# 네트워크 트래픽 특성을 이용한 개인정보유출 탐지기법

박 정 민[†] · 김 은 경[††] · 정 유 경[††] · 채 기 준[†††] · 나 중 찬[††††]


## 요 약

유비쿼터스 네트워크 환경에서 개인정보의 유출은 다양한 사이버 범죄를 야기하며 개인정보의 상품화로 프라이버시의 침해가 증가하므로 개인정보의 유출을 탐지하는 것은 매우 중요하다. 본 논문은 네트워크의 트래픽 특성을 기반으로 한 개인정보 유출 탐지 기법을 제안하고자 한다. 실제 대학망에서 정상 상태의 트래픽을 수집하여 트래픽의 특성을 분석함으로써 네트워크 트래픽이 자기유사성을 지님을 확인하였다. 개인정보의 유출을 시도하는 악성코드의 사전정보수집단계를 모사한 비정상적인 트래픽에 대하여 정상 트래픽에서의 자기유사성과의 변화를 살펴봄으로써 이상을 조기 감지할 수 있었다.


**키워드 : 개인정보, 개인정보유출, 자기유사성, 트래픽 특성**

# Detection of Personal Information Leakage using the Network Traffic Characteristics

Jung-Min Park[†] · Eunkyung Kim[††] · Yukyung Jung[††] · Kijoon Chae[†††] · Jungchan Na[††††]


## ABSTRACT

In a ubiquitous network environment, detecting the leakage of personal information is very important. The leakage of personal information might cause severe problem such as impersonation, cyber criminal and personal privacy violation. In this paper, we have proposed a detection method of personal information leakage based on network traffic characteristics. The experimental results indicate that the traffic character of a real campus network shows the self-similarity and proposed method can detect the anomaly of leakage of personal information by malicious code.


**Key Words : Personal information, Personal information leakage, Self-similarity, Traffic characteristics**

## 1. 서 론

인터넷의 발달과 휴대전화, 노트북, PDA와 같은 휴대용 단말기 사용의 보편화 등 정보 통신 환경의 발달은 인터넷 뱅킹, 사이버 주식 매매, 사이버 교육, 각종 전자 정부시스템 등 정보 기술의 활용을 가속화하였다. 인터넷과 정보통신 기술의 발달로 인해 언제 어디서나 네트워크에 접속하여 소통할 수 있는 유비쿼터스 환경에서 사람들은 자신의 의지에 상관없이 네트워크 공간에 놓이게 된다.

개인정보는 생존하는 개인을 식별할 수 있는 정보로 이름, 주소, 생년월일, 신용카드번호, 직업, 병력, 학벌, 고향, 주민등록번호, 직장명 등이 해당된다. 과거 단순히 신분정보로만 사용하던 개인정보는 디지털화로 인해 수집, 가공 및 활용이 용이해짐에 따라 오늘날 전자상거래, 고객관리, 금융거래 등에 활용됨으로써 개인정보를 이용하여 이윤을 추구하는 기업이 증가하고 있다. 개인정보는 정보서비스업체들에게는 중요한 정보인 동시에 자산이 되므로 부적절한 방법으로 취득하여 유용할 수 있으며 취급부주의나 고의로 개인정보를 유출하는 사례가 급증하고 있는 실정이다. 개인정보의 유출은 다양한 사이버 범죄를 야기하며 개인정보의 상품화로 프라이버시의 침해가 증가하므로 개인정보의 보안은 필수적이다.

일반적으로 개인정보는 크게 두 경로를 통하여 유출된다. 첫째는 정보 서비스업체 관리자가 부주의나 비양심적인 행동으로 개인정보를 유출할 수 있으며 둘째는 악의적인 해커가 네트워크의 취약점을 찾아서 유출할 수 있다. 전자는 법제도적 방법으로 해결하며 후자는 기술적인 방법으로 네트워크의 보안성을 높임으로써 해결해야 한다. 지금까지 개인정보의 보호에 관한 연구는 주로 법제도적 해결방안에 대한 연구를 수행해왔으며 기술적인 측면에서 개인정보의 보호에

※ 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업(IITA-2006-C1090-0603-0028) 및 한국전자통신연구원(ETRI)의 지원을 받아 수행되었음.
† 준 회 원 : 이화여자대학교 컴퓨터정보통신공학과 박사과정
†† 준 회 원 : 이화여자대학교 컴퓨터정보통신공학과 석사과정
††† 종신회원 : 이화여자대학교 컴퓨터정보통신공학과 교수
†††† 정 회 원 : 한국전자통신연구원 능동보안기술연구팀 팀장
논문접수 : 2007년 3월 6일, 심사완료 : 2007년 6월 12일

대한 연구는 미비하였다. 개인정보는 다양한 목적을 가지고 여러 방법으로 수집되며 일반적으로 합법적인 방법과 불법적인 방법으로 수집된다. 대부분의 기업, 공공기관, 웹 서비스 제공자들은 사용자에게 안전하고 질 좋은 서비스를 제공하기 위하여 합법적인 다양한 개인정보 수집 기술들을 이용하여 사용자의 정보를 수집한다. 그러나 이러한 사용자 정보 수집 기술들은 개인정보 침해도구로도 사용되며, 불법적인 개인정보 수집자들은 스파이나 내부자가 개인정보를 유출하거나 미리 만들어진 프로그램이나 공격 기술들을 이용하여 사용자의 개인정보를 불법적으로 수집하는 경우로 해킹 및 컴퓨터 바이러스를 이용한다. 즉, 해커가 트로이 목마나 백도어 같은 악성 프로그램을 이용하여 사용자 ID, 패스워드, 계좌번호 등의 정보를 빼내는 것이 가장 일반적인 수법이며 이런 유형의 피해는 계속 증가하고 있는 실정이다.

해킹, 웜에 의한 개인정보의 유출 방지에 관한 연구에 있어서 대규모의 네트워크(global network)에서 웜의 탐지 연구[1]는 있었지만 사내망에서 기술적인 방법에 의한 개인정보 유출의 방지에 대한 연구는 거의 없었다. 따라서 본 논문에서는 대규모뿐만 아니라 소규모의 망에서도 유용하며 네트워크 취약점을 이용하여 악성코드들이 침입함으로써 개인정보를 유출하고자 할 때 네트워크 트래픽의 자기유사성 및 통계적 특성 분석을 통해 공격 징후를 조기에 감지함으로써 개인정보의 유출을 방지하는 방법을 제안하고자 한다.

본 논문에서는 정상적인 상태에서 네트워크 트래픽을 조사 분석하여 트래픽의 특성을 살펴본 후, 개인정보를 유출하는 비정상 상태를 야기하는 악성코드들이 네트워크 트래픽에 어떤 영향을 끼치는지 트래픽의 특성 변화를 살핌으로써 해킹으로부터 개인정보를 보호할 수 있도록 한다. 특히 침입을 시도할 때 네트워크에 어떤 영향을 끼치는지 살펴보고자 네트워크 트래픽의 변화에서 나타나는 특성을 통하여 악성코드의 침입을 탐지하는 기법을 제안하고자 한다.

서론에 이어, 2장에서는 본 연구에 있어서 기반이 되는 개념인 자기유사성과 유사연구인 이상탐지에 대하여 간략하게 살펴본다. 3장에서는 실제 대학망의 트래픽을 측정 분석한 후, 본 연구에서 제안하고자 하는 개인정보의 유출을 탐지하는 기법을 설명한다. 4장에서는 개인정보의 유출을 야기하는 비정상 트래픽이 발생할 때의 트래픽을 모사하여 측정 분석함으로써 제안한 기법의 적합성을 확인하고 5장에서 결론을 맺는다.

## 2. 관련연구

### 2.1 자기유사성

#### 2.1.1 자기유사성

최근 연구에 따르면 네트워크의 트래픽이 포아송(poisson) 특성 대신에 자기유사성이 나타남을 주장하고 있으며, 네트워크 트래픽에서의 자기유사성에 관한 연구는 크게 LAN 트래픽이 자기유사성을 가짐을 보인 연구[2]와 인터넷 웹 트래픽이 자기유사성을 지님을 실제 고속망에서 측정한 WAN에서의 연구[3]인 두 가지로 분류된다.

자기유사성[4,5,6,7]이란 차원 상 서로 다른 확대 비율이나 서로 다른 스케일에서 보았을 때 동일하게 보이거나 행동하는 자기유사한 현상이다. 수학적으로 설명하면, 급수를 다수 다중화시켰을 때도 새로운 급수들은 원래 급수와 같은 자기상관함수를 가지는 특성을 말한다. 이것을 수식으로 나타내면 다음과 같다. 데이터의 급수를 $X=(X_t : t=0,1,2,...)$라고 하고, m개의 데이타를 묶은 데이터를 $X^{(m)}=(X_k^{(m)} : k=1,2,3,...)$라고 정의한다. m개로 그룹화 된 데이터의 급수는 원래 데이터의 급수 X를 m 만큼의 크기로 중복되지 않게 값들을 선택한 후, 선택된 값들의 평균값을 내서 구한다. 만약 X가 자기유사성을 가진다면 X의 자기상관함수 r(k)는 식 (1)과 같다.

$$r(k)=E[(X_t-\mu)(X_{(t+k)}-\mu)] \quad (1)$$

r(k)는 모든 m에 대한 급수 $X^{(m)}$에 대해서도 같은 자기상관함수를 가지며, 이와 같은 특성은 급수들이 분포적인 자기상관성(distributionally self-similar)을 가진다고 한다. 시간에 관계된 급수를 표현하기 위해 자기유사(self-similar) 모델을 사용하며 자기유사 모델은 급수의 자기유사 정도를 하나의 파라메터로 나타낸다. 이 파라메터는 급수의 자기상관함수의 감쇄율을 나타내며 이 파라메터는 허스트(hurst) 파라메터라고 한다. 허스트 파라메터는 식 (2)와 같이 정의된다.

$$H=1-\beta/2 \quad (2)$$

즉, 허스트 파라메터 H는 통계적인 현상의 지속성에 대한 척도이고 확률과정의 장기간 종속에 대한 척도이다. 허스트 파라메터 H의 범위는 0.5 < H < 1 이며, H가 1에 가까워질수록 자기유사성의 정도가 높은 것을 의미한다. 자기유사성을 분석하기 위해서 자기상관모델의 파라메터인 허스트 파라메터 H를 구한다.

#### 2.1.2 자기유사성 측정방법

자기유사성을 측정하는 다양한 방법은 크게 시간기반(time-based) 방법과 주파수기반(frequency-based) 방법의 두 가지로 분류된다[8]. 본 논문에서는 허스트 파라메터를 구하는 방법으로 시간기반방법 중의 하나인 누적분산 방법(aggregate variance)과 주파수기반 방법 중의 하나인 피리오도그램(periodogram) 방법[5]을 이용하여 자기유사성을 측정하였으므로 이 두 가지 방법에 대해 살펴본다.

가. 누적분산(aggregate variance) 방법

이 방법은 관찰하고자 하는 데이터의 자기유사성을 그래프로 보여주는 방법이다. m개의 데이터를 하나로 묶어서 그 묶음의 평균값의 확률과정을 $X^{(m)}=(X_k^{(m)} : k=1,2,3,...)$라고 하면, 이 때 $X_k^{(m)}=\frac{1}{m}\sum_{i=km-(m-a)}^{km} X_i$, $X^{(m)}$의 분산들은 다음 식 (3)과 같이 정의된다.

$$Var(X^{(m)}) = \frac{1}{n}\sum_{k=1}^{n}(X^{(m)} - \overline{X^{(m)}})^2 = E[X^{(m)^2}] - \overline{X^{(m)^2}} \quad (3)$$

만약 이 확률과정이 자기유사성을 가진다면, 위의 식에서 나온 분산들은 다음의 식 (4)를 만족한다.

$$Var(X^{(m)}) = \frac{Var(X)}{m^{\beta}} \quad (4)$$

이 때 $\beta$값과 허스트 파라메터는 $H=1-\beta/2$인 일차함수의 관계를 갖는다.

위의 식에서 양변에 log를 취하면 다음 식 (5)가 된다.

$$\log[Var(X^{(m)})] \approx \log(Var(X)) - \beta\log(m) \quad (5)$$

log[Var(X)]는 m에 독립적인 상수이기 때문에 log-log 그래프에서 $Var(X^{(m)})$과 m을 나타내면 $-\beta$의 기울기를 가지는 직선이 나온다. 이 기울기 값이 - 1과 0 사이를 나타내면 자기유사적 특성을 가지는 것이다. $\beta$값을 이용해서 허스트 파라메터를 구하며 $\beta$값이 0에 가까워질수록 자기유사성 정도가 높은 것을 의미한다.

나. 피리오도그램(periodogram) 방법

이산시간으로 얻은 k 개의 표본 $X_0$, $X_1$, ..., $X_{k-1}$ 이 이산 시계열일 때, 이 시계열의 이산 푸리에 변환(discrete fourier transform)은 다음 식 (6)과 같다.

$$\hat{X_k}(f) = \sum_{m=0}^{k-1} X_m e^{-j2\pi fm} \quad (6)$$

$x_k(f)$의 크기를 제곱한 것은 주파수 f에서의 에너지를 나타낸다. 이 에너지를 전체 시간 k로 나누면, 주파수 f에서 파워의 추정값을 얻을 수 있으며 이 값을 식으로 나타내면 식 (7)과 같다.

$$\hat{p_k} = \frac{1}{k}|\hat{x_k}(f)^2| \quad (7)$$

위의 식을 달리 표현하면 식 (8)과 같이 된다.

$$I_N(w) = \frac{1}{2\pi N}\left|\sum_{k=1}^{N} x_k e^{jkw}\right|^2 \quad (8)$$

이것을 피리오도그램(periodogram) 또는 강성(intensity) 함수라 한다. 스펙트럼 밀도 $S(w)$는 자기상관함수 $r(k)$의 푸리에 변환쌍이므로 자기상관함수를 이용하여 스펙트럼 밀도 $S(w)$를 식 (9)와 같이 계산한다.

$$S(w) \sim \frac{1}{|w|^{\gamma}}, \quad \text{as} \quad w \to 0, 0 < \gamma < 1 \quad (9)$$
$$\log[S(w)] \sim -\gamma\log[|w|]$$

이때 $\log[S(w)]$와 $w$의 log-log 그래프를 그리면 기울기가 $-\gamma$인 직선이 나오며 $\gamma$을 이용하여 식 (10)과 같이 허스트 파라메터를 구한다.

$$-\gamma = \beta - 1 = 1 - 2H \quad (10)$$

2.2 이상 탐지

침입 탐지 기법은 대표적으로 알려진 침입의 형태를 모델화하여 그 모델과 일치하는 행위를 침입으로 간주하는 오용 탐지(Misuse Detection)와 정상적인 행위를 모델링하여 이에 위반하는 행위를 침입으로 간주하는 이상 탐지(Anomaly Detection)가 있다. 오용 탐지는 모델링된 침입 행위만을 탐지하며 변형되거나 새로운 형태의 공격에 취약하다는 단점을 가지며 오용 탐지를 위한 공격 유형을 분석하고 오용 탐지 규칙 등의 인코딩 작업에 시간과 비용이 많이 소요되는 문제점을 갖고 있다.

이상 탐지에서 가장 중요한 요소는 정상적인 행위에 대한 모델링이므로 대부분의 연구가 이러한 정상적인 패턴을 모델링하기 위한 기법에 집중하여 이루어지고 있다. 이상 탐지 방법으로는 엔트로피 통계와 카이제곱 통계 방법을 사용한 통계적인 방법[9], 데이터 마이닝기법을 이용한 방법[10], 신경망 기반의 침입 탐지 시스템[11]이 제안되었다. 대부분의 이상 탐지 기법은 일반적으로 이상 탐지의 행위에 대한 모델링이 어렵고 사용하는 알고리즘으로 인해 실시간으로 탐지하기가 매우 어렵다. 왜냐하면 방대한 크기의 데이터를 처리해야 하므로 효율적인 자료구조와 인덱스 기법이 필요하고, 오늘날 네트워크의 동적 특성으로 인해 정상적인 요청 및 서비스를 유형화하기가 어렵기 때문이다. 일정 시간 간격동안 정상인 것처럼 보이는 것이 새로운 컨텐츠에서 비정상으로 분류되거나 또는 그 반대의 경우가 발생하게 된다. 따라서 지금까지 제안되었던 기법들이 실제 침입 탐지에 적합하지 않은 경우가 많다.

## 3. 개인정보유출 탐지기법

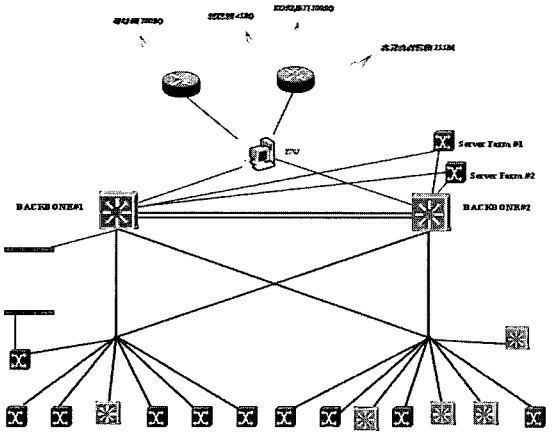
웜과 같은 악성 코드를 이용하여 개인정보의 유출을 시도할 때 대상을 찾기 위하여 매우 짧은 시간 안에 과도한 트래픽을 생성하게 된다. 이는 통계량 단위시간 당 평균 트래픽 양(volume)인 AVG(t, $t+t_0$)이나 MAX > $MAX_0$ 과 같은 최대 트래픽의 양(MAX)을 비교함으로써 구할 수 있다. 그러나 단기간에 트래픽의 통계값이 일시적으로 커질 수 있으므로 본 논문에서는 허스트 파라메터가 통계적 현상의 지속성에 대한 척도이고 확률과정의 장기간 의존성의 길이에 대한 척도이므로 네트워크 트래픽에 있어서 자기유사성의 변화로 탐지하고자 한다. 즉 자기유사성을 보이는 네트워크 환경에서 비정상적인 다수의 트래픽 증가는 트래픽의 자기유사성에 영향을 미칠 수 있다는 것에 중점을 두어 비정상적인 네트워크 트래픽에 의해 야기되는 자기유사성의 변화

를 분석한다. 이를 위해 네트워크 사용자의 정상활동을 모델링하고 이미 정의되어 있는 허스트 파라메터 및 통계량과 비교함으로써 비정상 트래픽을 탐지한다.

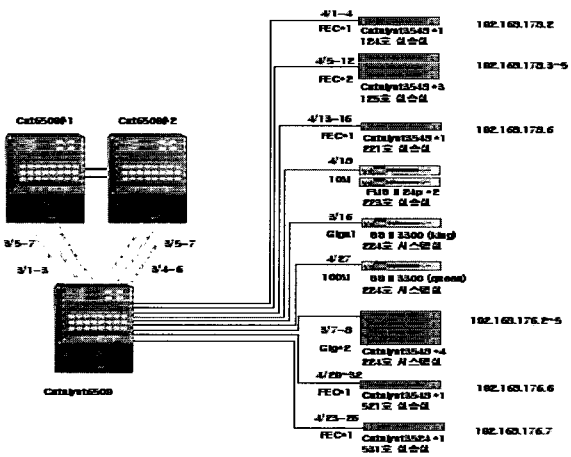
### 3.1 대학망에서 트래픽의 수집

본 연구에서는 개인정보의 유출을 야기하는 비정상적인 트래픽의 탐지에 앞서 네트워크에서 정상 트래픽을 모델링하고 특성을 살펴보기 위하여 네트워크 분석기 Anypa LAN-Sr[12]을 사용하여 본교 아산공학관 전체와 실습실에서 네트워크 트래픽을 측정하였다. 전체 네트워크의 구성도는 (그림 1)과 같으며 (그림 2)는 아산공학관의 네트워크 구성도를 나타낸다.

네트워크 트래픽은 네트워크 분석기를 이용하여 실제 데이터를 수집하였다. 호스트별 발생 패킷량, 송신지와 수신지 주소별 패킷량 등과 같은 자세한 정보를 수집하여 실험에 활용하는 것이 가능하지만 본 연구의 경우, 개인정보를 보호하기 위해 자세한 개인정보를 이용하거나 본의 아니게 개인정보를 엿보는 모순적 상황에 직면하기 때문에 네트워크



(그림 1) 전체 네트워크 구성도



(그림 2) 아산공학관 네트워크 구성도

의 종합 통계량만을 실험에 활용하였다. 대학망 실험실의 네트워크 트래픽 특성을 고려하여 오전 9시부터 오후 8시까지 각 프로토콜별 누적 패킷량과 전체 패킷량을 조사한 후 실험 데이터 세트를 구성하였다. 네트워크 규모에 따른 트래픽의 특성 변화를 고찰하기 위하여 실험 대상을 중간규모 네트워크로써 공학관 전체와 소규모 네트워크로 공학관의 컴퓨터 실습실로 한정하는 두 가지 경우에 대하여 실험하였다. 대규모의 글로벌 네트워크와 달리 중소규모 네트워크에서는 발생 패킷량이 많지 않으므로 단위시간(time unit)을 일 분과 한 시간으로 정하여 누적 패킷량을 측정해서 실험 데이터 세트를 구성하였다.

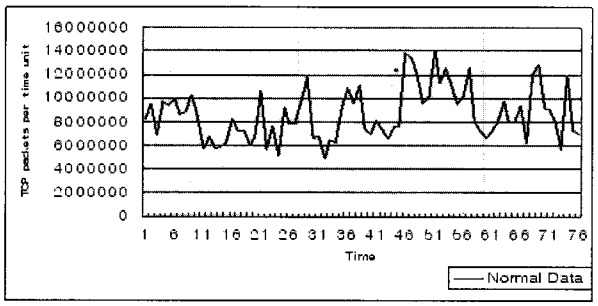
### 3.2 트래픽 측정 결과 및 특성 분석

본 연구는 윈도우 사이즈를 64로 하여 단위시간 당 새로운 데이터가 한 개씩 유입된다는 가정으로 실험하였다. 트래픽 수집결과 TCP 패킷이 발생 패킷량의 대부분을 차지하여 실험에는 단위시간당 발생된 전체 패킷량과 TCP 패킷량을 이용하였다. 64개의 데이터를 기본으로 하여 새로운 데이터가 추가될 때마다 자기유사성을 나타내는지 허스트 파라메터를 알아보고 데이터 세트에 대한 최소값(MIN), 최대값(MAX), 분산값(VAR), 평균값(AVG)과 같은 통계값의 변화를 살펴보았다.

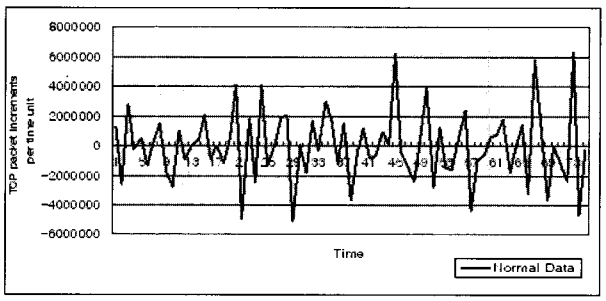
#### 3.2.1 중간규모 네트워크의 트래픽 측정 및 특성 분석

가. 통계값 분석

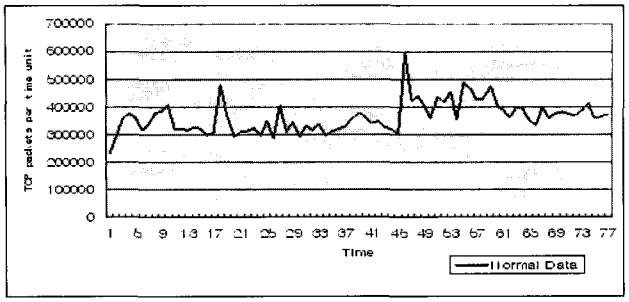
중간규모 네트워크의 측정 대상인 공학관에서 수집한 트래픽을 측정 단위시간별로 나타내면 다음과 같다.



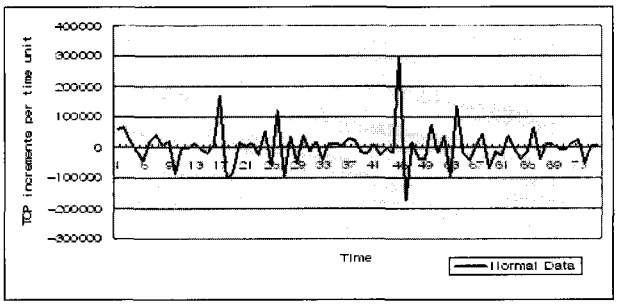

(그림 3) 공학관 - 단위시간당(1시간) TCP 패킷량




(그림 4) 공학관 - 단위시간당(1시간) TCP 패킷증가량




(그림 5) 공학관 - 단위시간당(1분) TCP 패킷량




(그림 6) 공학관 - 단위시간당(1분) TCP 패킷증가량

〈표 1〉 공학관 - 트래픽 데이터의 통계값(단위시간:1시간)

| | MAX | MIN | AVG | VAR |
|---|---|---|---|---|
| TCP 패킷량 | 14110683 | 4888670 | 8648785.0 | 7.37e+11 |
| TCP 패킷증가량 | 6286319 | -5074476 | -15895.4 | 5.83e+11 |

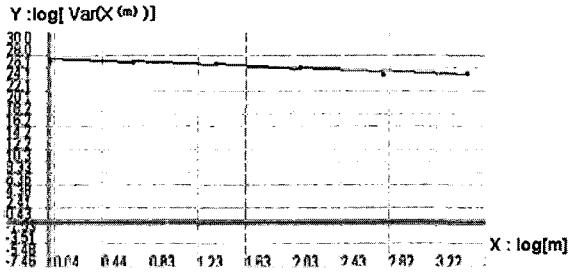
〈표 2〉 공학관 - 트래픽 데이터의 통계값(단위시간:1분)

| | MAX | MIN | AVG | VAR |
|---|---|---|---|---|
| TCP 패킷량 | 600894 | 232592 | 363756.4 | 7.37e+11 |
| TCP 패킷증가량 | 301381 | -177475 | 1823.8 | 5.83e+11 |

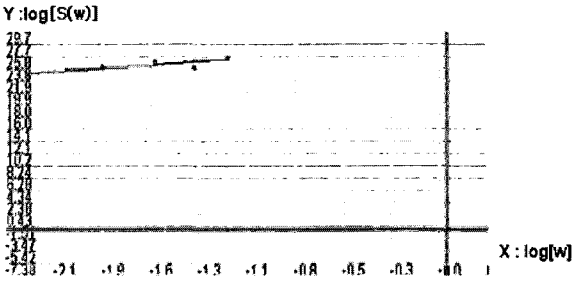
트래픽을 수집한 단위시간별로 네트워크 트래픽의 통계값을 표시하면 단위시간이 한 시간인 경우는 <표 1>과 같고 단위시간이 일 분인 경우는 <표 2>와 같다.

나. 트래픽과 자기유사성의 관계

측정된 네트워크 트래픽에 대하여 자기유사성을 측정하는 방법별로 자기유사성을 살펴보았다. 정상적인 네트워크 상태에서 수집된 트래픽에 자기유사성이 존재하는지 확인하기 위하여 누적분산방법과 피리오도그램 방법을 이용하여 자기유사성 척도인 허스트 파라메터를 구하였다. 2장에서 설명한 바와 같이 누적분산 방법에서 log[Var(X)]는 m에 독립적인 상수이므로 log-log 그래프에서 $Var(X^{(m)})$과 m을 나타내면 $-\beta$ 의 기울기를 가지는 직선이 나온다. 이 기울기 값이 -1과 0사이를 나타내면 자기유사성을 가짐을 의미한다. $\beta$값을 이용해서 허스트 파라메터 값을 구할 수 있으며 $\beta$값이 0에




(그림 7) 누적분산 방법의 결과




(그림 8) 피리오도그램 방법의 결과

가까워질수록 자기유사성의 정도가 높은 것을 의미한다. (그림 7)은 첫 번째 TCP 데이터 세트에 대하여 누적분산 방법을 이용하여 허스트 파라메터를 계산한 것이다. 피리오도그램 방법의 정의에 따라 계산된 log[S(w)]와 w를 log-log 그래프로 그리면 기울기가 $-\gamma$인 직선이 나오며 이를 통해 허스트 파라메터를 구할 수 있다. (그림 8)은 허스트 파라메터를 구하기 위하여 첫 번째 TCP 데이터 세트에 대하여 피리오도그램 방법을 이용하여 log-log 그래프를 나타낸 것이다.

이러한 방법을 통해 시간에 대해 연속적인 데이터로 이루어진 트래픽 데이터 세트에 대해 데이터 세트를 구성하는 단위시간별로 허스트 파라메터를 구하여 표로 나타내면 <표 3>과 <표 4>와 같다. 트래픽 측정 간격에 있어서 단위시간이 한 시간인 허스트 파라메터값에 비해 단위시간이 일 분인 경우, 자기유사성을 판단하는 허스트 파라메터값이 0.9에 근접하거나 0.9보다 큰 값을 보이므로 단위시간을 작게 하였을 때 자기유사성의 정도가 강하게 표현되며, <표 3>과 <표 4>에서 보는 바와 같이 $0.5<H<1$을 만족하며 네트워크 트래픽은 자기유사성을 지님을 알 수 있다.

〈표 3〉 허스트 파라메터 (단위시간: 1시간)

| | TCP/ 누적분산 방법 | TCP/ 피리오도그램 방법 |
|---|---|---|
| Dataset[1] | 0.776 | 0.753 |
| Dataset[2] | 0.784 | 0.765 |
| Dataset[3] | 0.825 | 0.766 |
| Dataset[4] | 0.795 | 0.784 |
| Dataset[5] | 0.799 | 0.734 |

〈표 4〉 허스트 파라메터 (단위시간: 1분)

| | TCP/ 누적분산 방법 | TCP/ 피리오도그램 방법 |
|---|---|---|
| Dataset[1] | 0.859 | 0.820 |
| Dataset[2] | 0.838 | 0.916 |
| Dataset[3] | 0.869 | 0.945 |
| Dataset[4] | 0.878 | 0.936 |
| Dataset[5] | 0.898 | 0.958 |

3.2.2 소규모 네트워크에서 트래픽 측정 및 특성 분석

가. 통계값 분석

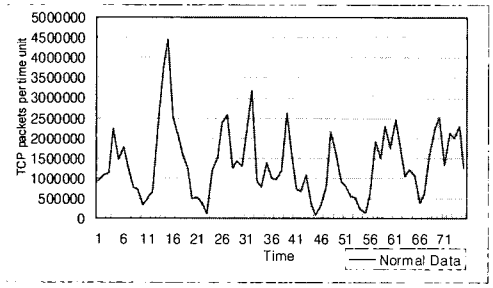
소규모 네트워크인 실습실에서 트래픽 수집은 발생 트래픽의 양이 많지 않아서 중간규모 네트워크의 트래픽 수집과는 달리 단위시간을 한 시간으로 하여 측정하였다. 자기유사성의 특성이 서로 다른 확대 비율이나 서로 다른 스케일에서 보았을 때 동일하게 보이거나 행동하는 현상이고 앞서 중간규모 네트워크 결과에 따르면 단위시간이 한 시간일 때에 비해 일 분 간격일 때 자기유사성이 더 강하게 나타났으므로 단위시간을 길게 하여 자기유사성을 보이면 더 작은 단위시간에 대해서 자기유사성을 더 강하게 보일 것으로 예측할 수 있다. 실습실에서 수집된 트래픽에 대하여 단위시간당 TCP 패킷량, 단위시간당 TCP 패킷의 증가량, 단위시간당 전체 패킷량, 단위시간당 전체 패킷의 증가량을 그림으로 표시하면 각각 (그림9)~(그림12)와 같으며, 네트워크의 트래픽 데이터에 대한 통계값은 <표 5>와 같다.

〈표 5〉 트래픽 데이터의 통계값

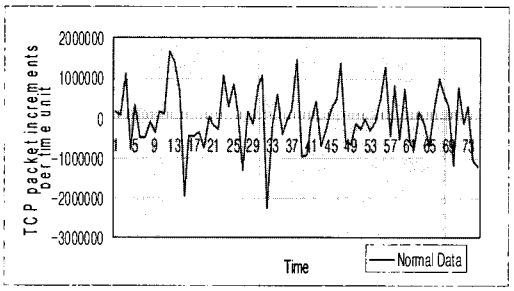
| | MAX | MIN | AVG | VAR |
|---|---|---|---|---|
| TCP 패킷량 | 4451621 | 81916 | 1376833.4 | 7.36e+11 |
| TCP 패킷 증가량 | 1670725 | -2232708 | -12486.5 | 5.83e+11 |
| 전체 패킷량 | 4537137 | 101994 | 1419069.0 | 7.54e+11 |
| 전체 패킷 증가량 | 1684178 | -2241191 | -12788.6 | 5.98e+11 |

나. 트래픽과 자기유사성과의 관계

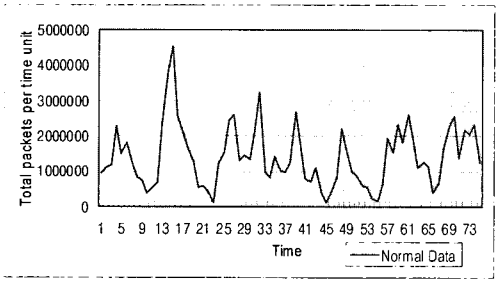
(그림 13)은 측정한 데이터세트에 대하여 누적분산 기법에 의해 log-log 그래프를 보인 것이며 (그림 14)는 피리오도그램 기법으로 log-log 그래프를 보인 것이다.



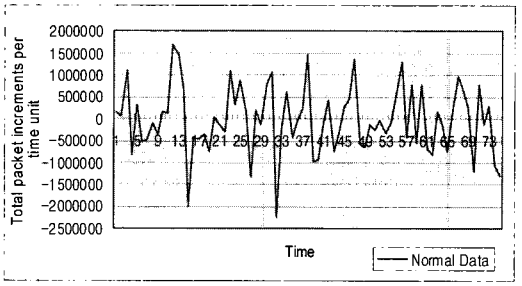

(그림 9) 실습실-TCP 패킷량



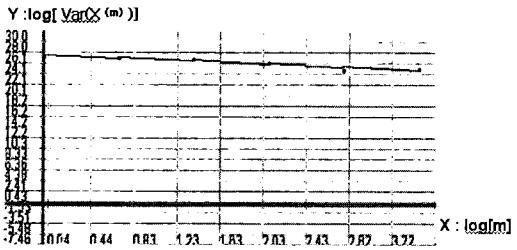

(그림 10) 실습실-TCP패킷의 증가량



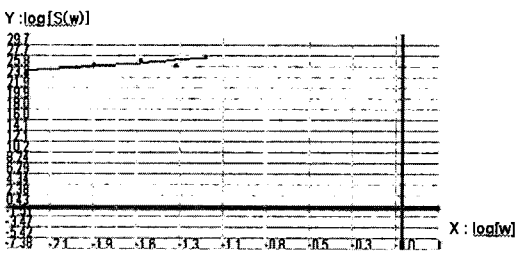

(그림 11) 실습실-전체 패킷량




(그림 12) 실습실-전체 패킷의 증가량




(그림 13) 누적분산 방법에 따른 허스트 파라메터




(그림 14) 피리오도그램 방법에 따른 허스트 파라메터

〈표 6〉 허스트 파라메터

| | TCP / 누적분산 방법 | TCP/ 피리오도그램 방법 | 전체 / 누적분산 방법 | 전체 / 피리오도그램 방법 |
|---|---|---|---|---|
| DataSet[1] | 0.612 | 0.683 | 0.611 | 0.681 |
| DataSet[2] | 0.619 | 0.660 | 0.618 | 0.654 |
| DataSet[3] | 0.628 | 0.640 | 0.627 | 0.634 |
| DataSet[4] | 0.641 | 0.688 | 0.640 | 0.682 |

지금까지 수집 분석한 여러 개의 트래픽 데이터에 대한 허스트 파라메터값을 나타내면 <표 6>과 같다. <표 6>에 나타난 결과에 따라 본 연구에서 수집조사된 트래픽 데이터들이 네트워크 규모가 작은 경우에도 자기유사성을 가지고 있음을 확인할 수 있다. 이 값은 공학관 전체를 대상으로 했을 때보다 자기유사성을 나타내는 척도인 허스트 파라메터값이 다소 낮지만 여전히 [0.5<H<1]의 관계를 만족한다. 네트워크의 규모가 작아짐에 따라 수집된 트래픽 데이터의 양이 상대적으로 적어서 작은 트래픽양의 변화로도 큰 영향을 미치므로 중간규모 네트워크의 대상 환경인 공학관에서 측정한 값에 비해 허스트 파라메터가 다소 낮게 나온 것으로 해석된다. 이 결과로써 트래픽 데이터 세트가 달라지더라도 시간에 대하여 연속적인 데이터이고, 동일한 네트워크 환경에서 측정된 데이터라면 허스트 파라메터에 있어서 다소간의 변화가 있을지라도 [0.5<H<1]의 관계를 만족하는 자기유사성이 있음을 볼 수 있다.

## 4. 비정상적 트래픽과 자기유사성의 관계

### 4.1 비정상적인 트래픽의 생성 및 분석

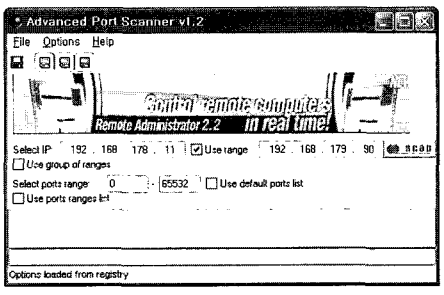
본 연구에서는 제안하는 기법의 적합성을 실험하기 위하여 예로써 개인정보가 악성코드에 의해 유출되는 경우, 악성코드의 탐지에 대해서 알아본다. 악성 코드는 여러 호스트를 감염시키기 위해 취약한 시스템을 탐색하는 과정을 거치며 그 기간 동안 비정상적인 트래픽이 발생된다. 비정상적인 트래픽이 발생할 때 이를 자기유사적 특성과 트래픽 함수의 통계량 변화를 이용하여 의미 있는 시간 내에 탐지함을 보이기 위해 다음과 같은 환경에서 실험을 수행하였다.

개인정보유출 기술들은 공격성을 가진 웜이나 바이러스, 악성 봇, 트로이목마 등의 악성코드 및 프로그램들로 구현된다. 개인정보유출기술 중에서 가장 많은 부분을 차지하는 웜은 바이러스와 비슷하나 전파와 실행이 누군가의 도움없이 독립적으로 수행될 수 있는 코드이다. 웜은 일반적으로 다음과 같은 네 단계를 거쳐 수행된다[13]. 첫 번째 단계는 공격할 대상을 선정하는 단계(target selection)이며, 두 번째는 특정한 취약점을 이용하여 공격 대상을 약탈하는 단계(exploitation), 세 번째는 자신의 공격 메커니즘을 가지고 실제로 공격하는 단계(attack)이다. 공격당한 호스트에서 자신의 코드가 실행된 후에는 네 번째 단계인 감염된 호스트가 공격 호스트가 되어 자신의 공격 코드를 복사하여 다시 전파시키는 작업(code propagation)을 실행한다. 네트워크에서 웜에 의한 트래픽의 변화를 살펴봄에 있어서 앞서 기술한 네 단계 중 두 번째와 세 번째 단계의 동작은 하나의 호스트에서 일어나며 네 번째 단계는 이미 웜에 감염되어 전파하는 단계에 해당되므로 본 연구에서는 고려하지 않았다. 즉, 악성코드가 네트워크 상에서 공격 대상을 찾을 때 실제적으로 트래픽에 영향을 주므로 첫 단계인 공격 대상을 찾는 방법만을 고려하였다.

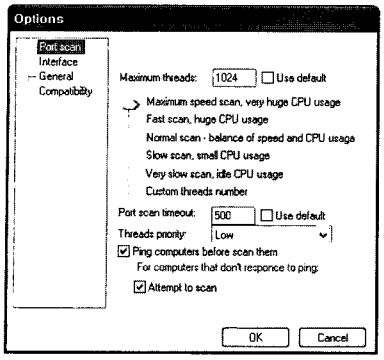
공격 대상을 찾는 전략은 여러 가지가 있다. 스캐닝을 통하여 공격 대상을 무작위로 선정하는 방법과 메타서버(metaserver)의 공격취약성을 갖는 호스트 목록, 미리 생성되어 있는 hit 목록, 또는 이론적으로 계산된 대상 목록을 사용하는 내/외부 대상 목록, 직접 접속하여 전파하는 방법 등이 있다. 목록을 사용하는 방법은 빠르지만 목록을 작성하는 작업이 먼저 이루어져야 하며 이는 네트워크의 변화에 영향을 미치지 않으므로 본 연구에서는 가장 일반적으로 사용하는 스캐닝을 통한 공격 대상 선정을 기반으로 연구를 진행하였다.

웜을 전파하기 위하여 포트 스캔이 웜 코드 내부에 구현되어야 하나, 본 연구에서는 본교의 네트워크를 대상으로 실험하였으므로 실제적인 공격 단계를 원하지 않으며 웜의 네 단계 동작 중 첫 번째 단계만 독립적으로 필요했기 때문에 상용 포트 스캔 도구를 사용하였다. 포트 스캐닝은 침입하고자 하는 시스템의 열려있는 포트를 알아보기 위한 기법으로 시스템에 직접적인 피해를 주지는 않지만 침입하고자 하는 시스템의 취약점 정보를 수집하는 첫 단계이다. 따라서 포트 스캐닝은 해킹의 징조라고 할 수 있다. 누군가 자신의 시스템에 침입하고자 한다는 것을 미리 안다면 해킹으로 인한 개인정보의 유출을 막는데 유용할 것이다. 스캔하고자 하는 IP를 학교 내의 주소로 제한하고 포트 스캔하면서 네트워크의 변화를 살펴보았다. (그림 15)은 포트 스캐닝을 위해 사용한 도구를 예시한 것이다. 또한 (그림 16)는 본 연구에서 악성코드가 출현하는 상태를 실험하기 위하여 포트 스캐닝을 할 때 사용한 옵션을 보인 것이다.

본 연구에서는 네트워크에서 비정상 트래픽을 모사하기 위해 앞서 기술한 바와 같이 포트 스캐너를 이용하였다. 윈도우 사이즈를 64 데이터 세트로 가정하였으므로 정상 데이터 64개로 이루어진 데이터 세트에서 시작하여 윈도우 내에 비정상 데이터 세트가 하나씩 추가될 때마다 허스트 파라메터와 트래픽 함수의 통계가 어떤 변화를 보이는지 살펴보았다.




(그림 15) Advanced Port Scanner




(그림 16) 포트 스캔시 사용옵션

비정상적인 상황에서 수집 구성된 데이터 세트는 (그림 17) 비정상상태에서 단위시간당 TCP 패킷량, (그림 18) 비정상상태에서 단위시간당 TCP 패킷의 증가량, (그림 19) 비정상상태에서 단위시간당 전체 패킷량, (그림 20) 비정상상태에서 단위시간당 전체 패킷의 증가량으로 나타난다. 또한 비정상상태에서 트래픽 데이터에 대한 통계값을 표로 나타내면 <표 7>과 같다.

### 4.2 자기 유사성 측정 방법별 탐지 결과 및 분석

포트 스캐닝을 통해서 만들어진 비정상적인 트래픽이 유입될 때 자기유사성의 변화는 각각 누적분산 기법과 피리오도그램 기법을 이용하여 허스트 파라메터값의 변화를 살펴보았다. <표 8>는 포트 스캐닝에 의한 비정상적인 트래픽이 유입되는 경우 변화하는 허스트 파라메터값을 나타낸다.

위의 표에 나타난 결과에서 볼 수 있는 바와 같이 피리오도그램 방법은 단위시간 2 이내에 비정상적인 트래픽임을 감지할 수 있으며 누적분산기법은 단위시간 4 이내에 비정상적인 트래픽임을 감지할 수 있다.

〈표 7〉 비정상상태에서 트래픽 데이터의 통계값

| | MAX | MIN | AVG | VAR |
|---|---|---|---|---|
| TCP 패킷량 | 4451621 | 81916 | 1469304.0 | 8.47e+11 |
| TCP 패킷 증가량 | 2024748 | -2232708 | -12486.5 | 6.21e+11 |
| 전체 패킷량 | 4537137 | 101994 | 1511054.0 | 8.67e+11 |
| 전체 패킷 증가량 | 2042510 | -2241191 | -12788.6 | 6.38e+11 |

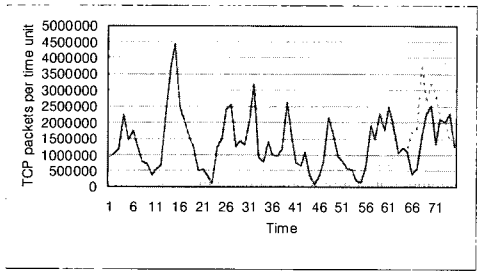
〈표 8〉 비정상 트래픽이 포함될 때 허스트 파라메터의 변화

| | TCP / 누적분산 | TCP/ 피리오도그램 | 전체 / 누적분산 | 전체 / 피리오도그램 |
|---|---|---|---|---|
| DataSet[1] | 0.612 | 0.683 | 0.611 | 0.681 |
| DataSet[2] | 0.617 | 0.644 | 0.616 | 0.641 |
| DataSet[3] | 0.601 | *0.479* | 0.600 | *0.475* |
| DataSet[4] | 0.594 | *0.442* | 0.593 | *0.438* |
| DataSet[5] | *0.473* | *0.409* | *0.472* | *0.404* |

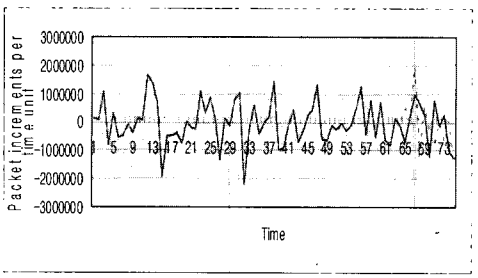
〈표 9〉 비정상 트래픽의 경우 TCP 트래픽 증가량 함수의 통계값 변화

| | MAX | MIN | AVG | VAR |
|---|---|---|---|---|
| DataSet[1] | 1670725 | -2232708 | -14632.6 | 6.01e+11 |
| DataSet[2] | 1670725 | -2232708 | -14407.4 | 5.91e+11 |
| DataSet[3] | 1670725 | -2232708 | -14189.2 | 6.10e+11 |
| DataSet[4] | 1670725 | -2232708 | -16327.9 | 6.13e+11 |
| DataSet[5] | *2024748* | -2232708 | -17397.1 | 8.51e+11 |

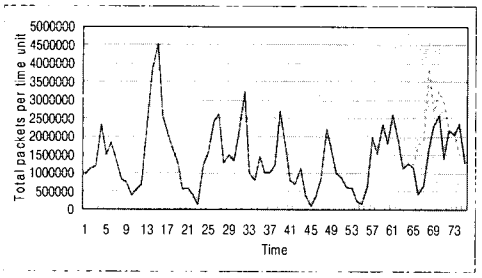
비정상적인 트래픽이 유입되는 경우 트래픽 함수에 어떤 영향이 있는지 살펴보기 위하여 허스트 파라메터와 트래픽 함수에 있어서 트래픽 증가량 함수의 통계값 변화를 <표 9>와 <표 10>에 나타내었다.



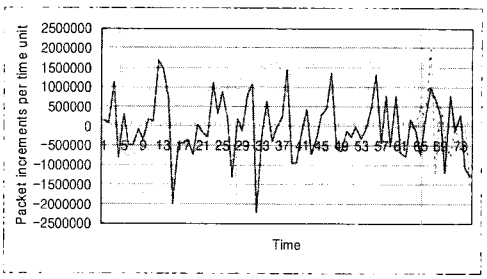

(그림 17) 비정상상태에서 단위시간당 TCP 패킷량




(그림 18) 비정상상태에서 단위시간당 TCP 패킷의 증가량




(그림 19) 비정상상태에서 단위시간당 전체 패킷량




(그림 20) 비정상상태에서 단위시간당 전체 패킷의 증가량

〈표 10〉 비정상 트래픽의 경우 전체 트래픽 증가량 함수의 통계값 변화

| | MAX | MIN | AVG | VAR |
|---|---|---|---|---|
| DataSet[1] | 1684178 | -2241191 | -14986.6 | 6.18e+11 |
| DataSet[2] | 1684178 | -2241191 | -17363.7 | 6.16e+11 |
| DataSet[3] | 1684178 | -2241191 | -18519.0 | 6.47e+11 |
| DataSet[4] | 1684178 | -2241191 | -36031.6 | 6.29e+11 |
| DataSet[5] | *2042510* | -2241191 | 37758.63 | 6.42e+11 |

<표 9>와 <표 10>에 나타난 결과에 따르면 윈도우가 움직일 때 평균과 분산이 변하는 것은 쉽게 관찰할 수 있으나 최대값과 최소값은 윈도우가 움직여도 변하지 않을 수 있음을 볼 수 있다. 그러나 비정상 트래픽이 발생되는 원리에 따르면 단위시간당 대량의 트래픽이 발생되므로 최대값의 변화에 따라 네트워크의 이상 증후를 예측할 수 있다. 결과에서 최대값은 단위시간 4에서 변화를 보이고 있으며 이는 앞서 자기유사성 특성에서 살펴보았던 누적분산기법에 의한 결과와 같은 시점이다. 트래픽의 증가량 함수만으로는 정확한 탐지가 어렵지만 자기유사함수와 함께 적용함으로써 피리오도그램 방법의 자기유사함수에서 단위시간 2 이내에 네트워크에 이상이 있음을 알려주는 알람을 발생시키고 누적분산 방법과 트래픽 증가량 함수에서도 이상이 감지되면 네트워크에 악성 트래픽이 발생되고 있다고 판단하는 충분한 근거가 될 것이다. 즉, 피리오도그램 방법, 누적분산 방법, 트래픽 증가량 함수, 이 세 가지 중 특정 하나의 방법에 의존하여 먼저 알람을 내기보다는 세 가지 방법을 모두 고려하여 알람을 발생시킴으로써 보다 정확하게 이상 현상을 탐지할 수 있다. 따라서 자기유사성을 지니고 있는 네트워크에서 단위시간당 트래픽 데이터를 누적하여 자기유사함수를 구하고 최대 4 단위시간이 경과한 후 이상 증후를 탐지할 수 있으므로 시간 효율성에 있어서도 의미 있는 결과라고 볼 수 있다.

## 5. 결론 및 향후 연구

본 논문에서는 개인 정보의 유출을 방지하기 위하여 네트워크의 트래픽을 수집 조사하여 트래픽의 특성을 이용한 탐지 기법을 제안하였다. 개인 정보를 유출시키는 대표적인 방법인 악성코드에 중점을 두어, 악성코드가 시스템에 침입하여 개인정보의 유출을 시도할 때 시스템의 취약점 정보를 수집하고 공격 대상을 탐색하는데 이러한 첫 단계에 가장 보편적으로 사용되는 기법이 포트 스캐닝이며 포트 스캐닝을 기반으로 하는 다양한 공격이 있으므로 본 연구에서는 개인 정보 유출의 시도를 나타내기 위한 모의실험 트래픽으로써 포트 스캐닝을 발생시켜 실험하였다.

정상 상태에서 수집된 네트워크 트래픽의 정보를 조사·분석한 결과 네트워크 트래픽이 자기유사성을 지님을 알 수 있었으며, 이 특성을 기반으로 비정상 상태의 네트워크 트래픽에 있어서 트래픽의 특성 변화를 살펴봄으로써 악성코드의 활동을 조기 진단하고 새로운 종류의 악성 코드 확산에도 대처할 수 있음을 알 수 있었다. 본 연구에서는 잘 알려진 포트 스캐닝 공격에 대하여 실제 대학망을 대상으로 실시하고 그 결과를 제시함으로써 제안된 기법의 타당성을 검증하였다. 제안한 방법은 자기유사성과 트래픽의 통계함수를 결합한 새로운 접근 방법이며, 새로운 형태로 출현하는 웜에 대응할 수 있으리라 기대된다. 또한 중소규모에서 탐지 가능성을 보임으로써 기존에 대규모의 네트워크에서 탐지기법과 달리 규모가 작은 네트워크 환경에서 내부자가 개인정보를 유출하고자 할 때 유용하다.

본 연구의 결과는 개인정보 보호 및 유출 방지를 위한 연구가 현재 법적 제도적인 측면에서는 활성화되고 있지만 기술적인 측면에서 시도가 부족하다는 점에서 초기 연구로 가치가 있으며, 개인정보를 암호화나 키 관리와 같은 단위 요소 기술의 측면에서 보호하는 것이 아니라 전체 시스템의 측면에서 트래픽의 변화를 고려하여 의미 있는 시간 내에 이상 탐지를 할 수 있으므로 개인정보의 유출을 미연에 방지하는데 도움을 줄 수 있을 것이다.

향후 연구 과제로는 본 연구에서 사용한 대학망의 트래픽은 계절별, 학기별, 실험기간 등과 같은 매우 다양한 요소에 의해서 변화하므로 장기간 관찰 수집한 트래픽을 토대로 다양한 변화요소를 반영하여 확장할 필요가 있으며 비정상상태에 대한 실험 또한 다양한 공격으로 확장 적용하여 실험할 필요가 있다. 또한 네트워크 트래픽 특성을 이용한 본 방법과 기존에 제안된 다른 기법들의 비교 연구를 통해 각 성능 척도에 따른 향상된 탐지 방법을 모색할 필요가 있다.

## 참 고 문 헌

[1] Y. Xin, B.-X. Fang, X.-C. Yun, and H.-Y. Chen, "Worm Detection in Large Scale Network by Traffic," Proc. of the 6th Intl. Conf. on Parallel and Distributed Computing, Applications and Technologies (PDCAT'05), pp.270-273, 2005

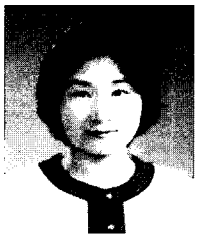
[2] W. Leland, W. Willinger, M. Taqqu and D. Wilson, "On the Self-similarity nature of Ethernet Traffic (Extended Version)," IEEE/ACM Transactions on Networking, Vol. 2(1), pp. 1-15, 1994

[3] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web Traffic: Evidence and Possible Causes," IEEE/ACM Trans. on Networking, Vol. 5(6), pp.835-846, 1997

[4] A. Popescu, "Traffic Self-similarity," Proc. of IEEE Intl. Conf. on Telecommunications, 2001

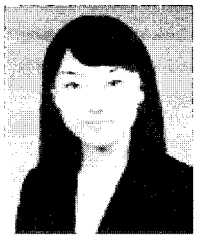
[5] R. Pacheco, J. Cesar and T. R. Deni, "Performance Analysis of Time-domain Algorithms for Self-similar Traffic," Proc. of IEEE Intl' Conf. on Electronics, Communications and Computers, 2006

[6] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," IEEE-ACM Transactions on Networking, Vol. 3(3), pp.226-244, 1995
[7] V. Paxson, "Fast Approximation of Self-similarity Traffic," Technical Report LBL-36750, Lawrence Berkeley Laboratory, 1995
[8] R. Kalden and S. Ibrahim, "Searching for Self-similarity in GPRS," LNCS Vol. 3015, pp. 83-92, 2004
[9] Laura Feinstein, Dan Schnackenberg Ravindra Balupari, Darrell Kindred, "Statistical Approaches to DDoS Attack Detection and Response," Proc. of The DARPA Information Survivability Conference and Exposition, 2003
[10] Wenke Lee, Salvatore J. Stolfo, "Data Mining Approaches for Intrusion Detection," Proc. of the 7th USENIX Security Symposium, pp.79-94, Jan. 1998
[11] Susan C. Lee, David V. Heinbuch, "Training a Neural-Network Based Intrusion Detector to Recognize Novel Attacks," IEEE Trans. on Systems, Man and Cybernetics, Vol. 31, No. 4, pp.294-299, 2001
[12] C & C Instruments Co. Ltd., http://www.cncinst.com/
[13] The CERIAS Intrusion Detection Research Group, "Digging for Worms, Fishing for Answers," Proc. of 18th Annual Computer Security Applications Conference, 2002



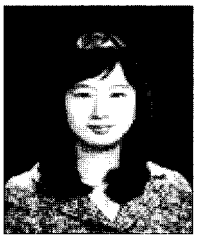
**박 정 민**

e-mail : pjm@kist.re.kr
1989년 이화여자대학교 전자계산학과 (이학사)
1991년 이화여자대학교 대학원 전자계산학과(이학석사)
1999년~현 재 이화여자대학교 컴퓨터정보통신공학과 박사과정
1991년~현 재 한국과학기술연구원 연구원
관심분야 : 네트워크 보안, 이동 IP, 센서 네트워크



**김 은 경**

e-mail : merilin@ewhain.net
2006년 이화여자대학교 컴퓨터학과 졸업(학사)
2006년~현재 이화여자대학교 컴퓨터정보통신공학과 석사과정
관심분야 : 네트워크 보안, 센서네트워크 보안, 유비쿼터스 네트워크 보안



**정 유 경**

e-mail : ykjung83@ewhain.net
2006년 이화여자대학교 컴퓨터학과 졸업(학사)
2006년~현재 이화여자대학교 컴퓨터정보통신공학과 석사과정
관심분야 : 네트워크 보안, 센서네트워크, 유비쿼터스 컴퓨팅

**채 기 준**

e-mail : kjchae@ewha.ac.kr
1982년 2월 연세대학교 수학과 학사
1984년 2월 미국 Syracuse University 컴퓨터학과 석사
1990년 2월 미국 North Carolina State University 컴퓨터공학과 박사
1990년 9월~1992년 2월 : 미국 해군사관학교 컴퓨터학과 조교수
1992년 3월~현 재 이화여자대학교 컴퓨터학과 교수
관심분야 : 네트워크 보안, 인터넷/무선통신망/고속통신망 프로토콜 설계 및 성능분석, 센서네트워크, 홈 네트워크, 유비쿼터스 컴퓨팅

**나 중 찬**

e-mail : njc@etri.re.kr
1986년 2월 충남대학교 계산통계학과 이학사
1989년 2월 숭실대학교 전자계산학과 공학석사
2004년 3월 충남대학교 컴퓨터과학과 이학박사
1989년~현 재 ETRI 능동보안기술연구팀 팀장
관심분야 : 실시간시스템, 네트워크 관리, 네트워크보안