마이닝을 이용한 이상트래픽 탐지 : 사례 분석을 통한 접근

(Detection of Traffic Anomalities using Mining : An Empirical Approach)

김 정 현 † 안 수 한 †† 원 유 집 ††† 이 종 문 †††† 이 은 영 ††††
(JungHyun Kim) (Soohan Ahn) (Youjip Won) (Jongmoon Lee) (Eunyung Lee)

요 약 본 논문에서는 실제 인터넷 백본으로부터 일주일간 캡쳐한 트래픽을 대상으로 기초 통계 분석을 하고, 여기서 발생한 이상트래픽을 분석한다. 이상트래픽은 국외에서 국내로 유입되는 UDP 기반 트래픽에서 나타났다. 트래픽 자료에 대한 탐색적 분석 결과 packets/sec 분포와 bytes/sec 분포에서 이상트래픽이 발생할 경우에 나타나는 새로운 형태의 특성이 발견되었다. 본 연구에서는 이러한 이상트래픽의 원인이 되는 플로우를 분류하기 위하여 자율학습(unsupervised learning) 방법의 하나인 분류분석(k-means clustering)을 이용하였으며, 분류된 플로우의 특성분석을 토대로 발생한 이상트래픽은 DoS 공격의 일종에 의한 것으로 결론지었다. 또한 본 연구에서는 이상트래픽의 원인이 되는 플로우의 존재 시점을 탐지하기 위하여 새로운 기법을 제시한다. 제시된 기법은 분포적합검정(goodness of fit test)의 한 방법인 Cramer-Von-Misses 검정에서 쓰이는 통계량에 바탕을 두고 있으며 1초 단위의 탐지기법이다. 제시된 기법의 응용 결과, 이상트래픽의 존재 시점으로 판단된 시점과 DoS 공격으로 판단된 플로우들의 시점이 일치함을 확인할 수 있었다.

키워드 : 이상탐지, 인터넷 보안, 인터넷 평가, 사례연구

Abstract In this paper, we collected the physical traces from high speed Internet backbone traffic and analyze the various characteristics of the underlying packet traces. Particularly, our work is focused on analyzing the characteristics of an anomalous traffic. It is found that in our data, the anomalous traffic is caused by UDP session traffic and we determined that it was one of the Denial of Service attacks. In this work, we adopted the unsupervised machine learning algorithm to classify the network flows. We apply the k-means clustering algorithm to train the learner. Via the Cramer-Von-Misses test, we confirmed that the proposed classification method which is able to detect anomalous traffic within 1 second can accurately predict the class of a flow and can be effectively used in determining the anomalous flows.

Key words : Anomaly Detection, Internet Security, Internet Measurement, Empirical Study

1. 서 론

2000년 2월에 미국 대형 포털 사이트들이 DDoS (Distributed Denial of Service) 공격에 의해 서비스가 중단되는 사건이 발생했고, 2001년 1월에는 비슷한 공격으로 인해 마이크로소프트의 네임서버(name server infrastructure)가 다운되었다[1]. 국내에서는 MS-SQL 서버의 취약점을 이용한 슬래머웜(Worm.SQL.Slammer)으로 인해 국내의 인터넷망 전체가 마비되는 사태가 벌어지기도 했다. 이것은 네트워크의 비정상적인 현상이 국가 전체에 큰 충격을 준 사건이었다. 초고속 통신망이 국가의 기간 구조 중에 매우 중요한 위치를 차지함에 따라서 네트워크의 이상이 사회에 미치는 파장이 더욱 커지게 될 것이다. 따라서 네트워크에 심각한 영향을 미치는 이상 트래픽(anomalous traffic)에 대한 연구는 국가 보안에 매우 지대한 역할을 할 것으로 판단된다. 현

† 비 회 원 : 한양대학교 전자통신컴퓨터공학부
sohankch@ece.hanyang.ac.kr
†† 비 회 원 : 서울시립대학교 통계학과 교수
sahn@uos.ac.kr
††† 종신회원 : 한양대학교 전자통신컴퓨터공학부 교수
yjwon@ece.hanyang.ac.kr
†††† 비 회 원 : 국가보안기술연구소 연구원

논문접수 2005년 10월 17일
심사완료 : 2006년 3월 23일

재 네트워크의 가장 큰 문제점으로, 첫째, 바이러스, 웜, 스파이웨어 등 새로운 형태의 이상트래픽이 계속적으로 발생하고 있다[1]. 둘째, 정상트래픽(normal traffic)과 이상트래픽을 정확히 구별할 수 있는 트래픽 특성 모델링에 대한 연구가 매우 미진하다[2,3]. 셋째, 대부분의 트래픽 특성화 연구가 이론적인 모델에서의 트래픽 특성화 작업으로 되어 있으며 실제 트래픽을 대상으로 한 검증 작업이 거의 없는 상황이다[4]. 인터넷에서 어렵지 않게 네트워크 공격 도구를 취득할 수 있는 현재 상황에서, 이러한 공격 도구가 발생시키는 이상트래픽에 대한 연구는 매우 시급하다고 할 수 있다. 특히, 초고속 통신망이 전 국토에 설치된 우리 나라의 트래픽 특성과 그렇지 않은 다른 나라의 트래픽 특성이 분명히 다르다는 점에 주목해야 한다. 이런 이유로 국내의 트래픽의 특성에 초점을 맞춘 본 연구의 가치는 매우 크다고 할 수 있을 것이다.

인터넷 트래픽의 특성에 대한 많은 연구가 지금까지 진행되어 왔다. 인터넷 백본이나 랜 등에 대한 트래픽의 통계적 특성을 규명하려는 많은 노력이 있었으며 초기에 인터넷 트래픽의 chaotic 성질에 대한 발견이 주목할 만 하다[5-8]. 최근 들어 백본 트래픽의 행태를 작은 시간 축에서 세밀하게 분석한 논문이 발표되었다[9]. 해당 논문에서는 백본망의 속도와 이에 영향을 받는 패킷 간의 간격을 연구하였다. 최근 인터넷에서 발생하는 "이상" 징후를 다양한 기법을 써서 발견하는 노력이 제시되었다. Barford et al.[10]은 인터넷 트래픽에서 플로우 수준의 이상 트래픽 특성을 연구하였다. 이들의 연구는 FLowscan이라는 공개소프트웨어를 이용하여 Netwflow에 의해 제공되는 CISCO 라우터 데이타를 이용하고 있다. Wavelet을 사용하여 네트워크 특정부분의 이상징후를 발견하는 방법[11]이 제시되었다. 정보이론(information theory)에서 자주 사용되는 엔트로피 개념을 이용하여 정보의 이상 징후 여부를 판단하고[4,12,13], 또는 플로우들을 분류하는 기법이 제시되었다[14]. 실제 네트워크 환경에서는 이상 징후가 정확히 정의될 수 없다. 따라서 이상 징후의 여부를 즉시 판단해 내는 것은 매우 어렵고 추상적인 작업이다. 그리고, 어느 수준까지를 이상 징후라고 판단해야 하는 지 역시 매우 어려운 작업이 아닐 수 없다. 위에서 언급한 기법들은 이상 징후의 특별한 정의 없이 매우 추상적으로 그들을 추려낼 수 있는 자율학습(unsupervised learning)에 근거한 기계 학습 기법을 제시하고 있다. 특정 형식의 공격 형태를 구체적으로 특성화하는 노력들이 최근 많이 등장하였다. 특히 DoS(Denial of Service) 공격은 매우 빨리 정확히 판명되어야 하기 때문에 이것의 통계적인 특성에 대한 분석은 큰 의미를 지닌다 하겠다[1,3,15]. Dos 공격 외에 네트워크에 피해를 입힐 수 있는 트래픽으로 인터넷 웜이 있다. 그러나 인터넷 웜이 발생시키는 트래픽은 양이 매우 적기 때문에 단순히 트래픽 양만을 감시하는 접근 방법에서는 이들을 발견해 내는 것이 매우 어렵다. 인터넷 웜의 확산을 전염병의 확산 속도를 예측하는 모델을 이용하여 유추하는 기법[16]이 제시되었다. 이를 온라인으로 발견해 내는 것은 쉽지 않은 작업이나 기존에 존재하는 웜의 트래픽을 분석하고 특성화하여, 현재 트래픽의 웜 존재 여부를 검사하는 접근방법은 웜 발견에 매우 유용하게 사용될 수 있다 하겠다[13].

본 논문은 논문의 목적과 철학적인 측면에서 Xu et al[14]이 발표한 논문과 상당한 부분을 공유한다고 볼 수 있다. 그러나, 구체적인 방법에서는 서로 상당히 다른 접근 방법을 취하고 있으며, 그로 인해 얻어지는 결과 역시 상호 보완적이라 할 수 있다. Xu et al.의 연구에서는 *srcIP*, *dstIP*, *srcPrt*, *dstPrt*에 대한 엔드로피(entropy)를 이용해서 마이닝 메트릭을 하였다. 본 논문에서는 마이닝의 메트릭으로 각 세션의 패킷 간격의 누적 분포함수를 이용하였다는 차이가 있다. 이는 많은 변수를 이용하여 트래픽을 분류하는 기법에 비해서 매우 효율적이라 할 수 있다. 특히 본 논문에서는 샘플링이 아닌 모든 패킷들을 수집해서 분석한 자료를 근거로 연구를 수행하였으므로 샘플링에서 발생할 수 있는 자료의 정확도 손실이 발생하지 않는다. Xu et al.을 포함한 대부분의 기존 연구는 상대적으로 큰 단위의 누적된 자료(5분에서 10분 단위의 자료 등)를 근간으로 분석하였다. 그러나, 인터넷의 속도 그리고 웜이나 DOS 공격이 네트워크에 심각한 영향을 미칠 때까지의 시간을 고려하면 큰 단위로 누적된 자료의 분석은 의미가 없을 수 있다. 본 연구에서는 1/10,000 초 단위시간까지 패킷의 시간을 분석함으로써 신속한 이상 트래픽 탐지에 대한 가능성을 확보하였다. 또한 Xu et al.을 포함한 기존 연구는 다른 유럽이나 미국의 네트워크 인프라에 대한 연구이다. 유럽이나 미국에서 발생하는 트래픽은 네트워크 인프라가 전 국토에 설치되어 있는 우리 나라와는 트래픽 특성이 다르다. 본 연구는 우리 나라의 트래픽에 대한 흔치않은 연구로써 매우 큰 가치가 있다고 할 수 있다. 본 연구를 기존 연구와 비교할 경우, 인터넷 이용과 관련된 트래픽의 (나라마다 다른) 문화적 측면까지도 살펴볼 수 있을 것이다.

본 논문은 다음과 같이 구성되었다. 2장에서는 인터넷 백본 트래픽(Internet backbone traffic)의 일주일간 기초 통계와 이 중에서 나타난 이상 트래픽을 간단히 살펴본다. 3장에서는 이상 트래픽의 특성을 자세히 분석하기 위해서 포트(port), 패킷수(packet count), 바이트량(bytes)을 기준으로 여러 가지 분석을 수행하였다. 4장

에서는 트래픽의 플로우 수준(flow level)에서 분석하였으며, 플로우를 이용한 이상 트래픽의 탐지 가능성을 살펴본다. 5장에서는 플로우의 특성과 분류분석을 다루었으며 6장에서는 이상트래픽의 발생 또는 존재유무를 탐지할 수 있는 기법을 제시하고자 한다.

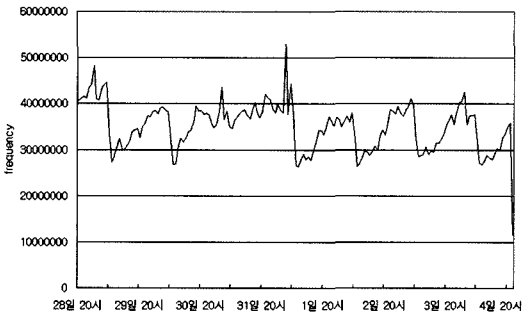
2. 미시적 인터넷 백본 트래픽의 거시적 분석

2.1 트래픽 수집 환경

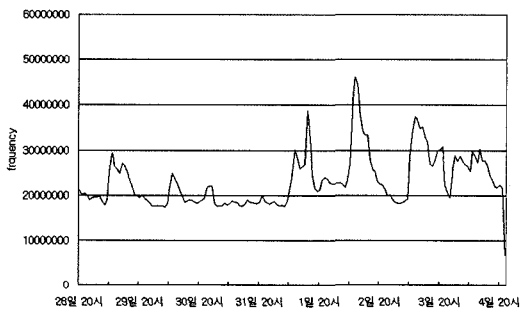
본 연구에서는 대용량의 네트워크 패킷 자료를 매우 세밀한 미시적 수준까지 수집하여 해당 네트워크 트래픽에 대한 통계적 특성, 성능적 특성 그리고 전체 서브넷에 미치는 영향까지고 분석할 수 있는 토대를 구축하였다. 2004년 10월 28일부터 11월 4일경까지 국내와 국외 사이를 통과하는 7일 분량의 트래픽 전체를 수집하였다. 수집된 총 데이타의 양은 압축된 후 1.5 TByte이다. 이 트래픽은 우리나라 인터넷 사용자와 외국 인터넷 사용자의 인터넷 이용 실태를 파악할 수 있는 중요한 데이타다. 측정한 망의 대역폭은 155Mbit/sec이며, 패킷은 헤더(header)를 44바이트를 캡쳐하였다. 캡쳐된 데이타는 0.1 usec(1/10,000,000)초 단위로 도착시간이 기록되어있다. 일주일 분량의 트래픽은 libpcap형태로 저장되었다[17]. 방대한 자료로부터 원하는 형태의 정보를 추출하는 작업은 트래픽 특성연구의 매우 중요한 축을 이룬다. 불행히도 현존하는 많은 트래픽 분석 소프트웨어들은, 예를 들어, Ethereal[18]은 작은 양의 거시적 정보들로부터 자료를 추출하도록 설계되었으며 따라서 Tera Byte 급의 자료를 고속으로 처리하는 것에는 적합하지 않다. libpcap형태로 저장된 데이타에서 각종 통계 자료를 추출하고 분석을 하기 위해서 본 연구에서는 자체적으로 트래픽 분석도구를 개발하였다. libpacp 은 데이타를 이진(binary)으로 저장하며, 이것을 처리하기 위해서는 변환작업이 필요하다. tcpdump[17]라는 유틸리티로 libpcap형태의 데이타를 읽을 수 있는 데이타로 변환 가능하다. 이러한 데이타를 데이타베이스화하기 위해서는 tcpdump의 결과물은 다시 파싱(parsing)해야 한다. 이러한 중복 처리의 문제점을 보완하기 위해서 본 연구에서는 DMC traff Mon을 개발하였다. 이 도구는 빠른 속도로 libpcap형태의 데이타를 원하는 형태의 텍스트 형태로 변환하는 기능을 가지고 있다. 본 연구에서 tcpdump와 awk스크립트를 이용해서 처리했을 때 140시간 정도 걸리던 작업이, DMC traff Mon을 이용했을 때는 23~24시간 정도로 시간이 단축되었다.

2.2 주간통계

그림 1은 일주일 간의 트래픽을, 1시간 단위로 지나간 패킷량(packet count or packet frequency)으로 나타낸 그래프이다. ingress 트래픽은 국외에서 국내로 유입되는 트래픽을 의미하며, egress 트래픽은 국내에서 국외로 유출되는 트래픽이다.



(a) ingress 트래픽



(b) egress 트래픽

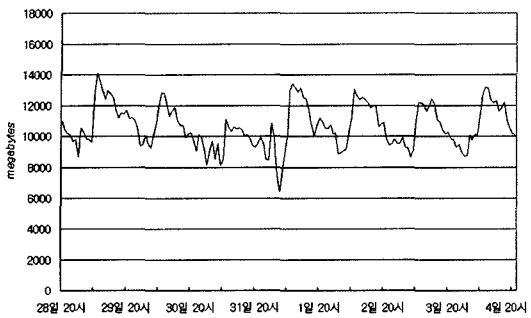
그림 1 일주일간 패킷량

그림 1(a)에서는 비교적 일정한 패턴이 반복적으로 나타나는 것을 볼 수 있다. 한 주기는 정확히 24시간에 해당한다. 그림 1(b)에서는 일정한 패턴이 나타나지 않는 것을 볼 수 있다. 우리나라 인터넷 사용자들이 국외의 인터넷 서비스를 매일 비슷한 패턴으로 사용하는 것을 알 수 있다. 국외 사용자들은 상대적으로 우리나라의 인터넷 서비스를 일정하게 사용하지는 않는 것을 알 수 있다.

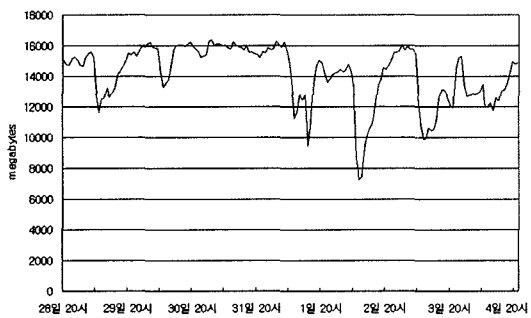
그림 2는 일주일 간의 트래픽을, 1시간 단위로 지나간 바이트량으로 나타낸 그래프이다. 그림 2(a)의 ingress 트래픽의 경우, 그림 1(a)와 유사하게 비교적 일정한 규칙성을 보이고 있다. 그림 2(b)의 egress 트래픽은 규칙성은 보이지 않고 있으나 전체적으로 바이트량이 ingress트래픽보다 크게 나타나고 있다. 다음 절에서는 ingress 트래픽에서 나타난 트래픽의 규칙성에 대해서 분석할 것이다.

2.3 일별 통계

하루 동안 트래픽량의 변화를 좀더 미시적인 측면에서 살펴보기로 한다. 그림 3은 각 시간대 별로 단위시간당 전달된 패킷의 개수(packets/hour)와 데이타의 양(megabytes/hour)을 나타내고 있다. 그림 3(a)는 각 일

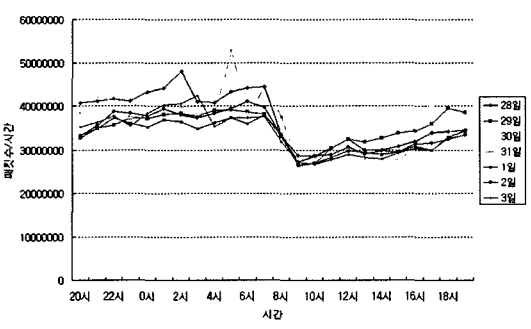


(a) ingress 트래픽

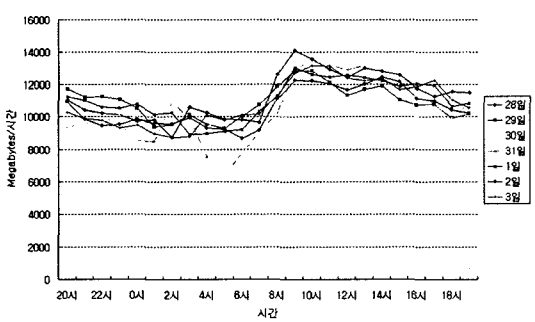


(b) egress 트래픽

그림 2 일주일간 바이트량



(a) packets/hour



(b) megabytes/hour

그림 3 ingress 트래픽의 일별 통계

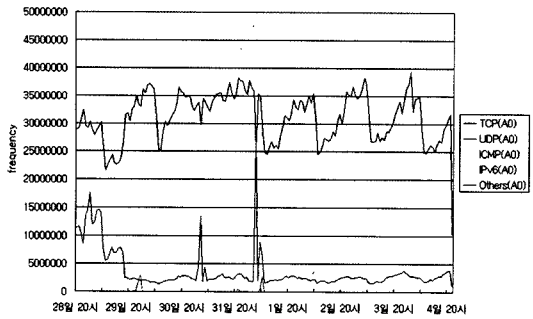
별로 나타나는 트래픽의 규칙성을 잘 보여주고 있다. 일반적으로 업무가 시작되는 아침 시간부터 패킷수(packets/hour)가 급격히 감소되는 형태가 나타나고 있다. 급격한 감소는 오전 7시 30분경부터 9시 30분경까지 나타나고 있다. 점심 시간인 12시부터 13시 사이에 단위시간당 패킷수(packets/hour)가 약간 증가하는 변화가 나타나고 있다. 이후에는 점차 packets/hour가 증가하고 있다. 이렇게 업무 시간의 특성을 반영한 packets/hour는 평일에 매일 반복되고 있다.

그림 3(b)에서는 그림 3(a)와 반대의 특성을 보이고 있다. 아침의 업무 시작 시간에 데이타의 양(megabytes/hour)가 급격하게 증가하며, 점심 시간인 12시부터 13시 사이에는 데이타의 양(megabytes/hour)이 약간 감소하는 변화가 나타나고 있다. 이후에는 점차 증가하고 있다. packets/hour와 megabytes/hour는 정반대로 나타나고 있다. 이에 대한 구체적인 해석을 위해서는 보다 심도 있는 연구가 필요하며 이에 대한 심층적인 해석은 본 연구의 범위에 포함되지 않는다. 그림 3에서 발생하는 특성의 가능한 원인은 업무 중에는 패킷의 크기가 큰 응용들을 주로 사용하지 않는가 하는 것이다. 이에 대한 일례로 업무가 없는 30일(일요일)에는 평일에 반복되는 패턴이 등장하지 않기 때문이다.

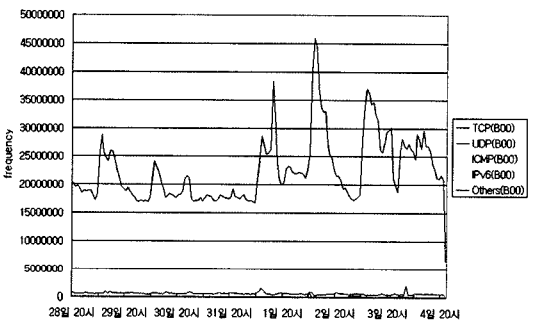
2.4 프로토콜별 통계

네트워크 트래픽의 특성을 연구하는데 있어서 각 프로토콜별 통계를 추출하는 것은 매우 중요한 의미를 가진다. 이의 가장 큰 이유는 각 프로토콜에 기반한 세션들은 매우 상이한 통계적 특성을 가지기 때문이다. 통계적 특성 중에는 패킷 간격들의 평균, 패킷 간격들의 분산, 그리고 3rd momentum 등이 있다. 일반적으로 각 TCP 세션은 혼잡제어 알고리즘에 의해 제어가 되기 단위시간 트래픽양이 진자운동(oscillating)한다. 이에 반해 UDP 세션의 경우는 혼잡제어 알고리즘이 존재하지 않으므로 보내는 측(sender)에 의해 해당 세션의 특성이 결정된다. 최근 들어 고속 통신망의 보급으로 인해 동영상의 실시간 전송 등 UDP에 기반한 멀티미디어 응용이 급격히 증가하였다. 그러나, 본 연구결과에 의하면 동영상 전송, 멀티미디어 회의 등 급격히 증가하는 멀티미디어 응용에도 불구하고 아직은 TCP에 기반한 세션이 전체 트래픽에서 주를 이루고 있음을 알 수 있다.

그림 4에서 보는 바와 같이 TCP와 UDP 트래픽이 주를 이루고 있음을 알 수 있다. UDP 트래픽은 egress보다 ingress가 많은 것을 알 수 있다. 국외의 서비스를 이용하기 위해서는 DNS(Domain Name System)의 요청이 필수적이다. 이러한 DNS는 UDP를 기반으로 서비스 된다. 마지막으로 그림 4(a) UDP 트래픽의 행태를 주의 깊게 보도록 하자. 31일부터 1일경 사이에 비 정상



(a) ingress 트래픽



(b) egress 트래픽

그림 4 프로토콜별 패킷량

적인 현상이 나타나고 있다.

2.5 이상트래픽 징후

그림 5에서는 시간대별 UDP 트래픽의 변화량을 구체적으로 시사하고 있다. UDP 트래픽이 비정상적으로 증가하는 구간을 period1, period 2, period3, period 4로 표시하였다. 이에 대한 비교분석을 위해서 정상트래픽이라고 생각되는 부분을 period 5를 정의 하였다. 이상트래픽을 자세히 분석하기 위해서 시간단위를 초단위로 바꾸어서 각각의 period별로 살펴보기로 한다(그림 6).

그림 6은 4개의 비정상 UDP 트래픽 구간을 확대하여 나타낸 것이다. 각 구간(period)에 대해서 2개의 그래프

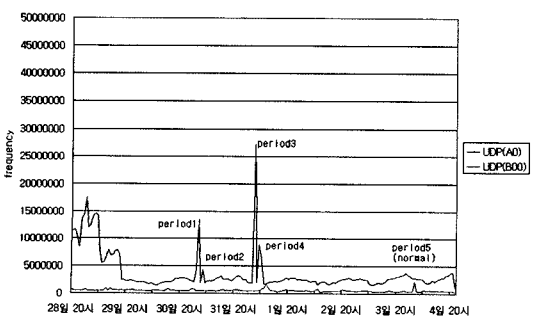


그림 5 ingress UDP 트래픽에서 나타난 이상트래픽 징후

를 가지고 있다. 이들은 각각 단위 시간당 전송된 패킷의 개수(packets/sec)와 바이트 양(byte/sec)이다. 미시적으로 관찰한 그래프는 그림 5에서 관찰되는 것과는 다른 형태를 보이고 있다. 이상 징후가 있는 부분에서의 UDP 트래픽은 처음에 트래픽이 급격히 증가한 후, 어느 정도 트래픽이 감소되어 일정 수준을 유지하다 정상상태로 돌아가는 특성을 보이고 있다.

관찰된 현상에 대한 구체적인 원인을 분석하는 것은 다수의 노드에서의 트래픽을 입체적으로 분석해야 가능하다. 본 연구에서와 같이 단일 노드에서의 패킷 트레이스를 통한 분석으로는 전체 서브넷 트래픽 특성변화의 원인을 규명하는 것은 매우 어렵다. 그림 5와 6에서 나타난 비정상적인 트래픽이 발생하는 구간에서는 특정 IP 주소에 집중적으로 패킷이 향하고 있었으며, 동일한 현상이 공통적으로 Period 1, 2, 3, 4에서 관측되었다. 이에 대한 보다 심층적인 분석은 4장에서 심도 있게 논의하도록 하겠다.

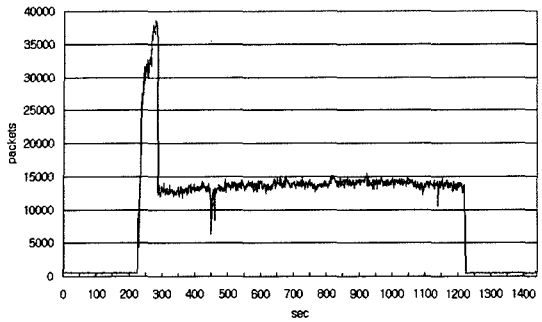
3. 이상트래픽 특성화

트래픽의 이상여부는 매우 검증이 어려운 주제이다. 근본적으로 트래픽의 이상여부를 정의하는 것이 불가능하기 때문이다. 그리고 어느 수준에서 보느냐에 따라 이상일수도 있고 또 정상으로 간주될 수도 있기 때문이다. Lakhina et. al. [13]에서는 이상 트래픽을 관찰자의 수준에 따라 분류하고 있다. 본 논문에서는 통합 트래픽 수준에서(aggregate traffic) 전체 량을 기준으로 이상여부를 판단한다. 그림 7에서 보는 바와 같이 트래픽의 양이 급격히 증가하는 짧은 구간들을 period 1, 2, 3, 4로 명기 했었다. 해당 구간의 미시적인 트래픽 변화의 특성화를 위해서 정상트래픽으로 간주되는 구간을 설정하여 이 부분과의 특성을 비교한다. 기초적인 통계 분석을 통해서 본 연구에 대상이 되는 트래픽에 나타난 이상트래픽을 자세하게 살펴볼 것이다.

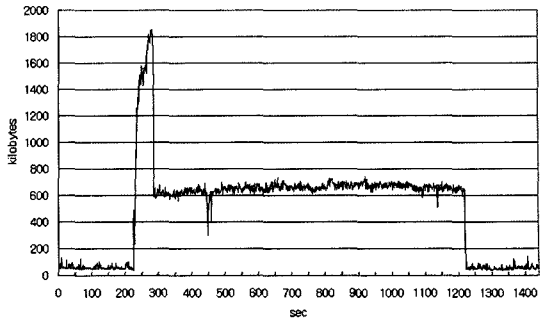
3.1 포트에 대한 분석

공격자는 대상 서버의 취약한 부분을 노리게 된다. 취약한 부분을 찾기 위해 대상 서버의 포트를 스캔하고, 특정한 포트를 대상으로 공격을 실행한다. 이상트래픽이 나타나는 구간에서는 대부분의 패킷들이 특정 목적 IP 주소(destination IP address)에 집중되어 있었다.

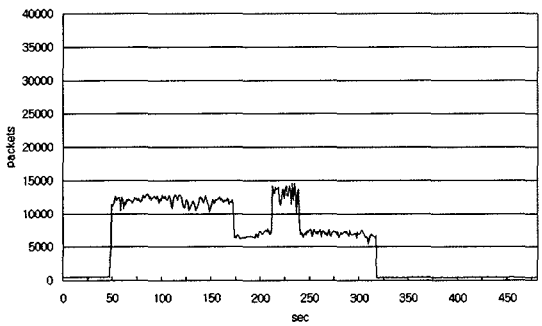
그림 7은 해당 목적 IP 주소로 향하는 트래픽을 포트별로 분류하여 나타낸 결과이다. period 1, 2, 3, 4의 이상트래픽에 대해서, 포트에 따른 트래픽 변화를 볼 수 있다. 각각의 이상트래픽에 사용된 목적 포트(destination port)를 보면 여러 개의 포트가 번갈아 나타나고 있는 것을 볼 수 있다. 53(DNS) 포트, 6667(IRC) 포트, 80(HTTP) 포트 등이 나타나고 있다. HTTP 프로토콜



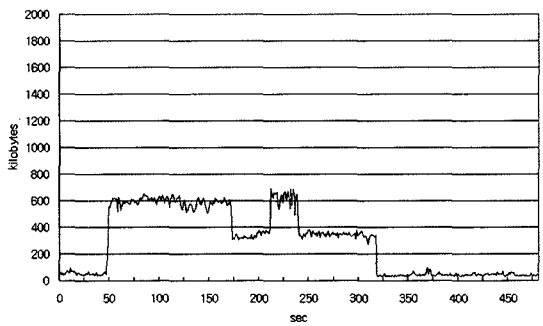
(a) period 1의 UDP 이상트래픽, packets/sec



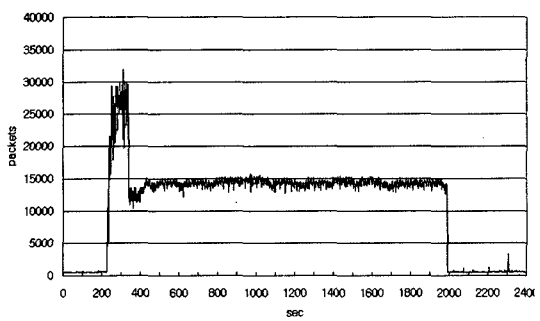
(b) period 1의 UDP 이상트래픽, bytes/sec



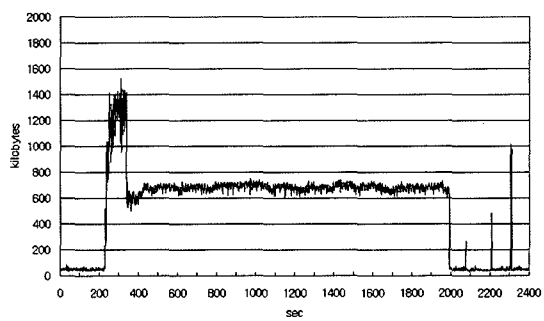
(c) period 2의 UDP 이상트래픽, packets/sec



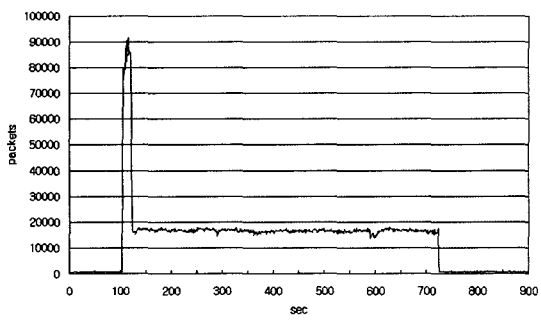
(d) period 2의 UDP 이상트래픽, bytes/sec



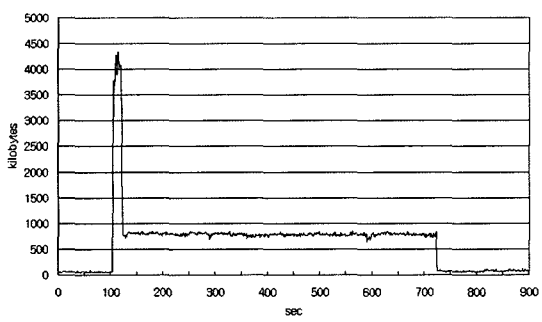
(e) period 3의 UDP 이상트래픽, packets/sec



(f) period 3의 UDP 이상트래픽, bytes/sec

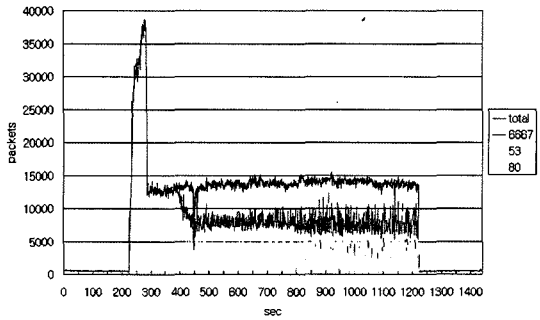


(g) period 4의 UDP 이상트래픽, packets/sec

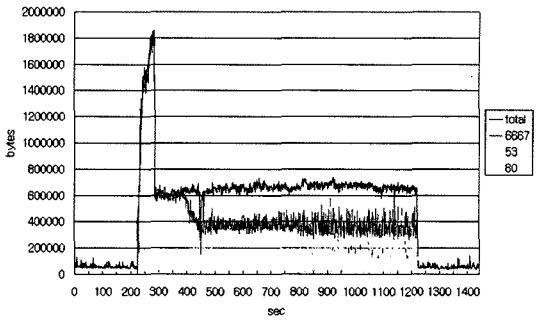


(h) period 4의 UDP 이상트래픽, bytes/sec

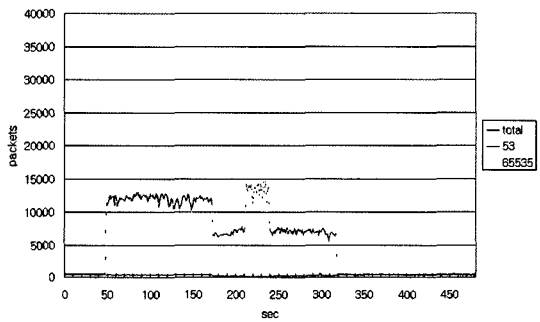
그림 6 발견된 UDP 이상트래픽 packets/sec와 bytes/sec



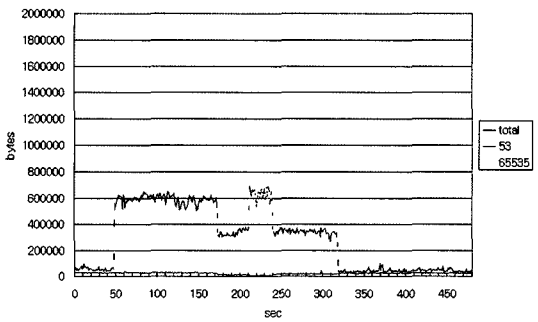
(a) period 1의 UDP 이상트래픽, packets/sec



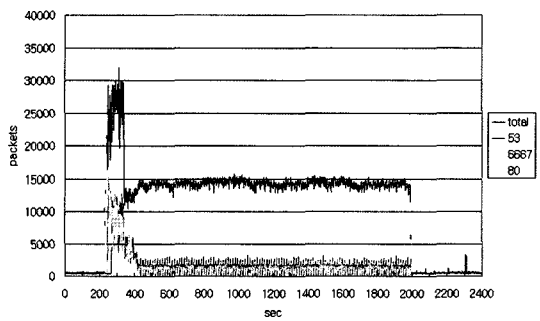
(b) period 1의 UDP 이상트래픽, bytes/sec



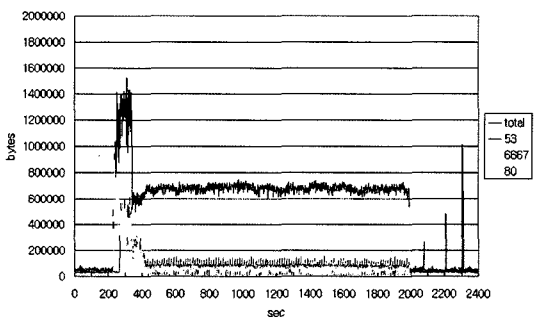
(c) period 2의 UDP 이상트래픽, packets/sec



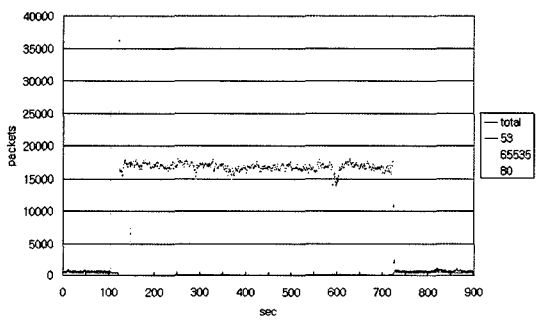
(d) period 2의 UDP 이상트래픽, bytes/sec



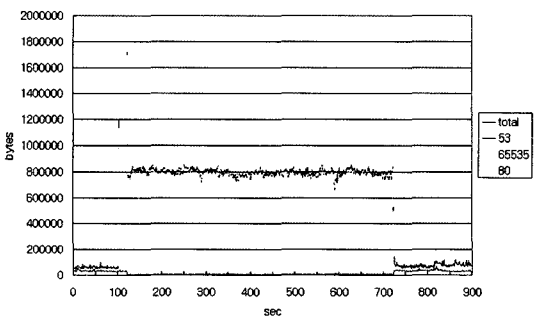
(e) period 3의 UDP 이상트래픽, packets/sec



(f) period 3의 UDP 이상트래픽, bytes/sec



(g) period 4의 UDP 이상트래픽, packets/sec



(h) period 4의 UDP 이상트래픽, bytes/sec

그림 7 UDP 이상트래픽에 대한 포트 분석

이 TCP에 기반해 있다. 따라서, HTTP를 담당하는 80 포트를 향하는 UDP 트래픽은 전형적인 이상트래픽의 일종이다. 발견된 각 포트들은 모두 DoS공격에서 일반적인 공격대상 포트이다. 그림 7에서 나타난 것과 같이, 여러 개의 포트가 나타나는 공격은 다중 포트 DoS 공격(multiple ports Denial-of-Service Attack)이며 전체 DoS 공격의 약 50%에 이른다[1].

3.2 패킷수(packet count)에 대한 분석

일반적으로 DoS 공격이 발생하면, 단위시간당 패킷수(packets/ms)가 급격하게 증가한다[1-3,15]. 이런 특징을 분석하기 위해 이상트래픽 period 1, 2, 3, 4와 정상트래픽 period 5에 대해서, 1ms당 지나간 패킷수(packets/ms)의 분포를 그림 8에 나타내었다. x축은 1ms당 패킷량을 나타내고 y축은 그에 해당하는 구간의 개수를 나타낸다.

그림 8의 각 그래프에서 이상트래픽의 특징을 발견할 수 있는데, 정상트래픽인 period 5(그림 8의 (e))에는 나타나지 않는 고주파 영역, 즉 단위시간당 지나간 패킷이 매우 많은 구간이 period 1, 2, 3, 4(그림 8의 (a), (b), (c), (d))에는 다수 존재함을 볼 수 있다. 이것은 DoS 공격의 특징인 단위시간당 패킷수(packets/ms)의 급격한 증가를 나타낸다.

따라서 가장 원시적이면서도 기본적인 DoS 공격 감지 방법은 단위시간당 패킷수의 변화를 측정하는 것이다[1]. 그림 8의 결과는 본 연구에서 발견한 이상트래픽이 DoS공격일 가능성이 높다는 것은 반증하고 있다.

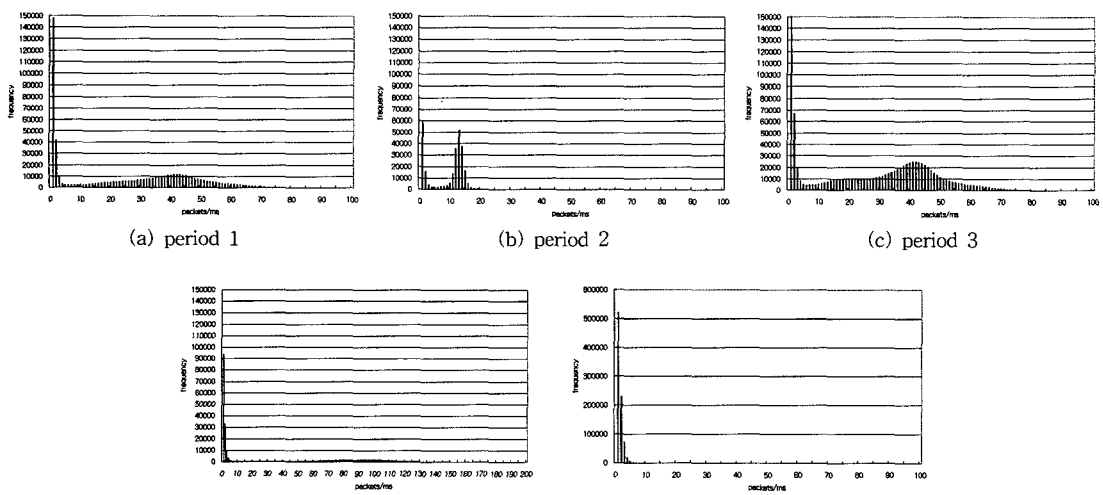
3.3 바이트량(byte count)에 대한 분석

공격이 발생하면, 단위시간당 패킷수(packets/ms)가 급격히 증가하는 만큼, 단위시간당 바이트량(bytes/ms)도 급격히 증가한다. 이런 특징을 분석하기 위해 이상트래픽 period 1, 2, 3, 4와 정상트래픽 period 5에 대해서, 1ms당 지나간 바이트량(bytes/ms) 분포를 그림 9에 나타내었다. x축은 1ms당 지나간 패킷 전체의 바이트량을 나타내고 y축은 그에 해당하는 구간의 개수를 나타낸다.

그림 9를 그림 8과 비교할 경우, 매우 재미있는 특징을 발견할 수 있다. 그림 9의 (a), (b), (c), (d)를 보면 그림 8의 (a), (b), (c), (d)에서 나타난 고주파영역과 유사한 형태가 나타나고 있다. 정상트래픽인 그림 9의 (e)에서는 이러한 형태가 나타나지 않고 있다.

더욱 주목할 것은 그림 9의 (a), (b), (c), (d)에서는 일정한 간격을 두고 갑자기 치솟는 빈도(frequency)가 나타나고 있다. 이 현상에 대해서 분석한 결과 동일한 크기의 패킷이 집중적으로 통과하고 있었기 때문이었다. 그림 9에서 발견한 균일 간격 빈도 현상을 분석하기 위해서 패킷 크기에 대한 분포를 분석하였다(그림 10).

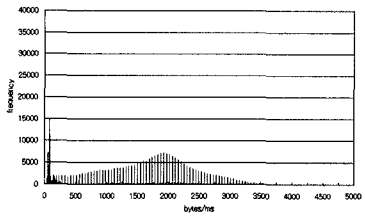
그림 10은 period 1, 2, 3, 4, 5에 대해서 각 period마다 지나간 패킷 크기(packet bytes)를 분포로 나타낸 것이다. x축은 패킷의 크기(size)를 나타내고, y축은 패킷의 발생 빈도를 나타내고 있다. 이상트래픽 period 1, 2, 3, 4의 경우(그림 10의 (a), (b), (c), (d)), 분포가 거의 하나의 직선으로 나타나고 있다. 반면에 비교대상인 정상트래픽 period 5(그림 10 (e))의 경우, 0~100바이트 사이에 집중적으로 분포하며 나머지 부분에서는 드문드문 분포하고 있다. 그림 10의 (a), (b), (c), (d)는 그림 10(e)와 분명하게 다른 형태를 보이고 있다. 그림

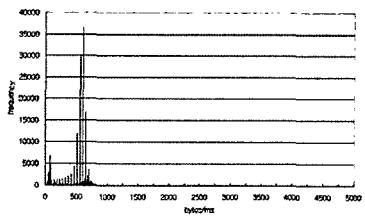


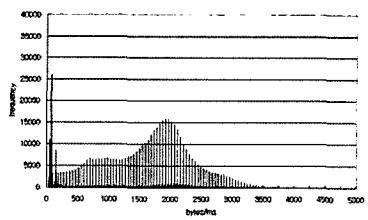
(a) period 1 (b) period 2 (c) period 3

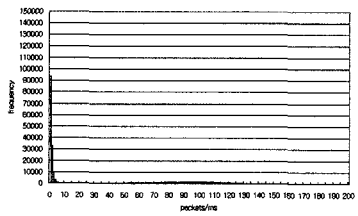
(d) period 4 (e) period 5 (정상트래픽)

그림 8 UDP 이상트래픽에 대한 packets/ms 분포


(a) period 1


(b) period 2


(c) period 3


(d) period 4

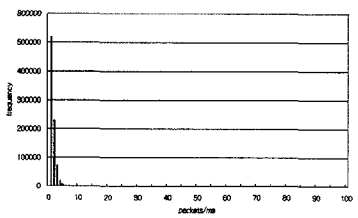
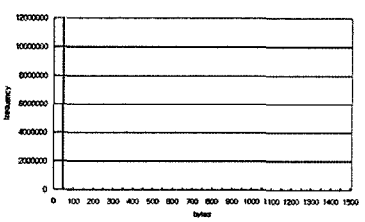
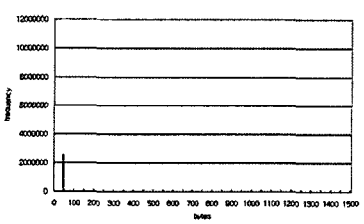
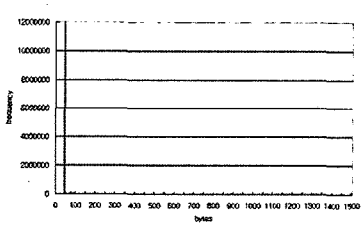
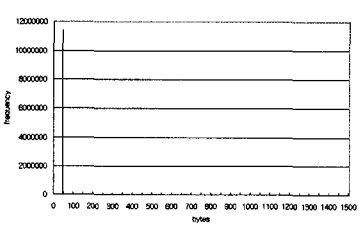

(e) period 5 (정상트래픽)

그림 9 UDP 이상트래픽에 대한 bytes/ms 분포


(a) period 1의 bytes distribution


(b) period 2의 bytes distribution


(c) period 3의 bytes distribution


(d) period 4의 bytes distribution

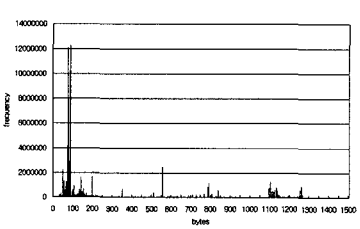

(e) period 5의 bytes distribution

그림 10 UDP 이상트래픽에 대한 패킷크기(packet size) 분포

10의 (a), (b), (c), (d)의 직선은 모두 43바이트 부분에서 나타나고 있다. 즉, 집중적으로 나타나고 있는 이상트래픽은 모두 43바이트의 패킷으로 구성되어 있다는 것을 알 수 있다. 그림 9에서 나타났던 균일 간격 빈도 현상은 43바이트의 배수의 간격으로 나타난 것이었다.

참고로, TCP에서 이러한 플로우는 ack플로우일 가능성이 높다. 긴 데이타 플로우가 발생하면 그에 대한 ack 패킷들이 모여 ack플로우를 형성하다. 이 경우 패킷의 크기는 40바이트 (ACK, FIN, RST 등) 또는 44바이트 (SYN)일 수 있다. 따라서 40바이트의 ack패킷으로 이루어진 ack플로우를 관찰할 수 있게 된다. 그러나 본 연구에서 발견한 이상트래픽은 UDP 기반이므로 TCP와는 다른 형태의 특성을 보인다.

4. 플로우 수준에서의 이상트래픽 특성

플로우 수준에서 트래픽을 분석하는 것은 쉽지 않다. 각 패킷 간의 상관관계를 일일이 비교하여 플로우를 판별해야 하기 때문이다. 뿐만 아니라 플로우의 정의를 어떻게 하느냐에 따라서 각기 다른 결과를 얻을 수 있다. 본 연구에서는 플로우를 동일한 source address, destination address, source port, destination port, protocol, 그리고 패킷간의 간격이 60초 이하인 패킷들의 연속으로 정의하였다[19].

플로우 수준의 분석을 위해 각 이상 트래픽을 데이타

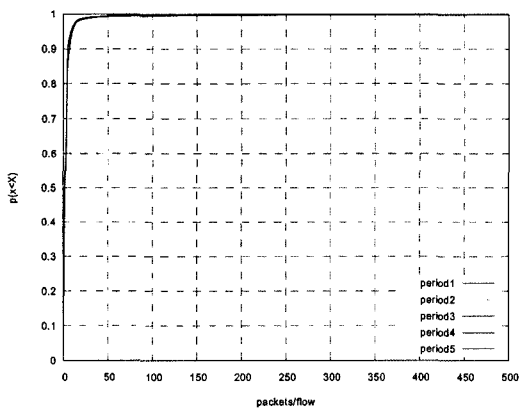
베이스화하여 처리한 후 [20], 분류된 플로우를 기준으로 분석했다. period 5는 이상트래픽과의 비교대상인 정상트래픽 구간이다. 플로우의 패킷수, 바이트, 지속시간 항목에 대한 누적분포를 이용한 분석은 5장의 이상트래픽 탐지 기법을 위한 기반 연구가 되었다. 기존 연구의 대부분은 샘플링 위주로 이루어지는 현실에서, 샘플링 없이 플로우 수준 분석을 수행한 본 연구는 확연하게 차별화 된다고 하겠다. 실제로 5장에서 누적분포에 대한 분류분석(clustering)을 수행하였고, 이를 통해 이상트래픽을 분류해 낼 수 있었다.

이 절에서는 플로우의 패킷수, 바이트량, 지속시간을 살펴보면서 플로우 수준에서의 이상트래픽과 정상트래픽의 차이를 살펴보도록 하겠다.

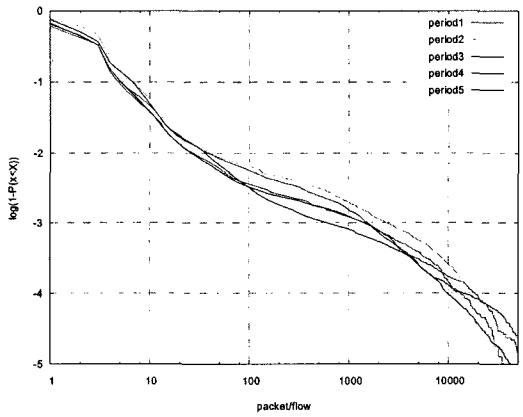
4.1 패킷수

그림 11은 플로우당 패킷수의 누적분포(cumulative distribution)를 도식적으로 표시하고 있다. 그림 11을 보면 TCP((a), (b))와 UDP((c), (d))의 간의 차이를 확인 할 수 있다. 특히 UDP 플로우에 대한 로그스케일 그래프 (d)에서 주목할 것은 period5과 period 1, 2, 3, 4의 차이가 명확하게 나타나고 있다는 점이다. 즉, 이상트래픽이 발생할 경우, 누적분포에서 헤비테일 분포(heavier tail distribution)가 나타남을 알 수 있다. 트래픽 분석에서 헤비테일 분포는 길고 패킷이 많은 플로우가 나타날 때 관찰된다. 본 연구에서 발견한 헤비테일 분포의 발생원인은 이상트래픽 때문이었다. 3.1절의 분석에서 발견한 이상트래픽을 다중포트 DoS 공격에 의한 것으로 결론지었었다. 플로우의 패킷수 속성 분석에서는 다른 바이트량 또는 지속시간 속성보다 정상트래픽과 이상트래픽의 명확한 차이를 확인할 수 있었다.

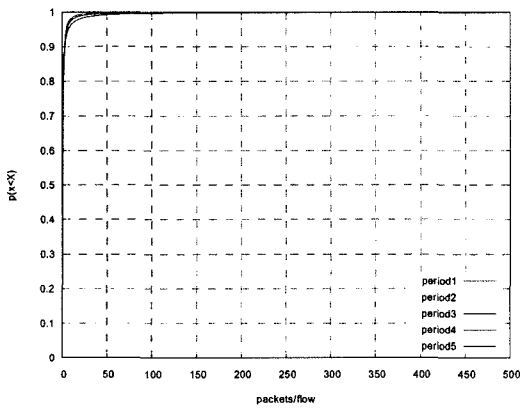
그림 11(a)와 (c)에서 누적분포를 살펴보면 10개 이하의 패킷으로 구성된 플로우가 90%를 넘어가는 것을 알 수 있다. 몇 시간 동안 지속되는 수 백만 개의 패킷으로 구성된 플로우도 존재하였지만, 매우 낮은 비율로 나타났다. 그렇기 때문에 로그스케일이 아닌 그래프에서는



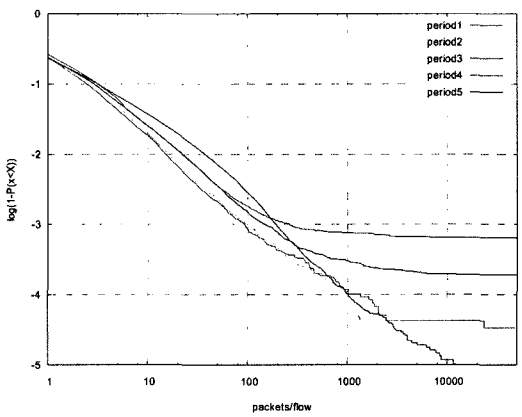
(a) TCP 플로우



(b) TCP 플로우(log scale)



(c) UDP 플로우



(d) UDP 플로우(log scale)

그림 11 패킷수/플로우 누적분포

특별한 차이점을 볼 수 없었다. 극도로 많은 패킷수를 포함한 플로우의 경우, 의도적이든 아니든 네트워크에 많은 영향을 주게 된다. 이러한 플로우 자체도 흔하지 않은 뿐만 아니라, 만약에 나타난다면 이상트래픽의 형태로 나타날 가능성이 높다.

TCP SYN flood의 경우, 1초당 약 500개의 패킷만으로도 서버를 다운시킬 수 있으며 1초당 약 14,000개의 패킷으로는 DoS공격에 내성이 있는 방화벽(firewall)이나 서버도 다운될 수 있다[21]. 따라서 서버에 급격하게 패킷이 몰리는 플로우가 발생할 경우, 적절한 완충장치가 필요하다. 예를 들면 ISP(Internet Service Provider)가 이상트래픽을 발견하면, 중간에서 이상트래픽을 버리는 방법이 있을 수 있다. 물론 정확도가 높은 이상트래픽 분류 방법이 존재할 경우 실행 가능하다.

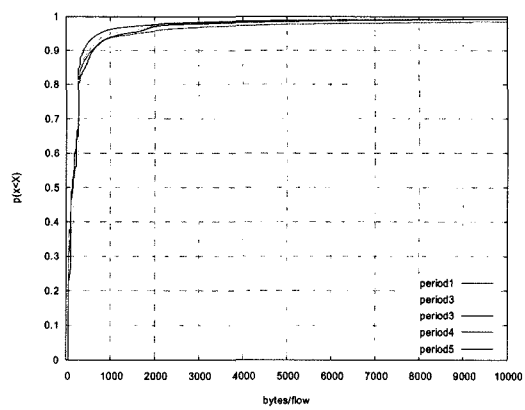
4.2 바이트량

그림 12는 플로우당 바이트량의 누적분포를 도식적으로 표시하고 있다. 그림 12(a), (c)에서는 정상트래픽과 이상트래픽의 명확한 차이를 확인할 수 없었다. 그러나 로그스케일인 그림 12(b), (d)에서는 그림 12와 유사한 헤비테일분포를 볼 수 있었다. 그러나 그 차이가 패킷수만큼 큰 차이를 볼 수는 없었다. 따라서 본 연구에서 발견한 이상트래픽의 경우, 플로우의 바이트량보다는 플로우의 패킷 수에서 더 명확한 차이를 발견할 수 있었다.

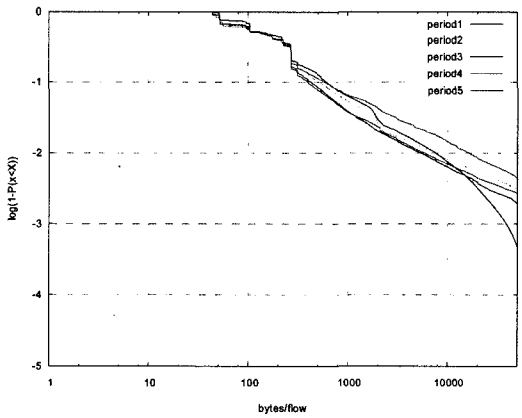
그림 12의 (a), (c)의 누적분포에서 약 500바이트 이하의 바이트량으로 구성된 플로우가 90%정도인 것을 알 수 있다. TCP의 플로우가 UDP플로우에 비해 큰 편임을 확인 할 수 있다. 플로우가 비교적 큰 용량일 경우, 파일 전송 관련 플로우일 가능성이 높다. 하지만, 플로우의 전체바이트량/전체패킷수의 값이 적을 경우, DoS 공격일 가능성을 배제할 수 없다.

4.3 지속시간

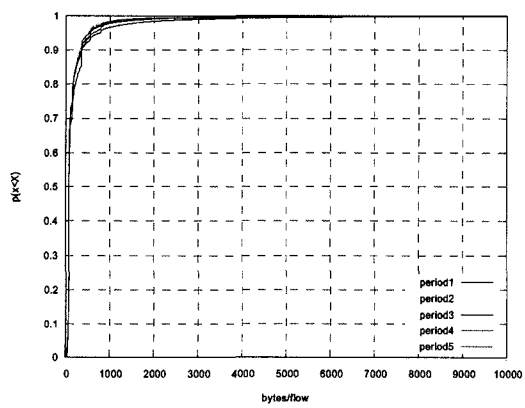
그림 13은 플로우가 시작되는 시간과 끝나는 시간을 기준으로 플로우 지속시간의 누적분포를 도식적으로 표시하고 있다. 그림 (a), (c)에서 큰 차이를 확인할 수 있



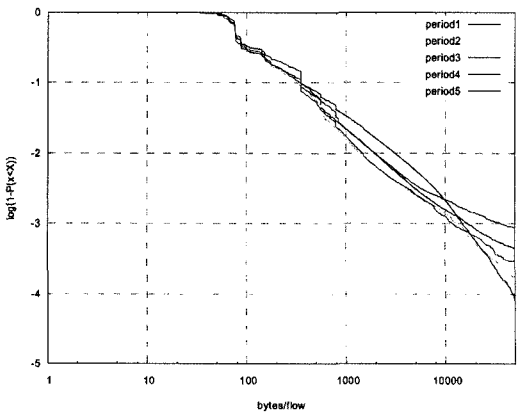
(a) TCP 플로우



(b) TCP 플로우(log scale)

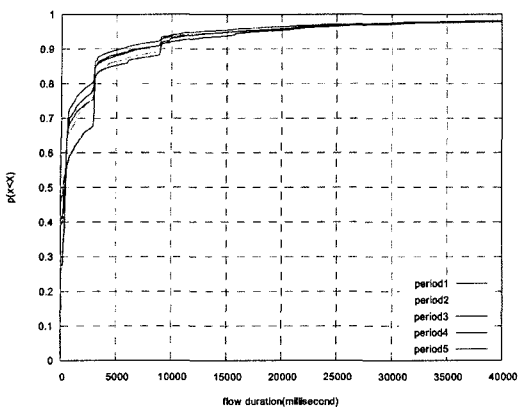


(c) UDP 플로우

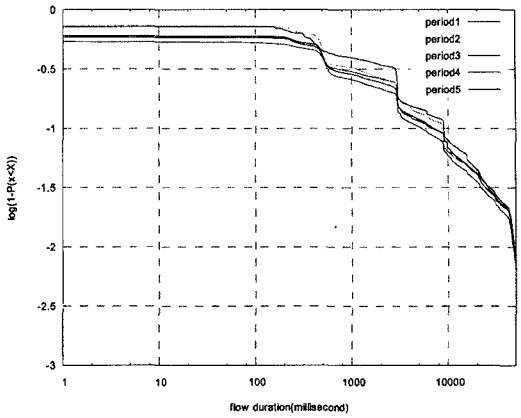


(d) UDP 플로우(log scale)

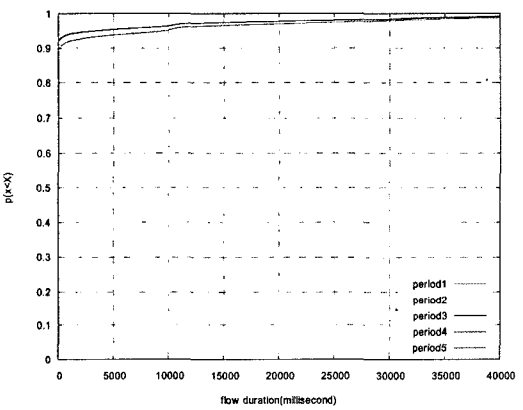
그림 12 바이트량/플로우 누적분포



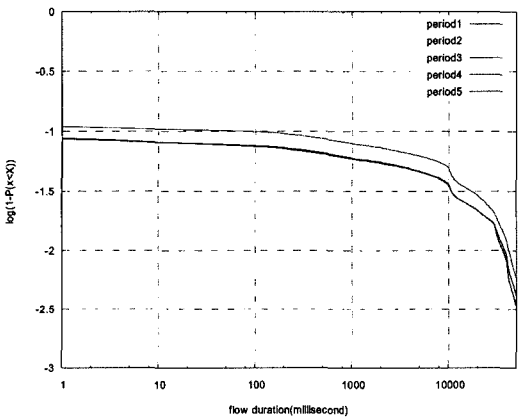
(a) TCP 플로우



(b) TCP 플로우(logscale)



(c) UDP 플로우



(d) UDP 플로우(logscale)

그림 13 플로우 지속시간 누적분포

다. 특히 TCP의 플로우 지속시간 누적분포인 그림 13 (a)에서 각 period마다 다른 결과를 보이고 있는 것은 TCP의 특성 때문이다. TCP트래픽은 시간대에 따라서 다른 트래픽량을 보여주기 때문이다. 이것은 앞의 2.3절에서 살펴보았었다. 그림 13(c), (d)에서 재미있는 현상을 발견할 수 있다. 이상트래픽을 포함한 구간 4개와 정상트래픽을 포함한 구간 1개가 존재함에도 2개의 누적분포로 나타나고 있다. 이것은 이상트래픽의 정도가 다르기 때문이다. period 1, 3은 비교적 큰 이상플로우들이 발생했으며, period 2, 4에서는 상대적으로 작은 이상플로우들이 발견되었다. 그리고 발생한 이상플로우의 개수도 period 1, 3과 period 2, 4는 크게 차이가 났다. 이상플로우의 공격 강도가 낮을 경우, 정상트래픽과 비슷하게 나타났다. 이 분석 결과는 5장에서 플로우의 패킷수와 상호보완적으로 사용될 수 있었다.

그림 13(a), (c)의 누적분포에서 약 0.5초 이내로 지속되는 플로우가 62%정도로 나타나고 있으며, 약 3초 이내로 유지되는 플로우가 85%정도인 것을 볼 수 있다. 대부분의 DoS 공격은 수십 분이며 많아도 1시간 이내로 나타나고 있다. 오랫동안 지속되는 플로우의 경우, 파일 전송 중일 가능성이 있지만 DoS 공격을 의심해 볼 필요가 있다. 본 연구에서 발견되었던 이상플로우들의 경우 수 십분 정도 지속되었으며 많아도 1시간을 넘지 않았다.

5, 6장에서는 2, 3, 4장에서 살펴봤던 트래픽 특성화 연구를 기반으로 1초단위로 이상트래픽을 분류할 수 있는 방법을 제안할 것이다.

5. 플로우 특성과 분류 분석

5.1 플로우 도착 시간

앞의 "4. 플로우 수준에서의 특성"에서 설명했듯이, 플로우는 source address, destination address, source port, destination port, protocol이 동일한 패킷의 연속으로 정의된다. 플로우 분석은 인터넷 트래픽 분석에 있

어서 중요한 방법 중의 하나이며 이상 트래픽의 발생시 그 진원지를 추적하고 대처하는데 유용한 도구를 제공할 수 있다. 플로우 분석은 플로우 들의 발생시점, 플로우의 지속시간, 플로우 내에 포함되는 패킷들의 양 또는 크기 등에 대한 분석을 필요로 하며 그림 14는 연속된 플로우들 간의 도착 간격에 대한 누적상대돗수 분포함수와 플로우들 간의 도착간격에 대한 평균(average)을 모수로 갖는 지수분포(exponential distribution)와의 Q-Q plot을 보여주고 있다.

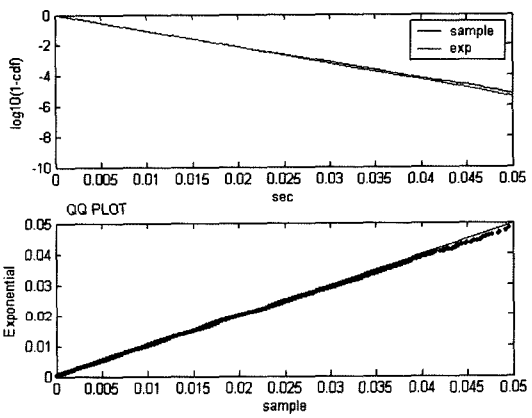


그림 14 Q-Q Plot

Q-Q plot은 누적상대돗수 분포함수와 지수분포함수를 각각 $\hat{F}, E$라고 하면 주어진 확률 p에 대하여 $(\hat{F}^{-1}(p), E^{-1}(p))$ 값을 도시한 그래프로 분포함수들이 서로 같은 경우 직선의 형태를 나타낸다. 따라서, 위의 그림 14에서 보면 플로우 도착 간격에 대한 분포는 지수분포로 잘 적합이 됨을 알 수 있으며, 이로부터 플로우의 도착과정(arrival process)은 포아송 확률과정(poisson process)으로 모형화 할 수 있음을 알 수 있으며 더 나아가서 임의 시점에서 존재하는 플로우의 개수(현재 패킷을 생성하는 플로우의 개수)는 M/G/∞ 큐잉 모형으로 적합할 수 있음을 알 수 있다. 이는 단위시간당 도착하는 패킷의 수와 양에 대한 모형의 수립에 아주 유용한 도구를 제공할 수 있다. 또한 들어오는 플로우의 개수가 이상적으로 증가하는 지의 여부를 판단하고자 하는 경우에도 이 결과는 유용한 도구를 제공할 수 있다.

5.2 플로우 분류 분석

앞 절에서 나타난 이상트래픽의 현상을 규명하기 위해서는 이상트래픽의 원인이 되는 플로우들을 분류하고 분류된 플로우의 특성분석이 필요하다 하겠다. 다음 그림에서는 몇 개의 선택된 소스에서 발생되는 트래픽의 형태를 나타낸다. 여기서 x 축은 상대적 시간을 나타내며 각 바늘은 각각의 플로우 내에서의 패킷들의 도착시간을 나타낸다.

그림 15에서 볼 수 있듯이 각 플로우 내에서의 패킷 도착간격(packet interval)은 각기 다른 형태를 보이며 나아가서 첫 번째의 형태를 갖는 플로우에 의하여 단위시간당 패킷수의 증가가 야기 됐음을 추측할 수 있다. 이로부터 하나의 플로우를 관측했을 때, 이 플로우가 공격인지 아닌지를 분류하는 분석이 필요하다 할 수 있겠다. 이를 위해서는 공격의 특성을 갖는 플로우들을 하나의 그룹으로 묶을 수 있어야 하며, 또한 이를 위한 변수의 개발이 선행되어야 한다.

본 논문에서는 이를 위하여 각 플로우의 패킷간 도착간격에 대한 누적 상대돗수 분포함수(empirical cumulative distribution function, empirical cdf)를 이용하여 이들 함수간의 거리(Euclidean distance)를 측도로 하는 군집분석(clustering analysis, [23])을 행하였다.

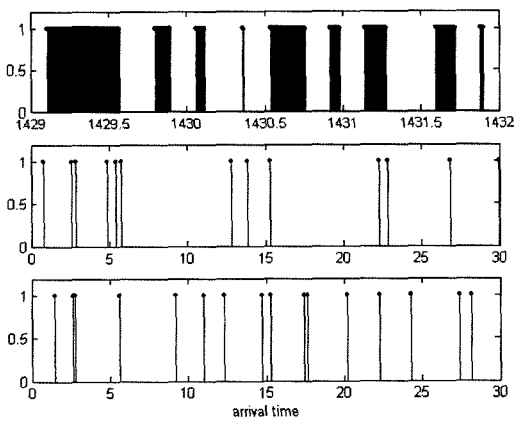


그림 15 플로우의 형태

즉 각 플로우에서 변수

$$f=(f_1,\cdots,f_{81}),\ \ f_i=f(x_i),\ x_i=10^{-6+(i-1)0.1}$$

를 정의하고, 단 여기서

$$f(x_i)=\frac{\text{number of inter - arrival times less than } x_i}{\text{number of inter - arrival times of a flow}},$$

이들간의 거리(euclidean distance)를 이용한 군집분석을 행하였다. 참고로 군집분석은 SAS 통계패키지를 이용하였다. 군집의 개수를 3개 이상으로 제한하였을 경우 공격의 특성을 가지는 플로우들은 하나의 군집으로 분류 되었으며 다음의 그림 16은 군집의 개수를 3개로 하였을 때의 각 군집에 속하는 플로우에서의 패킷간 도착간격에 대한 상대누적도수 분포함수를 나타낸다.

그림 16을 보면, 군집 1(group 1)에 해당하는 플로우

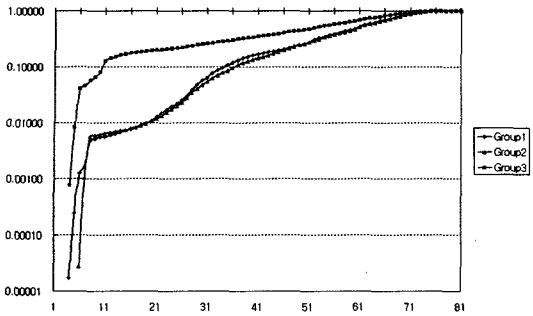


그림 16 플로우에 대한 군집 분석 결과

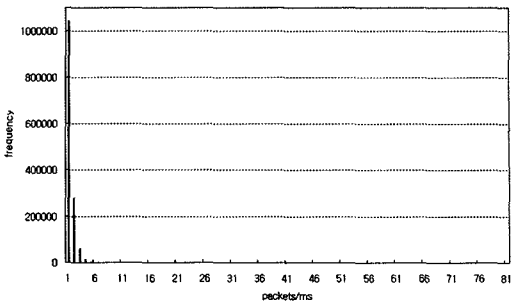


그림 17 공격이 없을 때의 히스토그램

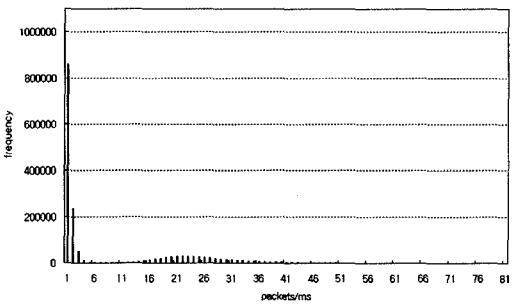


그림 18 공격이 있을 때의 히스토그램

는 패킷도착간격이 매우 작음을 알 수 있으며 이는 곧 단위시간 당 많은 양의 트래픽을 생성함을 알 수 있다. 실제로 군집 1에 속하는 54개의 플로우 중 53개가 공격의 형태를 띰을 알 수 있었다. 이러한 플로우의 분류결과는 새로운 플로우가 관측되었을 때, 그 플로우가 공격의 형태를 보이는 플로우인지 아닌지를 판단할 수 있는 매우 유용한 도구를 제공할 수 있다.

6. 이상트래픽 탐지

6.1 공격에 의한 트래픽 특성 변화

인터넷에서 공격은 TCP SYN flood attack, UDP flooding, Smurf Attack 등 다양한 종류가 있으며 이러한 공격은 트래픽 특성에 변화를 가져온다[22]. 앞 절에서 보인 바와 같이 이상트래픽의 원인이 되는 플로우들은 매우 조밀한 패킷간 도착간격을 보여주며 이는 단위시간당 패킷수에 심각한 변화를 야기할 수 있다. 본 연구에서는 단위시간당 들어오는 패킷수의 분포를 이용하여 앞 절에서 나타난 이상 플로우의 발생시점과 존재시점을 탐지하는 기법을 제시하고자 한다.

먼저 단위시간당 들어오는 패킷수에 대한 변화를 고찰한다. 다음 그림 17은 공격이 없을 때(period 5)의 단위시간당(1ms) UDP 트래픽 양에 대한 히스토그램이다. 그러나 인터넷 망 공격이 있을 때의 분포를 보면 그림 18(period 3)와 같이 히스토그램의 변화가 일어났음을 볼 수 있다.

그림 18을 그림 17의 히스토그램과 비교하면 하나의 고주파 영역이 새롭게 생겼음을 관측할 수 있다. 즉, 공격 성향의 트래픽이 들어오는 경우에, 들어오는 패킷수가 정상적인 상태에서보다 많은 단위시간 구간이 늘어나는 것을 뜻하며, 상대적으로 정상적인 경우의 보통의 패킷수가 들어오는 단위시간 구간은 상대적으로 감소하게 됨을 알 수 있다. 이로부터 공격의 성격을 갖는 트래픽에 의해 발생되는 패킷들은 매우 짧은 시간에 몰려서 들어오는 성질을 가지고 있음을 추측할 수 있으며, 이는 이상 트래픽의 탐지에 있어서 단위시간당 패킷수는 중요한 변수가 될 수 있음을 보여준다.

6.2 단위 시간당 지나간 패킷에 대한 분포적합

그림 19는 트래픽이 정상적이라고 생각되는 Period 5의 시간구간에서의 단위 시간당 지나간 패킷수에 대한 분포와 적합된 기하분포(geometric distribution)의 분포함수를 나타낸다.

여기서 기하분포는 다음과 같이 주어지는 분포함수로

$$P[X=k]=p(1-p)^k,\ k=0,\ 1,\ 2,\ \cdots,$$

여기서 X는 단위시간당 들어오는 패킷수를 나타내는 확률변수이고, 모수 p에 대한 추정량 $\hat{p}$은 최대우도방법(maximum likelihood method, 15)을 이용하여 추정되었다. 즉,

$\hat{p}$ = (Period 5에서 단위시간 구간들에서의 패킷수에 대한 average)$^{-1}$

그림 19에서 보는 바와 같이 기하분포가 단위시간당 지나간 패킷수에 대한 분포에 적절하게 적합됨을 알 수 있으며, 이는 이상트래픽의 발생여부를 판단하는데 유용한 도구를 제공할 수 있다. 즉, 이상트래픽이 발생할 경우 앞 절에서 언급한 바와 같이 단위시간당 들어오는 패킷수의 변화를 가져올 것이고 따라서 이의 분포는 정상적인 상태에서의 분포와 다른 모습을 보일 것이며 이는 곧 적합된 기하분포와도 다른 모습을 보일 것임을

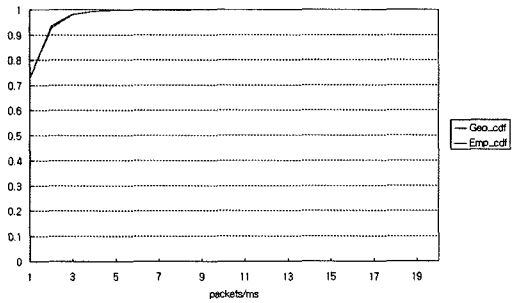


그림 19 단위 시간(ms)당 지나간 패킷수에 대한 분포함수

예측할 수 있다. 이러한 분석 결과를 이용하여 다음 절에서는 적합된 기하분포와의 분포적합검정(goodness of fit test)을 이용한 이상트래픽의 발생시점을 탐지하는 기법을 소개하고자 한다.

6.3 이상트래픽 탐지 기법

먼저 이상트래픽의 발생과 패킷수 분포의 변화를 살펴보자. 각각의1초 동안의 시간구간에서는 1000개의 1ms 길이의 단위시간구간으로 나뉘어지며 각각의 단위시간 구간에서 관측된 1,000개의 패킷수를 이용하여 표본분포함수 (empirical distribution function)를 구할 수 있다. 즉, $X_{t,i}$를 $(t+(i-1)*0.001)$sec과 $(t+i*0.001)$sec 사이의 1ms 시간구간에서의 패킷수라고 정의하면 앞 절에서 언급한 바와 같이 정상트래픽에서의 $X_{t,i}$의 분포는 모수 $\hat{p}$을 갖는 기하분포를 따른다고 가정할 수 있다. 따라서 정상트래픽에서의 $X_{t,i}$의 표본분포함수는 모수 $\hat{p}$을 갖는 기하분포와 비슷한 형태를 띠게 되며 이상트래픽에서의 표본 분포함수는 기하분포함수와 차이를 보이게 될 것이다. 이 절에서는 $X_{t,i}$의 표본분포함수와 모수 $\hat{p}$을 갖는 기하분포의 차이를 설명할 수 있는 거리함수로서 Cramer-Smirnov-Von Mises 검정통계량을 제시하고 이의 확률적 특성을 이용하여 이상플로우 또는 이상트래픽의 발생여부와 존재여부를 탐지할 수 있는 기법을 제시한다.

그림 20에서는 이상트래픽이 발생하면서 생기는 패킷수의 분포변화를 살펴보기 위하여 연속된 초(sec) 단위 시간구간에서의 패킷수의 히스토그램의 변화를 나타내었다. 그림 20은 period 3의 1427부터 1430초 구간의 단위시간당 패킷수(packets/ms) 분포를 나타내고 있다.

그림 20에서 볼 수 있듯이 히스토그램의 형태가 정상적인 구간 (Period 5)에서의 형태를 보이다가 어느 순간부터 고주파 영역이 나타난 형태를 보인다. 즉, 패킷수가 갑자기 늘어난 시간구간이 많아짐을 의미하며 이 형태는 적합된 기하분포와는 상이한 모습을 보이게 된다.

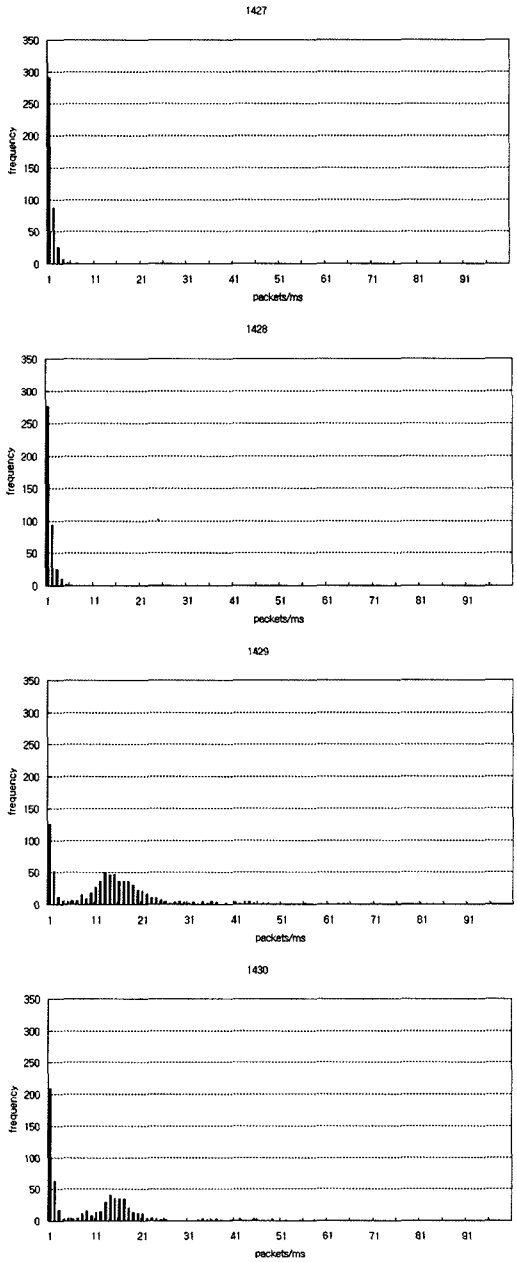


그림 20 단위 시간(ms)당 지나간 패킷수에 대한 분포함수 (period 3의 1427~1430초 구간)

이로부터 본 연구에서는 이상트래픽의 발생시점은 패킷수의 분포가 기하분포와 상이한 모습을 보이는 시점이라고 가정하고 제안된 탐지기법은 이러한 상이한 모습을 보이는 시점의 판단 여부를 매 1초 시간 구간에서 얻어지는 ms당 지나간 패킷수에 대한 분포와 적합된 기하분포와의 분포적합 검정(test of goodness of fit)의

유의확률을 이용하여 결정하는 것이다.

본 연구에서 사용된 분포적합 검정통계량(goodness of fit test statistic)은 Cramer-Smirnov-Von Mises 검정통계량이며[24] 다음과 같이 정의된다.

$$W^2 = \frac{1}{n}\sum_{i=1}^{n}(F_0(x_i) - \hat{F}(x_i))^2 \quad (1)$$

여기서 $n(=1000)$은 자료의 수를 뜻하며 F_0는 기하분포함수 $\hat{F}$은 자료부터 구해지는 표본분포함수 즉 $X_{t,i}$의 표본 분포함수를 나타낸다. 이 검정에서의 귀무가설은 "단위시간(1ms) 동안의 패킷수는 적합된 기하분포를 따른다"이며, 유의확률은 Anderson과 Darling[24]에 의해 주어진 확률표를 이용한다. 즉, 이 방법은 위의 통계량에 대하여 주어진 확률표를 이용하여 유의확률(p-value)을 구하고 유의확률이 정해진 값보다 (일반적으로 0.05 또는 0.01) 작으면 패킷수의 분포가 주어진 기하분포가 아니라는 결론을 내리는 것이라 할 수 있다.

그림 21은 실제 인터넷 트래픽 자료에(그림 5에서의 *period* 3) 위의 검정통계량을 이용하여 유의 확률을 구한 것을 그린 것이다.

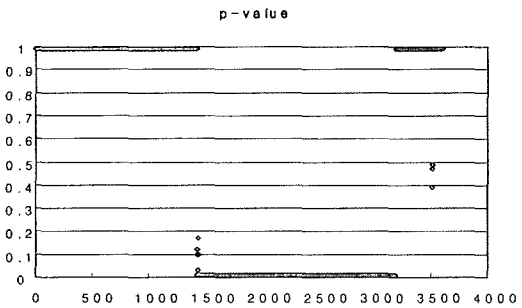


그림 21 유의확률. Cramer-Smirnov-Von Mises Test

본 논문에서는 유의확률이 0.01보다 작아지는 시점을 이상 트래픽의 발생 시점이라고 하였다. 본 논문에서 제안된 방법은 실시간(1초)으로 이상여부를 판단할 수 있으므로 이상트래픽에 대한 신속한 대응을 가능하게 할 수 있으며 계산량이 적어 실제 시스템에 적용하는데 부하가 걸리지 않는 장점을 가지고 있다.

7. 결 론

본 연구의 목적은 초고속 인터넷의 트래픽을 분석하고 여기서 발생되는 이상트래픽의 특성을 추출해내어, 앞으로 발생할 이상트래픽의 탐지와 방어 방법에 대한 기초를 제시하려는 것이다. 본 연구에서 사용한 데이타는 국내와 국외 사이를 통과하는 트래픽을 샘플링 없이 1/10,000,000초의 정밀도를 가진 장비로 캡쳐하였다. 이로 인해, 기존 연구와는 달리 미시적 관심에서 트래픽을 심도있게 분석할 수 있었다. 따라서 본 연구는 국내와 국외의 사용자간 인터넷 이용실태를 포함한 이상트래픽의 실제 사례를 파악할 수 있는 흔치않은 연구 결과라는 점에서 매우 큰 가치가 있다고 할 수 있다. 대량의 트래픽은 다루기 어렵기 때문에 기존 연구에서는 일반적으로 수 시간에서 수 일을 샘플링 된 형태로 캡쳐하여 분석하였다. 반면에 본 연구에서는 일주일간의 트래픽을 샘플링 없이 캡쳐하였으며, 자체적으로 대량의 트래픽 분석을 위한 전용 소프트웨어를 개발하여 분석하였다.

본 연구에서는 캡쳐된 트래픽 분석 중에 발견된 이상트래픽을 중점적으로 분석하였으며, 이러한 실제 사례를 기반으로 이상트래픽의 빠르게 탐지하기 위한 통계적 기법을 제시하였다. 기존 연구에서는 트래픽을 다루기 쉽게 하기 위해서 5~10분 정도의 시간 동안 모인 트래픽 데이타의 통계정보를 분석하였다면, 본 연구에서는 마이닝 기법의 일종인 분류분석을 1초단위로 트래픽에 적용하여 신속한 이상트래픽 감지 가능성을 보였다. 즉 본 논문에서 제안한 기법을 이용하면 1초 단위의 트래픽 분석으로 이상트래픽인지 아닌지를 판단할 수 있는 것이다. 이것은 수준 높은 수학적 기법을 통해 수동적 방법(passive method)으로 이상트래픽을 사후에 찾아낼 수 있는 기존 연구들과는 확실하게 차별화되는 점이다.

본 연구에서 이상트래픽의 특성화를 위해 포트별, 패킷수, 패킷바이트량을 기준으로 분석하였다. 특히 이 특성화에서 밝혀낸 packets/ms 분포에서 나타나는 고주파 영역, bytes/ms 분포에서 나타나는 균일 간격 빈도 현상 등은 앞으로 계속될 이상트래픽의 탐지와 방어 알고리즘 개발에 필요한 매우 중요한 특성으로 사용될 것이다.

향후 과제로는 실시간으로 이상트래픽의 발생을 탐지할 수 있는 기법의 개발이다. 본 논문에서 제안한 통계적 기법을 통해서 이상트래픽을 신속히 분류해 낼 수 있었지만, 대량의 트래픽에 대해서 실시간으로 이상트래픽 탐지를 수행하기에는 무리가 있을 것으로 예상된다. 실시간으로 이상트래픽을 탐지하는 기법이 개발될 경우, 상당수의 네트워크 공격 도구는 무용지물이 될 것이다.

참 고 문 헌

[1] D. Moore, G. M. Voelker, and S. Savage, "Inferring *Internet Denial-of-Service Activity*," *presented* at The 2001 USENIX Security Symposium, 2001.

[2] M. Thottan and C. Ji, "Anomaly Detection in IP Networks," IEEE Transactions on Signal Processing, vol. Vol. 51, No. 8, 2003.

[3] A. Hussain, J. Heidemann, and C. Papadopoulos,

"A Framework for Classifying Denial of Service Attacks," presented at SIGCOMM, 2003.
[4] A. Lakhina, M. Crovella, and C. Diot, "Characterization of Network-Wide Anomalies in Traffic Flows," presented at The ACM/SIGCOMM Internet Measurement Conference, 2004.
[5] W.~Leland, M.~Taqqu, W.~Willinger, and D.~Wilson, "On the Self-Similar Nature of Ethernet Traffic," in Computer Communication Review: ACM, 1992, pp. 203-213.
[6] W. Walter, T. Murad, and S. Robert, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Srouce Level," in Proceedings of SIGCOMM '95. Cambrideg, MA, USA, 1995, pp. 100-113.
[7] A. C. G. A.~Feldmann and W.~Willinger, "Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic," ACM Computer Communication Review, vol. 28 %8 Sept 1998, pp. 42-55, 1998.
[8] V. Paxson, "An analysis of using reflectors for distributed denial-of-service attacks," presented at ACM SIGCOMM Computer Communication Review, 2001.
[9] Z.-L. Zhang, V. J. Ribeiro, S. Moon, and C. Diot, "Small-Time Scaling Behaviors of Internet Backbone Traffic: An Empirical Study," presented at IEEE INFOCOM, 2003.
[10] P. Barford and D. Plonka, "Characteristics of Network Traffic Flow Anomalies," presented at ACM Internet Measurement Workshop '01, San Francisco. CA, USA, 2001.
[11] M. Crovella and E. Kolazcyk, "Graph Wavelets for Spatial Traffic Analysis," presented at IEEE INFOCOM, 2003.
[12] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing Network-Wide Traffic Anomalies," presented at ACM SIGCOMM 2004, 2004.
[13] A. Lakhina, M. Crovella, and C. Diot, "Mining Anomalies Using Traffic Feature Distributions."
[14] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Profiling Interent Backbone Traffic: Behavior Models and Applications," presented at ACM SIGCOMM, Philadelphia, Pennsylvania, USA, 2005.
[15] L. Feinstein, D. Schnackenberg, R. Balupari, and D. Kindred, "Statistical Approaches to DDoS Attack Detection and Response," presented at The DARPA information Survivability Conference and Exposition (DISCEX'03), 2003.
[16] C. Z. Cliff, G. Weibo, T. Don, and G. Lixin, "Monitoring and Early Detection of Internet Worms," IEEE/ACM Trans. on Networking.
[17] tcpdump/lipcap, "TCPDUMP public repository," in http://www.tcpdump.org.
[18] Ethereal, "The world's most popular network protocol analyzer," in http://www.ethereal.com/.
[19] Sprint, "Packet Trace Analysis," in http://ipmon.sprint.com/packstat/packetoverview.php.
[20] MySQL, "The World's Most Popular Open Source Database," in http://www.mysql.com.
[21] T. Darmohray and R. Oliver, ""Hot Spares" For DoS Attacks," in http://www.usenix.org/publications/login/200-7/apropos.html. ;login:, 2000.
[22] R. R. Panko, Corporate Computer and Network Security: Prentice Hall, 2004.
[23] R. A. Johnson and D. W. Wichern, Applied multivariate statistical analysis, 5 ed: Prentice Hall, 2002.
[24] Anderson and Darling, "Asymptotic theory of certain goodness of fit criteria based on stochastic process," Annals of Mathematical Statistics, vol. 23, pp. 193-212, 1952.

김 정 현

2004년 8월 명지대학교 컴퓨터공학과 졸업(학사). 2004년 8월~현재 한양대학교 전자통신컴퓨터공학부 석박사통합과정. 관심분야는 인터넷 보안, 운영체제, 멀티미디어

안 수 한

1992년 2월 서울대학교 계산통계학과 졸업(학사). 1994년 2월 서울대학교 계산통계학과 졸업(석사). 2000년 2월 서울대학교 통계학과 졸업(박사). 2001년~2003년 AT&T Labs-Research, Post-Doc, Consultant. 2004년 3월~현재 서울시립대학교 통계학과 조교수. 관심분야는 Fluid Flow model, Queueing, Telecomunication Network 등

원 유 집

1990년 2월 서울대학교 계산통계학과 졸업(학사). 1992년 2월 서울대학교 전산과학과 졸업(석사). 1997년 8월 University of Minnesota 졸업(박사). 1997년 9월~1999년 2월 Intel 연구원. 1999년 3월~현재 한양대학교 전자통신컴퓨터공학부 부교수. 관심분야는 운영체제, 컴퓨터네트워크, 성능평가

국가보안기술 연구소 소속 저자(이종문, 이은영)는 국가보안기술연구소 측의 요청에 의해 저자약력을 생략함.