논문 19-44-05-17 The Journal of Korean Institute of Communications and Information Sciences '19-05 Vol.44 No.05
https://doi.org/10.7840/kics.2019.44.5.914

플로우 연관성 기반의 연속적 그룹핑을 통한 악성 트래픽 탐지 방법

박 지 태◆, 백 의 준*, 이 민 섭*, 신 무 곤*, 김 명 섭°

Detection of Attack Traffic Using the Sequential Grouping Based on Flow Correlation

Jee-Tae Park◆, Ui-Jun Baek*, Min-Seob Lee*, Mu-Gon Shin*, Myung-Sup Kim°

요 약

오늘 날 비약적으로 증가하는 네트워크 환경에 따라서 악성 트래픽의 공격도 점점 정교해지고 복잡해지고 있다. 이러한 공격에 대한 피해를 줄이고 예방하기 위해서는 악성 트래픽에 대한 정확한 분석이 필요하다. 네트워크 트래픽 분석 방법 중 가장 널리 알려진 방법으로 시그니처 기반 분석 방법과 기계 학습 기반 분석 방법이 있다. 이 두 가지 방법은 모두 높은 정확성과 탐지율의 장점이 있지만 과정이 복잡하고 요구 조건을 충족 시켰을 때만 가능하다는 단점이 있다. 그래서 최근에는 플로우에 대한 통계적, 헤더 정보를 바탕으로 연관성을 계산하고, 연관성 값을 바탕으로 탐지를 하는 방법이 연구되고 있다. 여기서 통계적 정보로 패킷 크기 등이 있으며, 헤더 정보로는 플로우의 출발지, 도착지 IP 주소와 포트번호, 프로토콜로 구성 된 플로우의 5-tuples 정보를 사용한다. 하지만 기존의 두 가지 정보 모두 사용하는 방법에서 플로우의 통계적 정보를 구할 때 많은 시간과 비용이 들기 때문에 실제 환경에서 적용하기 어렵다는 단점이 있다. 따라서 본 논문에서는 플로우의 헤더 정보만으로 플로우 간의 연관성을 계산하여, 연관성을 기준으로 악성 트래픽을 탐지하는 방법에 대해 제안한다. 본 논문의 타당성을 검증하기 위해서 실제 악성 트래픽을 사용하여 실험을 진행 하였으며, 플로우의 두 가지 정보를 사용하는 이전 방법과 비교한 결과 탐지율에서 5~30% 정도 향상 된 결과가 나타났다

Key Words : Flow Correlation Index, Traffic Analysis, Network Management, Traffic Classification

ABSTRACT

Today, the network environment is dramatically increasing, and the attack of malicious traffic is getting more sophisticated and complicated. For the accurate analysis of malicious traffic, it is necessary to reduce and prevent damage to such attacks. The most widely known methods are Signature-based analysis and Machine Learning-based analysis. Both of these methods have the advantages of high accuracy and detection rate, but they are disadvantageous only when the process is complicated and the requirement is met. Recently, a method of calculating the flow correlation based on the flow statistical and header information and detecting with the correlation value has been studied. The statistical information includes packet size, and the header information

※ 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업과(NRF-2018R1D1A1B07045742), 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2018-0-00539-002,블록체인의 트랜잭션 모니터링 및 분석 기술개발)

◆ First Author : Department of Computer and Information Science, Korea University, pjj5846@korea.ac.kr학생회원
° Corresponding Author : Department of Computer and Information Science, Korea University, tmskim@korea.ac.kr, 종신회원
* Department of Computer and Information Science, Korea University, {pb1069, chenlima2, tm0309}@korea.ac.kr, 학생회원
논문번호 : 201812-390-B-RN, Received December 17, 2018; Revised March 6, 2019; Accepted March 20, 2019

includes the source and destination IP address, port number, and protocol of the flow. However it takes a lot of time and money to obtain statistical information in the real environment. Therefore, in this paper, we propose a method to detect attack traffic by calculating flow correlation based on header information. In order to verify the validity of this paper, we conduct several experiments with real attack traffic and the detection rate was improved by 5~30% compared with the previous method.

Ⅰ. 서 론

오늘 날 비약적으로 발전하는 초고속 인터넷으로 인해 네트워크 환경이 크게 증가하였고, 대용량의 트래픽이 빈번하게 발생하고 있다. 이에 다양한 응용과 악성 행위의 트래픽이 급격하게 발생하고 있다[1]. 이러한 다양한 악성 트래픽이 발생함에 따라 피해도 커지고 있다. 예를 들면, 2017년 워너 크라이 랜섬웨어가 전 세계적으로 발생함에 따라 많은 국가와 기업의 금전적 피해가 있었다. 그리고 해커 혹은 특정 나라에서 자신의 이익을 위하여 상대 국가와 기업에 공격을 감행하여 특정 서버 및 시스템을 마비시키거나 기밀 파일을 유출하는 사례도 적지 않다.

이에 많은 국가와 기업에서는 효과적으로 악성 행위를 막고 예방하기 위해서 악성 트래픽에 대한 정확한 분석에 대한 연구를 진행되고 있다[1]. 현재 가장 보편적으로 연구되어 지고 있는 방법으로 트래픽의 시그니처를 기반으로 패턴 분석을 통한 트래픽 분석 방법과 기계 학습을 이용한 트래픽 분석 방법이 있으며[5,6], 두 가지 방법 모두 높은 탐지율을 가진다는 장점이 있다.

하지만 시그니처 기반의 트래픽 분석 방법은 처리 과정이 복잡하고 시간이 오래 걸리며, 기계 학습 기반의 트래픽 분석 방법 또한 높은 성능을 기대하기 위해서 많은 양의 정확한 학습 데이터가 필요하다는 단점이 존재한다. 따라서 최근에는 기존의 분석 방법 이외의 새로운 분석 방법들이 다양하게 연구되어 지고 있다.

그 중 한 가지로 플로우의 통계적 정보와 헤더 정보를 사용하여 유사 플로우를 분석하는 방법이 있다. 플로우의 통계적 정보에는 패킷 길이, 패킷 도착 시간 간격 등의 정보로 구성되어 있으며, 헤더 정보에는 플로우 IP 주소, 프로토콜, 포트번호 등의 정보로 구성되어 있다. 이 방법은 비교 할 두 플로우의 통계적, 헤더 정보를 사용하여 두 플로우의 연관성 수치적으로 계산하고, 이를 바탕으로 연관 플로우만 분류하는 방법이다. 하지만 이 방법 역시 높은 성능을 지니지만 플로우의 통계적 정보를 얻는데 많은 노력과 비용이 든다는 단점이 있다. 따라서 본 논문에서는 기존의 플로우의 통계적 정보와 헤더 정보 모두 사용하는 분석 방법을 개선하여 헤더 정보만을 이용하여 플로우 간의 연관성 값을 계산한 다음, 그 값을 바탕으로 연관된 악성 플로우를 탐지하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 본 장의 서론에 이어, 2장에서는 트래픽 분석에 관련된 연구에 대해 설명하고, 제안하는 시스템의 키워드에 대해 설명한다. 3장에서는 제안하는 시스템 구조와 시스템에 사용된 알고리즘에 대해 설명한다. 4장에서는 제안한 방법을 검증하기 위한 실험과 실험 결과에 대해 기술한다. 마지막으로 5장에서는 결론으로 제안하는 방법에 대한 전반적인 내용과 향후 연구에 대해 언급한다.

Ⅱ. 관련 연구

서론에서 언급한대로, 가장 널리 알려진 트래픽 분석 방법으로 시그니처 기반 분석 방법과 기계학습 기반 분석 방법이 있다.

시그니처 기반의 분석에는 여러 가지 방법이 존재하는데, 그 중에서도 가장 널리 알려진 방법은 페이로드 시그니처 기반의 분석 방법이다[2-4]. 페이로드 시그니처 기반의 분석 방법은 여러 방법들에 비해 높은 분석률과 정확도를 지닌다는 장점이 있다[3]. 페이로드 시그니처를 생성하기 위해서는 추출하고자 하는 응용의 트래픽을 수집 한 다음, 수집 된 트래픽의 페이로드를 직접 비교하며 공통적으로 발생하는 패턴을 찾는 방법이다. 즉, 시그니처 추출 과정과 수작업으로 이루어지기 때문에 실시간으로 변하는 트래픽에 대하여 유연하게 대처 할 수가 없으며, 많은 노력과 시간이 필요하다는 단점이 있다.

이러한 한계를 극복하기 위해 시그니처를 빠르고 신속하게 생성하는 시그니처 자동 생성 방법에 대해 연구되고 있다. 이 연구는 패킷의 페이로드 내용을 기반의 시그니처 추출 과정을 자동적으로 생성하여 기존의 단점을 해결한다. 하지만 이 방법 역시 트래픽을 직접 수집하여야하기 때문에 시간이 오래 걸리며, 수집 된 트래픽은 대체적으로 단기간으로 수집 된 트래

픽이기 때문에 하나의 시그니처를 오랫동안 사용할 수 없다.

최근 가장 널리 연구되어 지고 있는 방법 중 하나로 기계학습을 이용한 분석 방법이 있다[7-9]. 기계학습을 이용한 방법은 트래픽 분석하기 전 미리 학습한 데이터를 바탕으로 분석하는 방법이다. 2017년 구글의 인공지능 알파고와의 대결을 이후로 인공지능의 기계학습에 대한 관심이 높아졌으며, 이에 기계 학습을 여러 연구 분야에 적용 시키는 연구가 활발해 졌다. 기계 학습을 이용한 트래픽 분석 방법은 학습 된 데이터에 대해서 높은 분석률과 정확도로 분석이 가능하다는 장점이 있다[7,8]. 그리고 텐서플로(Tensorflow)와 같은 오픈 소스 라이브러리가 있기 때문에 시그니처 기반의 분석 방법에 비해 누구나 더 쉽게 사용 할 수 있다. 하지만 학습 된 데이터가 잘못 된 데이터 일 경우 학습에 대한 성능이 많이 떨어지며, 새로운 트래픽이 나타났을 경우 능동적으로 대처하기 어렵다는 단점이 있다.

최근에는 기존의 분석 방법 이외의 방법들이 많이 연구되고 있다. 그 중 한 가지로 하나의 플로우의 정보를 바탕으로 연관성을 수치적으로 계산한 다음, 그 값을 이용하여 탐지 방법이 연구되어 지고 있다. 여기서 사용되는 정보로 패킷 길이, 패킷 도착 시간 간격 등의 플로우 통계적 정보와 IP 주소, 포트 번호 등의 플로우 헤더 정보가 있다. 이 방법을 구현한 시스템은 별도의 추가 작업 없이 작동 가능하며, 높은 성능을 가진다는 장점이 있다. 하지만 실제 환경에서 플로우의 통계적 정보를 구하기 위해서는 많은 비용과 노력이 필요하기 때문에 이 시스템을 실제 환경에서 구현하기에는 어렵다는 단점이 있다.

Ⅲ. 시스템 설계

본 논문에서는 기존의 플로우 통계적 정보와 헤더 정보를 모두 사용하는 방법 대신 헤더 정보만을 사용하여 탐지하는 방안에 대해 제안한다.

본 논문에서 제안하는 시스템을 설명하기 위해서, 먼저 몇 가지 항목에 대한 정의가 필요하다. 본 장에서 정의 되는 내용은 제안하는 시스템의 핵심이 되는 내용으로 플로우 연관성 지표(Flow Correlation Index), 시드 정보(Seed Information), 가이드라인(Guideline)으로 구성되어 있으며, 설명 후 전체 시스템 구조 및 알고리즘에 대해 설명한다.

먼저 시드 정보는 서론에서 언급한 탐지를 시작 할 때, 하나의 시드 플로우에 대한 정보로, 플로우의 5-tuples 정보로 구성되어있다. 여기서 플로우의 5-tuples 정보는 출발지 도착지의 IP주소와 포트번호, 그리고 프로토콜에 대한 정보로 구성되어 있으며, 플로우는 네트워크 환경 내의 트래픽 패킷들 중 5-tuples의 정보가 같은 패킷들의 집합을 뜻한다. 그리고 트래픽 내의 여러 플로우 중에 시드 정보와 일치하는 플로우를 시드 플로우라고 지정한다. 즉, 시스템 내의 플로우의 연관성 값은 시드 플로우를 기준으로 다른 플로우와의 특징 값 차이를 통하여 계산 된다.

그리고 서론에서 언급한 플로우 연관성 지표는 두 플로우 간의 연관성을 수치적으로 계산한 값이다. 이 값을 기준으로 두 플로우 간의 유사한 정도를 판단하고, 값이 기준 값보다 큰 경우 탐지 하게 된다. 이전

Connectivity Features	Explanation	Function	Range
ST	Start Time	$a_{ST}(f_x,f_y)=1-\sqrt{\frac{dist(f_x,f_y)}{maxdist(F)}}$	0~1
SIP	Source IP Address	$a_{IP}(f_x,f_y)=\left(\frac{prefixlenAddr(f_x,f_y)}{32}\right)^2$	0~1
DIP	Destination IP Address		
SPT	Source Port Number	$a_{PT}(f_x,f_y)=\left(\frac{prefixlenPort(f_x,f_y)}{16}\right)^2$	0~1
DPT	Destination Port Number		
PROT	L4 Protocol	$a_{PROT}(f_x,f_y)=\begin{cases}0: f_x.PROT \neq f_y.PROT\\1: f_x.PROT = f_y.PROT\end{cases}$	0 or 1
Ratio of Forward Packet (RP)	Forward Packet Ratio	$1-The\ Difference\ of\ Forward\ Packet\ Ratio\ in\ Flow\ (f_x,f_y)$	0~1
Ratio of Forward Byte (RB)	Forward Byte Ratio	$1-The\ Difference\ of\ Forward\ Byte\ Ratio\ in\ Flow\ (f_x,f_y)$	0~1
$Conn(f_x,f_y)=w_{1\_ST}\times a_{ST}(f_x,f_y)+w_{2\_IP}\times a_{IP}(f_x,f_y)+w_{3\_PT}\times a_{PT}(f_x,f_y)+w_{4\_PROT}\times a_{PROT}(f_x,f_y)+w_{4\_PROT}\times a_{PROT}(f_x,f_y)+w_{4\_PROT}\times a_{PROT}(f_x,f_y)$ $(where, 0\le a_i(f_x,f_y)\le 1,\ \sum_{i=1}^{4}w_i=1)$			0~1

그림 1. 플로우 연관성 지표 계산하는 방법
Fig. 1. Calculating Method of Flow Correlation Index

연구에서 플로우 연관성 지표는 플로우의 통계적 정보특성으로 구성 된 유사성 지표(Similarity Index)와 헤더 정보로 구성된 연결성 지표(Connectivity Index)로 이루어져있다. 하지만 본 논문에서는 플로우의 통계적 정보를 배제하고 헤더 정보만을 사용하기 때문에 플로우의 헤더 정보를 통해 계산 된 연결성 지표가 플로우 연관성 지표 값이 된다. 연결성 지표 값을 계산 할 때 사용되는 헤더 정보로는 플로우의 시작 시간, IP 주소, 포트 번호, 프로토콜과 패킷 수와 바이트 수의 정보가 있다. 연결성 지표는 그림 1과 같이 각각의 특징 값에 해당하는 값과 비중(Weight)을 곱한 값을 더해서 나오는 값이며, 플로우 연관성 지표 값은 최종적으로 나오는 연결성 지표 값이 된다. 이때, 비중 값은 각 특징 값에 대한 가중치 값으로 0~1 사이의 수로 구성된다. 비중 값과 기준 값의 경우 가이드라인에 정의 되어있으며, 두 값을 결정하는 알고리즘에 대해서는 3장에 자세하게 설명한다.

가이드라인은 시드 정보로 생성된 시드 플로우와 다른 플로우 간의 연관성 지표를 계산한 다음, 그 값을 그룹핑 할 것인지 판단하는 기준 값이 정의 되어 있는 파일이다. 각 플로우 간의 연관성 지표 값이 기준 값보다 크면 그룹핑을 진행하고, 그렇지 않으면 정상 플로우로 판단하여 그룹핑을 진행하지 않는다.

여기서 그룹핑이란 두 플로우 간의 연관성 값에 따라 분류하는 작업이다. 가이드라인의 경우 사전에 악성과 정상 플로우가 구분 되어있는 데이터에서 기준 값 설정 알고리즘(Threshold Setting Algorithm)을 통해 정해진 기준 값과 비중 값을 바탕으로 만들어진다.

즉, 본 논문에서 제안하는 분류 방법은 플로우 간의 연관성 지표를 계산하고, 지표 값과 가이드라인의 기준 값과 비교를 통해 그룹핑 여부를 판단하고, 이를 통해 악성 트래픽을 분류한다.

3.1 전체 시스템 구조

전체 시스템의 구조는 그림 2와 같이 되어있다. 그림 2의 전처리(Pre-Processing) 모듈에서 시스템의 입력으로 악성과 정상이 구분 되어 있는 트래픽 데이터가 들어온다. 이때, 트래픽의 파일 형식은 pcap이며, 해당 파일을 전처리 과정을 거쳐 fwp 포맷으로 만든다. 여기서 fwp 파일은 패킷 단위의 pcap파일을 플로우 단위 형태로 변환 하여 저장 한 파일이다. 다음으로 악성 플로우에 대해 시드 파일을 생성하는데, 시드 파일이란 어떤 플로우에 대한 5-tuples 정보를 추출하여 텍스트 파일 형태로 저장한 파일이다. 즉, 전처리 과정에서 시드 파일은 악성으로 구분 된 트래픽 플로우만 입력으로 생성된다.

이러한 전처리 과정을 거친 후에 처음에 변환 된 트래픽 파일(fwp)과 시드 파일을 가지고 가이드라인

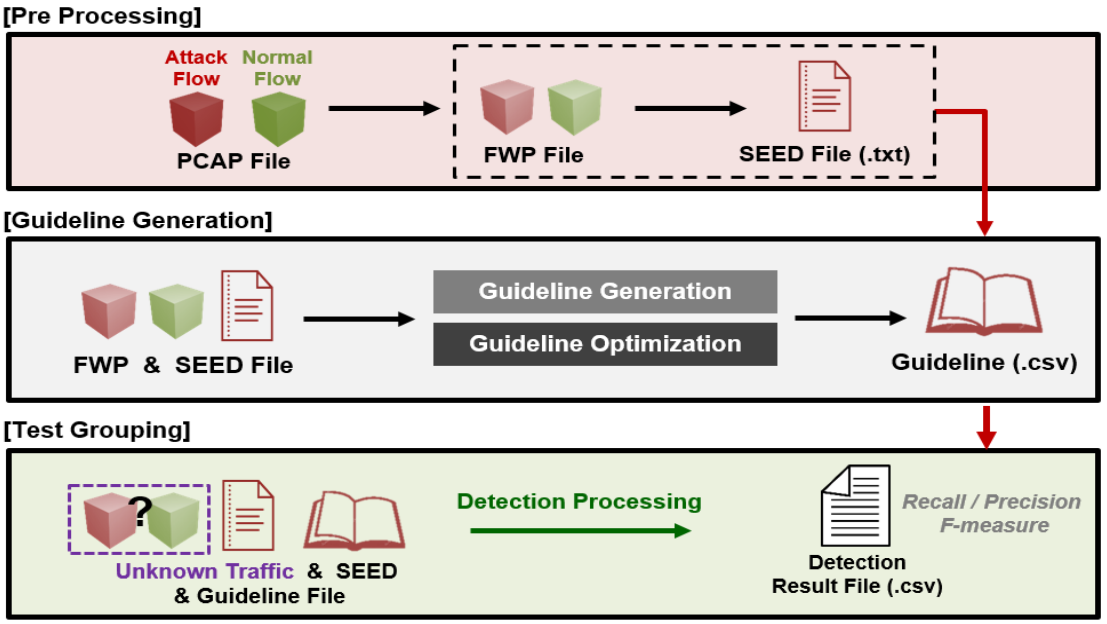

그림 2. 전체 시스템 구조
Fig. 2. Entire System Structure

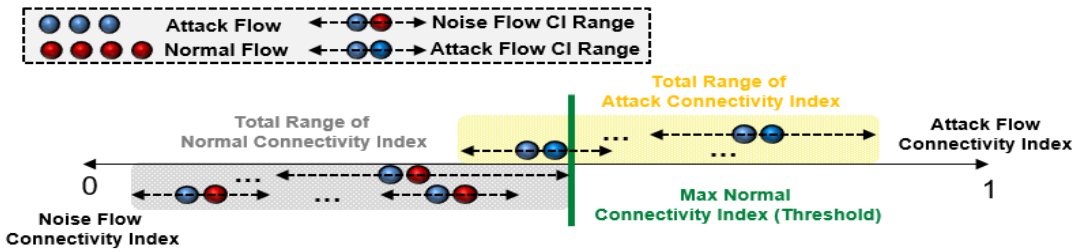

그림 3. 기준 값 설정 알고리즘 적용
Fig. 3. Applying Threshold Setting Algorithm

을 생성하며, 가이드라인 생성은 크게 가이드라인 생성과 최적화로 두 파트로 구성되어있다.

가이드라인 생성(Guideline Generation) 모듈은 생성된 각각의 시드 파일별로 기준 값 설정 알고리즘에 의해서 생성된다. 시드 파일 하나를 입력으로 해당 시드 정보에 맞는 플로우를 찾고, 해당 플로우와 다른 플로우 간의 연관성 값을 계산한 다음, 기준 값 설정 알고리즘에 의해 기준 값을 설정하고 이를 가이드라인에 저장한다.

가이드라인 최적화는 각각의 시드 별로 생성된 여러 가이드라인을 가지고 임의로 그룹핑을 진행한 다음 도출된 탐지율과 탐지정확도를 반영하여 최적화를 진행한다. 먼저 생성된 각 가이드라인의 기준 값 별로 정확하게 탐지한 플로우 비율을 계산 한다. 탐지한 플로우의 비율은 임의의 그룹핑을 통해 탐지 된 악성 플로우의 수와 탐지 정확도를 곱한 값이다. 다음으로 계산 된 비율을 반영해서 기준 값을 새로 계산 하고, 생성된 모든 가이드라인의 새로 계산 된 기준 값의 평균을 최적화된 가이드라인 기준 값으로 지정한다. 이 과정을 거치면 하나의 악성 트래픽에서 하나의 최적화 된 가이드라인이 나온다.

마지막으로 생성 된 가이드라인과 시드 파일, 트래픽 파일을 가지고 테스트 그룹핑을 진행한다. 이 때 트래픽 파일은 전처리 과정, 가이드라인 생성 모듈과 다르게 입력된다. 왜냐하면 실제 네트워크에서는 수집 된 정상과 악성 트래픽이 섞여있는 상태로 입력되고 분석되기 때문이다. 하지만 본 논문에서 진행한 실험에서는 실제 네트워크에서 수집 된 트래픽을 구하기 어렵기 때문에 전처리 과정에서 이전 모듈에서 사용한 정상과 악성 트래픽을 구분 하지 않고 입력으로 넣는다. 즉, 실제 네트워크 환경과 유사하게 실험하기 위해 전처리 및 가이드라인 생성 모듈에서는 정상과 악성 트래픽을 구분하여 입력으로 넣고 테스트 그룹핑 모듈에서는 이를 구분 하지 않은 상태로 넣는다.

그룹핑이 완료되면 탐지한 결과와 실제 구분 된 정상 및 악성 플로우와 비교를 통해서 탐지 결과가 나온다. 탐지 결과에는 탐지율(Recall), 탐지 정확도(Precision)와 두 가지의 값을 바탕으로 F-measure 값이 나온다. 탐지율 값은 전체 악성 플로우 중 탐지 된 플로우의 비율을 나타내며, 전체 악성 플로우 수에서 탐지 된 플로우 수를 나눈 값을 백분위로 나타낸다. 탐지 정확도는 탐지한 악성 플로우 수 중 정확하게 탐지한 악성 플로우 수를 나눈 값을 백분위로 나타낸다.

본 논문에서 제안하는 분석 시스템의 알고리즘으로 크게 기준 값 설정 알고리즘(Threshold Setting Algorithm), 다중 시드 & 가이드라인 선택 알고리즘(Multiple Seed & Guideline Selection Algorithm)로 구성 되어있으며, 각 알고리즘에 대한 설명은 다음과 같다.

3.2 시스템 알고리즘

3.2.1 기준 값 설정 알고리즘

기준 값 설정 알고리즘은 악성 플로우 탐지에 가장 핵심이 되는 알고리즘으로, 적용하는 과정은 그림 3과 같이 진행 된다. 마지막 모듈에서 그룹핑 탐지를 진행할 때, 높은 탐지율과 탐지 정확도가 나오기 위해서는 기준 값을 잘 설정해야 한다. 먼저 하나의 시드 정보로 부터 일치하는 시드 플로우를 정한다음, 그 플로우와 다른 플로우 간의 플로우 연관성 지표 값을 계산한다. 이때의 플로우 연관성 지표 값은 그림 2와 같이 일정 범위를 가지며, 기준 값은 정상 플로우와의 플로우 연관성 지표 값 중 가장 큰 값으로 정한다. 왜냐하면 그림 3에서처럼 정상 플로우의 FCI의 최댓값으로 설정해야 탐지 정확도 100%를 보장할 수 있기 때문이다. 만약 이 값보다 적은 값으로 설장하게 되면, 탐지율은 늘어나지만 정상 플로우가 같이 탐지되기 때문에 탐지 정확도는 낮아지게 된다. 분석 시스템의 구성상 이러한 과정을 여러 번 반복하기 때문에 정상 플

로우가 조금이라도 섞이면 이 후 그룹핑에 의한 탐지 과정에서는 좋지 않은 결과가 나오게 된다. 따라서

3.2.2 다중 시드 & 가이드라인 선택 알고리즘

이전의 분석 시스템에서는 하나의 시드 파일과 하나의 가이드라인만을 사용하여 그룹핑을 진행한다. 하지만 실제 네트워크 환경에는 여러 종류의 악성 트래픽이 존재하고 각 악성 트래픽 마다 특성이 서로 다르다. 따라서 한 종류의 악성 트래픽의 시드 파일과 가이드라인을 가지고 다른 악성 트래픽의 탐지하기에는 한계가 있다.

이러한 문제점을 해결하기 위해서 제안하는 시스템에 다중 시드 & 가이드라인 선택 알고리즘을 적용하였다. 먼저 다중 시드 선택 알고리즘의 경우 하나의 시드 파일을 사용 했을 경우에는 탐지 하지 못한 악성 플로우가 생길 수 있으며, 그 결과 탐지율이 낮게 나온다. 이때 탐지 하지 못한 악성 플로우를 다른 시드 파일을 입력으로 함께 사용 할 경우 탐지 정확도는 100% 유지하면서 탐지율도 높일 수 있다. 예를 들어 그림 4와 같이 악성 플로우 A, B, C, D있을 때, 시드 A를 사용 했을 때 플로우 A와 연관 플로우 A, B, C를 탐지 가능하고 시드 C를 사용 했을 때 연관 플로우 C, D를 탐지 가능할 때, 시드A, C를 함께 사용하면 연관 플로우 A,모든 악성 플로우를 탐지 가능하다.

다음으로 다중 가이드라인 선택 알고리즘의 경우 악성 트래픽 종류 별로 특성이 다르기 때문에 이러한 특성을 하나의 가이드라인으로 만족시키기 어렵다. 따라서 그림 5과 같이 각 악성 트래픽 종류 별로 가이드라인을 만든 다음, 같이 적용을 하면 이러한 문제점을

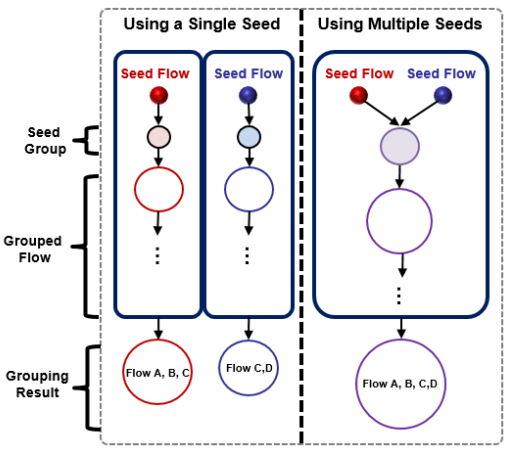

그림 4. 다중 시드 선택 적용 방법
Fig. 4. Multiple Seed Selection Algorithm

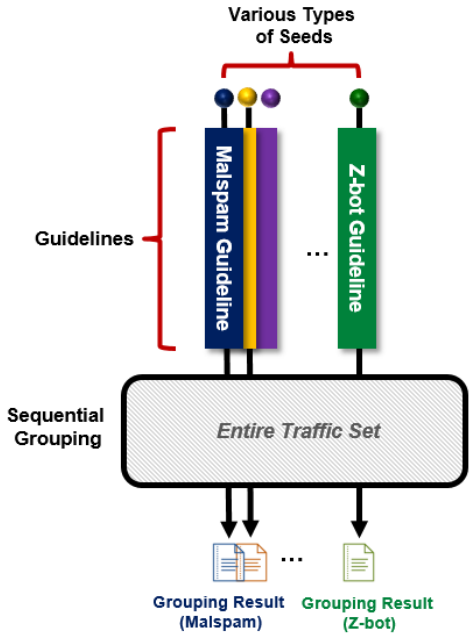

그림 5. 다중 가이드라인 선택 알고리즘 적용 방법
Fig. 5. Multiple Guideline Selection Algorithm

해결 가능하다. 즉, 다중 가이드라인 선택 알고리즘은 미리 여러 종류의 악성 트래픽에 대해 가이드라인을 만들어야 실제 적용이 가능하다.

Ⅳ. 실 험

본 논문에서는 플로우의 헤더 정보와 연속적 그룹핑을 이용한 악성 트래픽 탐지 방안을 제안하였고, 제안하는 시스템의 타당성을 검증하기 위하여 실제 트래픽을 사용하여 실험을 진행하였다. 실험에는 네 가지의 악성 트래픽을 사용하였으며, 실험에 사용된 트래픽에 대한 정보는 표 1과 같다.

실험에서 성능 평가 방법으로는 표 2와 같이 탐지율(Recall), 탐지 정확도(Precision), F-measure을 사용하였다. 여기서 F-measure는 탐지율과 탐지 정확도 값을 객관적으로 비교하기 위한 수치이며, F-measure(MS)의 경우 다중 시드 선택 알고리즘을 사용한 F-measure 결과이다. 다중 가이드라인 선택 알고리즘의 경우 현재 여러 종류의 악성 트래픽에 대해 가이드라인을 생성 할 만큼의 트래픽을 구 할 수 없기 때문에 향후 연구로 진행 할 계획이다.

제안하는 시스템의 성능을 객관적으로 검증하기 위하여 기존의 유사성 지표 값, 연결성 지표 값을 모두

표 1 실험에 사용한 정상 및 악성 트래픽 정보
Table 1. An Information of Normal & Attack Traffic

Normal Traffic			
Application	Size		
	Flow	Packet	Byte
Chrome (web)	491	60,266	46,852,995
Skype	576	107,534	95,635,596
Torrent	614	99,737	94,989,389
KaKaoTalk (messanger)	844	53,655	49,471,332
Attack Traffic			
Attack Method	Size		
	Flow	Packet	Byte
Malspam	71	18,055	15,167,725
Godzilla-Loader	69	1,862	1,358,410
Z-bot	23	1,232	1,229,269
Dreambot	100	3,615	6,564,613

사용한 이전의 시스템과 제안하는 연결성 지표 값만을 사용한 시스템과의 성능을 비교하였으며, 실험의 결과는 다음 표 2에 나타나 있다.

먼저 헤더, 통계적 정보 모두 사용하는 이전 시스템(Previous FCI System)의 경우 다중 시드 선택 알고리즘을 사용하지 않기 때문에 하나의 시드를 사용 하는 경우만 실험을 진행하였다. 반면에 본 논문에서 제안하는 헤더 정보만을 이용한 시스템(Proposed FCI System)의 경우 하나의 시드를 사용했을 경우와 다중 시드 선택 알고리즘을 적용한 경우의 두 가지로 실험 진행하였다.

이전 헤더, 통계적 정보 모두 사용하는 시스템의 경우 탐지 정확도는 100%로 나타났지만, 탐지율의 경우 60~85%의 결과가 나타났다. 반면, 제안하는 헤더 정보만을 사용한 시스템의 경우 탐지 정확도는 100%로 기존과 차이가 없었지만, 탐지율의 경우 95~100%의 성능을 보였다. 특히 다중 시드 선택 알고리즘을 사용했을 경우, 탐지율 값을 최대 100%까지 향상 시킬 수 있었다. 즉, 실험에 사용한 데이터의 경우 기존의 모델에 비해 탐지율을 5~30% 향상 시킬 수 있었다.

V. 결론 및 향후 연구

본 논문에서는 기존의 악성 트래픽 분석 방법 외에 새로운 분석 방법을 제시 하였다. 플로우의 헤더 정보를 이용하여 각 플로우 간의 연관성을 수치상으로 계

표 2 탐지 실험 결과
Table 2. Result of Detection Experiments

Detection Test Result						
Application		**Measurement**	**Result**			
			Previous FCI System		**Proposed FCI System**	
Attack	**Normal**		**Flow**	**Packet**	**Flow**	**Packet**
Malspam	Chrome (web)	Recall (%)	89.74	99.89	94.87	99.98
		Precision (%)	100	100	100	100
		F-measure	94.59		97.37	
		F-measure (MS 2)	-		100	
Godzilla-Loader	Skype	Recall (%)	68.12	95.27	97.1	97.53
		Precision (%)	100	100	100	100
		F-measure	81.03		98.53	
		F-measure (MS 3)	-		100	
Z-bot	Torrent	Recall (%)	82.61	99.35	100	100
		Precision (%)	100	100	100	100
		F-measure	90.48		100	
		F-measure (MS 3)	-		100	
Dreambot	KaKaoTalk (messanger)	Recall (%)	64.71	86.52	100	100
		Precision (%)	100	100	100	100
		F-measure	78.57		100	
		F-measure (MS 2)	-		100	

산하고, 그 값을 기준으로 그룹핑을 진행하여 악성 트래픽을 탐지하는 방법이다. 제안한 방법은 실제 악성 트래픽을 사용한 실험을 통해 타당성을 검증하였으며, 이전의 시스템의 탐지 방법과 비교하였을 때 높은 정확도와 탐지율로 분석 할 수 있었다. 게다가 이전 시스템의 문제점인 플로우의 통계적 정보가 없이 플로우의 헤더 정보만을 사용하여 악성 트래픽에 대해 높은 탐지율로 분석이 가능하다.

하지만 실험에 사용한 악성 트래픽의 종류는 4가지로 본 논문의 타당성을 검증하기 위해서는 아직 부족한 점이 있다. 그리고 시스템 동작 할 때 진행되는 가이드라인을 생성 부분에서 각각의 특징 값을 계산 할 경우에 기존의 알고리즘으로는 많은 시간이 걸린다는 단점이 있다.

따라서 향후 연구로는 더 다양하고 많은 양의 악성 트래픽을 구하여 실험을 진행 할 예정이다. 특히 여러 종류의 악성 트래픽을 수집하여 본 논문에서 제안한 다중 가이드라인 선택 알고리즘을 적용 실험을 진행할 예정이다. 그리고 실제 네트워크 내에서 수집 된 트래픽을 구할 수 있다면, 수집 된 트래픽을 본 시스템에 적용하여 실험 할 예정 이다. 이후에 기존의 가이드라인 생성 알고리즘을 개선 시켜 더 빠르고 효율적으로 생성 할 수 있도록 할 예정이다.

References

[1] M.-S. Kim, Y. J. Won, and J. W.-K. Hong, “Application-level traffic monitoring and an analysis on IP networks,” *ETRI J.*, vol. 27, pp. 22-42, 2005.

[2] S. H. Yoon, J. S. Park, Baraka D. Sija, M. J. Choi, and M. S. Kim, “Header signature maintenance for internet traffic identification,” *Int. J. Network Management*, vol. 27, no. 1, Jan. 2017.

[3] J. S. Park, S. H. Yoon, and M. S. Kim, “Software architecture for a lightweight payload signature-based traffic classification system,” in *Proc. Int. Conf. Traffic Monitoring and Anal. Workshop*, pp. 136-149, 2011.

[4] J. S. Park, J. W. Park, S. H. Yoon, Y. S. Oh, and M. S. Kim, “Development of signature generation system and verification network for application level traffic classification,” in *Proc. KIPS Conf.,* pp. 1288-1291, Pusan, Korea, Apr. 2009.

[5] N. F. Huang, G. Y. Jai, H. C. Chao, Y. J. Tzang, and H. Y. Chang, “Application traffic classification at the early stage by characterizing application rounds,” *Inf. Sci.*, vol. 232, pp. 130-142, 2013.

[6] C. A. Catania and C. G. Garino, “Automatic network intrusion detection: Current techniques and open issues,” *Comput. and Electrical Eng.*, vol. 38, no. 5, pp. 1062-1072, Sep. 2012.

[7] S. H. Lee and M. S. Kim, “Application traffic classification using TensorFlow machine learning tool,” in *Proc KICS Winter Conf.*, pp. 224-225, Korea, Nov. 2009.

[8] S. Zander, T. Nguyen, and G. Armitage, “Automated traffic classification and application identification using machine learning,” in *Proc. IEEE Ann. Conf. Local Computer Networks*, pp. 250-257, Sydney, NSW, Australia, 2005.

[9] T. Nguyen and G. Armitage, “A survey of techniques for internet traffic classification using machine learning,” in *Proc. IEEE Commun. Surv. and Tuts.,* vol. 10, no. 4, Fourth Quart. 2008.

박 지 태 (Jee-Tae Park)

2017년 : 고려대학교 컴퓨터정보학과 학사

2017년~현재 : 고려대학교 컴퓨터정보학과 석박사통합과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-8515-6164]

백 의 준 (Ui-Jun Baek)

2018년 : 고려대학교 컴퓨터정보학과 학사

2018년~현재 : 고려대학교 컴퓨터정보학과 석박사통합과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-4358-7839]

신 무 곤 (Mu-Gon Shin)

2019년 : 고려대학교 컴퓨터정보학과 학사

2019년~현재 : 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0003-3703-3319]

이 민 섭 (Min-Seob Lee)

2018 : 고려대학교 컴퓨터정보학과 학사

2018년~현재 : 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0003-4854-9521]

김 명 섭 (Myung-Sup Kim)

1998년 : 포항공과대학교 전자계산학과 학사

2000년 : 포항공과대학교 전자계산학과 석사

2004년 : 포항공과대학교 전자계산학과 박사

2006년 : Dept. of ECS, Univ of Toronto Canada

2006년~현재 : 고려대학교 컴퓨터정보학과 교수

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크

[ORCID:0000-0002-3809-2057]