GENESIS: Internet Disk P2P 트래픽 탐지를 위한 시그너춰 자동 생성 방안

(GENESIS: An Automatic Signature-generating Method for Detecting Internet Disk P2P Application Traffic)

이 병 준 † 윤 승 현 † 이 영 석 ††
(Byungjoon Lee) (Seunghyun Yoon) (Youngseok Lee)

요 약 다량의 네트워크 대역폭을 소모하는 P2P 응용 프로그램 트래픽을 차단하기 위해, 학내망 혹은 기업망의 방화벽에는 상례적으로 P2P 트래픽 차단 규칙들이 등록되고 있다. 하지만 포트 번호만을 사용하는 단순한 차단 규칙들은 'Port Hopping' 등의 기법으로 방화벽을 우회하거나, HTTP 기반 인터넷 디스크 서비스 등으로 위장된 P2P 응용의 트래픽은 차단해 내지 못한다. 이러한 트래픽을 올바르게 식별하고 차단하기 위해서는 페이로드 시그너춰(payload signature) 기반의 패킷 식별 방법을 사용하여야 하며, 현재 상당수의 IDS 시스템들이 이를 지원한다. 하지만 이 방법은 정확도가 높고 간단하게 적용될 수 있는 반면, 시그너춰를 찾는 작업 자체의 난이도가 높아서 시그너춰의 목록을 최신 상태로 유지하는 것이 어렵다. 그러므로 이 방법이 효율적으로 운용되기 위해서는 패킷의 페이로드(payload)로부터 시그너춰를 자동 추출하는 방안이 필요하다. 본 논문에서는 인터넷 디스크 형태로 서비스되는 P2P 응용 프로그램의 시그너춰를 자동 추출하는 방안을 소개하고, 해당 방안을 충남대학교 학내망에 적용한 사례를 보인다.

키워드 : P2P, 시그너춰, 트래픽 측정, 플로우, IDS

Abstract Due to the bandwidth-consuming characteristics of the heavy-hitter P2P applications, it has become critical to have the capability of pinpointing and mitigating P2P traffic. Traditional port-based classification scheme is no more adequate for this purpose because of newer P2P applications, which incorporating port-hopping techniques or disguising themselves as HTTP-based Internet disk services. Alternatively, packet filtering scheme based on payload signatures suggests more practical and accurate solution for this problem. Moreover, it can be easily deployed on existing IDSes. However, it is significantly difficult to maintain up-to-date signatures of P2P applications. Hence, the automatic signature generation method is essential and will be useful for successful signature-based traffic identification. In this paper, we suggest an automatic signature generation method for Internet disk P2P applications and provide an experimental results on CNU campus network.

Key words : P2P, signature, traffic measurement, flow, IDS

1. 서 론

P2P 응용이 점유하는 트래픽 양이 증대됨에 따라, 많은 기업망/학내망 관리자들은 P2P 응용에 의해 발생하는 트래픽을 차단하기 위해 널리 알려진 P2P응용 관련 포트(TCP/UDP)를 차단하는 규칙을 방화벽에 추가하여 대처하고 있다. 하지만 이런 P2P 응용에 의해 발생하는 트래픽의 유입을 효과적으로 차단하기 힘든데, 이는 최근의 P2P 응용 프로그램들이 방화벽을 우회하기 위해 'Port Hopping'과 같은 기법을 도입하고 있는 등 활발하게 진화하고 있기 때문이다. 또한, 최근 국내에는 '인터넷 디스크 P2P'라고 불리는 새로운 종류의 P2P 응용 프로그램들이 등장하고 있는데, 이런 P2P 응용 프로그램은 인터넷 디스크 서비스(Internet Disk Service) 형태로 가입자에게 제공되기 때문에 일반적인 HTTP 응

· 본 연구는 정보통신부 및 정보통신연구진흥원의 대학IT연구센터 지원 사업의 연구 결과입니다. (IITA-2005-(C1090-0502-0020))
† 비 회 원 : 한국전자통신연구원 NCP기술팀 선임연구원
bjlee@etri.re.kr
shpyoon@etri.re.kr
†† 정 회 원 : 충남대학교 전기정보통신공학부 컴퓨터전공 교수
lee@cnu.ac.kr
(Corresponding author임)
논문접수 : 2006년 7월 3일
심사완료 : 2007년 4월 24일

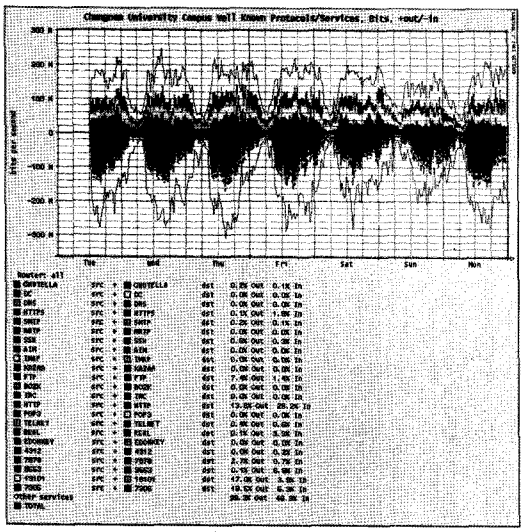


그림 1 CNU-KOREN 트래픽(포트 기준. 2005.7.20)

용 트래픽과 구별하기가 쉽지 않다는 문제를 가지고 있다. 따라서 이러한 트래픽을 정확하게 식별하는 동시에 일반적인 HTTP 트래픽과 구별하기 위해서는, 시그너춰(signature)에 근거한 내용 기반 트래픽 모니터링(content-based traffic monitoring) 시스템이나 IDS를 사용하여야 한다.

그림 1은 충남대학교와 외부망(KOREN)을 연결하는 주 링크상에서 양방향으로 수집된 트래픽을 포트 기준으로 분류한 것이다. eDonkey 등의 P2P가 사용하는 포트들을 방화벽에서 차단하였음에도 다량의 미식별 트래픽들이 발생하고 있음을 알 수 있다. 그 중 상당수가 7305, 9553, 19101 포트를 상당한 기간 동안 사용하고 있는 바, 새로운 포트를 사용하는 새로운 P2P 응용이 등장한 것으로 볼 수 있는 바, 수집된 트래픽에 대한 off-line 분석 작업 및 인터넷을 통한 자료 조사를 통해 이들 포트가 어떠한 인터넷 응용에 의해 사용되는 것인지 확인하고자 하였다. 그 결과, 이들 포트들 중 일부가 인터넷 디스크 서비스를 사용하고자 할 때 설치되는 ActiveX 컨트롤에 의해 사용되고 있음을 확인하였다. 이러한 컨트롤은 가입자가 인터넷 디스크 서비스를 사용하고자 할 때 자동적으로 설치되는 것으로, 그 중 몇몇은 P2P 클라이언트 응용 프로그램과 동일한 기능을 수행하는 것으로 확인되었다. 다른 가입자의 하드 디스크로부터 파일을 다운받거나, 다른 가입자가 자신의 하드 디스크로부터 파일을 다운받을 수 있도록 지원하는 것 등이 이에 해당한다.

이상의 분석과정이 보여주듯이, 인터넷 디스크처럼 상당한 양의 트래픽을 유발하는 인터넷 응용 프로그램의 등장은 망 관리의 부담을 상승시킨다. 통상적으로, 기업망이나 학내망의 외부 접속 구간에는 유해 트래픽의 유입을 차단하기 위한 IDS 장비들이 부설된다. 이러한 장비들은 웜이나 바이러스같은 특정 부류의 트래픽을 차단하는 역할을 수행할 뿐 아니라, 특정한 종류의 인터넷 응용으로부터 발생하는 트래픽의 유입 양을 제한하기 위한 용도로 사용될 수 있다. 하지만 이러한 IDS 장비들의 이론적인 근간이 되는 내용 기반 감지 기법(content-based detection technology)은 그 특성상 장비 관리자의 부담을 추가로 증가시킨다. IDS 장비의 입력으로 사용되는 시그너춰의 목록을 최신상태로 유지시키기 위해서는 인터넷 트래픽을 수집하고 분석하여 시그너춰를 찾아내는 작업이 반드시 수반되어야 하기 때문이다. 이러한 작업은 그 특성상 많은 시간을 소요하며, 또한 복잡도도 높은 것으로 알려져 있다. 인터넷 웜이나 바이러스에 대해서는 시그너춰를 자동 탐색하는 알고리즘이 몇 가지 제안되어 있으나, P2P 응용 프로그램에 일반적으로 적용할 수는 없는, 제한적인 알고리즘들이다.

따라서 본 논문에서는 인터넷 디스크 형태의 P2P 응용 프로그램으로부터 발생되는 플로우들을 빠른 시간 안에 식별해 내고, 그로부터 시그너춰를 자동적으로 추출해 내는 시스템인 GENESIS(System for GENErating SIgnatureS)를 제안한다. GENESIS는 raw traffic dump로부터 웹 기반 P2P 응용으로부터 발생한 것으로 추정되는 트래픽들만을 추출하고, 해당 추출 결과로부터 시그너춰들을 생성해 낸다. GENESIS에 의해 생성된 시그너춰들은 IDS를 포함하는 트래픽 차단/제한(traffic filtering/shaping) 시스템의 입력으로 사용될 수 있다.

상기 방안을 기술하는 데 있어, 본 논문은 다음 순서를 따른다. 2절에서 관련 연구들을 설명하고, 3절에서 GENESIS의 구조 및 그 구현에 사용된 주요 알고리즘들에 대해서 설명한다. 현재 GENESIS가 인터넷 디스크 P2P에 대해서만 시그너춰 생성이 가능하도록 구현되어 있으므로, 인터넷 디스크 P2P 응용 프로그램의 동작 방식에 대해서도 간략하게 설명할 것이다. 4절에서는 GENESIS를 실제로 수집된 패킷들에 대해서 적용한 결과를 보일 것이고, 5절에서는 GENESIS의 확장성에 대해서 약술할 것이며, 6절에서는 이 논문의 내용을 요약하는 동시에, 제안한 방법이 갖는 단점에 대해서 설명하고 향후 연구 방향을 간략히 제시하도록 할 것이다.

2. 관련 연구

대부분의 트래픽 차단(packet filtering) 시스템, 그리고 새롭게 제안되고 있는 트래픽 모니터링 툴에는 시그너춰에 기반을 둔 트래픽 감지 및 분류기능이 포함되어 있다. 통상적으로, 시그너춰(signature)는 패킷 페이로드(payload)에서 추출된 패턴을 의미하는 것으로, 특정한

응용에 의해 생성된 트래픽 플로우나 패킷을 해당 응용에 속한 것으로 판정하기에 충분할 만큼 빈번히 관측되어야 한다. 하지만 최근 이러한 시그너춰의 전통적인 용법이 확장되어, 다양한 종류의 시그너춰들이 등장하고 있다. 최근 문헌에 등장하는 시그너춰는 대략 다음의 세 가지 범주, 즉 (1) 페이로드 시그너춰(payload signature), (2) 통신 패턴 시그너춰(communication-pattern signature), 그리고 (3) 통계적 시그너춰(statistical signature)의 세 가지로 나누어 볼 수 있다.

페이로드 시그너춰는 IDS와 같은 내용 기반 트래픽 식별/차단 시스템에 의해 주로 사용된다. [1]에 제안된 방법은 페이로드 시그너춰를 사용하여 인터넷 트래픽을 어떻게 분류할 수 있는지를 보여주는 좋은 사례이다. 하지만 앞서 언급한 바와 같이, 페이로드 시그너춰를 최신 상태로 유지하기 위해서는 망 관리자가 상시적으로 오프라인 분석 작업을 해야 한다. 이러한 분석 작업에 드는 비용이 크기 때문에, 페이로드 시그너춰의 올바른 운용에 드는 오버헤드는 높다고 볼 수 있다. 이 오버헤드를 줄이기 위해, 시그너춰를 자동생성하기 위한 방안들이 제안되어 있다. [2]는 인공 지능 기법 중 하나인 기계 학습법(machine learning)을 사용한 방안을 제안하고 있다. 하지만 이 방안은 시그너춰를 생성하기 위한 입력으로 사용되는 인터넷 트래픽 트레이스(traffic trace)가 어떠한 인터넷 응용 프로그램에 의해 생성된 것인지 반드시 알고 있어야 한다는 단점을 가지고 있어, 수집된 트래픽 트레이스에서 특정한 부류의 응용 프로그램에 의해 생성된 트래픽 패킷들만을 추출한 뒤 해당 패킷들에 대해 시그너춰를 검출하는 용도로는 사용하기 부적합하다. 이 이외에도 시그너춰의 자동 생성에 관해 다루고 있는 다양한 논문들이 있으나[3-5] 웜이나 바이러스로부터 생성된 트래픽으로부터 시그너춰를 추출하는 문제에 대해서만 다루고 있어, 그 통신 형태가 보다 복잡한 일반적 형태의 인터넷 응용 프로그램들에 적용하기에는 무리가 있다.

통신 패턴 시그너춰는 인터넷 응용 프로그램의 통신 패턴을 시그너춰로 사용하는 것으로, 서로 다른 부류의 인터넷 응용 프로그램은 서로 다른 통신 패턴을 갖는다는 가정에 기반하고 있다. [7]에서는 'graphlet'이라는 이름의 통신 패턴 시그너춰를 제안하고 있다. graphlet을 사용하면 트래픽을 분류하는 작업이 패킷의 페이로드를 검사하지 않고도 이루어질 수 있다는 장점이 있다. 하지만 새로운 graphlet을 만들어 내는 비용이 페이로드 기반 시그너춰에 비해 낮다고 볼 수 없을 뿐 아니라, 최근 Skype의 경우에서 확인된 바대로[11], 실제로는 VoIP 응용이지만 하부 프로토콜로는 eDonkey 프로토콜을 사용하는 등 통신 패턴만을 고려할 경우 완벽한 트래픽 분류가 이루어 질 수 없는 경우가 있어, 페이로드 시그너춰에 대한 완벽한 대안이라고 볼 수는 없다. 또한 통신 패턴은 트래픽이 완전히 수집된 뒤, 다시 말해 트래픽이 인터넷을 통해 유통된 사후에야 검출이 가능한 정보이므로, graphlet과 같은 형태의 시그너춰는 인터넷 트래픽에 온라인으로 적용하기에는 적합하지 않다.

통계적 시그너춰는 트래픽에 대한 통계적 지식을 시그너춰로서 사용하는 것이다[8,9]. 통계적 시그너춰의 일부로 사용되는 정보들로는 플로우의 평균 패킷 사이즈 및 평균 지속시간, 플로우를 구성하는 패킷 간의 도착 간격 시간(inter-arrival time) 같은 것들이 있다. 통계적 시그너춰 또한 통상 플로우에 대해서 정의되는 바, 통신 패턴 시그너춰와 마찬가지로 온라인으로 적용되기에는 무리가 있다고 볼 수 있다. 플로우가 완전히 종료된 이후에야 그 플로우에 대한 통계적 정보를 활용할 수 있기 때문이다.

표 1은 인터넷 트래픽을 분류 혹은 식별하는 데 있어서 어떠한 시그너춰가 적합한지를 보여준다. 이 표에서 응용 식별(application identification)은 어떤 플로우나 패킷이 어떤 인터넷 응용 프로그램에 의해 발생한 것인지를 결정하는 작업을 지칭하며, 트래픽 분류(traffic classification)는 플로우나 패킷들을 그 응용 프로그램에 의해 분류하는 일반적인 방법을 지칭하는 것이다. 따라서 트래픽 분류의 경우, 해당 분류작업의 결과로는 트래픽 트레이스를 구성하는 플로우들이 대략적으로 어떠한 응용 프로그램들에 의해 만들어진 것인지를 나타내는 값이 산출된다. 예를 들어, 몇 종류의 P2P 응용 프로그램들이 발생한 트래픽은 트래픽 분류 작업의 결과로 "P2P" 응용 프로그램들에 의해 만들어 진 것으로 판정될 수 있다. 이는 트래픽 분류 작업에 의한 판정 결과가 반드시 해당 트래픽을 발생시킨 응용 프로그램을 정확하게 결정해 낼 수는 없을 수도 있다는 것을 의미한다. 본 논문에서는 응용 식별 문제에 초점을 맞추고 있다.

표 1 트래픽 분석 작업 및 시그너춰 유형간 관계

	응용 식별	트래픽 분류
시그너춰 종류	페이로드 시그너춰	페이로드 시그너춰 통신 패턴 시그너춰 통계적 시그너춰

3. GENESIS 구조

앞서 언급했던 바와 같이, GENESIS를 설계하는 데 있어서의 목표는 시그너춰를 자동적으로 생성하는 것이다. GENESIS에 의해 추출된 시그너춰는 Wise<TrafView>[10]나 [1]에서 제안된 시스템과 같은, 응용 식별이나 트래픽 분류를 목적으로 하는 시스템의 입력으로 사용될

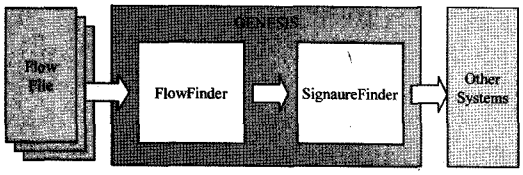


그림 2 GENESIS 구조

수 있다. 또한 GENESIS에 의해 생성된 시그너춰는 IDS와 같은 트래픽 유입 감지/차단 시스템의 입력으로 사용될 수도 있다.

그림 2는 GENESIS의 구조를 도시한 것이다. 도시된 바에 따르면, GENESIS는 FlowFinder와 Signature-Finder의 두 부분으로 구성된다.

FlowFinder는 트래픽 플로우를 일정한 속성에 따라 분류한다. 이 속성으로는 통신 패턴과 같은 것이 사용될 수 있다. 따라서 FlowFinder의 입력으로는 플로우 레코드로 구성되는, 일련의 트래픽 파일들이 주어지게 되고, FlowFinder의 실행 결과로는 공통의 속성을 갖는 플로우들만이 모여 이루어진 일련의 트래픽 파일이 만들어지게 된다. FlowFinder의 주목적이 트래픽 분류이기 때문에, FlowFinder를 구현하는 데 있어서 현존하는 다양한 분류방법이 사용될 수 있다. 즉, [7]에서 제안하는 분류방법이 FlowFinder의 구현에 사용될 수 있다. 이 때 고려해야할 유일한 제약조건은 각각의 플로우 레코드에 포함되는 패킷 레코드들이 페이로드를 포함하여야 한다는 것뿐이다. 본 논문에서는 실험을 위해 간단한 형태의 FlowFinder를 직접 구현하여 사용하였다. 이 Flow-Finder는 인터넷 디스크 P2P에 의해 발생한 것으로 의심되는 플로우들을 찾아내어 그 플로우들을 일정한 특성에 따라 분류한다.

SignatureFinder는 FlowFinder가 생성한 각각의 파일로부터 시그너춰를 추출한다. 따라서 SignatureFinder의 입력은 하나의 플로우 파일이며, 그 출력은 추출된 시그너춰의 목록이다. 시그너춰들은 그 적중도(coverage) 값에 따라 내림차순으로 정렬되어 출력된다.

3.1 FlowFinder

앞서 언급한 바와 같이, 현재 구현된 FlowFinder는 인터넷 디스크 P2P 응용만을 그 대상으로 하고 있다. FlowFinder의 구현을 위해, 대한민국에서 사용되는 인터넷 디스크 P2P 응용 프로그램에서 발생한 트래픽을 오프라인으로 분석하였다. 현재 널리 사용되고 있는 인터넷 디스크 서비스들을 사용하기 위해서는 먼저 인터넷 디스크 서비스 제공업체의 포털 사이트에 접속하여야 한다. 사용자가 로그인하는 순간 사용자의 데스크탑에 ActiveX 컨트롤이 설치되며, 사용자는 이 ActiveX 컨트롤을 통해 파일을 공유하거나 다운로드 받는 등 인터넷 디스크 서비스를 이용할 수 있게 된다. 로그인 한 사용자는 관심 있는 파일을 소유한 사용자를 '친구'로 등록한 뒤, 해당 사용자가 친구 등록을 승인한 다음 그 사용자가 가지고 있는 파일들을 다운 받을 수도 있고, 아니면 누구나 다운로드 할 수 있는 공개 파일들을 다운받을 수도 있으며, 자신이 소유한 파일들을 인터넷에 올릴 수도 있다. 인터넷에 파일을 업로드 하는 절차를 제외한 나머지 절차들이 어떤 순서로 이루어지는지를 그림 3에 요약하였다.

상기의 서비스를 이용하는 사용자는 사용자간의 파일 공유가 인터넷 디스크 상에서 이루어진다고 생각하지만, 구현에 따라서는 파일이 인터넷 디스크 상에 업로드가 끝난 뒤에는 그 공유 과정이 P2P 응용 프로그램과 유사한 과정을 통해 이루어지는 경우도 있다. 사용자가 인터넷 디스크로부터 파일을 다운로드 받기 시작하면, 그 사용자는 파일 소유자 및 현재 그 파일을 다운로드 받고 있는 사용자들을 연결하는 P2P 세션(session)에 자동적으로 참여하게 된다. 파일을 다운로드 받고 있는 사용자가 없는 경우에는 파일 다운로드는 인터넷 디스크로부터 직접적으로 이루어지지만, 파일을 다운로드 받고 있는 다른 사용자가 있는 경우에는 해당 P2P 세션에의

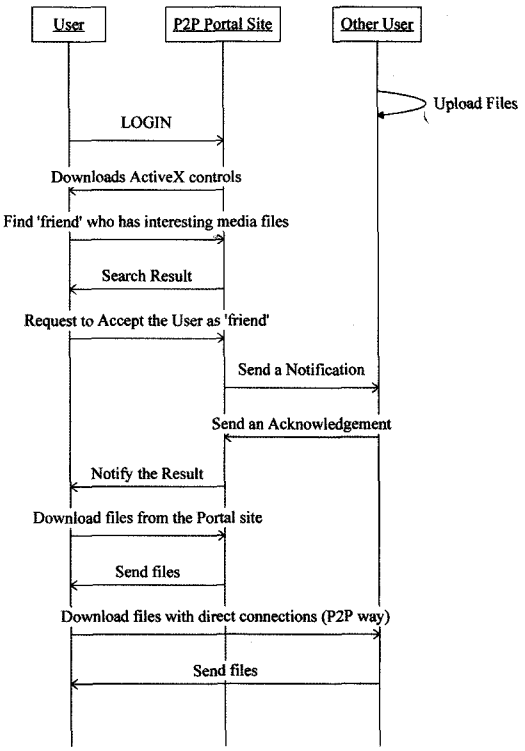


그림 3 인터넷 디스크 P2P 응용 개체간의 상호작용 다이어그램(Interaction Diagram)

신규 참가자는 해당 파일의 내용 중 아직 다운로드 받지 못한 부분을 P2P 세션 내의 다른 사용자로부터 전송 받게 된다. 이러한 기법은 인터넷 디스크에 대해 발생하게 되는 엄청난 양의 다운로드 요구를 효율적으로 부하분산하기 위해 도입된 것이나, P2P 세션의 크기가 커지면 결국 일반 P2P 응용과 마찬가지로 인터넷 대역폭을 과도하게 점유하게 된다는 문제를 낳는다.

상술한 웹 기반 P2P 응용 프로그램의 특성에 근거하여, 인터넷 디스크 P2P에 의해 발생한 것으로 추정되는 트래픽의 플로우를 수집하는 데 사용되는 규칙을 만들었다. 그림 4에 도시된 바와 같이, 인터넷 디스크 P2P 응용 프로그램에 참여하는 개체는 두 가지로 분류할 수 있다. 하나는 웹 포탈 서버이고, 다른 하나는 인터넷 디스크 서비스를 사용하는 사용자이다. 따라서 해당 개체들 간에 교환되는 트래픽을 그 방향 및 참여 개체에 따라 구별하기 위한 규칙은 네 가지가 만들어 질 수 있다 (표 2). 예를 들어, 어떤 플로우의 송신자 IP 주소가 인터넷 디스크 P2P 웹 포탈의 IP 주소 중 하나이고 그 목적지 주소는 W 범주에 속하는 플로우에 의해 사용된 송신자 주소들 중 하나일 경우, 해당 플로우는 X 범주에 속하는 플로우이다. 그 경우 해당 플로우는 송신자 IP 주소와 일치하는 이름을 갖는 디렉터리 아래에 X.genesis라는 이름의 파일에 추가되어 저장된다. FlowFinder는 이와 같은 과정을 통해 인터넷 디스크 P2P 응용 프로그램에 의해 발생된 것으로 추정되는 플로우들을 분류하여 저장하며, 그 과정에서 특정한 인터넷 디스크 포털 사이트와 연관된 트래픽의 대략적인 식별 또한 가능하다.

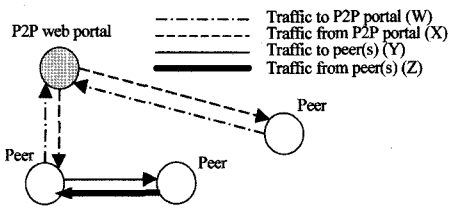


그림 4 웹 기반 P2P 응용 프로그램 개체간의 트래픽 흐름

표 2 플로우 분류 규칙

범주	송신자 IP 주소	수신자 IP 주소
W	P2P 웹 포털의 IP 주소가 아닐 것	P2P 웹 포털의 주소
X	P2P 웹 포털의 주소	W 플로우들에 사용된 송신자 IP주소 중 하나
Y	W 플로우들에 사용된 송신자 IP주소 중 하나	P2P 웹 포털의 IP 주소가 아닐 것
Z	P2P 웹 포털의 IP 주소가 아닐 것	W 플로우들에 사용된 송신자 IP주소 중 하나

유의할 것은 상기의 분류 규칙이 의미를 가지기 위해서는 웹 기반 P2P 포털 사이트의 주소가 필요하다는 점이다. 이 IP 주소들을 찾는 과정은 다음 절차에 의거하여 자동화될 수 있다: (1) TCP 프로토콜을 사용하고 목적지 포트 번호가 80번이며 SYN 패킷이 포함된 플로우로부터 <송신자 주소 X, 목적지 주소 Y>의 순서쌍을 추출한다. (2) 각각의 순서쌍에 대해서, X와 Y를 제외한 기타 주소간에 교환된 데이타 트래픽 플로우들을 검색한다. (3) 주소 Y에 대해, (2)의 과정을 통해 검색된 데이타 트래픽이 지속적으로 관측될 경우, Y는 인터넷 디스크 P2P 포털 사이트의 IP 주소일 가능성이 매우 높다.

(1), (2), (3)의 과정을 거치면 인터넷 디스크 P2P 응용 프로그램의 웹 포털 사이트 IP 주소일 가능성이 높은 주소들만을 식별해 낼 수 있으나, 현재로서는 FlowFinder의 구현을 단순화하기 위해 대한민국에서 사용되고 있는 인터넷 디스크 서비스 업체들의 IP 주소를 직접 수집하여 입력으로 사용하고 있으며, 추후 확장할 예정이다.

3.2 SingnatureFinder와 시그너춰 추출 과정

SignatureFinder는 각각의 *.genesis 파일에 대해서 두 단계로 동작한다. 첫 번째 단계에서는 플로우 파일들에 저장되어 있는 모든 패킷들의 모든 페이로드 내의 모든 바이트들에 대한 통계 정보를 기록하고, 두 번째 단계에서는 해당 정보를 사용하여 시그너춰를 생성한다. 한편, SignatureFinder를 구현하는 데 있어서 다음의 두 가지를 가정하였다.

(1) 페이로드의 첫 B 바이트(B = 64) 까지를 검사하면 충분히 시그너춰를 찾을 수 있다.

(2) P2P 응용 플로우 중 데이타 플로(본 논문에서는 100개 이상의 패킷으로 구성되며, 그 패킷들의 평균 크기가 1000이상인 플로우를 지칭하는 것으로 가정) 추정되는 플로우의 경우에는, 그 패킷들 중 첫 P개 (P = 10)만 검사하면 충분하다.

상기한 B와 P 값의 타당성을 검증하기 위해, TCP/IP 헤더를 제거한 모든 페이로드에 대해서, 해당 페이로드를 구성하는 각각의 바이트 값의 빈도수를 검사하였다. 이를 위해, 페이로드의 i번째 위치에 그 값이 j인 바이트가 나타나는 빈도를 저장할 2차원 배열을 구성하였다 ($0 \le i \le 1500$, $0 \le j \le 255$). 12시간 분량의 패킷으로부터 이 2차원 배열을 구성한 뒤, 각각의 i 값에 대하여, (i,0), (i,1), … (i,255) 번 원소들 간의 최대값 max(i), 최소값 min(i), 평균값 avg(i)를 구하였다. 그리고 그 값들을 토대로, 페이로드의 i 번째 바이트에 signature가 존재할 가능성을 나타내는 값 PSE(i)를 다음과 같이 계산하였다.

$$PSE(i) = (max(i)-avg(i)) / max(i) - avg(i) / max(i) \quad (1)$$

PSE(i) 값의 범위는 [-1,1]이다. 그 값이 1에 근접하면 해당 위치 i에서 특정한 값을 갖는 바이트가 다른 값보다 훨씬 빈번하게 나타난다는 것을 의미한다. 그림 5에 12시간 분량의 패킷들의 페이로드로부터 계산된 PSE(i)의 그래프를 보였다. 그래프에 나타난 대로, 첫 64 바이트 이내의 PSE(i) 값이 다른 구간에 비해 높으며, PSE(i)의 값은 해당 구간 이후에는 감소하여 평형 상태에 도달하게 된다.

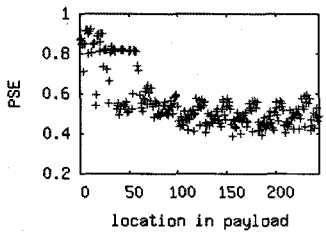
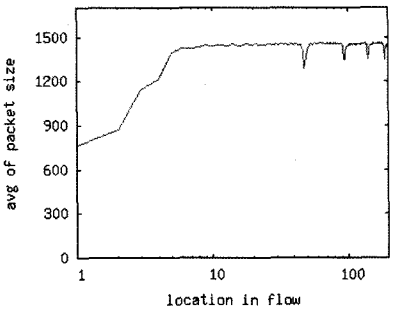


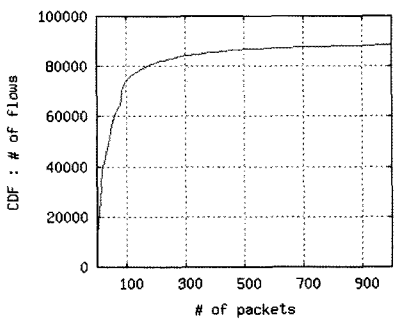
그림 5 PSE(i) 그래프
(2005년 11월 12일 18:00~11월 12일 06:00)

두 번째 가정을 검증하기 위해서는 다음과 같은 방법을 사용하였다. 우선, 모든 데이타 플로우에 대해서 그에 속한 패킷들의 평균 크기를 플로우 내의 각 위치에 대해서 계산하였다. 그 결과는 그림 6(a)에 제시되어 있는데, 통신 프로토콜의 운용에 관련된 컨트롤 패킷(control packet)이라고 볼 수 있는 작은 크기의 패킷들이 플로우 앞쪽 10개 이내에 집중되어 있음을 확인할 수 있다. 아울러, 모든 데이타 플로우의 크기(플로우에 속한 패킷의 개수)를 조사하였다. 그 결과를 그림 6(b)에 CDF 형식으로 보였다. 대부분의 플로우들(78.56%)이 100개 미만의 패킷들을 갖는 짧은 플로우들이었으나, 해당 플로우들에 의해 점유되고 있는 트래픽의 양은 12.22%에 불과하였다. 반면, 100개 이상의 패킷들을 갖는 플로우들이 점유하고 있는 트래픽의 양은 87.78%에 달했다. 이는 통상적으로 알려져 있는 바와 같이, 인터넷 회선의 대부분의 대역폭은 그 지속시간이 긴 플로우들에 의해 점유되고 있다는 사실과도 부합하는 것이다. 따라서 앞서 제시한 두 번째 가정은 유효하다고 할 수 있다.

상기 관측 결과에 의거하여 SignatureFinder의 1 단계 수행 과정의 알고리즘을 설계하였다(알고리즘 1). 이 알고리즘의 목적은 목적은 패킷의 페이로드들로부터 2차원 배열 형태의 통계 정보를 생성하는 것이다. SignatureFinder는 이 정보를 packet_counters와 flow_byte_counters의 두 2차원 배열에 저장한다. packet_counters



(a) 평균 패킷 크기



(b) 플로우 크기에 대한 CDF
그림 6 평균 패킷 크기와 플로우 분포

의 (i,j)번 원소는 페이로드의 i번째 바이트에 j라는 값(0≤j≤255)이 나타난 빈도를 기록하며, flow_byte_counters의(i,j) 번 원소는 페이로드의 i번째 위치에 j라는 값을 갖는 패킷이 속한 모든 플로우의 총 량(바이트 단위)을 기록한다.

알고리즘 1. SignatureFinder의 1단계 알고리즘

```
DECLARE_GLOBAL_VARIABLES
 var packet_counters[64][256];
 var flow_byte_counters[64][256];
 var total_flows_in_bytes := total_packets := 0;
PROCEDURE_BEGIN
 *.genesis file안의 모든 플로우 F에 대해
   var fsz = 플로우 F의 크기 (byte 단위)
   var tmp_matrix[64][256];
   total_packets += 1;
   total_flows_in_bytes += fsz;

   F가 데이타 플로우면 bound := 10.
   그렇지 않은 경우, bound := (F 내의 패킷 개수);

   F 내의 첫 'bound'개 패킷들에 대해서
     페이로드 내의 모든 바이트에 대해
       i := 해당 바이트의 위치;
```

```
        j := 해당 바이트의 값;
        packet_counters[ i ][ j ] += 1;
        tmp_matrix[ i ][ j ] := 1;

    for i where 0 <= i < 64
      for j where 0 <= j < 256
        tmp_matrix[i][j]의 값이 0보다 크면
          flow_byte_counters[i][j]+= fsz;
PROCEDURE_END
```

제 2 단계에서, SignatureFinder는 플로우 파일을 처음부터 다시 분석하여, 1 단계에서 생성한 통계 정보를 사용하여 시그너춰들을 만들어 낸다. 이 과정을 그림 7에 보였다.

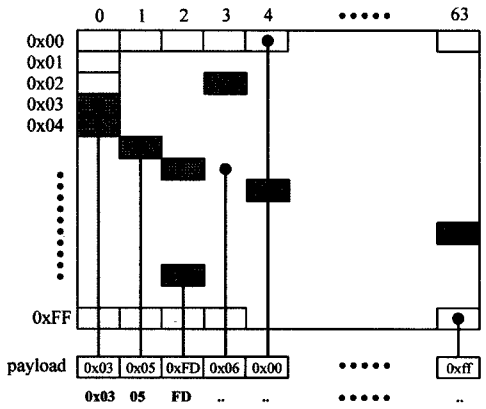


그림 7 시그너춰 생성 과정

그림 7의 상단에 등장하는 2 차원 배열('sigmatrix')는 1 단계를 거쳐 수집한 통계 정보를 사용해 만들어진다. 이 배열의 (i,j)번 원소의 값은 페이로드의 i번째 바이트의 값이 j인 패킷이 10 이상이고, 그런 패킷을 갖는 플로우의 크기의 총 합이 해당파일 내 전체 플로우의 90% 이상일 경우에만 1로 설정되며, 그 이외의 경우에는 0으로 설정된다. 그림 6에서 회색으로 표시된 배열 원소들은 그 값이 1로 설정된 것들이다.

SignatureFinder는 플로우 내의 모든 패킷들의 페이로드를 순서대로 검사한다. 페이로드내의 i번째 바이트의 값이 j라고 할 때($0 \le i \le 63$), 각각의 패킷의 페이로드에 대해서 sigmatrix[i][j]가 0이 아닌 모든 j를 순서대로 모아 시그너춰를 구성하는 것이다. 따라서 패킷을 검사할 때 마다 시그너춰가 만들어지게 되는데, 시그너춰가 없는 경우는 다음 패킷으로 진행하고, 있는 경우에는 시그너춰 리스트에 집어넣는다. 생성된 시그너춰가 시그너춰 리스트에 이미 존재하는 경우에는 해당 시그너춰의 통계량만을 갱신한다. 이 통계량으로는 다음과 같은 것이 있다.

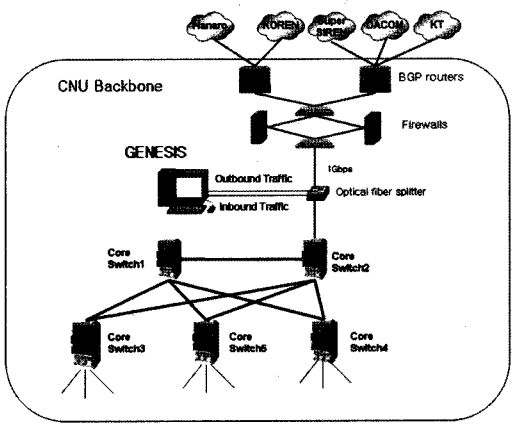


그림 8 측정 환경

- 해당 시그너춰에 의해 식별되는 플로우의 총 개수
- 해당 시그너춰에 의해 식별되는 패킷의 총 개수
- 해당 시그너춰에 의해 식별되는 패킷들의 총량(바이트 단위)
- 해당 시그너춰에 의해 식별되는 플로우들에 포함된 패킷들의 개수의 총합
- 해당 시그너춰에 의해 식별되는 플로우의 크기(바이트 단위)의 총합

제 2 단계가 끝나면 시그너춰 리스트에 저장된 시그너춰들이 상기 통계량 중 다섯 번째 통계량에 해당하는 값의 역순으로 정렬되어 출력된다.

4. 실 험

GENESIS 시스템에 대한 실험은 충남대학교 학내망을 외부망과 연결하는 주 링크 상에서 이루어졌다. 이 링크는 1Gbps Ethernet 링크이며, 최대 사용량은 300Mbps 정도이다. 이 링크에서 12시간 분량의 트래픽을 수집하였다(2005년 11월 11일 18:00~2005년 11월 12일 06:00). 이 수집된 트래픽 상에서 GENESIS 시스템을 구동하여 시그너춰를 생성하였다. 본 실험에 있어서는 표 3의 웹 포털 주소들을 사용하였다.

표 3 인터넷 디스크 포털 URL 목록

URL
http://www.clubbox.co.kr
http://idisk.megapass.net
http://diskpot.chol.com
http://disk.dreamwiz.com
http://www.hotdisk.co.kr
http://www.lenthard.com
http://www.folderplus.com
http://www.edisk.com

표 4에는 상기 링크에서 수집된 트래픽에 대한 기본적 통계를 보였다. 관측된 트래픽 중 대부분은 TCP 트래픽이었으며, 사용된 송신자 측 포트번호와 수신자 측 포트번호를 그 사용량에 따라 순위를 매긴 결과는 표 4와 같다. 표 4에 드러난 바와 같이, 상위 10개의 포트 중 well-known port는 80과 25뿐이며, 다량의 데이타 플로우가 19101, 10101, 9553등의 포트에 연관되어 있음을 확인할 수 있다. 이 포트들은 앞서 서론에서 언급하였던 다량의 미식별 트래픽 관련 포트들과 중복되는 것들이다.

상기 트래픽으로부터 GENESIS가 추출해 낸 시그너춰들은 표 5에 요약되어 있다. 생성된 시그너춰와 오프라인 분석 결과를 비교한 결과, 현재로서는 CLUBBOX와 IDISK만이 인터넷 디스크 P2P 응용의 통신 패턴을 따르고 있음을 확인할 수 있었다. CLUBBOX와 IDISK를 제외한 다른 인터넷 디스크 서비스에 대해서는 그 사용량이 너무 적어(10Mbytes 미만) 의미있는 결과를 확인할 수 없었다. 한편, 표 5와 표 4를 대조해 보았을 때, GENESIS가 특정 인터넷 디스크를 사용하는 과정에서 점유되는 TCP 포트 번호들을 정확하게 발견해 내고 있음을 확인할 수 있다.

표 4 트래픽 점유율 상위 10개 포트 번호들

(a) 송신자 측 포트 번호 상위 10개

포트	플로우 수	패킷 수	바이트 수
80	7,650,206	280,949,804	340,833,743,679
19101	50,773	106,817,848	152,183,241,926
10101	3,801	14,638,104	17,029,899,750
9553	11,501	8,740,681	13,013,854,641
9575	122,064	8,782,216	11,756,617,597
554	10,440	6,772,154	9,748,092,702
8080	81,916	9,808,597	6,830,629,533
6881	108,682	6,402,313	5,423,845,716
9090	12,017	3,482,729	5,114,327,645
8629	159,221	4,628,615	4,779,302,193

(b) 수신자 측 포트 번호 상위 10개

포트	플로우 수	패킷 수	바이트 수
80	6,083,126	202,252,175	23,860,631,350
19101	11,383	10,655,279	12,609,578,925
10101	460	6,277,010	8,822,307,210
9553	60,614	143,804,974	6,055,251,622
9575	351,220	5,498,767	4,593,762,494
554	336	2,641,703	3,767,074,691
8080	117,349	6,726,289	2,817,319,214
6881	88,355	6,046,770	2,778,507,800
9090	62,228	15,516,376	2,660,240,535
8629	647,256	6,318,813	2,574,938,800

GENESIS가 만들어 낸 시그너춰들은 시그너춰 기반 트래픽 분석 시스템인 Wise<TrafView> 분석 서버[10]에 적용되었다. Wise<TrafView>는 포트 기반의 단순 분석 방법과 페이로드 시그너춰 기반의 분석 방법을 결합하여 인터넷에서 사용되는 응용 프로그램을 식별해 낸다. 상기 시그너춰를 Wise<TrafView>에 적용하기 위해서는 그 설정 파일 analysis.conf를 그림 9와 같이 바꾸어야 한다. CLUBBOX_DATA와 IDISK_DATA의 두 port_rep_name 구조체를 추가하고, 해당 구조체들의 대표 포트 번호로 19101과 10101, 그리고 9553을 부여하였다. 또한, 19101이나 10101이 송신자 측 포트 번호로 사용되었을 경우에는 시그너춰 '0x0000'이 페이로드의 지정된 위치에 나타나야만 CLUBBOX_DATA 응용에 속한 것으로 판정되도록 하였고, 그 역방향으로 나타나는 플로우는 해당 판정 결과에 의거하여 자동적으로 CLUBBOX_DATA 응용에 속한 것으로 판정될 수 있도록 하였다. IDISK_DATA에 대해서도 유사한 변경 작업을 시행하였다.

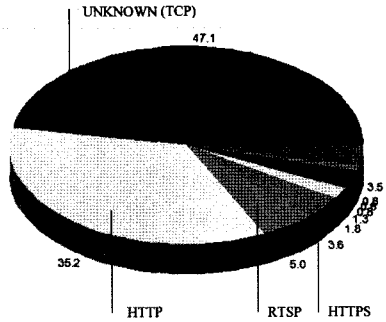
```
application INTERNET_DISK_P2P {
   port_rep_name CLUBBOX_DATA port 19101,10101 \
     protocol TCP {
       src_disc_pattern=="0x0000" in pkt 0-10 \
         at byte 6-7
       decision_group UP dir up
       decision_group DOWN dir down
   }
   port_rep_name IDISK_DATA port 9553 protocol TCP {
       src_disc_pattern=="0x0000" in pkt 0-10 \
         at byte 5-6
       decision_group UP dir up
       decision_group DOWN dir down
   }
}
```

그림 9 수정된 analysis.conf (부분)

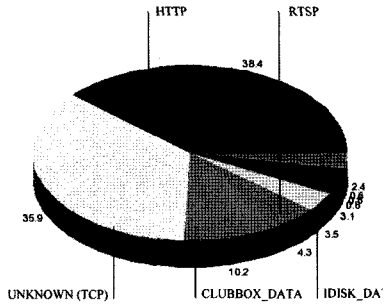
표 5 GENESIS에 의해 만들어진 시그너춰

Coverage = 해당 시그너춰에 의해 식별된 플로우의 총량 (byte) / 해당 파일 내의 모든 플로우의 총량 (byte)

웹 포탈 IP 주소	시그너춰 파일명	시그너춰(HEX): 첫 12byte	주로 사용된 포트 번호	Coverage
203.238.140.70 (www.clubbox.co.kr)	DC.signature	....0000	19101, 10101 (TCP, source)	86%
	WS.signature	48545450 2f312e31 20..30..	80 (TCP, source)	99%
	WC.signature	47455420 2f.....	80 (TCP, destination)	99%
211.62.15.20 (idisk.megapass.net)	DS.signature	....0000..	9553 (TCP, source)	81%
	WS.signature	48545450 2f312e31 20323030	80 (TCP, source)	99%
	WC.signature	47455420 2f.....	80 (TCP, destination)	99%



(a) 시그너춰 적용 전 (2006년 2월 22일)



(b) 시그너춰 적용 후 (2006년 3월 13일)

그림 10 시그너춰 적용 전후 인식율 비교

Wise<TrafView>에 시그너춰를 적용하여 실험한 결과를 그림 10에 제시하였다. 그림 10에서 확인할 수 있는 바 대로, GENESIS가 추출한 시그너춰를 적용한 이후 미식별 트래픽의 비율이 47.1%에서 35.9%로 낮아졌으며, clubbox 서비스와 idisk 서비스에 의해 점유된 트래픽 비율은 각각 10.2%와 3.5%로 확인되었다. 이 결과와는 별도로, 수집된 트래픽에 대한 오프라인 분석 작업을 시행하여, GENESIS가 자동 검출한 시그너춰들이 CLUBBOX와 IDISK의 두 인터넷 응용과 연관된 플로우의 페이로드에 등장하는 것 또한 확인하였다.

5. GENESIS의 확장성

본 논문에서는 GENESIS를 인터넷 디스크 P2P 응용 프로그램에 적용한 결과만을 기술하였으나, GENESIS는 다른 인터넷 응용 트래픽으로부터 시그너춰를 추출하기 위한 용도로도 사용될 수 있다.

SignatureFinder의 구현에 사용된 알고리즘이 특정한 인터넷 응용 프로그램에 종속되어 있지 아니하기 때문에, GENESIS는 다른 종류의 인터넷 응용 프로그램으로부터 발생한 트래픽을 식별하기 위한 시그너춰를 탐색하는 용도로도 사용될 수 있다. 즉, 그림 11에 보인 바와 같이 새로운 종류의 FlowFinder를 추가하여, 해당

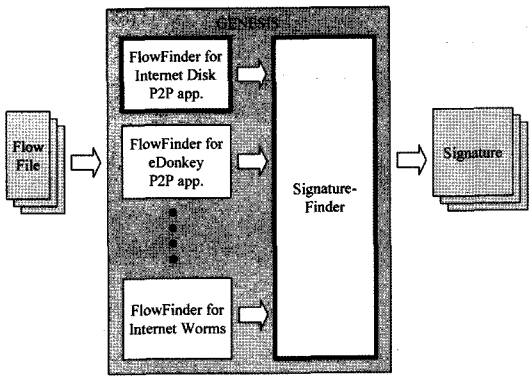


그림 11 GENESIS의 확장성

인터넷 응용 프로그램으로부터 발생한 것으로 추정되는 후보 트래픽을 분류하는 작업만을 새로 정의하면 된다. 따라서 BLINC[7]와 같은 시스템을 사용하여 FlowFinder를 새롭게 정의하거나 확장하는 것도 가능해 진다. 그 경우 BLINC 시스템을 FlowFinder로 사용하고, 새롭게 graphlet을 추가하는 형태로 FlowFinder를 확장해 나갈 수 있다.

6. 결 론

P2P 응용 프로그램에 의해 점유되는 대역폭이 증가함에 따라, P2P 응용에 의해 발생하는 트래픽을 감지하는 것의 중요성도 증대되어 왔다. 하지만 P2P 응용 프로그램이 고정된 포트 번호를 사용하던 종래 방법에서 탈피해 동적으로 포트를 선택하는 방향으로 진화함에 따라, 시그너춰 기반 탐지 기법의 실용성이 주목받기 시작하였다. 하지만 시그너춰를 생성하고 유지하는 작업은 어려울 뿐 아니라 많은 시간을 소요하는 힘든 작업이다. 따라서 시그너춰를 자동 생성할 수 있는 방법이 절실하게 요구된다. 본 논문에서는 웹 기반 P2P 응용 프로그램을 감지하고 그 시그너춰를 자동 추출해 주는 방법을 제안하고, 그 구현 결과물인 GENESIS 시스템에 대해서 소개하였다. 그 실험 결과는 시그너춰 기반 트래픽 분석 시스템인 Wise<TrafView>를 통해 검증하였다. 본 논문이 제시하는 방안 중 SignatureFinder 구현에 사용된 알고리즘 및 방법은 웹 기반 P2P 응용 프로그램에 제한된 방안은 아니며, FlowFinder 구현에 사용된 알고리즘을 확장하여 새로운 형태의 P2P 응용 프로그램 플로우를 검출하게 되면, 다양한 형태의 P2P 응용 프로그램으로부터 시그너춰를 추출하는 데 사용될 수 있다. 따라서 본 논문에서 설명한 분야의 P2P 응용 프로그램을 식별하는 것 이외에도, 다양한 형태의 P2P 응용 프로그램 식별에 사용될 수 있을 것이다.

참 고 문 헌

[1] S. Sen, O. Spatscheck, and D. Wang, "Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures," WWW, 2004.

[2] P. Haffnet, S. Sen, O. Spatscheck, and D. Wang, "ACAS: Automated Construction of Application Signatures," ACM SIGCOMM, 2005.

[3] S. Singh, C. Estan, G. Varghese, and S. Savage, "Automated Worm Fingerprinting," OSDI, 2004.

[4] J. Newsome, B. Karp, and D. Song, "Polygraph: Automatically Generating Signatures for Polymorphic Worms," IEEE Symposium on Security and Privacy, 2005.

[5] H. Kim and B. Karp, "Autograph: Toward Automated, Distributed Worm Signature Detection," 13th USENIX Security Symposium, 2004.

[6] T. Karagiannis, A. Broido, M. Faloutsos, and K. C. Claffy, "Transport Layer Identification of P2P Traffic," ACM Internet Measurement Conference, 2004.

[7] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," ACM SIGCOMM, 2005.

[8] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-Service Mapping for QoS: A Statistical Signature-based Approach to IP Traffic Classification," ACM Internet Measurement Conference, 2004.

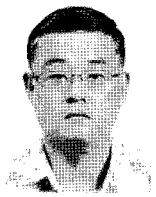
[9] W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," ACM SIGMETRICS, 2005.

[10] T. Choi, S. Yoon, H. Chung, J. Park, B. Lee, S. Yoon, and T. Jeong, "Flow-based Application-aware Internet Traffic Monitoring and Field Trial Experiences," APNOMS, 2005.

[11] Luca Deri, "Open Source VoIP Traffic Monitoring," SANE 2006.

이 병 준

1996년 서울대학교 컴퓨터공학과 학사
1998년 서울대학교 컴퓨터공학과 석사
2001년~현재 한국전자통신연구원 선임연구원. 2005년~현재 충남대학교 전기정보통신공학부 박사과정. 관심분야는 네트워크 관리, 통신 프로토콜 개발 등



윤 승 현

1991년 성균관대학교 산업공학과 학사
1993년 성균관대학교 산업공학과 석사
1997년 성균관대학교 산업공학과 박사
1997년~현재 한국전자통신연구원 선임연구원. 관심분야는 인터넷 트래픽 엔지니어링, 트래픽 측정 및 분석, 네트워크 제어, 시스템 성능분석 등

이 영 석

1995년 서울대학교 컴퓨터공학과 학사
1997년 서울대학교 컴퓨터공학과 석사
2002년 서울대학교 컴퓨터공학부 박사
2002년~2003년 University of California, Davis 방문연구원. 2003년~현재 충남대학교 전기정보통신공학부 컴퓨터전공 조교수. 관심분야는 차세대 인터넷, IPv6, 인터넷 트래픽 측정 및 분석