소까지의 네트워크 대역폭도 비교적 대용량이어야 한다. 이것은 인터넷 측정 연구가 많은 비용이 들 수밖에 없게 만드는 이유가 된다. 트래픽 관리 시스템의 연산 능력이 뛰어나더라도, 대상이 되는 트래픽 데이터의 크기가 주기억 장치 (Main Memory)의 용량 (Capacity)을 크게 넘어서고 있으므로 처리 시간이 오래 걸릴 수밖에 없다. 사실상 대용량 트래픽 처리 과정은 CPU보다는 I/O 성능에 영향을 많이 받게 된다. 본 연구에서 사용한 (샘플링 되지 않은) 트래픽 데이터의 경우, 수백 기가바이트 단위의 데이터가 하루 동안 캡쳐되었다. 일주일 동안에는 테라바이트 단위의 트래픽이 저장되었다. 이렇게 엄청난 양의 데이터는 간단한 연산을 적용하는데도 많은 시간을 소모하게 된다. 일례로, 본 연구에서 단순히 1.5 테라바이트의 원본 트래픽 데이터를 읽고 변환하여 DBMS (Database Management System)에 저장하기 위한 과정에만 24시간 이상 걸렸다. 효과적인 연구를 위해서는 반드시 샘플링이 필요하다.

다. 데이터 신뢰성 (Data Reliability)

앞에서 설명했듯이 개인 사생활 정보를 가진 트래픽을 얻는 것은 쉽지 않다. 또한 데이터를 얻게 되더라도, 보유한 데이터를 통한 연구 결과를 어떻게 일반화 할 수 있나는 문제가 있다. 예를 들면, 일반적으로 연구자들은 대학 캠퍼스의 트래픽을 주로 연구용 데이터로 사용한다. 대학생들의 인터넷 사용 습관과 일반인의 인터넷 사용습관은 일부 차이가 있을 가능성이 있다. 이는 캠퍼스 트래픽 데이터로 인한 연구 성과는 인터넷 트래픽의 특별한 경우 (Special Case)의 결과일 수도 있다는 점이다. 또다른 예로, 다양한 링크가 모여 인터넷을 구성하고 있음에도, 대부분의 연구는 단일 링크의 데이터에 대해서만 트래픽을 연구한 것들이다. 이러한 상황에

한계를 느끼고 다중링크에 대한 트래픽 연구가 많은 주목을 받기도 했다.

4. 트래픽 샘플링

앞에서 살펴봤듯이, 인터넷 측정 연구과정에서는 다양한 어려움들이 존재하고 있다. 우선 해결되어야 할 당면과제는 트래픽 용량을 어떻게 감당할 것인가이다. 현실적인 해결책은 트래픽을 샘플링하는 것이다. 본 연구에서는 주로 패킷 샘플링과 관련 기법에 대해서 다룰 것이다. 구현과 적용이 비교적 쉬워 많이 사용되는 샘플링 기법으로는 단순 랜덤 샘플링 (Simple Random Sampling), 규칙적 샘플링 (Systematic Sampling), 층화 샘플링 (Stratified Sampling)이 있다[8, 11~12, 18].

가. 단순 랜덤 샘플링 (Simple Random Sampling)

모집단 (Population)을 구성하는 개체에 번호를 부여한 후, 추출할 개체의 양만큼 랜덤 번호를 생성하다. 생성된 랜덤 번호에 해당하는 개체를 모집단에서 추출한다. 이 방식은 매우 간단하고 적용이 쉬운 샘플링 기법이다. 대체로 모집단이 작은 경우 유용하다. 일반적인 트래픽 샘플링에서 표준적인 기법으로 많이 사용된다.

나. 규칙적 샘플링 (Systematic Sampling)

모집단에서, 일정한 간격에 한 번씩 개체를 추출한다. 이러한 특성 때문에 규칙적 샘플링을 간격 샘플링 (Interval Sampling)이라고도 한다. 간격의 설정 방식에 따라서 계수 기반 규칙적 샘플링 (Count-based Systematic Sampling)과 시간 기반 규칙적 샘플링 (Time-based Systematic Sampling)으로 나눌 수 있다.

계수 기반 규칙적 샘플링은 다음의 식(1)과 같은 트리거 (Trigger)를 사용한다. 개체의 추출은 정수의 주기 $N > 0$에 발생한다.

$$i_n = nN + i_0 \quad (1)$$

시간 기반 규칙적 샘플링은 다음의 식(2)와 같은 트리거를 사용한다.

$$\tau_n = nT + \tau_0 \quad (2)$$

규칙적 샘플링은 트리거가 동작할 때마다 해당 순서의 개체를 추출한다.

다. 층화 샘플링 (Stratified Sampling)

모집단에서, 동일한 속성을 가지는 개체들을 묶어 여러 층 (Strata)을 생성하고, 각 층으로부터 일정한 개체를 추출하는 방법이다. 층화 샘플링은 트래픽을 한 덩어리로 보고 단순한 샘플링 기법을 적용하는 것보다는 효과적일 것으로 예상된다. 트래픽은 다양한 프로토콜 (TCP, UDP, ICMP 등)의 특성에 따라 그 양과 형태가 다르기 때문이다.

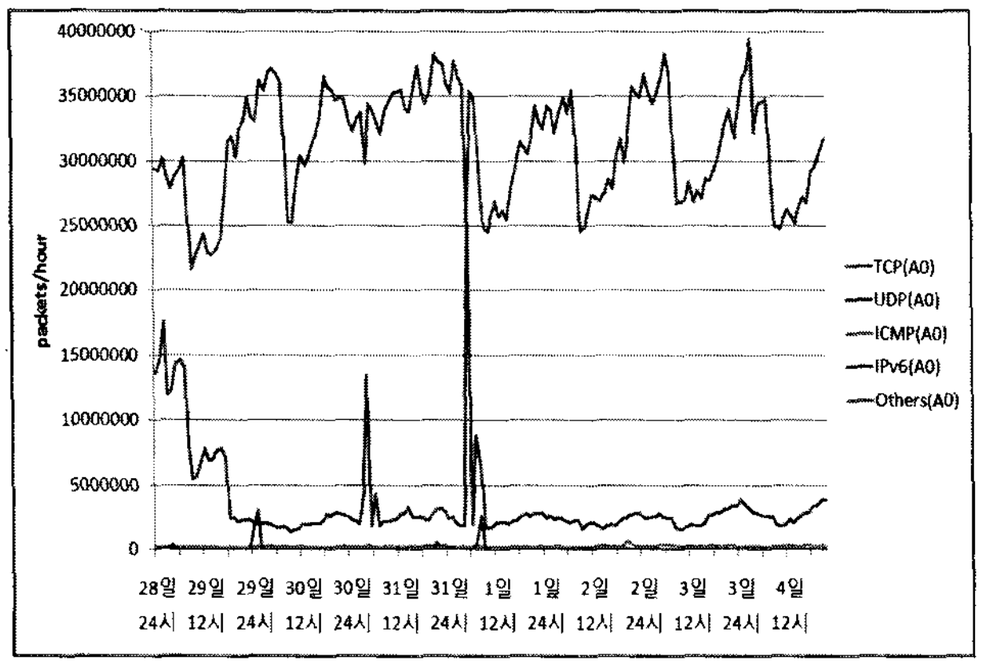
다. 샘플링의 강도 (Sampling Intensity)

1/10, 1/100, 1/1000 등[8, 11~12, 18]이 많이 사용되며, 대체적으로 통계학자들은 1/10정도가 적절하다고 한다. 샘플링의 강도에 따라 추출된 데이터의 신뢰도가 크게 달라진다. 실제 인터넷 측정 연구에서는 1/100정도도 사용되고 있다. 어떤 연구에는 1/65536의 샘플링에서 비교적 정확한 연구가 가능하다고 주장하기도 한다. 그러나 짧은 시간 단위 (Small Time Scale)의 연구에서는 너무 강한 샘플링은 적절하지 않다. 예컨대, 1초 단위의 트래픽 분석이 필요한 웜바이러스 탐지에서, 대상 링크에서 1초에 패킷이 3700개 발생했다고 가정하겠다. 이 경우 1/1000의 강도로 샘플링을 하면 3, 4개의 패킷이 남는다. 샘플링 후에 남은 3, 4개의 패킷을 가지고는 어떠한 분석이나 판단이 가능하지 않다. 다음 장에서는 이러한 샘플링의 강도에 따라서 어떠한 영향이 발생하는지 살펴볼 것이다.

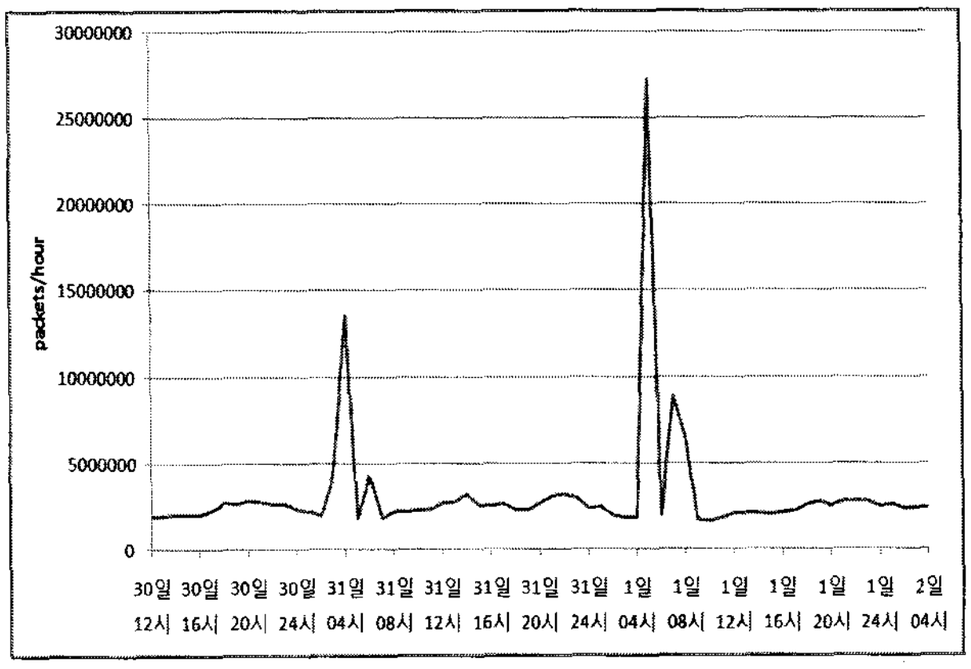
Ⅲ. 트래픽 샘플링 기법의 평가

1. 데이터 (Data)

본 연구에서 사용된 트래픽은 우리나라 백본망의 155Mbps인 한 링크에서 일주일간 (2004년 10월 29일부터 11월 4일까지) 캡쳐된 것이다. 대부분의 연구에서에서 일반망이 아닌, 대학 캠퍼스망에서 캡쳐된 트래픽을 주로 사용한다. 본 연구에서 사용한 트래픽은 우리나라의 일반망에서 캡쳐된 흔치 않은 데이터다. 캡쳐된 원본 트래픽 (Raw Traffic)은 pcap 포맷[19]이며, 샘플링되지 않은 약 1.5테라바이트 용량의 데이터다. 대용량의 트래픽을 다루기 위해서, pcap 포맷의 데이터를 텍스트 데이터로 변환하여 데이터베이스로 관리하고 있다. 본 연구에서 사용한 트래픽 관리 유틸리티는 직접 개발한 것들이며, 그 집합을 "DMC_Traff_Mon"이라고 부른다. "DMC_Traff_Mon"은 libpcap[19] 기반이며, 본 연구의 목적에 맞게 최적화되어 빠른 속도로 데이터를 추출하고 변환한다. 또한 IP address와 같은 민감한 데이터는 암호화하는 기능도 포함하고 있다.



(a) 전체 트래픽



(b) 이상이 탐지된 UDP 트래픽

그림 1. 일주일 간의 트래픽 데이터
Fig. 1. Traffic Data for a week.

표 1. 분석에 사용된 구간
Table 1. Periods for Analysis.

이름	시작	끝	기간
Period1	Sun Oct 31 03:55	Sun Oct 31 04:19	24 min
Period2	Mon Nov 1 20:00	Mon Nov 1 20:30	30 min

연구에서 사용된 트래픽 데이터는 UDP flooding 공격과 다양한 웜바이러스 트래픽을 포함하고 있다. 본 연구에서는 이러한 이상 트래픽의 측정에 샘플링이 어떤 영향을 미치는지 중점적으로 분석할 것이다. 그림 1은 본 연구에서 사용한 일주일간의 트래픽 데이터를 packets/hour 그래프로 나타내었다.

그림 1 (a)를 살펴보면, 31일 새벽과 1일 새벽에 이상 현상이 발견되고 있다. 그림 1 (b)는 발견된 이상 UDP 트래픽 (Anomalous UDP Traffic)을 나타낸 것이며, 분석 결과 다중 포트 (Multi-port) UDP flooding 공격으로 판단되었다. 이렇게 판단하는 근거는 초당 10,000개 이상의 패킷이 특정한 호스트의 3개 포트에 집중되고 있기 때문이다. 이는 한 호스트가 정상적인 서비스를 받거나 제공하기 어렵게 할 만한 트래픽을 받고 있는 것이며, DoS (Denial-of-Service) 공격을 받고 있다고 판단할 수 있다[2, 17]. 본 논문에서 연구에서 비교에 사용한 구간을 표 1에 정리하였다. Period1은 발견된 4개의 UDP flooding 공격 트래픽 중에서 대표로 선택한 것이며, Period2는 정상 구간을 임의로 선택한 것이다.

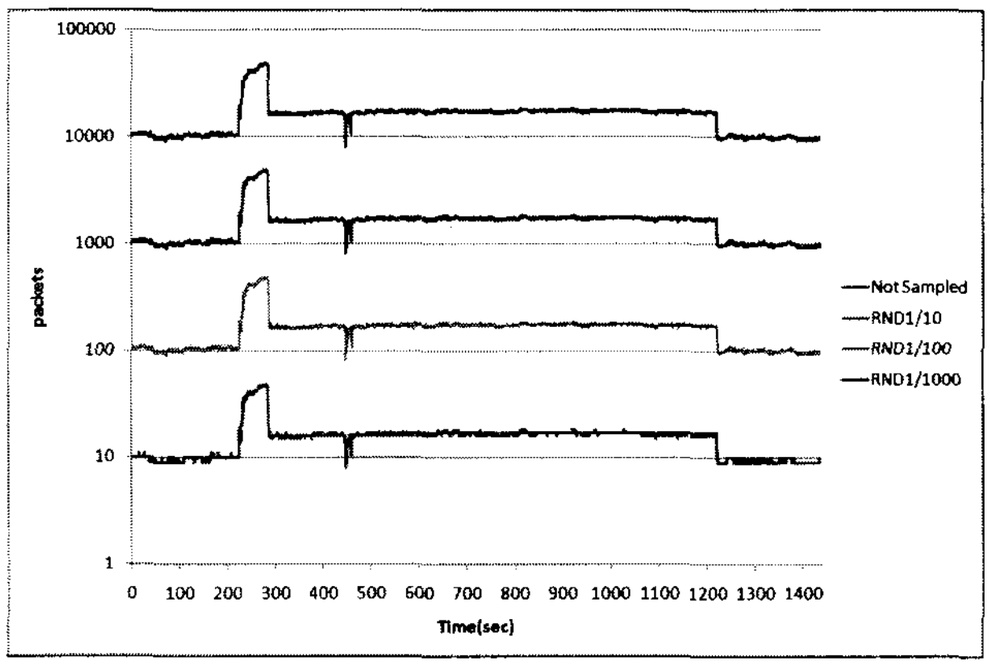
2. 트래픽 크기 (Traffic Volume)

트래픽 크기의 분석은 인터넷 측정에서 가장 기본이 된다. 본 절에서는 1/10, 1/100, 1/1000 단위로 트래픽 샘플링의 결과를 비교 분석한다. 본질적으로 트래픽을 샘플링은 손실과정 (Lossy Process)이므로 트래픽의 크기가 줄게 된다. 트래픽 크기 비교를 손쉽게 하기 위해서, 편의상 log 스케일 그래프를 사용할 것이다. 짧은 시간 안에 공격을 판단하고 대응하는 것이 매우 중요하다[7, 17]. 그러므로 본 연구에서 주로 1초를 시간 단위 (Timebin)로 샘플링이 트래픽에 끼치는 영향을 살펴볼 것이다.

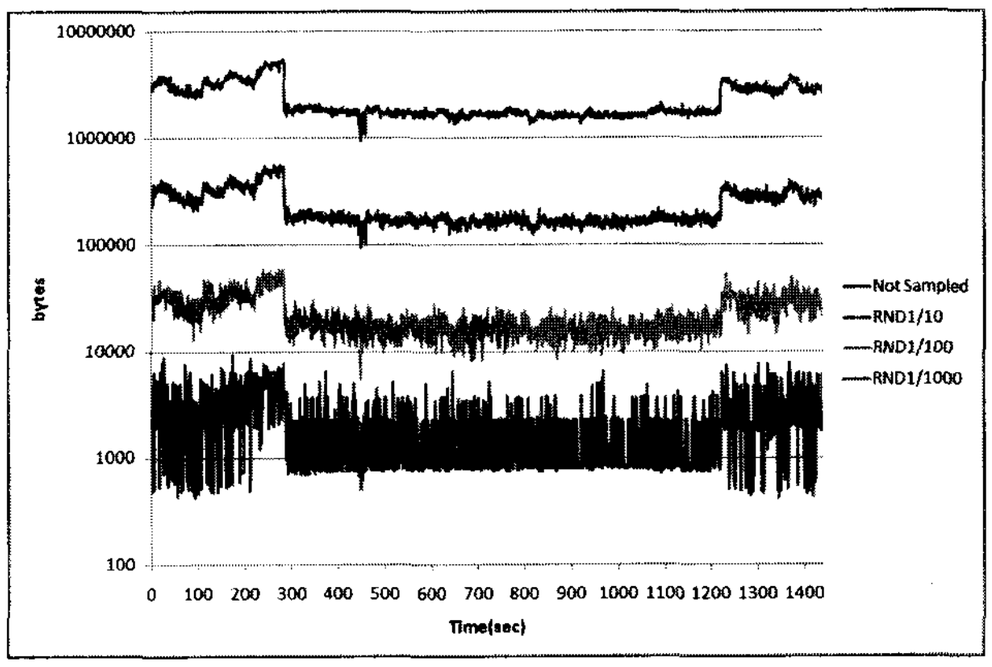
가. 단순 랜덤 샘플링의 결과

단순 랜덤 샘플링은 매우 흔하게 사용되는 방법이며 이전 연구[8]에서 트래픽 샘플링 기법의 표준으로 소개되기도 했다. 그림 2는 원본 데이터와 샘플링된 데이터에 대한 packets/sec 및 bytes/sec 그래프를 log 스케일로 나타낸 것이다.

그림 2 (a)에서 볼 수 있듯이, packets/sec 그래프에서 원본 트래픽과 단순 랜덤 샘플링을 적용한 데이터간의 차이를 찾기 쉽지 않다. 시간에 따라 변하는 트래픽의 흐름을 잘 반영하기 때문으로 볼 수 있다. 반면에, bytes/sec 그래프인 그림 2 (b)에서는 샘플링의 강도에 따라서 트래픽의 왜곡 (Distortion)이 점점 더 강해지는 것을 확인할 수 있다. 패킷 크기 (Packet Size)는 균일 (Uniform)하지 않기 때문에, 일정량의 패킷이 샘플링되면, bytes/sec의 값이 치우친(Biased) 형태를 보이게 되기 때문이다. 패킷 수 (Packet Count)만큼이나 패킷 크기는 인터넷 측정에서 매우 중요하다. 그림 2를 통해서, 단순 랜덤 샘플링은 packets/sec와 같은 패킷 수 분석에 유용함을 알 수 있다. 하지만, 패킷 크기 분석을 위

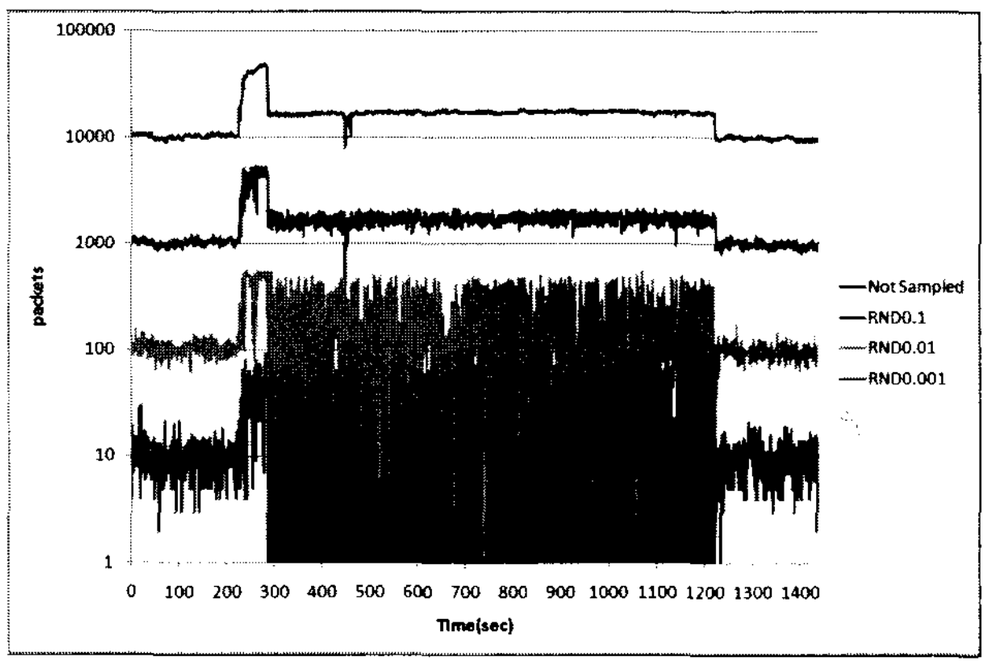


(a) packets/sec

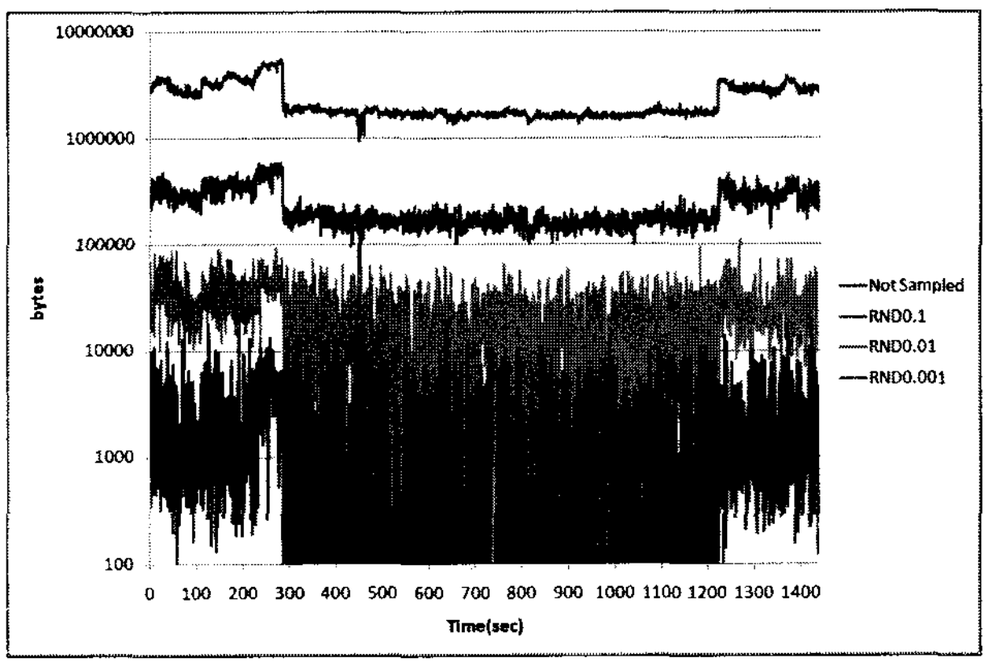


(b) bytes/sec

그림 2. 단순 랜덤 샘플링 결과의 트래픽 크기
Fig. 2. Traffic Volumes on Simple Random Sampling Results.



(a) packets/sec



(b) bytes/sec

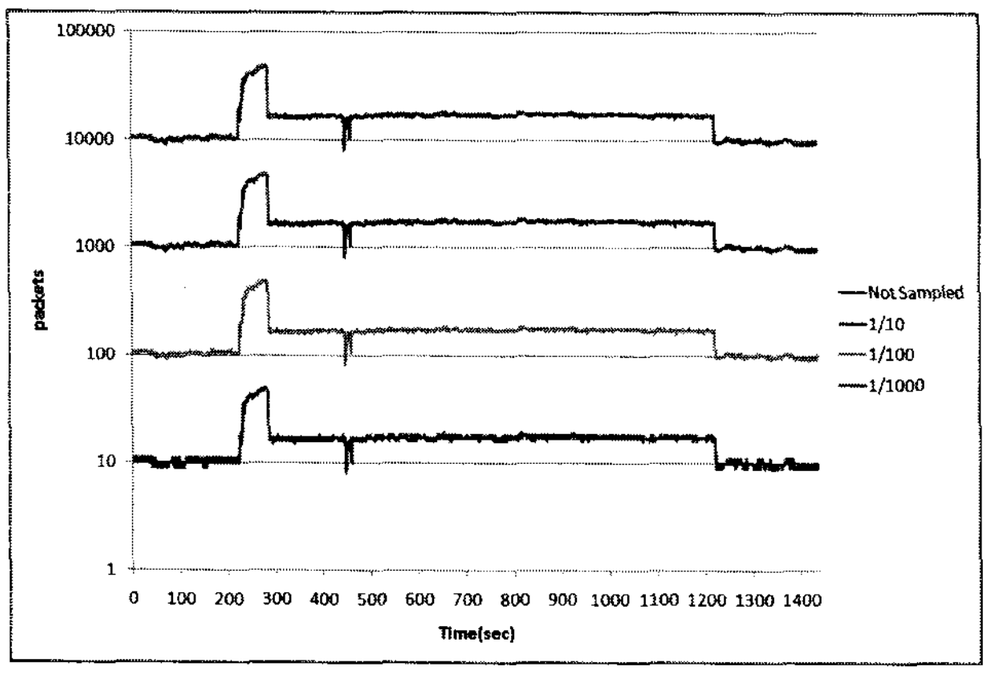
그림 3. 시간 기반 단순 랜덤 샘플링 결과의 트래픽 크기
Fig. 3. Traffic Volumes on Time-based Simple Random Sampling Results.

해서는 가능한 단순 랜덤 샘플링의 강도를 1/100 이내로 설정하는 것이 좋다고 볼 수 있다.

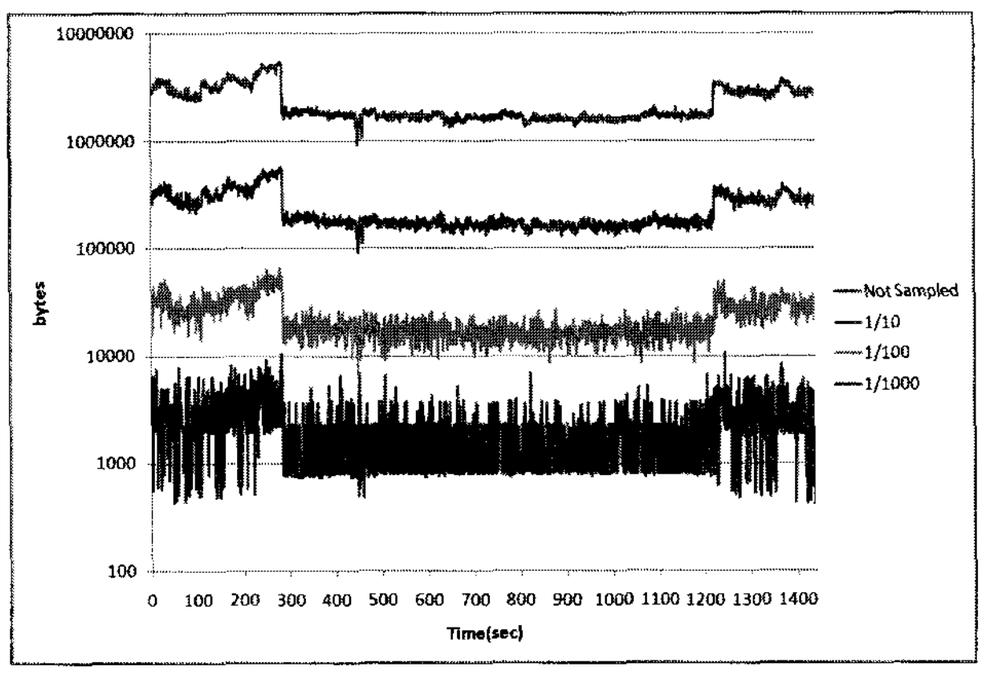
그림 3에서는 시간 기반 단순 랜덤 샘플링(Time-based Simple Random Sampling)의 결과를 packets/sec와 byte/sec 그래프로 나타내었다. 그림 3 (a)에서 볼 수 있듯이, 매초마다 임의로 범위의 트래픽을 추출한, 시간 기반 단순 랜덤 샘플링의 성능은 좋지 않은 것을 확인할 수 있다. 1/10을 넘는 샘플링 강도가 가해지면 심하게 왜곡이 발생한다. 이러한 양상은 그림 3 (b)에서도 명확히 나타나고 있다. 시간에 따라 크기 변화가 일어나는 트래픽의 특성을 전혀 반영하고 있지 못하다. 이를테면, 10분의 트래픽 중에서 1분의 트래픽을 임의로 뽑아 전체 트래픽을 대표하게 하는 형태는 적절하지 않다는 것이다. 특히 짧은 시간 단위의 분석에서 시간 단위 샘플링 강도가 강해지면, 패킷이 존재하지 않는 구간을 선택하게 되는 경우가 발생한다. 예를 들어, 1초 범위의 트래픽에서 임의로 0.001초를 선택하여 트래픽을 추출하면 패킷이 존재하지 않는 0.001초를 선택하게 되는 경우가 발생한다. 이러한 빈 구간의 선택은 트래픽을 크게 왜곡시킨다.

나. 규칙적 샘플링의 결과

규칙적 샘플링 (Systematic Sampling)도 많이 사용되는 기법이며, 특히 계수 기반 규칙적 샘플링 (Count-based Systematic Sampling)이 많이 사용된다. 그림 4는 계수 기반 규칙적 샘플링의 결과를 나타낸 것이다. 그림 4를 그림 2와 비교하면 계수 기반 규칙적 샘플링의 결과는 단순 랜덤 샘플링의 결과와 유사한 것을 알 수 있다. 그림 4 (a)와 그림 2 (a) 그리고 그림 4 (b)와 그림 2 (b) 모두 유사하다. 그림 2와 4가 유사하다는 것

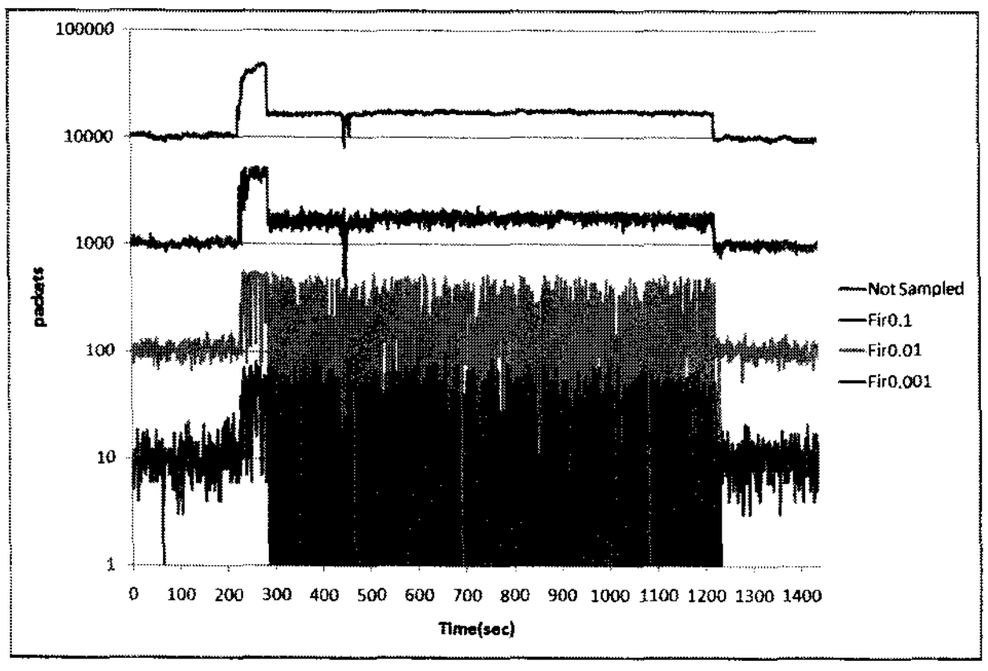


(a) packets/sec

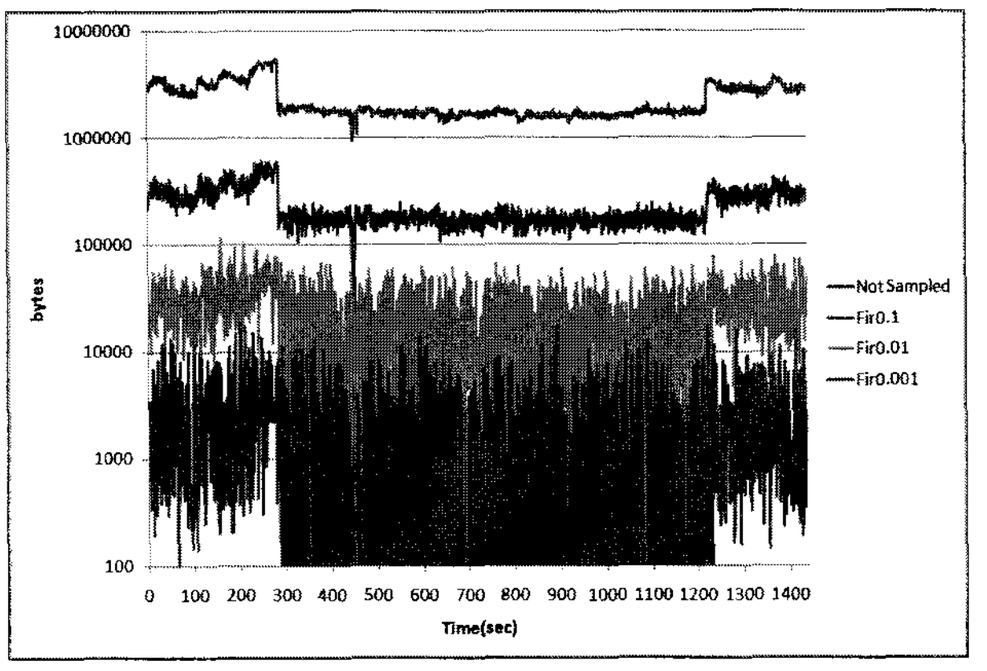


(b) bytes/sec

그림 4. 계수 기반 규칙적 샘플링 결과의 트래픽 크기
Fig. 4. Traffic Volumes on Count-based Systematic Sampling Results.



(a) packets/sec



(b) bytes/sec

그림 5. 시간 기반 규칙적 샘플링 결과의 트래픽 크기
Fig. 5. Traffic Volumes on Time-based Systematic Sampling Results.

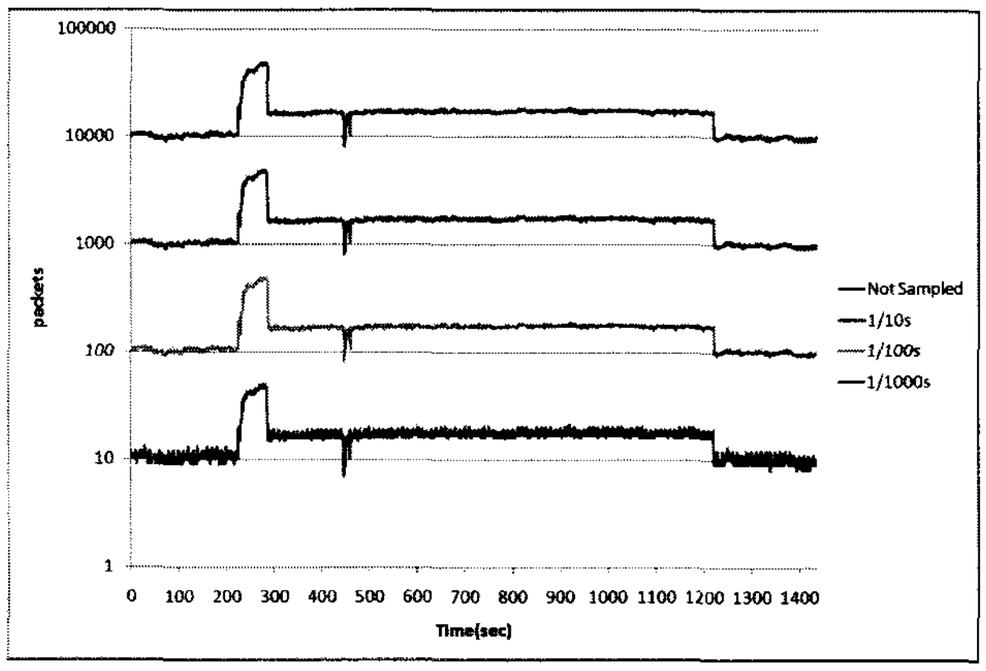
은, 시간에 따른 트래픽 변화를 단순 랜덤 샘플링과 계수 기반 규칙적 샘플링 모두가 잘 반영할 수 있는 기법이라는 것을 뜻한다.

반면, 그림 5는 시간 기반 규칙적 샘플링 (Time-base Systematic Sampling)의 결과 그래프다. 매초마다 일정한 양의 트래픽을 일정하게 추출하는 방식이다. 그림 5와 그림 3은 유사하게 왜곡된 모습을 보여주고 있다. 그림 5 (a)와 (b) 모두에서 확인할 수 있듯이, 1/10 정도의 샘플링 강도만 사용가능하며 그 이상의 샘플링 강도에서는 왜곡이 너무 심하게 나타나고 있다. 시간 기반 규칙적 샘플링도 시간 기반 단순 랜덤 샘플링과 같이, 시간에 따른 트래픽의 흐름을 잘 표현하고 있지 못하다.

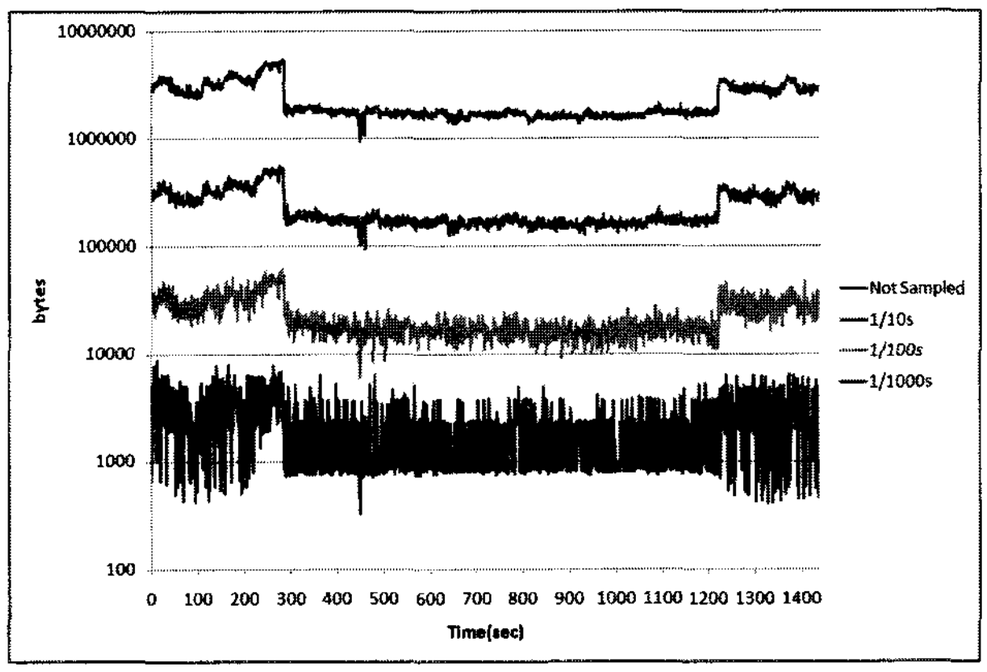
패킷 수 단위의 샘플링 방법인, 단순 랜덤 샘플링이나 계수 기반 규칙적 샘플링은 무난한 결과를 보여주는 것으로 확인되었다. 두 기법 모두 샘플링 강도가 1/100 일 때까지는 짧은 시간 단위의 인터넷 측정에서도 매우 유용하고 사용할 있다고 볼 수 있다.

다. 층화 샘플링의 결과

층화 샘플링 (Stratified Sampling)에서는 샘플링 대상인 모집단의 특성을 분석하여, 모집단을 의미있는 몇 개의 층 (Strata)으로 나눈다. 나누어진 층에서 각각 일정한 샘플링을 실행하는 것이 층화 샘플링이다. 인터넷 트래픽은 TCP, UDP와 같은 전송층 프로토콜 (Transport-Layer Protocol) 의 특성에 따라서 크게 달라진다. TCP와 UDP는 완전히 다른 패러다임 (Paradigm)을 가지는 프로토콜이기 때문이다. 특히 인터넷 트래픽에서 TCP 기반 패킷과 UDP 기반의 패킷은 균일하게 나타나지 않는다. 트래픽의 비율도 TCP

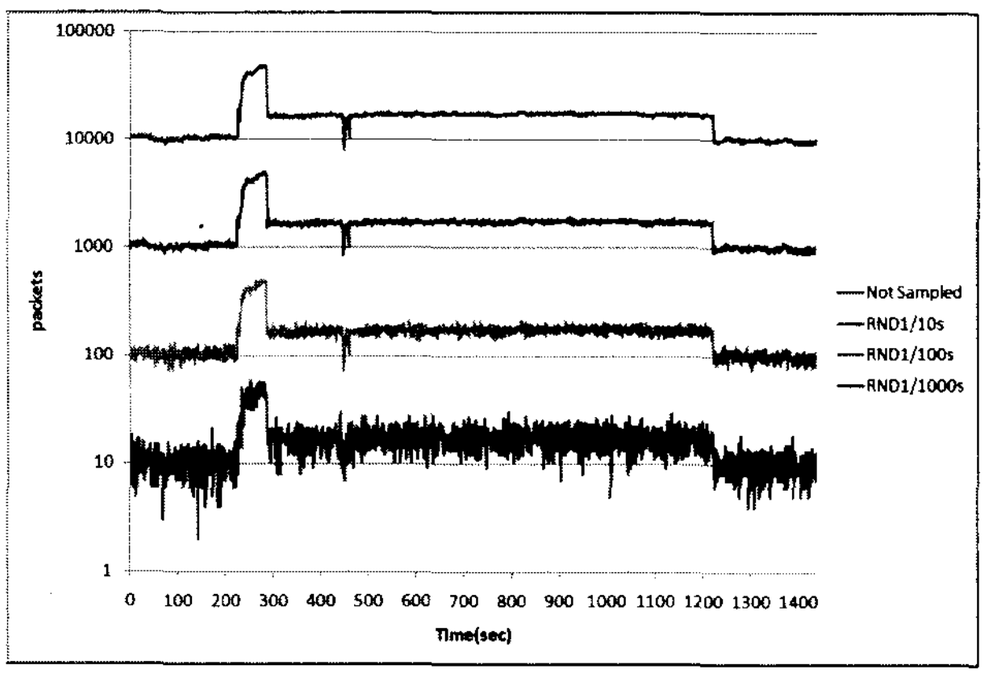


(a) packets/sec

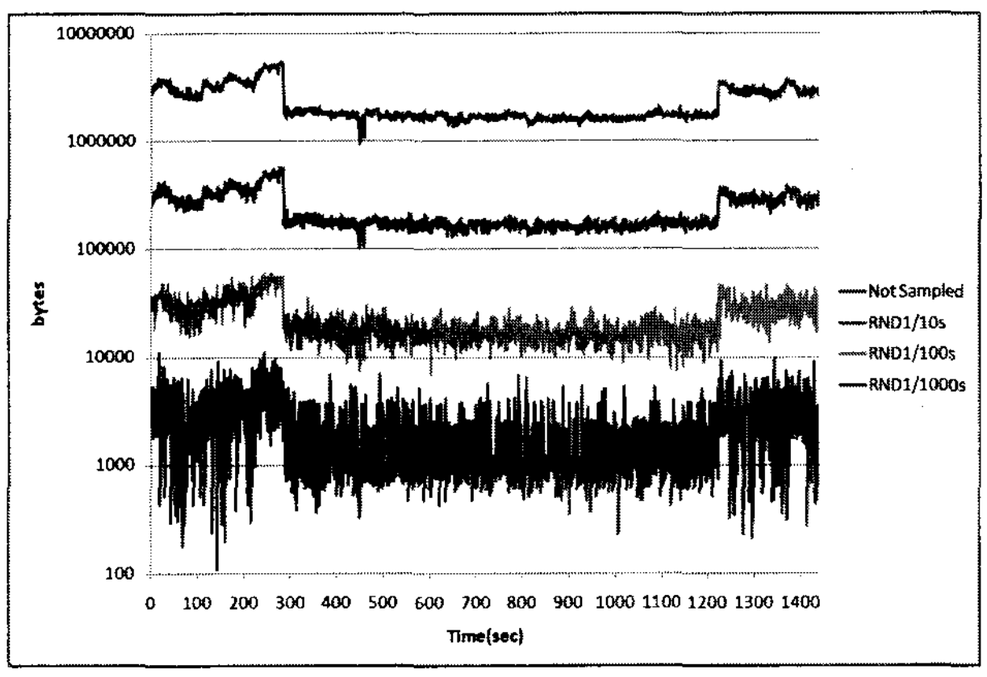


(b) bytes/sec

그림 6. 규칙적 샘플링을 이용한 층화 샘플링 결과의 트래픽 크기
Fig. 6. Traffic Volumes on Stratified Sampling Results using Systematic Sampling.



(a) packets/sec



(b) bytes/sec

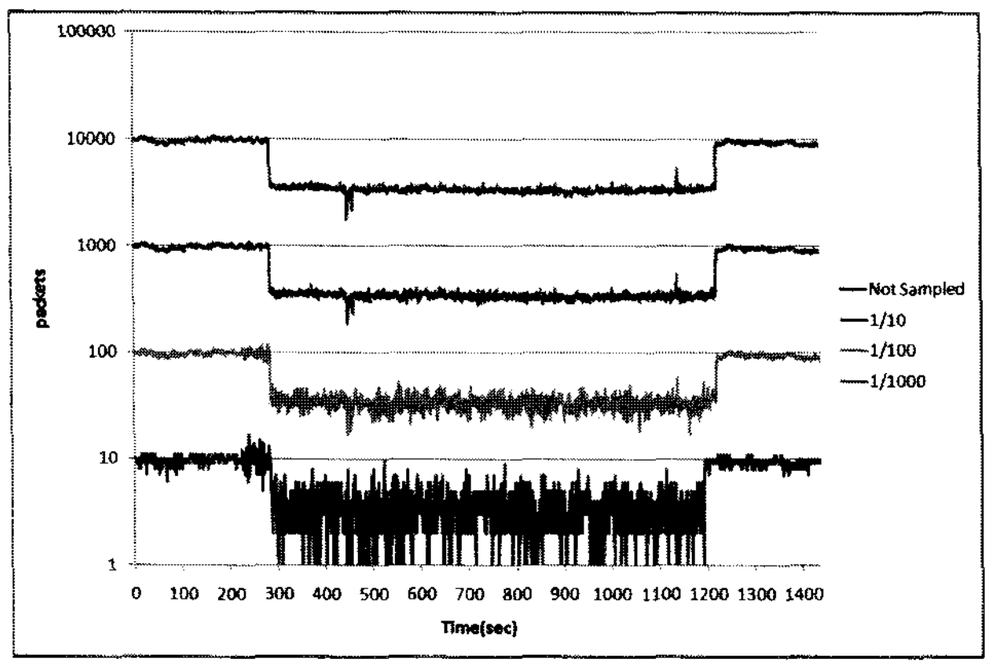
그림 7. 단순 랜덤 샘플링을 이용한 층화 샘플링 결과의 트래픽 크기
Fig. 7. Traffic Volumes on Stratified Sampling Results using Simple Random Sampling.

기반 트래픽이 UDP기반 트래픽보다 훨씬 크게 나타난다[17]. 이러한 트래픽 특성을 반영하면 더욱 좋은 결과를 보여줄 것으로 예상된다. 따라서 TCP, UDP, Others와 같이 3개의 층으로 트래픽을 나누어 규칙적 샘플링을 실행하였다(그림 6).

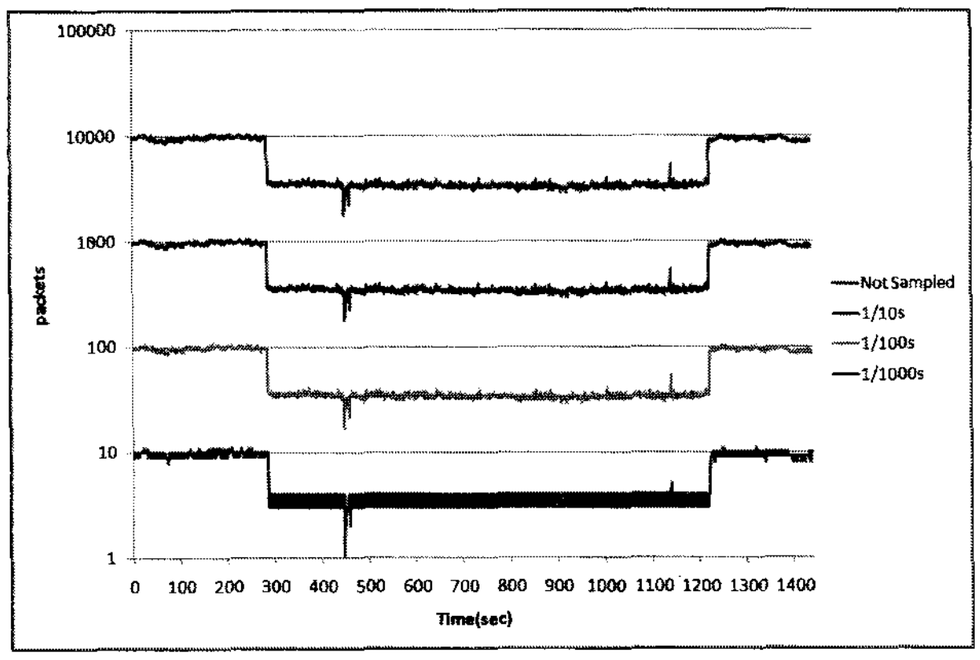
그림 6은 계수 기반 규칙적 샘플링을 이용한 층화 샘플링의 결과다. 그럼에도 그림 6은 그림 2이나 그림 4와 유사한 결과를 보여준다. 층화 샘플링도 시간에 따른 트래픽의 흐름을 잘 유지하고 있다.

그림 7은 단순 랜덤 샘플링을 이용하여 층화 샘플링을 실행한 결과이다. 그림 6 (a)에 비해 그림 7 (a)의 1/1000 샘플링 강도 결과가 더 좋지 않게 나타나고 있다. 규칙적 샘플링은 일정한 간격으로 샘플링 되지만, 단순 랜덤 샘플링을 층화 샘플링에 활용할 경우 성능이 약간 떨어짐을 알 수 있다. 그러나 1/100 정도의 샘플링 강도에서는 여전히 유용하다.

층화 샘플링은 의미있는 층을 나누는 것이 핵심이다. 층화 샘플링은 TCP 및 UDP 와 같은 전송층 프로토콜 단위로 트래픽을 모니터링하는 인터넷 측정 연구에서 적절하다고 볼 수 있다. 그림 8 (a)와 (b)에서 확인할 수 있듯이, 층화 샘플링은 계수 기반 규칙적 샘플링 보다 더욱 정확하게 트래픽 변화를 유지하고 있다. 그림 8과 같이 TCP 기반 트래픽을 보았을 때 명확한 차이를 확인할 수 있다. 그림 8 (a)는 샘플링 강도에 따라서 트래픽 크기의 모양이 왜곡되어가지만, 그림 8 (b)에서는 샘플링 강도가 강해지더라도 거의 트래픽 크기가 왜곡되지 않고 있다. 이러한 양상은 UDP 기반 트래픽을 나타낸 그림 9에서도 확연히 나타난다. 그림 9 (a)보다 층화 샘플링 결과인 그림 9 (b)에서 샘플링 강도에 따른 트래픽의 왜곡이 (특히 양쪽의 정상구간에서) 적게 나

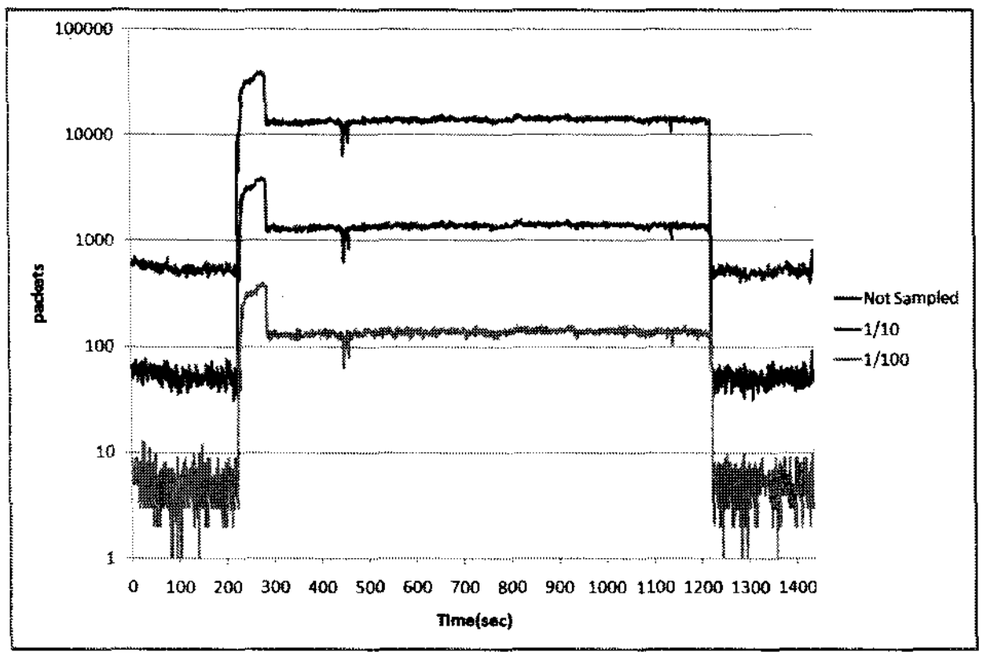


(a) 계수 기반 규칙적 샘플링 (packets/sec)

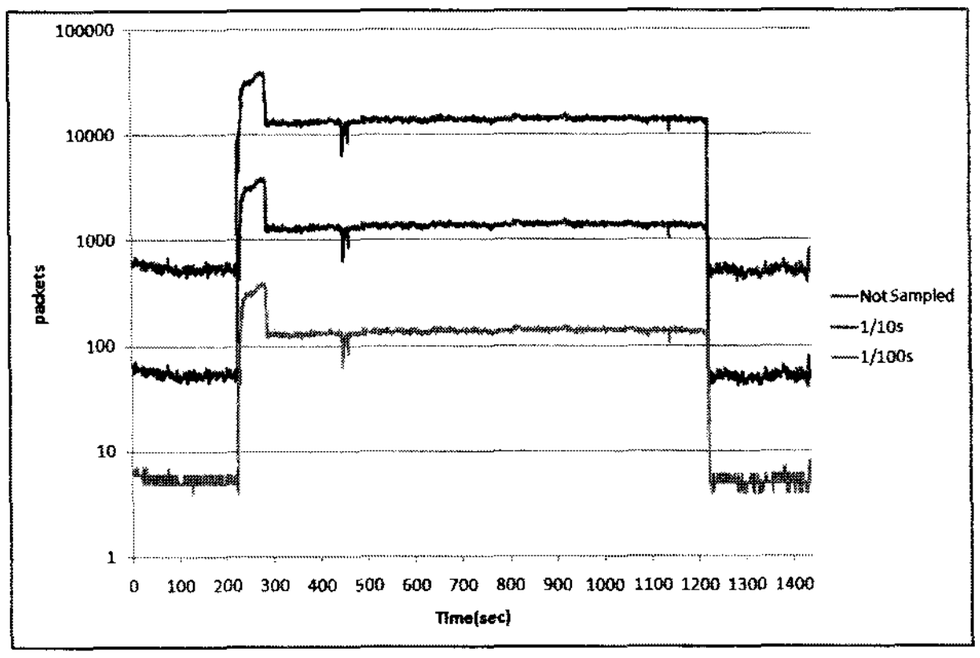


(b) (계수 기반 규칙적 샘플링을 이용한) 층화 샘플링 (packets/sec)

그림 8. TCP 트래픽에서, 층화 샘플링과 계수 기반 규칙적 샘플링 비교
Fig. 8. In TCP Traffic, Comparing Stratified Sampling Results with Count-based Systematic Sampling.



(a) 계수 기반 규칙적 샘플링 (packets/sec)



(b) 층화 샘플링 (packets/sec)

그림 9. UDP 트래픽에서, 층화 샘플링과 계수 기반 규칙적 샘플링 비교
Fig. 9. In UDP Traffic, Comparing Stratified Sampling Results with Count-based Systematic Sampling.

타나고 있다. 트래픽이 갑자기 증가하는 구간 (약 200초부터 1200초까지)에서는 샘플링 방식이나 강도와 관계없이 트래픽의 흐름을 잘 유지하고 있다.

참고로, 그림 8에서 약 290초 구간부터 갑자기 TCP 트래픽이 낮아지는 현상은 UDP 기반 트래픽에서 DoS 공격의 일종인 UDP flooding 공격이 발생했기 때문이다. 그림 9에서는 약 240초정도 구간에서 UDP 트래픽의 갑자기 증가하고 있다. 1초당 약 12,000부터 19,000개 패킷으로 구성된 트래픽이 한 호스트를 향하고 있었다. 이 정도의 DoS 공격은 방화벽이 무력화될 수 있을 정도의 트래픽이라 볼 수 있다[2].

라. 분석

수동적 방법의 목적인 트래픽 특성 분석에는 가능한 낮은 트래픽 샘플링 강도가 필요하다. 샘플링 강도가 강해질수록 트래픽의 왜곡이 점점 심해지므로, 데이터에 대한 신뢰도가 낮아지게 된다. 특히 짧은 시간 단위 분석에서는 1/100보다 강한 샘플링은 신뢰할 수 없는 분석 결과를 보일 것으로 예상된다. 인터넷 측정 연구에서 사용되는 일반적인 샘플링 강도도 1/10에서 1/100 정도이므로 적절하다고 볼 수 있겠다. 단, 샘플링 방식에 따른 왜곡정도도 다르게 나타났다. 주로 패킷 수 단위의 규칙적 샘플링이나 단순 랜덤 샘플링이 좋은 성능을 보였다. 그러나 트래픽은 TCP, UDP 등과 같은 전송층 프로토콜에 의해 영향을 많이 받는다. 이러한 특성을 감안한 층화 샘플링은 실제 인터넷 측정에서 매우 효과적인 것을 확인하였다. 실제 인터넷 측정은 TCP, UDP와 같은 프로토콜 단위로 많이 이루어지기 때문이

다. 트래픽 크기를 급격히 변화시키는 트래픽 크기 이상 (Traffic Volume Anomaly)은 샘플링의 영향이 크지 않았다. 샘플링이 되더라도 많은 양의 악의적 트래픽 (Malicious Traffic)이 대부분 남아있기 때문이다. 따라서 인터넷 측정의 목적이 트래픽 크기 이상을 찾는 것이라면, 비교적 강한 샘플링 강도 (1/100나 1/1000정도)도 유용할 것으로 분석되었다. 특히 네트워크 시스템은 보호하는 방화벽조차도, 많은 양의 이상 트래픽에 노출되면 DoS 공격을 그대로 받게 된다. 따라서 적당한 샘플링을 통한 공격 탐지는 유용하다.

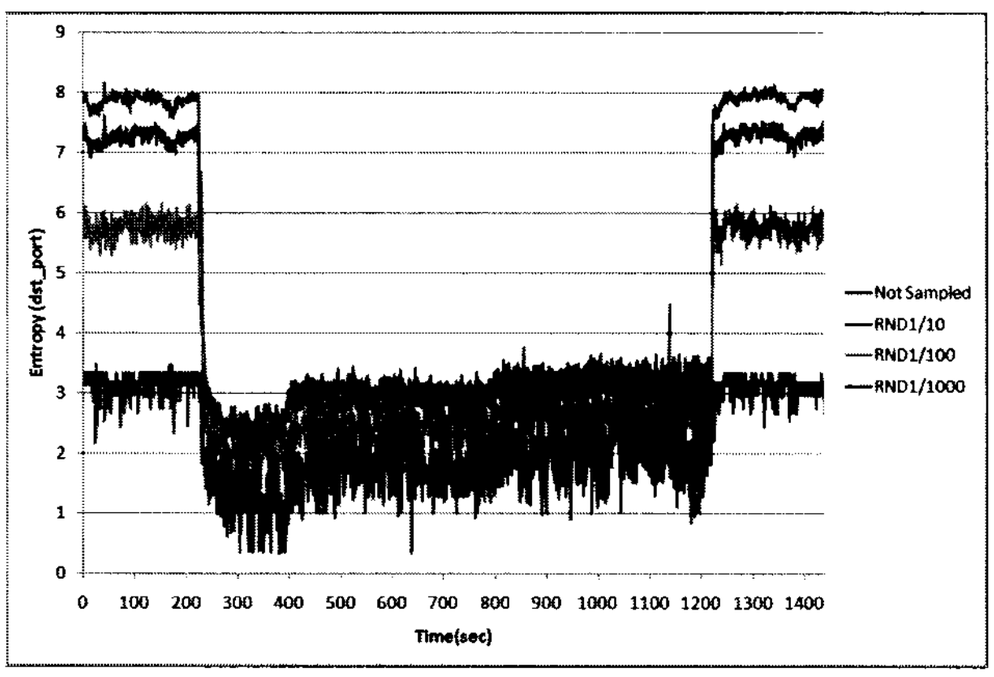
3. 엔트로피 (Entropy)

엔트로피 (Entropy)는 정보이론 (Information Theory) 분야에서 널리 쓰이는 데이터의 복잡성에 대한 측정 규준 (Metric)이다. 인터넷 측정 분야에서는 이상 트래픽을 분석하고 탐지하는데 유용하게 사용될 수 있다[17, 21]. 특히 DoS 또는 DDoS 공격이 발생시키는 트래픽은 source IP address ,destination IP address나 source port, destination port의 복잡성(Complexity)을 변화시킨다. 이러한 복잡성의 변화는 엔트로피의 급격한 증가 또는 감소로 나타나게 된다. 일반적으로 엔트로피는 다음의 식 (3)과 같이 나타낸다[20].

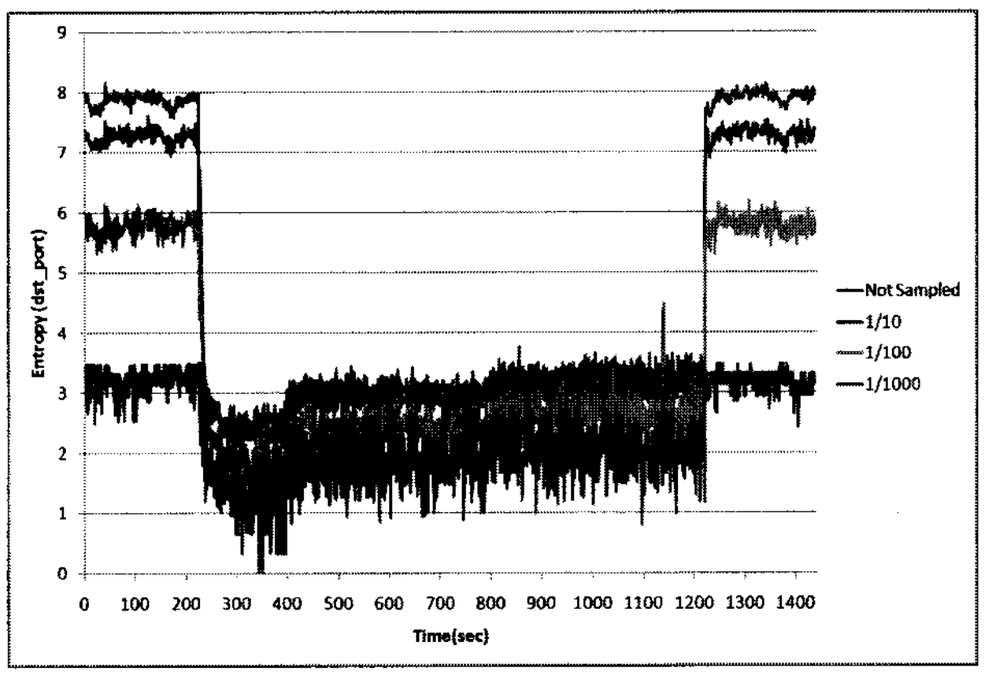
$$H(X) = -\sum p_i(x)\log_2 p_i(x) \tag{3}$$

$p_i(x)$를 정의하는 것이 매우 중요하다. 인터넷 측정 분야에서는 다음의 식 (4)와 같이 $p_i(x)$를 정의하여 사용할 수 있다[17].

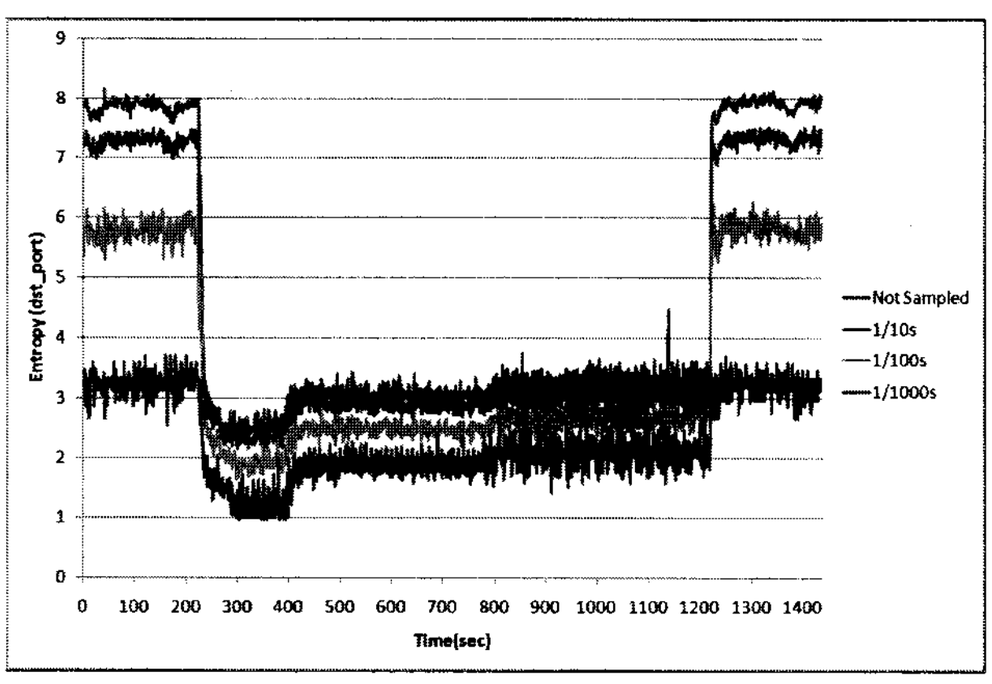
$$p_i(x) = \frac{Packet\ Count\ of\ [Distinct\ Property]}{Packet\ Count} \tag{4}$$



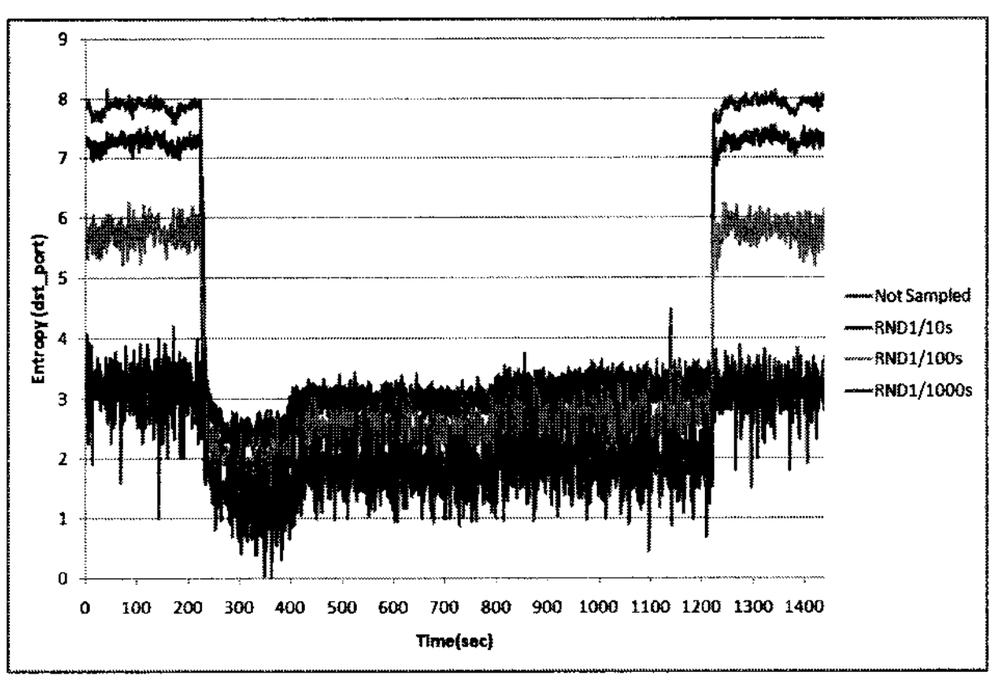
(a) 단순 랜덤 샘플링



(b) 계수 기반 규칙적 샘플링



(c) (규칙적 샘플링을 이용한) 층화 샘플링



(d) (단순 랜덤 샘플링을 이용한) 층화 샘플링

그림 10. 샘플링 결과의 엔트로피 비교
Fig. 10. Comparison on Entropy of Sampling Results.

일정한 범위(본 연구에서는 1초)의 트래픽에서 Distinct Property에 각각 i로 번호가 매겨지며 ($i = 0, 1, 2, \ldots, n$), i마다 식 (4)와 같은 $p_i(x)$를 구하여 식 (3)의 $H(X)$를 구하면 해당 구간의 엔트로피를 얻게 된다. Distinct Property로는 source IP address, destination IP address, source port, destination port등과 같은 패킷 헤더 (Packet Header)의 속성들을 설정할 수 있다.

본 연구에서는 destination port를 Distinct Property로 선택하여 entropy/sec 그래프를 각 샘플링 기법별로 나타내었다. 사용한 샘플링 기법은 단순 랜덤 샘플링, 계수 기반 규칙적 샘플링, 층화 샘플링 (규칙적 샘플링과 단순 랜덤 샘플링 기반)을 사용하였다. 이외의 기법들을 사용할 경우, 원본 트래픽과의 샘플링 결과의 차이가 비교적 많이 발생하므로 제외하였다.

가. 전체 트래픽

앞에서 살펴보았던 이상 트래픽 발생 구간을 그대로 이용하여 entropy/sec 그래프를 그림 10와 같이 나타내었다. 그림 10 (a), (b), (c), (d)에서 공통적으로, x축 230초 부근에서 급격히 감소되는 엔트로피를 확인할 수 있다. 원인은 UDP flooding 공격이 발생하여, 트래픽의 크기와 내부 구성을 급격히 변화시켰기 때문이다. 특정한 호스트의 특정한 포트로 트래픽이 몰리기 때문에, 정상 트래픽의 비율이 갑자기 줄게 된다. 이는 트래픽에서 destination IP address / port의 다양성에 영향을 크게 미친다. 이러한 급격한 변화는 엔트로피의 급격한 감소를 초래한다.

엔트로피는 샘플링 강도에 영향이 크지 않다는 것은 기존 연구에서 밝혀졌다[11]. 본 연구에서 분석한 결과도 동일하게 나타났다. 앞 장에서, 트래픽 크기는 샘플링에 워낙 큰 영향을 받기 때문에 편의상 log 스케일 그래프를 사용하였다. 그러나 엔트로피는 샘플링 강도에 영향을 비교적 적게 받으므로 그림 10과 같이 log 스케일이 아닌 그래프로 표현 가능하다.

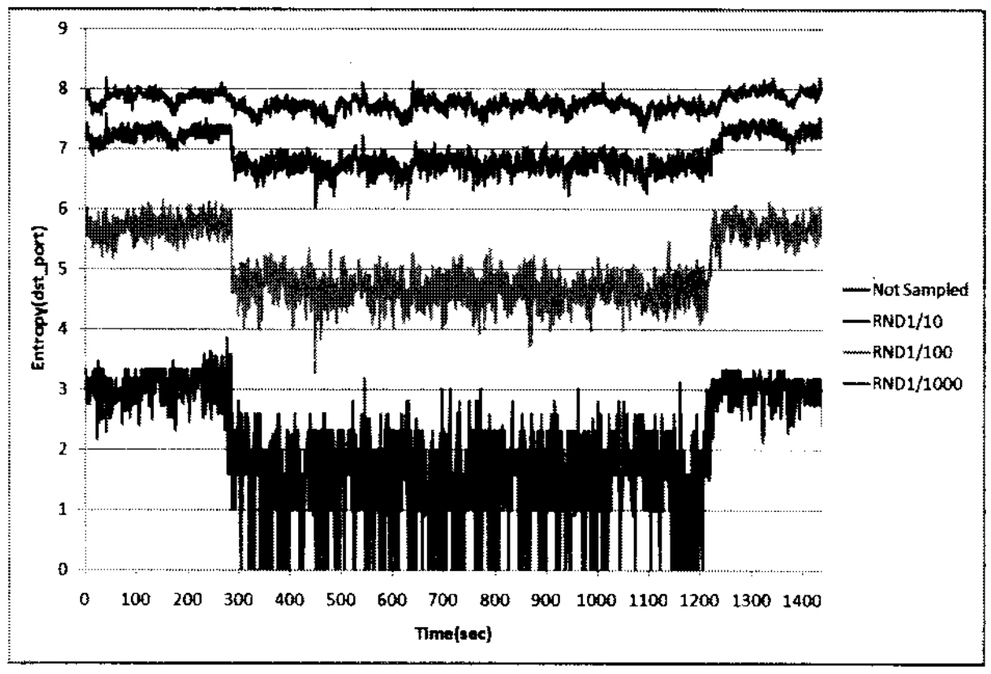
샘플링 강도에 따라서 엔트로피가 약간 떨어지는 현상은 분명하다. 샘플링의 강도가 강해질수록 트래픽 내부의 destination IP address / port의 복잡성이 감소한다. 즉 다양성이 떨어진다. 이러한 복잡성의 감소는 엔트로피의 감소를 수반한다. 그림 10의 그래프들에서 공통적으로, 정상 트래픽의 엔트로피인 0에서 200초구간과 1250에서 1400초 구간의 감소는 매우 크다. 비정상 구간 (UDP flooding 공격 발생 구간)인 230에서 1200초 구간의 엔트로피 감소는 정상 구간에 비해 상대적으로 미약한다. 비정상 구간에서는 이미 트래픽의 대부분이 악의적 트래픽 (Malicious Traffic)에 의해 destination IP address / port의 복잡성이 감소된 상태다. 즉, 비정상 구간의 트래픽은 트래픽의 크기는 급격히 증가 했으나, 트래픽의 복잡성은 감소한 상태인 트래픽이다. 이러한 비정상 구간의 엔트로피는 정상 구간의 트래픽에 비해 샘플링의 영향을 매우 적게 받는 것으로 분석할 수 있다. 이러한 이상 트래픽과 샘플링의 특성은 앞의 트래픽 크기 분석에서도 발견되었다.

트래픽 크기 분석에서 무난한 성능을 보여주었던 샘플링 기법들(그림 10 (a), (b), (d)) 모두 엔트로피 분석에서도 좋을 결과를 보여주었다. 특히 그림 10 (c)의 (규칙적 샘플링을 이용한) 층화 샘플링 기법은 가장 뛰어난 성능을 보여주고 있다. 층화 샘플링 기법은 샘플링 강도가 강해지더라도 엔트로피의 변화를 다른 기법에 비해 잘 유지하고 있다.

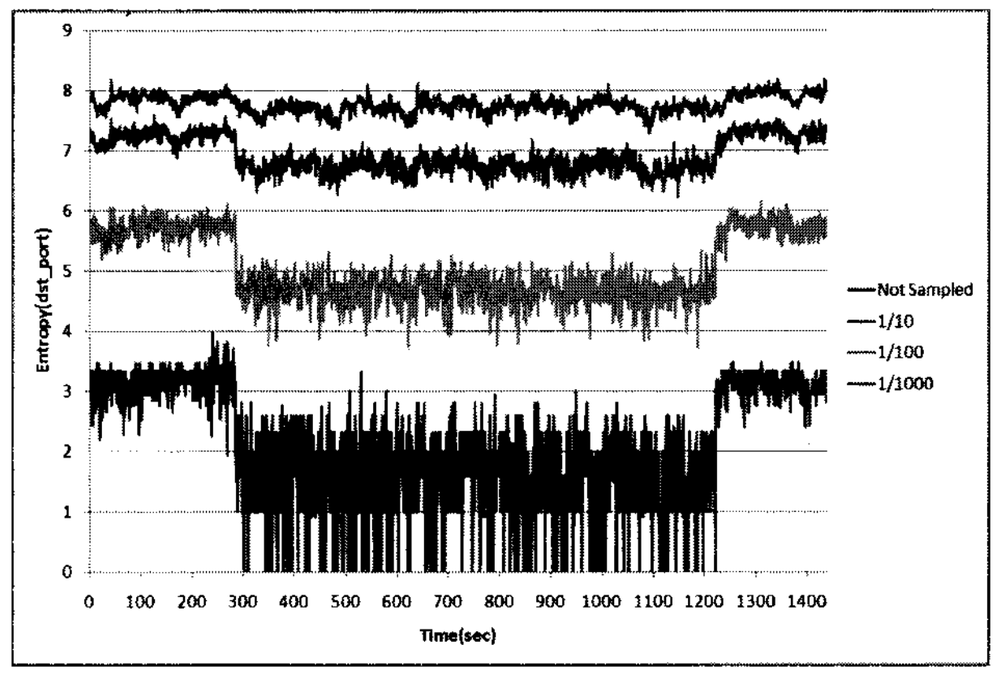
나. TCP 트래픽

그림 11은 TCP 기반 트래픽에 대한 entropy/sec 그래프를 나타낸 것이다. 그림 11에서 볼 수 있듯이 샘플링되지 않은 TCP 기반 트래픽의 엔트로피는 약 8정도를 계속 유지 하고 있다. 그러나 샘플링 강도가 강해지면. UDP flooding 공격이 발생한 비정상 구간과 거의 일치하는 부분에서 감소하는 엔트로피를 확인할 수 있다. 1/10정도만 샘플링 되더라도, 정상 구간에서 7부터 7.5정도의 엔트로피가 나타나며, 비정상 구간에서 6부터 7정도의 엔트로피가 나타난다. 이러한 엔트로피의 차이는 샘플링 강도에 따라 크게 나타난다. 샘플링은 실제 TCP 트래픽이 받은 UDP flooding 공격의 영향보다 더욱 큰 엔트로피 변화가 나타나게 한다. TCP 기반 원본 트래픽은 DoS 공격의 일종인 UDP flooding 공격으로 인한 병목 현상으로 약간의 트래픽 감소가 나타났다 (그림 8).

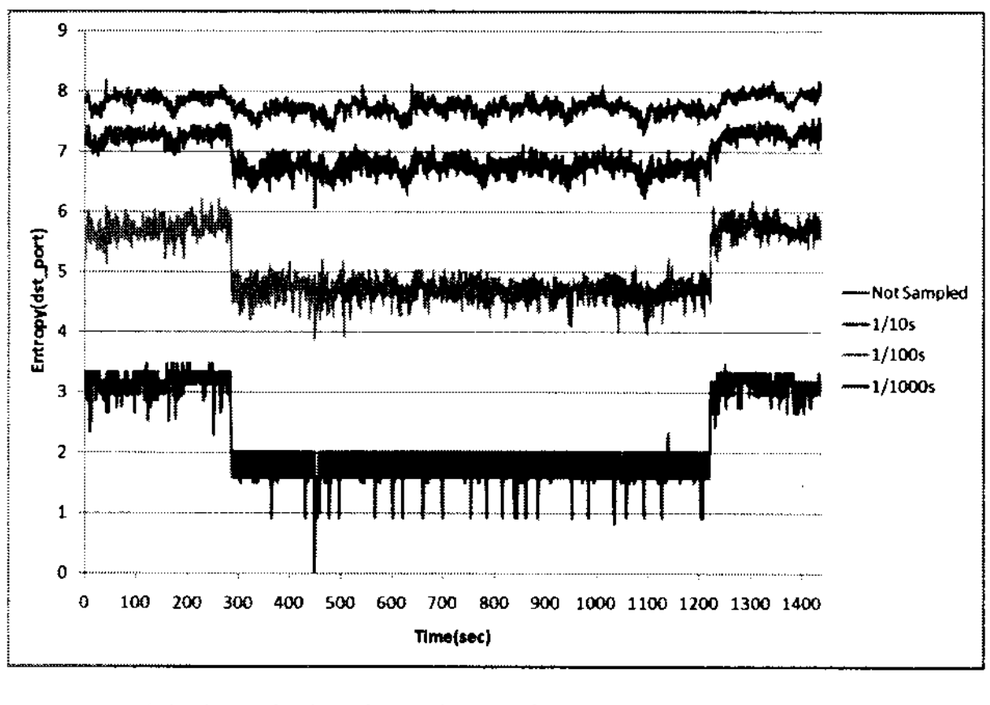
그러나 그림 11의 볼 수 있듯이 원본 트래픽의 엔트로피 자체는 거의 감소하지 않았다. UDP flooding 공격에 의한 TCP 트래픽 병목 현상은 TCP 트래픽을 약간 감소시키지만, TCP 트래픽 내부의 복잡도에 큰 영향을 주지 않음을 뜻한다. 그러나 샘플링 강도가 강해지면, 엔트로피의 감소가 왜곡되어 강하게 나타남을 확인할 수 있다. 이상 트래픽이 포함된 트래픽에서 샘플링 강도가 강할 경우, 엔트로피 분석이 잘못될 가능성이 있다.



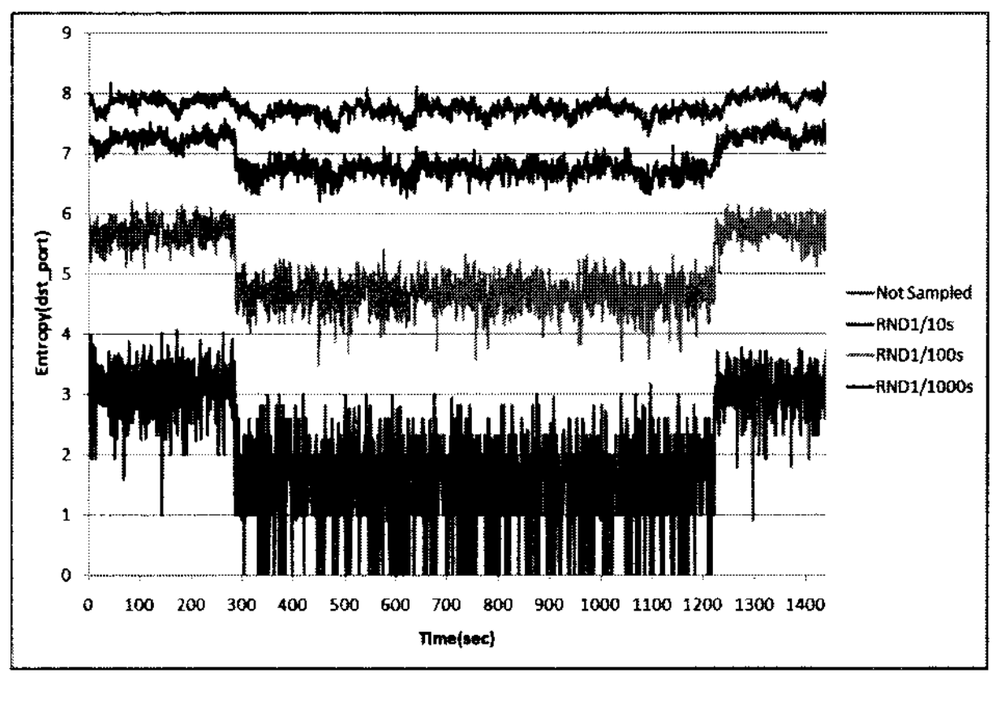
(a) 단순 랜덤 샘플링



(b) 계수 기반 규칙적 샘플링



(c) (규칙적 샘플링을 이용한) 층화 샘플링



(d) (단순 랜덤 샘플링을 이용한) 층화 샘플링

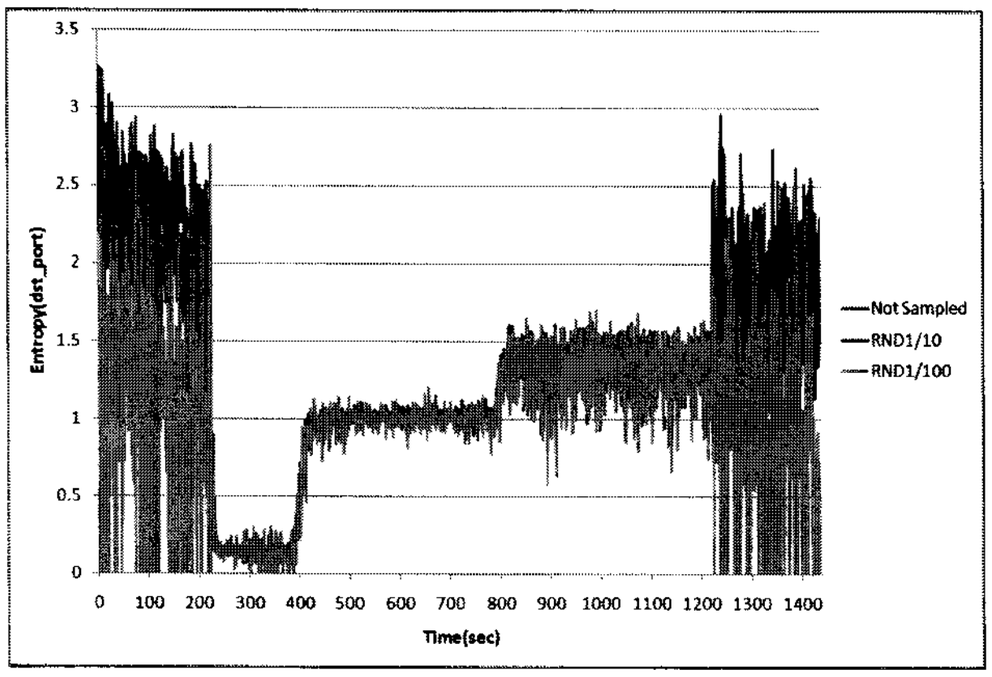
그림 11. TCP 트래픽 샘플링 결과의 엔트로피 비교
Fig. 11. Comparison on Entropy of Sampling Results in TCP Traffic.

층화 샘플링 결과에 엔트로피를 구한 그림 11 (c)에서는 1/1000에서도 매우 적은 왜곡을 보이고 있다. (물론 비정상 구간에서 엔트로피 감소가 강하게 보이는 현상은 그대로다) 이것은 전송층 프로토콜별로 층을 나누어 층화 샘플링하는 것이 적절함을 뜻한다. 층화 샘플링에서 사용한 규칙적 샘플링 자체가 트래픽의 흐름을 잘 반영하고 있는데다가, 일반적인 모니터링 단위인 전송층 프로토콜별로 샘플링을 하였기 때문에 좋을 결과를 보이고 있다. 즉, 층화 샘플링으로 인해 TCP든 UDP든 한쪽으로 샘플링 비율이 치우치지 않게 했기 때문에 결과가 좋은 것이다. 비슷하게 전송층 프로토콜별로 단순 랜덤 샘플링을 실행한 그림 11 (d)에서는 그림 11 (a), (b)보다 더 뛰어난 점을 찾기 어려웠다. 대체적으로 단순 랜덤 샘플링보다는 규칙적 샘플링이 인터넷 트래픽을 층화 샘플링하는 기법으로 더욱 적절함을 확인할 수 있다. 정리하면 랜덤한 추출보다는 규칙적인 추출이 대체적으로 더 좋을 결과를 기대할 수 있다는 것이다.

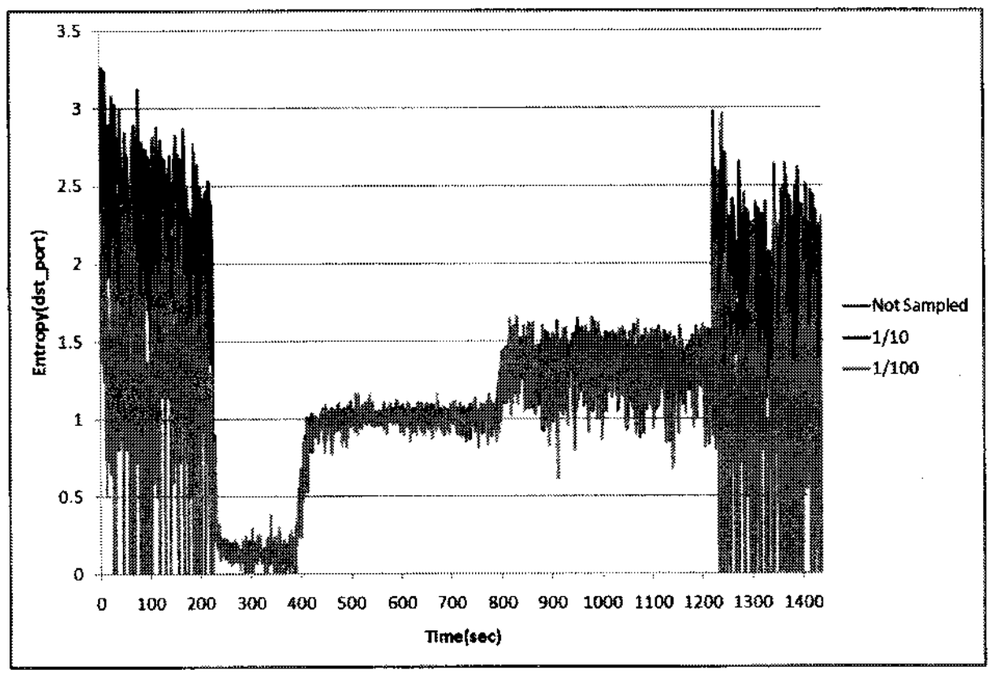
다. UDP 트래픽

그림 12는 UDP 기반 트래픽에 대한 entropy/sec 그래프를 나타낸 것이다. 그리고 UDP flooding 공격에 의한 비정상 트래픽이 약 200초부터 1200초 구간에서 발생하고 있다. 그림 12에서 공통적으로 흥미로운 엔트로피의 층(floor) 형태가 나타나고 있다. 이것은 UDP flooding공격이 어떤 destination IP address에 대해 하나의 destination port를 대상으로 하지 않기 때문이다. 하나의 호스트에 대해 총 3개의 destination port가 공격 대상이 되었다. 차례로 6667 (IRC), 80 (HTTP), 53 (DNS) 포트에 대한 트래픽이 발생하고 있다. 즉, 첫 번째 층에서는 6667 포트만 공격 대상이 되었으며, 두 번째 층에서는 6667과 80포트, 세 번째 층에서는 6667, 80, 53 포트가 함께 공격 되고 있다. 이러한 분석을 통해서 발생한 공격은 다중 포트 UDP flooding 공격으로 분석되었다.

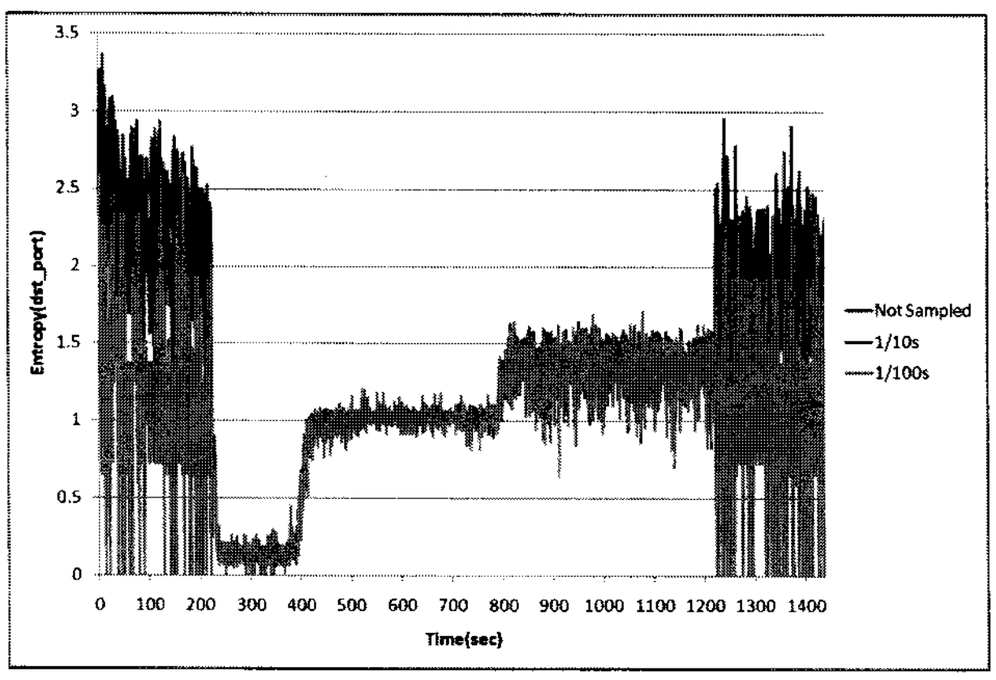
참고로, 샘플링 강도가 1/1000인 경우, 트래픽이 너



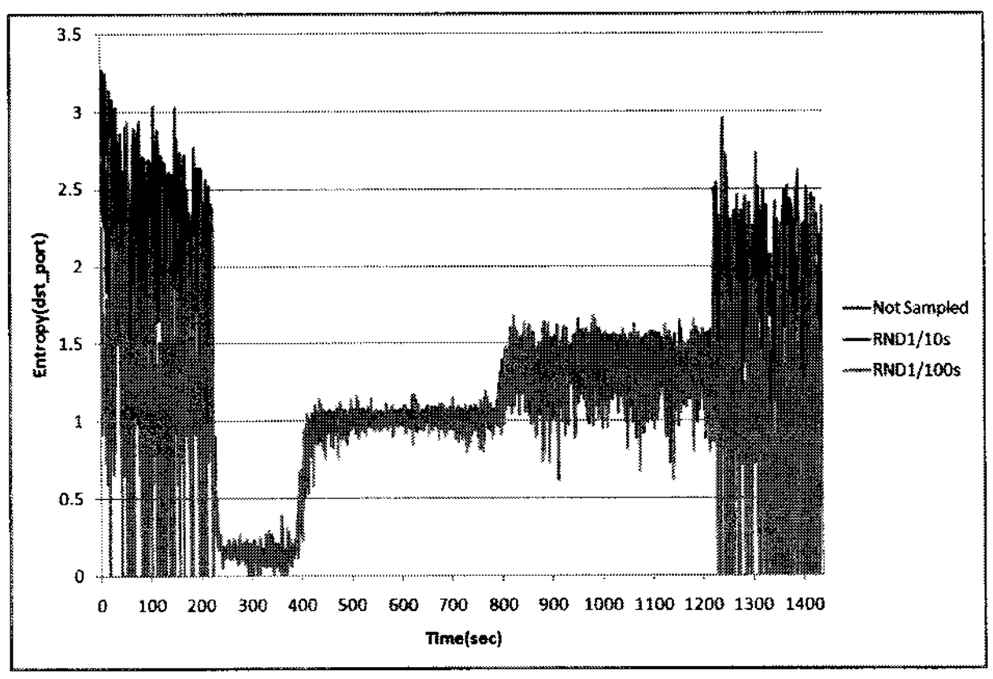
(a) 단순 랜덤 샘플링



(b) 계수 기반 규칙적 샘플링



(c) (규칙적 샘플링을 이용한) 층화 샘플링



(d) (규칙적 샘플링을 이용한) 층화 샘플링

그림 12. UDP 트래픽 샘플링 결과의 엔트로피 비교
Fig. 12. Comparison on Entropy of Sampling Results in UDP Traffic.

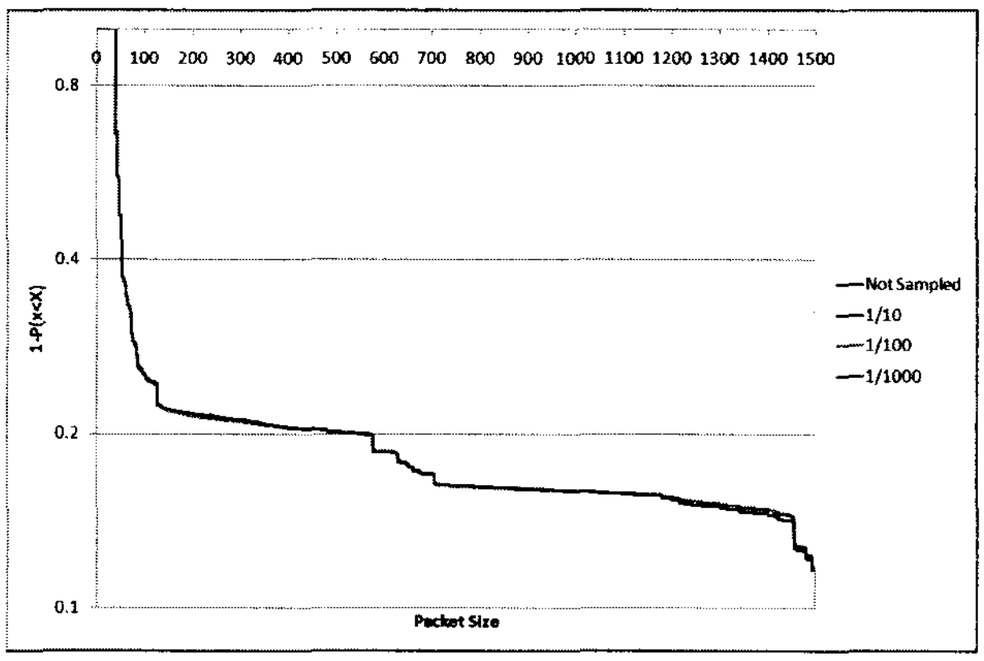
무 많이 손실되므로 1초단위의 UDP 트래픽에 대한 엔트로피는 의미가 없다. 즉, 1/1000의 샘플링을 적용했을 때, 어떤 판단을 할 수 있는 정도의 엔트로피를 얻지 못했다. 따라서 그림 12에서 1/1000의 그래프는 생략하였다.

라. 분석

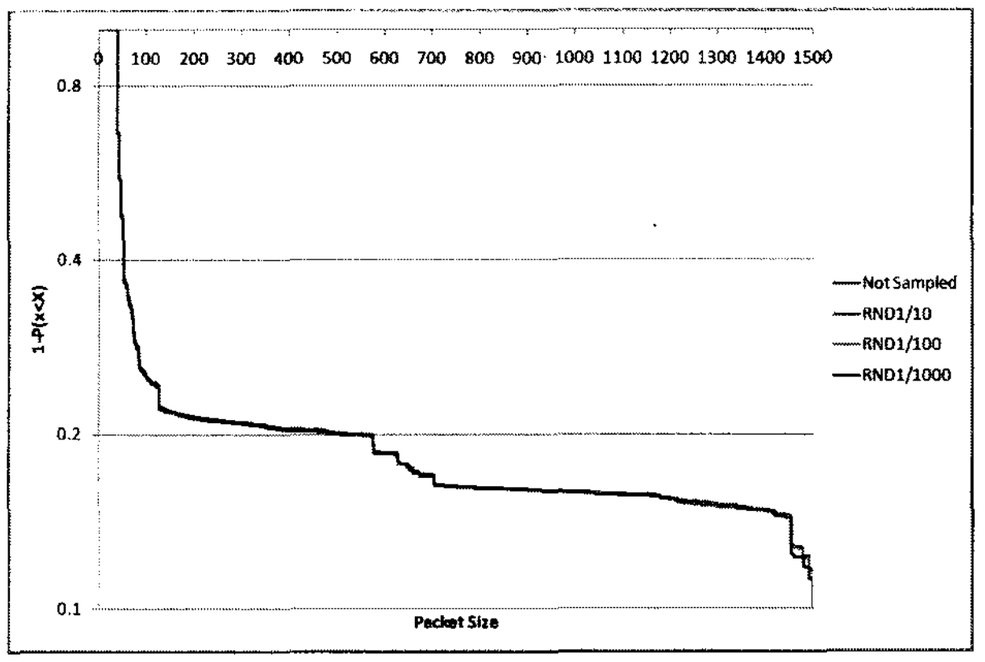
기존 연구에서 밝혀졌듯이, 본 연구의 결과에서도 엔트로피는 샘플링의 영향을 비교적 적게 받고 있다. 그러나 일부 새로운 특성을 발견하였다. 다중 포트 UDP flooding 공격에 의해, 병목 현상을 겪은 TCP 트래픽은 트래픽 크기에서 약간의 감소가 발생했다. 그럼에도 TCP 트래픽의 엔트로피는 변화가 거의 없었다. 유의할 것은, 샘플링된 TCP 트래픽에서 구한 엔트로피는 비정상 구간에서의 감소된 것이다. 이러한 감소의 정도는 샘플링 강도에 따라서 더욱 강해졌다. 원본 TCP 트래픽에서 내부 복잡도에 변화가 없었으나, 샘플링이 강할수록 마치 복잡도가 크게 달라진 것 같은 결과를 얻었다. 샘플링된 데이터에서 엔트로피 분석을 할 경우 반드시 고려해야 할 사항이다. 반면, 직접적으로 공격이 발생한 UDP 트래픽의 엔트로피는 샘플링과 상관없이 거의 동일한 엔트로피를 유지했다 (트래픽 크기 분석에서는, 트래픽 크기 자체는 줄었으나 트래픽 크기의 흐름을 유지했었다). 분석을 종합하면, 엔트로피 분석이 샘플링과 무관하게 이상 탐지 활용될 수 있음을 다시 확인할 수 있었다.

4. 패킷 크기 분석

패킷 크기 분포는 인터넷 측정에서 기본적인 측정 항목이며, 네트워크 구성요소 설계에 영향을 미치는 중요한 부분이다[1, 16]. 그러나 기존 샘플링 연구[11~12]에서는 패킷 크기와 관련된 분석을 간과하고 있다. 본 절에서는 샘플링이 패킷 크기 분포에 어떤 영향을 미치는지 살펴보겠다.

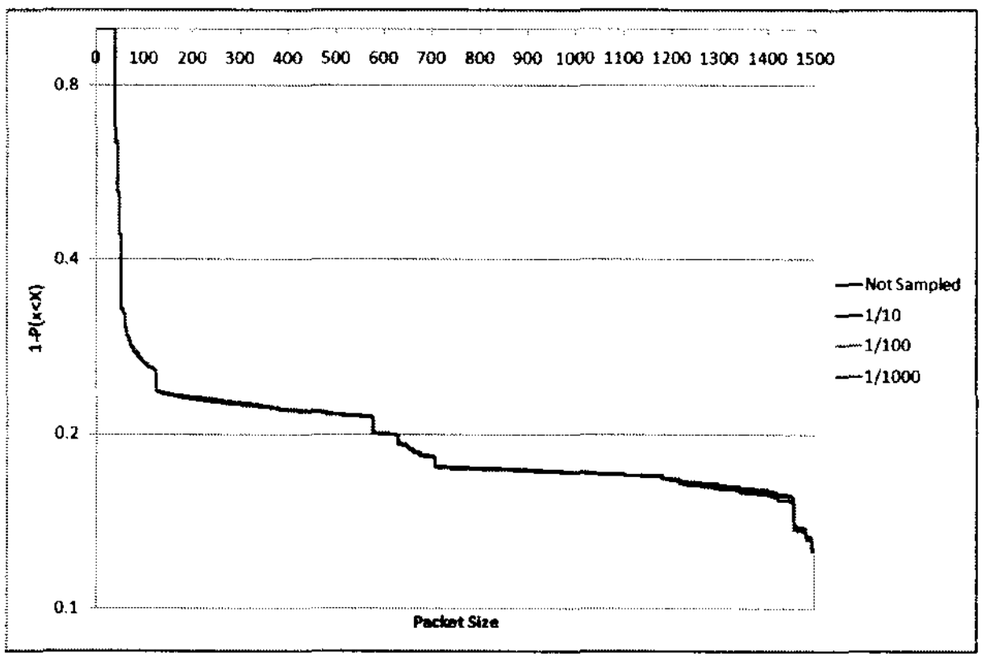


(a) 규칙적 샘플링 결과

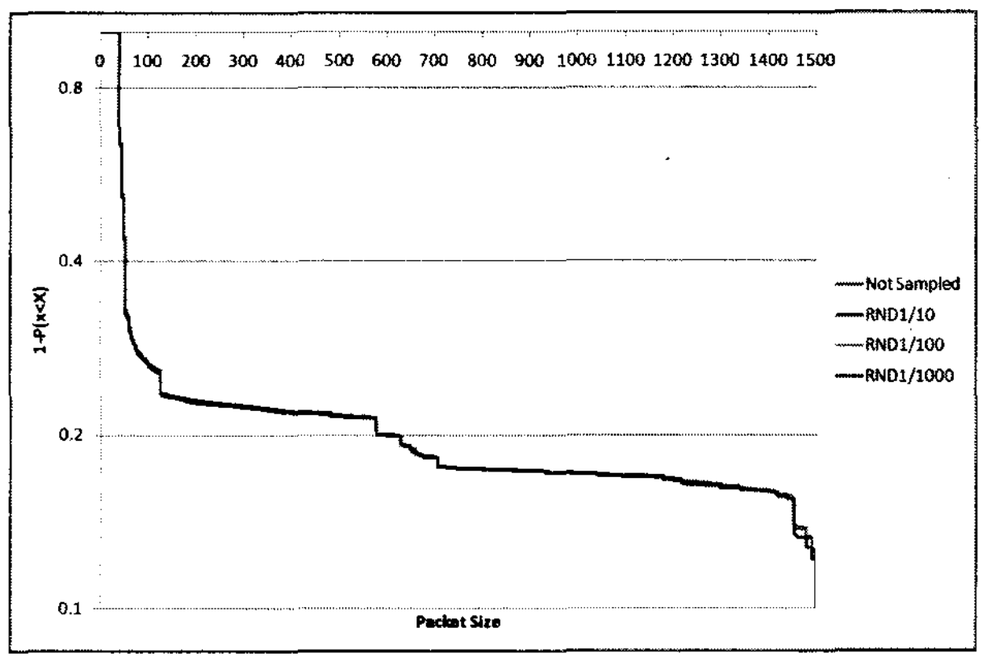


(b) 단순 랜덤 샘플링 결과

그림 13. 패킷 크기 분포 - 30분, 정상 구간
Fig. 13. Packet Size Distribution - 30 min, Normal Period.



(a) 규칙적 샘플링 결과



(b) 단순 랜덤 샘플링 결과

그림 14. 패킷 크기 분포 - 30분, 정상 구간, TCP
Fig. 14. Packet Size Distribution - 30 min, Normal Period.

가. 전체 트래픽

그림 1 (b)과 표 1에서 Period2 (정상 구간)에 해당하는 30분간의 트래픽을 대상으로 패킷 크기 분포를 그림 13와 같이 나타내었다. 그림 13에서 볼 수 있듯이, 샘플링 방식이나 샘플링 강도 (1/10, 1/100, 1/1000)와는 무관하게 원본과 샘플링된 트래픽의 패킷 크기 분포가 매우 유사하다.

본 연구 중에 미리 예상했던 것과는 달리 패킷 크기 분포는 샘플링에 매우 강한 모습을 보여주고 있다. 1/1000정도의 샘플링 강도에도 매우 훌륭하게 원본과 유사한 패킷 크기 분포가 나타나고 있는 것은 매우 인상적이다.

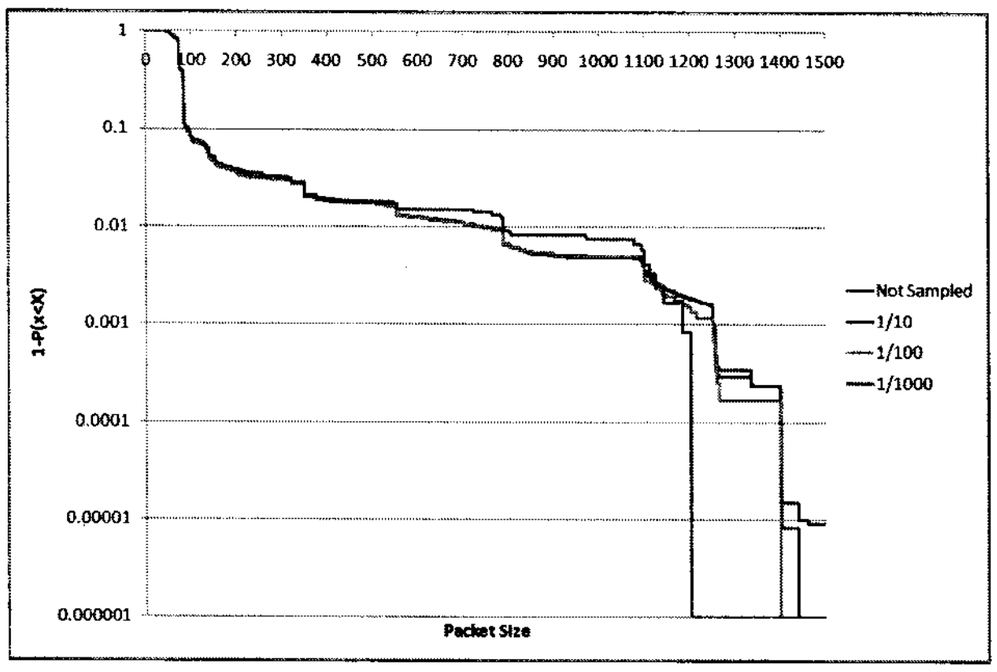
나. TCP 트래픽

그림 14는 TCP 트래픽의 패킷 크기 분포를 나타낸 것이다. TCP 트래픽의 패킷 크기 분포는 전체 트래픽의 패킷 크기 분포와 매우 유사하다. 전체 트래픽의 80-90%정도가 TCP 트래픽임을 생각해본다면 그림 13와 14가 유사한 것은 납득할 수 있다. TCP 트래픽의 패킷 크기 분포 역시 샘플링의 종류나 강도와는 상관없음을 확인할 수 있다.

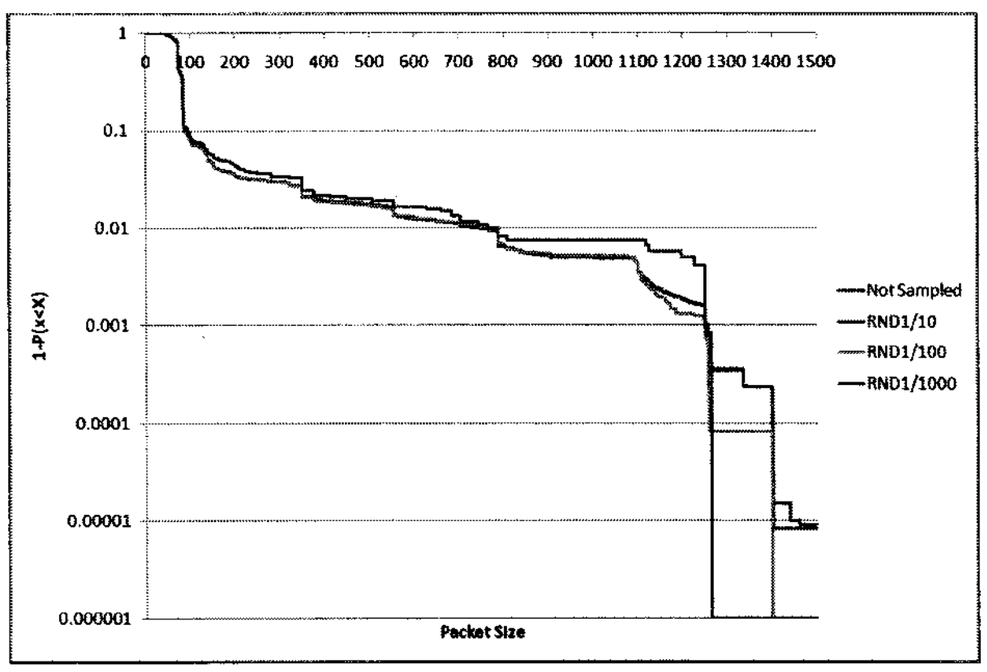
다. UDP 트래픽

그림 15는 UDP 트래픽의 패킷 크기 분포를 나타낸 것이다. UDP 트래픽의 패킷 크기 분포는 TCP 트래픽의 패킷 크기 분포와는 달리 약간의 차이가 감지된다. 전체적으로 봤을 때, 샘플링 강도에도 분포를 잘 유지하고 있으나, 1200바이트 이상의 분포는 약간의 차이가 있으나 크게 문제가 될 만한 부분은 아니다.

그림 15의 (a)와 (b)에서 y축의 0.1부터 0.01의 범위



(a) 규칙적 샘플링 결과



(b) 단순 랜덤 샘플링 결과

그림 15. 패킷 크기 분포 - 30분, 정상 구간, UDP
Fig. 15. Packet Size Distribution - 30 min, Normal Period.

를 보면 약간의 성능차이가 보인다. 규칙적 샘플링이 단순 랜덤 샘플링에 비해 좋은 결과를 보이고 있다.

라. 분석

패킷 크기 분포가 샘플링의 영향을 거의 받지 않는다는 것은 본 연구에서 밝혀진 사실이다. 심지어 1/1000의 강도로 샘플링된 트래픽의 패킷 크기 분포나 원본의 패킷 크기 분포를 구분하기 쉽지 않았다. 트래픽 크기 분석이나 엔트로피 분석에서 밝혀졌듯이, 이상 탐지에 샘플링이 유용한 만큼 패킷 크기 분포를 얻는데도 샘플링을 유용하게 활용할 수 있다. 덧붙여, 가능한 규칙적 샘플링을 사용하는 것이 좋은 결과를 쉽게 얻을 수 있다.

Ⅳ. 결 론

이제 인터넷은 일상생활에서 수도나 전화망과 같은 기간시설과 유사한 위치에 있다. 이러한 인터넷에서 예상치 못한 문제가 발생한다면, 우리 사회는 혼란에 빠지게 된다. 인터넷 측정 분야에서는 인터넷에서 발생되는 트래픽을 분석하여, 다양한 상황에 대처할 정보를 얻게 한다. 수동적 인터넷 측정에서 가장 큰 어려움은 대용량의 트래픽을 어떻게 효과적으로 관리하고 처리할 것이냐는 것이다. 이에 대한 현실적인 해결책으로 트래픽 샘플링이 사용되고 있다. 본 논문에서는 트래픽 샘플링이 인터넷 측정에 어떤 영향을 미치는지 검증하였다. 구현과 적용이 용이해서 많이 사용되는, 규칙적 샘플링, 단순 랜덤 샘플링, 층화 샘플링에 대하여 분석하였고, 샘플링 강도는 1/10, 1/100, 1/1000을 사용하였다. 연구 대상 트래픽 데이터로는 실제로 우리나라 백본망에서 캡쳐된 트래픽이 사용되었으며, DoS 공격과 같은 다양한 이상 트래픽을 포함하고 있다. 세부 분석 항목은 "트래픽 크기 분석", "엔트로피 분석", "패킷 크기 분석"이며, 다른 연구와는 달리 패킷 크기와 관련된 분석에서 새로운 사실을 발견한 것은 차별화된다.

트래픽 크기 분석에서, (주로 1초단위의 시간 단위로 분석할 경우) 대체적으로 1/100보다 작은 샘플링 강도가 유용하며, 샘플링 방식은 계수 기반 (Count-based) 샘플링 방식의 규칙적 샘플링이나 단순랜덤 샘플링 모두 유용함을 확인하였다. 더 좋은 결과를 얻기 위해서 TCP나 UDP와 같은 전송층 프로토콜단위로 층화 샘플링하는 것이 효율적으로 좋은 샘플링 결과를 얻을 수 있었다. 가능한 이상 트래픽 탐지에서는 샘플링이 사용하는 것이 유리한 것도 확인하였다.

엔트로피 분석에서, 이상 트래픽 탐지에 엔트로피 분석이 많이 활용되고 있다. 기존 연구들과 같이 본 연구에서도 엔트로피는 샘플링 강도에 영향을 매우 적게 받는 것을 확인하였다. UDP 기반 이상 트래픽으로 인해 병목현상을 겪은 TCP 트래픽이 샘플링 강도에 따라 잘못된 결과를 보임을 발견하였다. 이는 잘못된 분석을 가능하게 하므로 유의해야 한다.

패킷 크기 분석에서, 패킷 크기 분포는 샘플링의 방식이나 강도에 영향을 거의 받지 않는 것으로 밝혀졌다. 패킷 크기 분포는 인터넷 측정에 가장 기본이 되는 만큼, 샘플링과 패킷 크기 분포의 관계를 분석한 것은 가치가 있다. 패킷 크기 분포를 얻기 위해서 인터넷 측정을 할 경우, 1/1000정도도 무난하게 사용가능했다. 미세하지만 규칙적 샘플링이 단순 랜덤 샘플링에 비해 좋을 결과를 보였다.

본 연구의 결과는, 심도있는 검증없이 간단하게 구현되어 쓰였던 트래픽 샘플링 방식들을 실증적 데이터로 비교 분석하여 나온 것이다. 앞으로 계속 진행될 인터넷 측정 연구를 위한 기반 연구로서, 본 연구의 성과는 트래픽 샘플링을 위해 필수적으로 참고할만한 자료가 될 것이다.

향후 연구 과제로는 샘플링에 따른 플로우 (Flow)의 특성 변화를 분석하는 것과 최근에 발표된 트래픽 분산 그래프 (TDG, Traffic Dispersion Graph)를 이용한 샘플링된 트래픽의 분석을 목표로 하고 있다.

참 고 문 헌

[1] M. Crovella and B. Krishnamurthy, "Internet Measurement: Infrastructure, Traffic, and Applications", John Wiley & Sons, Ltd, 2006.

[2] D. Moore and G. M. Voelker and S. Savage, "Inferring Internet Denial-of-Service Activity", In Proc. of Usenix Security Symposium, pp. 9-22 Washington, DC, August 2001.

[3] J. Mirkovic and P. Reiher, "A Taxonomy of DDoS attack and DDoS defense Mechanisms", ACM SIGCOMM Computer Communication Review, Vol. 34, Issue 2, pp. 39-53, April 2004.

[4] R. R. Panko, "Corporate Computer and Network Security", Prentice Hall, 2004.

[5] CERT Advisory MS-SQL Server Worm, "http://www.cert.org/advisories/CA-2003-04.html", January 2003.

[6] CERT Advisory W32/Blaster worm, "http://www.cert.org/advisories/CA-2003-20.html", August 2003.

[7] D. Moore and V. Paxson and S. Savage and S. Staniford and N. Weaver, "Inside the Slammer worm", IEEE Security & Privacy, Vol. 1 issue 4, pp. 33-39, August 2003.

[8] Nick Duffield, "Sampling for Passive Internet Measurement: A Review", Statistical Science Vol. 19, No. 3, pp. 472-498, 2004.

[9] A. Lakhina and M. Crovella and C. Diot, "Characterization of Network-Wide Anomalies in Traffic Flows" In Proc. ACM Internet Measurement Conference, pp 201-206, Taormina, Sicily, Italy, October 2004.

[10] A. Soule and F. Silveira and H. Ringberg and C. Diot, "Challenging the Supremacy of Traffic Matrics", In Proc. ACM Internet Measurement Conference, pp 105-110, San Diego, California, USA, October 2007.

[11] D. Brauckhoff and B. Tellenbach and A. Wagner and M. May and A. Lakhina, "Impact of Packet Sampling on Anomaly Detection Metrics", In Proc. ACM Internet Measurement Conference, pp 159-164, Rio de Janeriro Brazil, October 2006.

[12] J. Mai and C. Chuah and A. Sridharan and T. Ye and H. Zang, "Is Sampled Data Sufficient for Anomaly Detection?", In Proc. ACM Internet Measurement Conference, pp 165-176, Rio de Janeriro, Brazil, October 2006.

[13] J. Xia and L. Gao and T. Fei, "Flooding Attacks by Exploiting Persistent Forwarding Loops", In Proc. ACM Internet Measurement Conference, pp 36-41, Berkeley, CA, USA, October 2005.

[14] ping, "http://en.wikipedia.org/wiki/Ping", Wikipedia

[15] RFC 1393, Traceroute Using an IP Option, "http://tools.ietf.org/html/rfc1393"

[16] W. John and S. Tafvelin, "Analysis of Internet Backbone Traffic and Header Anomalies observed", In Proc. ACM Internet Measurement Conference, pp 111-116, San Diego, California, USA, October 2007

[17] J. Kim and S. Ahn and Y. Won, "Mining An Anomaly: On The Small Time Scale Behavior of The Traffic Anomaly", In Proc. of IADIS International Conference WWW/Internet, Murcia, Spain, PP. 552-559, October 2006

[18] Juniper Traffic Sampling, "http://www.juniper.net/techpubs/software/junos/junos60/swconfig60-policy/html/sampling-overview.html"

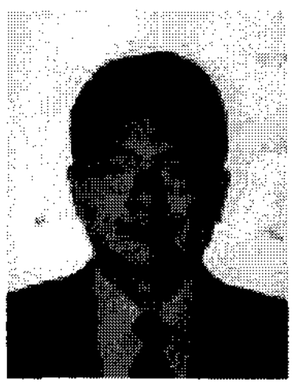
[19] TCPDUMP/LIBPCAP public repository, "http://tcpdump.org"

[20] T. M. Cover and J. A. Thomas, "Elements of Information Theory", Wiley Interscience, 1991.

[21] A. Lakhina and M. Crovella and C. Diot, "Mining Anomalies using Traffic Feature Distributions", ACM SIGCOMM Computer Communication Review, Vol 35, Issue 4, pp. 217-228, October 2005.

저 자 소 개

김 정 현(학생회원)
2004년 명지대학교 컴퓨터공학과 학사 졸업.
2008년 한양대학교 전자컴퓨터 통신공학과 박사과정수료.
<주관심분야 : 인터넷 측정, 이상 탐지>



안 수 한(비회원)
1992년 서울대학교 계산통계학과 학사 졸업.
1994년 서울대학교 계산통계학과 석사 졸업.
2000년 서울대학교 통계학과 박사 졸업.
2001년~2003년 AT&T Labs-Research, Post-Doc, Consultant.
2004년~현재 서울시립대학교 통계학과 부교수
<주관심분야 : Fluid Flow Model, Queueing, Telecommunication Network>

원 유 집(정회원)
1990년 서울대학교 계산통계학과 학사 졸업.
1992년 서울대학교 전산과학과 석사 졸업.
1997년 University of Minnesota 박사 졸업.
1997년~1999년 Intel 연구원.
1999년~현재 한양대학교 전자컴퓨터통신공학과 부교수.
<주관심분야 : 운영체제, 컴퓨터네트워크, 성능평가>