Skype 트래픽 분류에 관한 연구 [1]

이상우*, 정아주*, 이현신** 김명섭*,
*고려대학교 컴퓨터정보학과, **고려대학교 수학과
e-mail : {ilovejam, yesvery, oshyuns, tmskim}@korea.ac.kr

Research on Skype Traffic Classification

Sang-Woo Lee*, Ah-Joo Jung*, Hyun-Shin Lee**, Myung-Sup Kim*
*Dept. of Computer and Information Science, Korea University
**Dept. of Mathematics, Korea University

요 약

네트워크 관리자 입장에서 효율적인 네트워크 관리를 위해 응용 프로그램 별 트래픽 분류의 중요성이 커지고 있다. 응용 프로그램 별 트래픽 분류를 위해 signature 기반, machine learning 방법들이 제안되고 있지만 p2p 방식의 Skype 응용프로그램에 대한 적용결과는 그 신뢰성이 떨어지고 있는 것은 사실이다. 본 논문에서는 Skype 의 트래픽을 분류하기 위해 각 Client 마다 Skype application install 시 동적으로 변화하는 Port 를 알아내는 방법, UDP 패킷의 특정위치의 특정 signature, TCP signal flow 의 특정위치 패킷에 대한 payload 크기 등을 이용한 Skype traffic 분류 방법을 제안한다. 제안된 방법론은 학내 네트워크에 적용하여 그 타당성을 TMA 를 통해 검증하였다.

1. 서론

Skype application[1] 은 p2p 방식의 메신저로써 사용자간 채팅, 음성통화, 화상통화, 전화 교환망을 통한 일반 전화, 파일전송 등의 기능을 제공한다. 안정적인 통화품질 서비스 제공과 저렴한 가격은 오늘날 전 세계적으로 많이 사용하는 메신저로 만들었다. 하지만 Skype 의 트래픽들은 기본적으로 암호화[2]가 되어있고, install 시 동적 포트 할당, 일반적인 프로토콜[2]을 사용하지 않아 네트워크 관리자의 입장에서 Skype 의 트래픽 분류 기준을 정하기 어렵게 만들고 있다.

본 논문에서는 Skype application install 시 마다 달라지는 클라이언트의 포트를 알아내는 방법과 사용자 리스트를 만들어 트래픽을 분류하는 방법, UDP 패킷의 특정 signature 존재, signal[3] flow 의 2, 3, 4 번째 패킷의 byte 크기 등을 기반으로 네트워크 상의 Skype 의 트래픽을 분류하는 방법을 제안 한다.

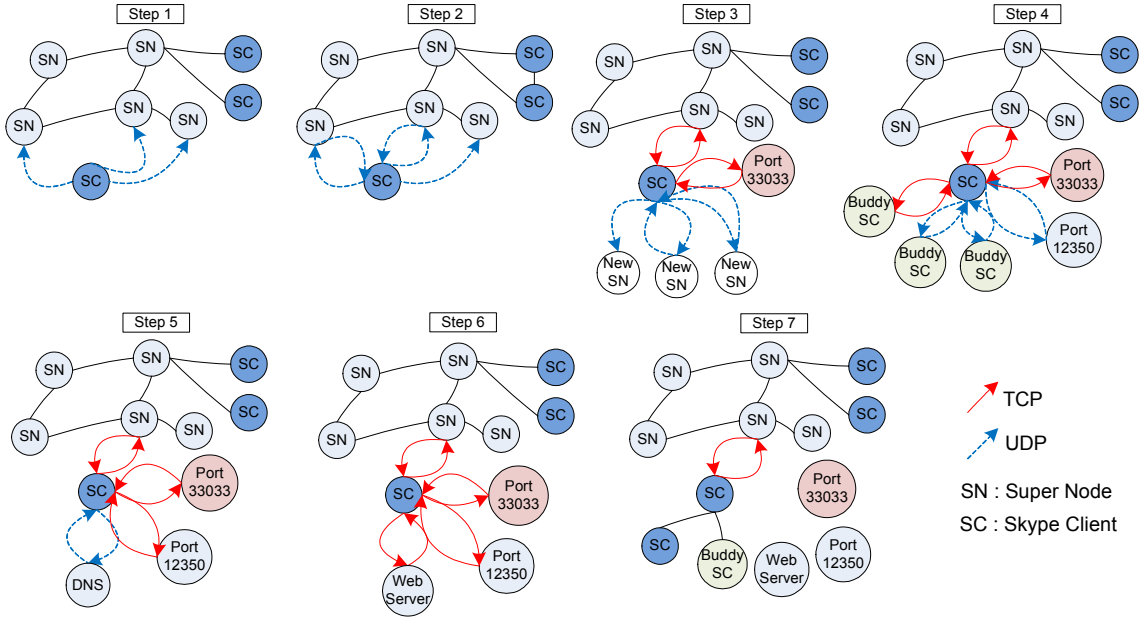
2 장에서는 Skype 의 로그인 단계들을 도식화 하여 순차적으로 보여주며, 3 장에서는 시스템 동작 환경 및 검증 시스템을 소개한다. 4 장에서는 자세한 분류 알고리즘을 flow chart 로 보여주며, 5 장 에서는 실제 학내 망에 분류 시스템을 적용 후 검증결과를 보여주며 다른 논문에서 제안한 방법론들과 비교한다. 6 장에서는 문제점 제시를 하며, 7 장에서는 결론 및 향후 연구과제를 기술한다. 본 논문에서 사용한 Skype application 의 version 은 3.8 과 4.0 을 사용하였다.

2. Skype 로그인 단계

Skype 의 로그인단계는 총 7 단계로 나누어 질 수 있으며 (그림 1)의 각각의 단계에서 Skype 의 로그인의 특성을 알 수 있다.

- Step1 – 기존의 SC(Skype Client)에 저장되어 있던 SN(Super Node)[4]의 리스트를 기반으로 하여 SC 가 SN 에게 UDP 패킷을 전송한다.
- Step2 – SC 에서 UDP 패킷을 보낸 SN 으로부터 reply UDP 패킷을 받는다.
- Step3 – reply 패킷이 온 SN 에게 TCP 연결을 맺는다. 복수의 SN 으로부터 reply 가 온 경우 reply 패킷을 보낸 1 번째 또는 2 번째 SN 을 선택하여 TCP 연결을 맺는다. TCP 연결을 맺은 SN 으로부터 new SN 리스트를 전송 받는다고 추정된다. 리스트를 받음과 동시에 UDP 패킷을 new SN 에게 전송한다. 이 때 reply 온 new SN(새로운 SN)에 한하여 최종 확정된 새로운 리스트를 저장한다고 추정된다. dstport(Destination Port : 목적지 포트번호) 33033 인 노드로 TCP 연결도 동시에 이루어진다.
- Step4 – SC 가 dstport 12350 인 노드로 UDP 패킷을 전송한다. 이 시점에서 SC 의 Skype 포트번호를 알 수 있으며 Skype 트래픽 분류 알고리즘에서 SC 의 IP 와 Port 를 기록한다. SC 와 buddy (Skype 친구등록)관계인 SC 들에게 UDP 패킷을 전송하며 만약 프로필 변경, 사진 변경 등 갱신 정보가 있을 때 buddy 관계의 SC 와 TCP 연결을 맺어 데이터를 전송 받는다.
- Step5 – DNS 서버로 query 를 전송하여 Skype web server 의 IP 를 받아온다. Step5 의 단계에서는

[1] 이 논문은 2007 년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2007-331-D00387)



(그림 1) Skype login 단계

dstport 12350 인 노드로 TCP 연결을 맺는 경우와 맺지 않는 경우 의 두 가지 경우가 생긴다.

1. Step4 에서 SC 가 이전에 로그인하였던 호스트에서 다시 로그인 하였을 경우
2. Step4 에서 SC 가 이전에 다른 호스트에서 로그인 을 하였고 이번 로그인 이 그전의 호스트와 다를 경우

1 번의 경우 step5 의 단계에서 dstport 12350 인 노드와 TCP 연결을 맺지 않는다.

2 번의 경우 dstport 12350 인 노드와 TCP 연결을 맺게 된다. 따라서 Step4 의 buddy SC 에게 UDP 패킷을 전송하는 단계는 한 단계 뒤로 미루어지게 된다.

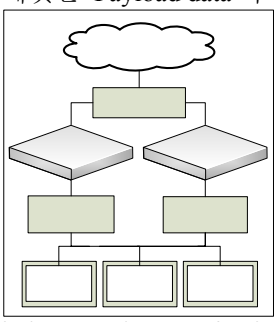
- Step6 – DNS 서버로부터 받은 IP 인 web server 와 TCP 연결을 맺고 version 을 체크한다.
- Step7 – dstport 33033 인 노드와 접속을 해제하며, step5 에서 dstport 12350 인 노드와 TCP 연결을 맺었을 경우 이 단계에서 접속을 해제한다.
- Step7 이후 – buddy SC 와 통신, 그 외 SC 와 통신, 로그아웃 할 경우 step3 단계에서 SN 과 맺었던 TCP 연결을 끊게 된다.

3. 트래픽 수집 및 검증 시스템

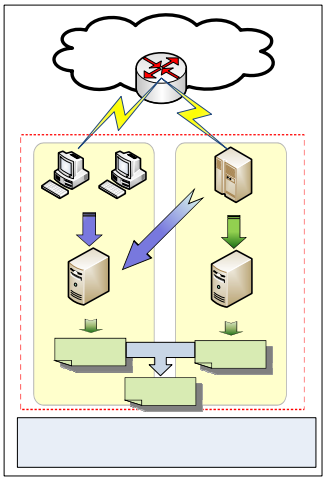
이 장에서는 Skype 트래픽을 수집하기 위해 구성한 트래픽 수집 장소 및 방법을 기술한다. 또한 TMA[5] (Traffic Measurement Agent) 를 이용하여 Skype 분류 알고리즘이 분류해낸 트래픽들의 검증방법을 소개한다.

3.1 트래픽 수집장소 및 방법

트래픽 수집은 본 연구실에서 개발한 실시간 트래픽 모니터링 시스템인 KU-MON[6]을 설치하여 라우터(외부 인터넷 망과 연결되어있는)와 두 대의 코어스위치 사이에 있는 2 개의 링크로부터 트래픽을 수집하였다. Flow 데이터는 일반적으로 사용되는 패킷 헤더의 5-tuple 정보(Source IP, Source Port, Destination IP, Destination Port, Protocol)가 동일한 단 방향(Uni-flow) 패킷들의 집합으로 정의한다. 또한 flow 의 Payload 가 있는 첫 10 개의 패킷은 Payload data 까지 수집하였다.



(그림 2) 인터넷 트래픽 수집 장소 및 방법



(그림 3) TMA 검증 시스템 구조

3.2 TMA 검증 시스템

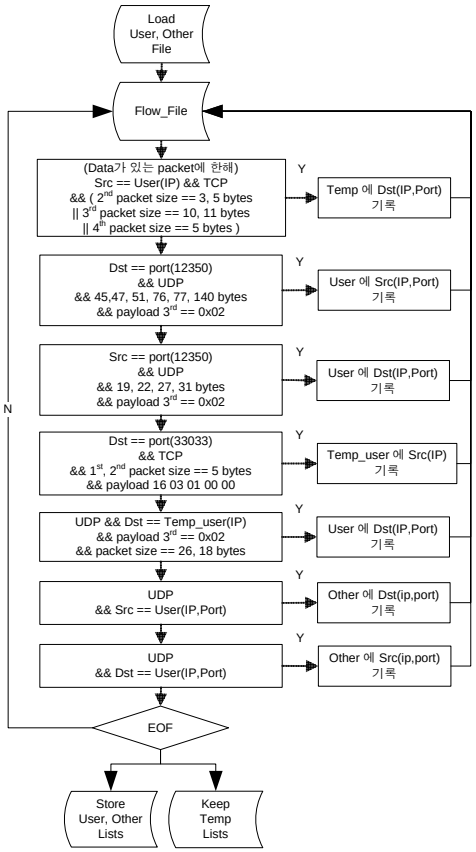
본 논문에서 사용된 검증 시스템은 학내 네트워크의 종단 호스트에 설치된 TMA 를 이용하여 트래픽을 응용프로그램 별로 분류하는 방법을 사용했다. TMA

는 해당 호스트의 현재 활성화된 소켓 정보를 토대로 TMA 정보(Process Name, Source IP, Source Port, Destination IP, Destination Port, Protocol)를 제공해 준다. 이를 이용하여 KU-MON 으로 생성한 패킷과 flow 데이터와 비교하여 해당 패킷과 flow 가 어떤 응용 프로그램에 의해 실행되었는지 판단할 수 있다.

본 논문에서는 (그림 3)과 같은 증명 시스템을 사용하였으며 TAS 에서 다음의 알고리즘이 Skype 의 트래픽을 분류하며, Answer result 와 비교하여 결과를 도출한다.

4. 알고리즘

본 논문에서는 Skype 트래픽을 분류하기 위해 두 단계의 알고리즘을 사용하였다.

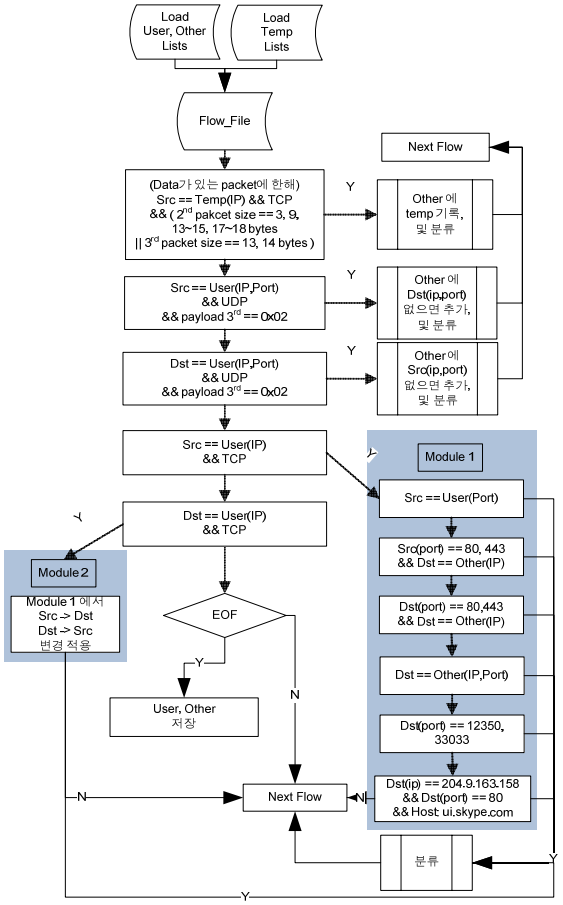


(그림 4.1) 1 단계 알고리즘

두 가지 알고리즘 모두 같은 flow 파일을 사용하며, flow 파일은 실시간으로 2 분전 학내의 모든 트래픽들이 저장 되고 있다.

Skype 의 트래픽을 분류하기 위해서는 다음 4 가지의 리스트가 구성 되어야 한다.

- User : 유추된 학내 Skype 사용자 IP, Port
- Other: User 와 연관되어 트래픽을 발생시키는 노드의 IP, Port
- Temp: Outbound 트래픽(학내 망에서 외부 인터넷 망으로 나가는 트래픽)에서 User 와 연관되어 트래픽을 발생시키는 Other 가 될 가능성이 있는 노드의 IP, Port
- Temp_user: Outbound 트래픽에서 User 가 될 가능성이 있는 노드의 IP



(그림 4.2) 2 단계 알고리즘

1 단계 알고리즘(그림 4.1) 에서는 학내 Skype 사용자의 IP 와 Port 를 기록(User), User 와 관련되어 Skype 트래픽을 일으키는 가능성(아직 확실하게 Other 라 결론 내릴 수 없는)이 있는 노드를 기록(Temp), User 와 연관되어 트래픽을 발생시키는 노드를 기록(Other), 또한 dstport 12350 이 항상 나오는 단계가 아닐 수 있으므로 로그인시마다 매번 TCP 연결을 맺는 dstport 33033 으로 User 를 추론(Temp_User)의 4 가지 리스트를 구성한다. Temp_User 리스트는 1 단계에서 사용되고 소멸되며 학내 사용자 IP 를 기록하고 있으며 Temp_User 에서 UDP 프로토콜로 Inbound 트래픽(외부 인터넷 망에서 학내 망으로 들어오는 트래픽)이 발생할 때 알고리즘을 거쳐 User 에 기록된다.

2 단계 알고리즘(그림 4.2)에서는 1 단계 알고리즘에서 기록된 4 가지 리스트를 기반으로 트래픽 들을 분류하며, 가능성이 있는 노드들(Temp)을 한번 더 검증하여 최종으로 Other 리스트에 추가한다. 또한 User 리스트를 기록하기 이전시점의 flow 들에 대해서 Other 를 다시 한번 리스트에 추가하는 작업을 수행한다. 위 단계에서 분류된 각각의 flow 들은 TMA 에서 수집된 데이터와 비교, 검증하여 분석률(Precision, Recall, F-measure)을 보여주며 TP(True Positive), FP(False Positive), FN(False Negative) 등으로 나타내어지게 된다.

5. 적용 및 검증결과

본 장에서는 제안한 해당 알고리즘의 정확도 (Precision, Recall)와 다른 논문에서 제안하는 Skype 트래픽 분류방법론과의 비교를 다루고자 한다. 다른 논문에서와 실험대상의 dataSet 자체가 다르고 계산하는 수식자체가 다르기 때문에 정확한 수치데이터 비교는 어렵지만 본 논문에서의 정답지 (TMA 기반 Data)와 학내 망 전체를 대상으로 실시간 트래픽 분류가 1 분마다 이루어지기 때문에 신뢰성은 어느 다른 방법론보다 높다고 생각한다.

<표 1> AdaBoost and C4.5 [7]

	AdaBoost		C4.5	
	DR	FP	DR	FP
UDP, TCP avg (Skype traffic 15%)				
Skype	0.956	0.20	0.96	0.05

<표 2> GA based Random Forest (RF) [8]

Classifier	Accuracy (%)
RF	96.22
SVM	62.75
C4.5	94.97

<표 1>과 <표 2>은 Machine Learning 기법을 이용한 분류 결과이다. <표 1>은 전체 트래픽 중 15%가 Skype 의 트래픽인 data set 을 실험하였으며 UDP 와 TCP 에 대해 각각 테스트 한 결과의 평균 accuracy 를 나타낸다. <표 2>는 Random Forest 기법을 이용한 분류 결과이다. 각각의 실험은 모두 training 을 위한 별도의 시간이 필요하나, <표 3>의 User Port 추출 방법은 별도의 training 시간도 필요 없으며 분류결과 또한 높은 정확도를 가지고 있다.

<표 3> Skype User Port 추출 기반 및 분류 결과

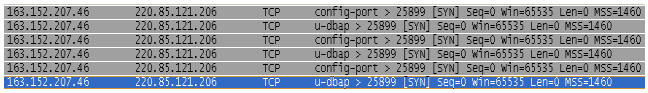
	09/02/02	09/02/05	09/02/16
TP (# of flow)	14812	13504	82199
FN (# of flow)	296	124	86
FP (# of flow)	2	0	4
Precision (%)	99.986	100.000	99.995
Recall (%)	98.040	99.090	99.895
F-Measure (%)	99.112	99.593	99.950

$$precision = \frac{TP}{TP+FP} \quad , \quad Recall = \frac{TP}{TP+FN}$$

$$F\text{-}measure = \frac{(1+\alpha)*precision*recall}{\alpha*precision+recall} \quad (\alpha : \text{alpha} = 0.8)$$

6. 문제점

본 논문에서 제안하는 방법론으로 분류할 수 없는 flow 들이 있다. (그림 5)에서의 FN 들은 호스트가 리스트에 등록되어 있지만 dstip 가 Other 리스트에도 없으며 TCP SYN 패킷 만 전송하는 경우이다. TMA 에서 Skype 의 Flow 라고 말하고 있지만 (그림 5)의 경우는 패킷의 특정바이트를 추정 할 수도 없으며 호스트의 포트 또한 5000 번 이하의 랜덤포트(OS 할당) 인 경우이다.



(그림 5) Wire Shark[9]에 나타난 SYN 패킷

또 한가지의 FN 종류로서는 TLS(Transport Layer Security) version 1 방식의 SSL(Secure Socket Layer) flow 들이다. 이 특정 flow 들은 데이터의 길이도 알 수 없으며 패킷의 내용조차 볼 수 없으므로 분석이 불가능한 경우이다. 나머지 FP 의 경우에 대해서는 signal flow 에 대해서 특정 위치의 payload 크기를 보고 후보자 모집과 그것에 대한 Inbound, Outbound 트래픽을 모두 보고 Skype 트래픽 이라 판단하지만 다른 특정 P2P application (특히 torrent)에서 FP 가 발생하는 상황이 있다.

7. 결론 및 향후 연구과제

특정 방법론(Signature-based, Port-based)만을 사용하지 않고 Skype 트래픽의 패턴을 분석하여 동적 port 를 찾아내어 사용자의 ip 와 port 의 리스트를 만들고, Signature-based, Port-based 방법론들을 혼합하여 트래픽들을 분류 할 수 있다는 점에 의미를 두었다.

특정 P2P application 의 FP 들이 발생 하였지만 최대 Precision(100.0%), Recall(99.8%) 이루는 만족할만한 결과를 내었다. 본 논문에서는 Skype 의 패턴 및 특성이 휴리스틱하게 분석되었고 그에 따른 알고리즘이 제안되었다, 하지만 향후 연구과제로서 p2p 기반의 동적 port 를 사용하는 특정 application 의 트래픽을 분류 할 수 있는 사용자의 port 를 잡아내는 시스템의 개발이 필요하다.

참고문헌

[1] Skype, Web Site, http://www.skype.com

[2] S. A. Baset, H. Schulzrinne, “An analysis of the skype peer-to-peer internet telephony protocol”, In Proceedings of IEEE INFOCOM'06.

[3] Dario Rossi, Marco Mellia, Michela Meo,“Understanding Skype signaling”, Computer Networks 53 (2009) 130–140.

[4] Yanfeng Yu, Dadi Liu, Jian Li, Changxiang Shen, “Traffic Identification and Overlay Measurement of Skype”, Computational Intelligence and Security, 2006 International Conference on, 3-6 Nov. 2006.

[5] 윤성호, 노현구, 김명섭, "TMA(Traffic Measurement Agent)를 이용한 인터넷 응용 트래픽 분류", 통신학회 하계종합학술발표회, 라마다플라자호텔, Jul. 2-4, 2008, pp.618.

[6] 박상훈, 박진완, 김명섭, "Flow 기반 실시간 트래픽 수집 및 분석 시스템", 정보처리학회 추계학술대회, 목포대학교, 전주, Nov. 9-10, 2007, pp. 1061.

[7] Angevine, D. Zincir-Heywood, A.N. “A Preliminary Investigation of Skype Traffic Classification Using a Minimalist Feature Set”, Availability, Reliability and Security, 2008. ARES 08. Third International Conference on 4-7 March 2008 Page(s):1075 – 1079.

[8] Li Jun Zhang Shunyi Xuan Ye Sun Yanfei, “Identifying Skype Traffic by Random Forest”, Wireless Communications, Networking and Mobile Computing, 2007.WiCom 2007. International Conference on, Publication Date: 21-25 Sept. 2007.

[9] Wire Shark, Web Site, http://www.wireshark.org/.