ISSN(Print): 2287-5891

http://dx.doi.org/10.3745/KTCCS.2012.1.1.055

Feature Selection with PCA based on DNS Query for Malicious Domain Classification

Sun-Hee Lim† · Jaeik Cho†† · Jong-Hyun Kim† · Lee, Byung Gil†††

ABSTRACT

Recent botnets are widely using the DNS services at the connection of C&C server in order to evade botnet's detection. It is necessary to study on DNS analysis in order to counteract anomaly-based technique using the DNS. This paper studies collection of DNS traffic for experimental data and supervised learning for DNS traffic-based malicious domain classification such as query of domain name corresponding to C&C server from zombies. Especially, this paper would aim to determine significant features of DNS-based classification system for malicious domain extraction by the Principal Component Analysis(PCA).

Keywords : DNS, Botnet, DDoS, Malware Detection, C&C Server

1. 서 론

현재 사이버 공간에서는 제3자의 개인정보를 획득하여 악용하거나, 불특정 다수를 향해 음란, 광고 메일을 유포하여 금전적 이익을 보거나, 또는 경쟁사의 정보화 기기의 서비스를 못하게 하는 등의 인터넷 상의 위협 요인들이 산재해 있다. 이러한, 전역네트워크에서의 대규모 공격들을 위해 봇넷(Botnet)을 형성하고 있다.

봇넷은 악성코드 봇(Bot)에 감염된 다수의 컴퓨터들이 네트워크로 연결되어 있는 형태로서 분산서비스거부공격(Distributed Denial of Service), 부정클릭(Click Fraudulence), 스팸(Spamming), 개인정보 유출(Identity Theft)와 같은 다양한 공격들이 60%이상 봇넷을 통하여 이루어지고 있고, 이러한 공격들은 단순한 사고가 아닌 자산 손실과 같은 경제적으로도 위협이 되고 있다.

봇넷의 구조는 Fig. 1과 같이 실제 감염된 봇들에게 공격 명령을 내리는 봇마스터(Bot Master), 명령 및 제어 메시지를 전달하는 C&C 서버, 악성코드 봇(Bot)에 감염되어 공격자로부터 제어를 받는 좀비(Zombie) PC들로 구성된다[1].

초기의 봇넷은 채팅 프로그램에서 많이 사용되었던 IRC(Internet Relay Chat) 프로토콜을 이용한 중앙집중형 구조의 봇넷 구조에서 웹 프로토콜 HTTP 봇넷 구조 혹은 분산형 명령/제어 방식, 즉 P2P 구조의 봇넷 및 하이브리드(Hybrid) 구조의 봇넷으로 진화되고 있다. 중앙집중형 구조의 봇넷에서는 하나의 C&C서버가 다수의 좀비들을 명령/제어하기 때문에 C&C탐지가 용이하고, 이러한 C&C서버가 탐지 및 차단 대응되면 다수개의 좀비들이 유실됨으로써 공격자에게는 커다란 손실을 가지게 된다. IRC 프로토콜과 같은 특정 프로토콜을 적용한 봇넷은 탐지 및 대응이 용이해지기 때문에 공격자들은 이러한 대응 기술들을 피하기 위해 비정상 행위분류가 어려운 HTTP 웹 프로토콜을 적용하거나, 모든 봇들이 C&C서버 역할이 가능한 분산형 구조 방식으로 탐지 및 대응이 어렵도록 진화되고 있다.

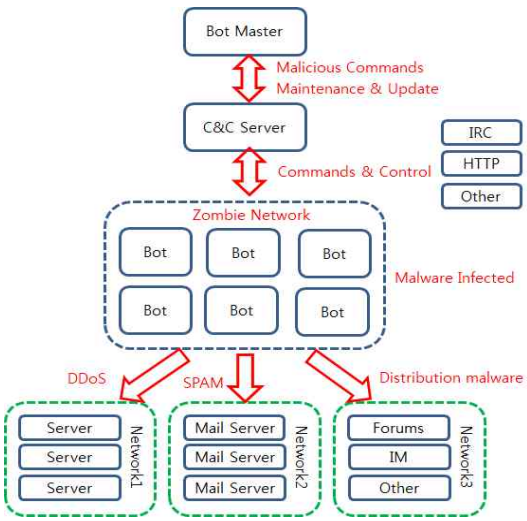

Fig. 1. Botnet Architecture and Attack Type

최근 봇넷은 프로토콜 및 구조적으로 진화했을 뿐만 아니라, C&C 서버 접속시 탐지 회피를 위하여 DNS 서비스를 이용하고 있다. 봇넷은 좀비 PC와 C&C서버가 명령/제어 메시지를 송수신 받기 위해 네트워크로 연결되어 있는 구조로써 기존의 C&C서버 IP주소를 직접 악성코드 봇 프로그램에 내포되어 접근하는 방법은 쉽게 탐지 및 차단이 가능하다. 이러한 대응을 회피하기 위해 도메인 주소를 통해 C&C서버

※ 본 연구는 방송통신위원회 정보보호 원천기술개발 사업의 일환으로 수행하였음.[2012/10912-06002, 전역적 협력기반의 통합보안제어 시스템 개발]
† 정 회 원 : 한국전자통신연구원 선임연구원
†† 정 회 원 : 삼성전자 연구원
††† 정 회 원 : 한국전자통신연구원 융합보안연구팀 팀장(책임)
논문접수 : 2012년 5월 16일
수 정 일 : 1차 2012년 8월 13일
심사완료 : 2012년 8월 13일
* Corresponding Author : Jaeik Cho(chojaeik@korea.ac.kr)

로 접근 방법으로 DNS 서비스를 적용하고, IP주소를 계속적으로 변경시키는 DDNS(Dynamic DNS) 혹은 Fast-Flux 기술을 적용함으로써 탐지 및 대응을 더욱 어렵게 하고 있다.

본 논문에서는 다양하게 진화된 봇넷 기술들 중 도메인 네임 서비스를 사용하는 봇넷 기술에 대한 탐지를 위해 DNS 쿼리 수집 및 분석(Passive Analysis)을 통한 정상적인 도메인 질의 트래픽과 좀비PC들이 C&C서버에 접속하기 위한 도메인 주소 질의 트래픽간의 특성(Characterization) 즉 분석성분(Feature) 도출을 목적으로 한다. DNS 트래픽의 여러 가지 정보에 정상 쿼리와 비정상 쿼리간의 특성을 나타내는 비중있는 정보를 선정하기 위한 방법을 제안한다. 제안하는 방법은 주성분 분석 PCA 방법을 통해 실제 DNS 트래픽 헤더 및 DNS 트래픽의 IP 패킷 헤더, UDP 헤더 정보에서 사용 가능한 모든 속성을 분석한다. 본 논문에서는 캠퍼스망에서 3단계에 걸친 실험을 통해 수집한 데이터를 이용하여 분석하였다.

본 논문은 2장에서 선행 연구된 봇넷 탐지 기술 및 DNS 서비스를 적용한 봇넷 기술에 대해 기술하고, 3장에서는 비정상도메인을 분류하기 위해 DNS트래픽 수집 및 주성분분석기술 기반의 분석성분을 추출에 대해 연구한다.

2. 봇넷 탐지 기술

2.1 선행 봇넷 기술 연구

1) DNS트래픽 기반의 그룹행위 분석을 통한 봇넷 탐지 기술[2]

봇넷에서 발생되는 DNS 트래픽과 정상 사용자들의 DNS 트래픽과는 호스트에 접근하는 소스 IP, 접속 활동 및 패턴, DNS 타입으로 특성화하고 있다. 봇넷을 탐지하기 위해 호스트에게 접근하려는 클라이언트는 특정 패턴을 가지는 소스 IP주소 특성을 가지며, 순간적인 데이터양의 증가, DDNS 사용을 비정상행위의 특성화로 분류하고 있다.

정상적인 사용자들은 DNS 트래픽 패턴으로 특정 도메인 주소에 접근하는 소스 IP의 패턴이 정의되어지지 않는 랜덤성을 가지며 연속적인 행위로서 행위를 그룹화하기 어렵다는 특징으로 봇넷의 비정상행위와 분류 기술한다.

2) BotSniffer/BotMiner[3][4]

BotSniffer은 중앙집중형 봇넷 구조에서의 C&C 행위를 탐지하기 위한 기술이다. 중앙집중형 봇넷은 주로 IRC 혹은 HTTP 프로토콜을 사용함으로써 BotSniffer 기술에서는 IRC 프로토콜과 HTTP 프로토콜의 특징을 기반으로 탐지한다. 하지만, 특정 프로토콜을 기반으로 하는 봇넷 탐지 기술에서는 그 외 프로토콜 및 구조를 가지는 봇넷을 탐지하기 어렵다는 한계를 가지고 있다.

BotMiner는 BotSniffer기술의 한계점을 해결하기 위해 프로토콜 및 토폴로지에 의존하기 않고 데이터 마이닝 기술을 통해 봇넷의 C&C를 탐지하는 기술을 제안한다. BotMiner의 선행기술에서는 “who is doing what” 호스트의 수행 행위 측면과 “who is talking to whom” 네트워크 유사 패턴 측면을 기반으로 군집화(Clustering)한다. 하지만, 네트워크 패턴 측면에서의 봇넷 유형으로 일시적인 데이터량 증가의 특징을 중심으로 네트워크 플로우 기반의 분석 방법으로만 적용한다는 한계를 가지고 있다.

2.2 DNS를 이용한 C&C 서버 접근 기술

봇넷 기술들은 탐지 및 대응을 어렵게 하기 위해 다양한 방법으로 구조 및 프로토콜 측면에서 진화되고 있고, 최근에는 C&C서버와 좀비 PC간의 DNS서비스를 사용하여 봇넷을 형성하고, 명령/제어 메시지를 전달하고 있다. 다음은 봇넷에서 DNS서비스를 사용하는 방법으로 두 가지가 있다.

1) DNS서비스로 C&C서버의 도메인네임 질의

좀비PC로 감염시키기 위한 악성코드 봇 프로그램에 C&C 서버의 주소를 직접 IP주소를 하드코딩하지 않고, 도메인 주소를 할당하여 DNS서비스를 통해 C&C서버의 도메인 주소에 대한 IP주소로 연결하는 방법이다. IP주소의 하드코딩은 IP주소의 특정 포맷에 대한 탐지가 쉬워지고, 연속되는 비정상 행위 패턴으로 인한 침입탐지시스템에서의 탐지 및 대응이 가능하다. 이러한 문제점을 해결하기 위해 진화된 봇넷 기술에서는 직접 IP 주소를 하드코딩하지 않고, 도메인 주소를 할당하여 DNS서비스를 악용하고 있다. 더불어, 도메인네임에 대응하는 IP주소가 계속 변경되는 DDNS(Dynamic DNS) 서비스 혹은 Fast-Flux기법을 악용함으로써 봇넷 탐지 및 대응이 어려워지고 있다[5][6].

2) DNS트래픽으로 명령/제어 전송

Feederbot은 DNS데이터에 명령/제어 메시지를 포함하여 전송한다. 이러한 방법은 명령/제어 메시지 전달 채널로 DNS 데이터를 사용함으로써 비정상 DNS 데이터를 구분하기 어려울 뿐만 아니라, DNS 데이터를 암호화하였을 경우는 탐지가 어려워진다. 또한, DNS서비스는 사용자들에게 인터넷 서비스로서 중요한 역할을 하기 때문에 DNS 데이터를 침입탐지 시스템에서 차단하기는 어렵다[7].

2.3 봇넷에서 DNS서비스 적용 구조

Fig. 2와 같이 악성코드인 봇 프로그램에 감염된 좀비PC들은 봇에 코드화된 도메인 주소를 DNS서버에 질의하고 DNS서버는 도메인 주소에 대한 IP 주소를 좀비PC들에게 응답한다. DNS서버는 좀비PC들의 비정상 도메인 주소에 대한 질의임을 판단하지 못하고 정상 서비스를 제공한다. C&C서버의 IP주소를 전달받은 좀비PC들은 IP주소가 캐쉬되어 있는 동안 C&C서버와 계속적으로 통신이 가능하다. 더불어, 최근의 봇넷 기술은 이러한 행위들이 네트워크 운영자에게 쉽게 탐지되지 못하도록 DNS 쿼리를 정상 사용자와 유사한 패턴의 형태로 랜덤하게 질의한다거나 DDNS 서비스 혹은 FastFlux 기술을 이용하여 도메인 주소에 IP주소

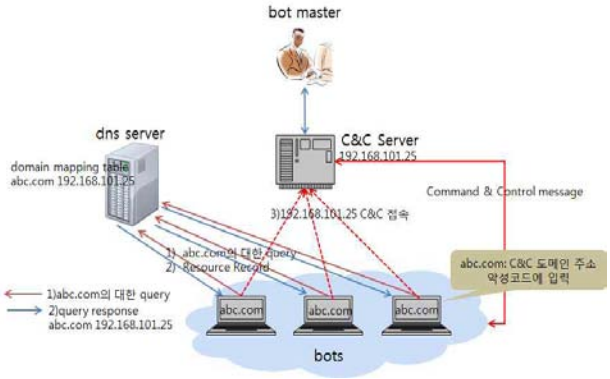

Fig. 2. Botnet architecture based on DNS Service

를 동적으로 빈번하게 변경시키는 서비스를 적용함으로써 탐지 및 대응을 어렵게 하고 있다[8][9].

3. 비정상 도메인을 분류하기 위한 DNS쿼리 수집 및 주성분 분석 기술

본 논문에서는 DNS서비스를 이용하여 좀비PC가 C&C서버에 접속하는 봇넷 기술에서 비정상 도메인 C&C 서버를 추출하기 위해 정상 DNS 트래픽과 비정상 즉 좀비 PC들이 질의하는 DNS 트래픽 분류를 목적으로 한다. 비정상 DNS 트래픽을 분류하기 위해 본 논문에서는 DNS 트래픽 수집 및 주성분 분석 기술인 PCA 기술을 기반으로 비정상 DNS 트래픽의 주성분을 추출한다.

3.1 DNS패킷 수집

비정상도메인을 탐지하기 위해 본 논문에서는 3가지 실험을 통해 DNS 데이터를 수집하였다.

1차 실험에서는 외부네트워크와의 연결을 단절시킨 상태에서 악성코드를 실행시키고 내부 네트워크에 수집되는 패킷을 수집하였다. 2차 실험에서는 외부 네트워크와 연결된 상태에서 악성코드를 실행시키고 외부의 통제되지 않은 C&C 서버와 통신하는 패킷을 수집하였다. 이 실험에서는 좀비 PC가 외부의 통제되지 않은 C&C서버에게 감염되었음을 알리는 패킷을 송신하고, 이 패킷을 수신받은 C&C서버가 좀비PC들을 자신들의 봇넷에 포함시킬 것을 고려하여 실험하였다. 3차 실험에서는 외부 네트워크와 연결된 상태에서 통제하에 있는 C&C 서버와 악성코드를 실행시키고 C&C 서버에서 좀비에게 공격 명령을 내리고 외부로 공격이 이루어지는 패킷을 수집하였다.

실험 결과의 수집 데이터양은 Table 1 과 같다.

1차 실험에서는 외부네트워크와의 연결이 단절되어 DNS 서버의 질의가 실험망 이상의 상의 DNS 서버로 질의되지 못하는 문제점을 가지게 되었고, 2차 실험에서 외부의 통제되지 않은 C&C서버의 접근이 거의 미비하였다. 본 실험에서의 3단계에 걸친 실험에서 3차 실험만이 DNS 트래픽 분석을 위해 효과적으로 이용가능 하였다.

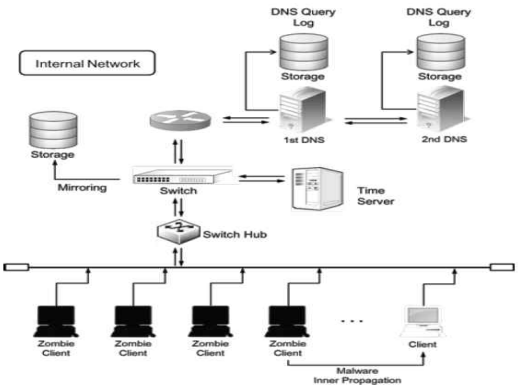

Fig. 3. closed network testbed

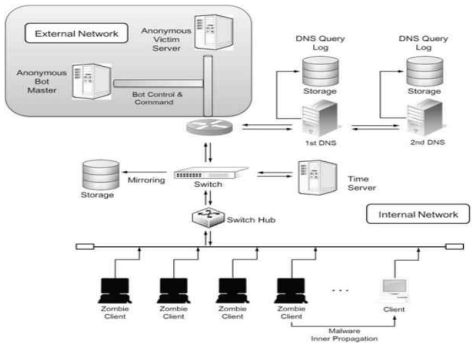

Fig. 4. uncontrolled C&C server in open testbed

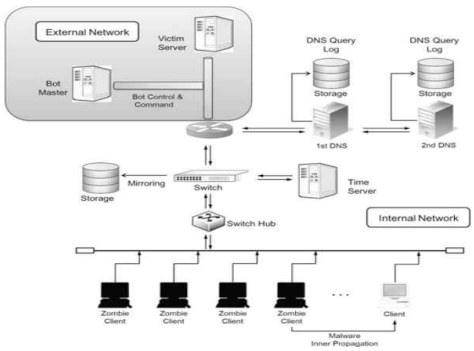

Fig. 5. controlled C&C server in open testbed

Table 1. Collected DNS Traffic

	1차 실험 패킷 수	2차 실험 패킷 수	3차 실험 패킷 수
1st DNS	633,170	839,718	2,421,275
2nd DNS	303,839	41,623	2,398,922

3.2 파라미터 구성을 위한 학습 및 실험 데이터 구성

파라미터 구성을 위하여 본 논문에서는 수집된 폐쇄망 데이터와 정상 DNS 쿼리 패킷 데이터를 혼용하였다. 수집된 폐쇄망 실험 데이터에서는 악성 데이터 중 DNS 패킷 데이터가 아닌 라우팅 관련 패킷, 이더넷 장치의 브로드캐스팅 메시지를 제외한 데이터에서 1차 DNS로 유입되는 데이터를 선별 하였다. 선별된 데이터 중 커널, 혹은 전송에서 발생한

프라그먼트 패킷은 소거하였으며, 학습 단계의 공격 데이터로 선분류 하였다.

정상 데이터의 경우 일반적인 SSL 암호화 통신 및 허가된 클라이언트의 호출에만 응답하는 구조를 이용하는 Skype사의 DNS 데이터를 이용하였다. 이러한 Skype사의 DNS 데이터는 봇넷에 감염되지 않은 거의 순수한 DNS 트래픽이다. 일반적인 송신/수신의 데이터에서 DNS 쿼리 데이터인 DNS 단에 수신되는 패킷만을 구분하여 사용하였으며, 마찬가지 커널 및 전송에서 발생한 프라그먼트 패킷은 소거하였다.

전체 공격 및 정상 데이터에서 분석의 정확도를 기대하기 위하여 10회 라운딩 학습을 시도 하고, 10회 실험을 반복하였으며, 도출된 결과 오차범위 3%이내 검증으로 정확도를 확인하였다.

학습과 실험에 사용된 데이터는 데이터 분류에서 주로 사용하는 비율인 7:3으로 임의 선택 되었으며, 학습과 실험 두 단계의 데이터 또한 공격과 정상 비율을 임의 선택으로 7:3 비를 구성하였다.

3.3 비정상 도메인 분류를 위한 DNS쿼리 기반의 주성분 분석기술

1) 주성분분석(Principal Components Analysis)

주성분 분석은 고차원 데이터로부터 데이터의 구조를 밝히거나, 데이터의 차원을 낮추는데 많이 이용되는 다변량 통계 분석 방법이다. 이는 상관행렬(Correlation Matrix)의 고유벡터(Eigenvector)를 계산하여 높은 값을 찾아내는 방법이다.

주성분분석은 데이터를 표현하는데 최적으로 사용될 수 있는 k개의 n차원 직교벡터(Orthogonal Vector)를 찾는다(단, k ≤ n), 즉, 입력 데이터를 분산이 최대가 되는 축으로 변환하는 것으로, 이 새로운 차원에서의 데이터의 벡터들을 주성분(Principal Components)이라고 한다. 원본 데이터가 매우 작은 차원의 공간으로 투영되어지므로 데이터의 차원 축소가 이루어진다[10].

이때, 분산이 작은 성분들을 제거함으로써 데이터의 차원(Dimensionality Reduction)을 줄이는 동시에 분석성분 추출(Feature Selection), 데이터에 포함되어 있는 잡음을 제거(Noise Filter)할 수 있다.

주성분분석을 구하기 위해 입력 데이터를 데이터 셋(DataSet)하고, 이에 대한 상관 행렬(Correlation Matrix)을 구한다.

$$\text{확률벡터 } X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \text{ Xi : 확률 변수}$$

$$X\text{의 상관행렬 } \rho = \begin{pmatrix} \rho_{1,1}\ \rho_{1,2} \cdots \rho_{1,n} \\ \rho_{2,1}\ \rho_{2,2} \cdots \rho_{2,n} \\ \ddots \\ \rho_{n,1}\ \rho_{n,2} \cdots \rho_{n,n} \end{pmatrix},$$

$$\rho_{i,j} = \frac{cov(X_i, X_j)}{\sigma_i \sigma_j} = \frac{E[(X_i - \mu_i)(X_j - \mu_j)]}{\sigma_i \sigma_j},$$

cov : 공분산벡터, E: 기대값
σ_i, σ_j : X_i, X_j의표준편차, μ_i, μ_j : X_i, X_j의중간값

이러한, 상관행렬의 고유값(Eigenvalue), 고유벡터(Eigen vector)값을 계산하여 높은 차순의 고유값을 찾는다.

2) DNS 트래픽의 성분 추출(Feature Selection)

주성분 분석은 대응되는 성분을 이용하여 데이터의 주성분 및 성분의 성향을 분석할 수 있는 방법이다. 본 논문에서는 구성된 학습 및 실험 데이터를 수량화하고 수량화된 데이터를 이용하여 주성분 분석을 시도하였다.

본 논문에서는 DNS 트래픽 기반의 정상 쿼리와 비정상 쿼리의 성향을 분석하기 위해 DNS 트래픽에서 획득할 수 있는 최대한의 정보들 DNS 쿼리 헤더 데이터 및 DNS 쿼리 데이터의 IP 패킷 정보들을 수집하여 Table 2와 같이 DNS 쿼리 및 IP패킷 속성에 해당되는 모든 분석성분들을 테스트하였다.

Table 2. Collection data for feature

번호	계층	속성벡터	설명
1	IP	Diff	이전 DNS쿼리와 시간차
2	IP	Version_Length	IP패킷 버전 및 헤더길이
3	IP	ToS	IP 패킷의 Type of Service
4	IP	Total_LEN	ip 데이터그램의 총 길이
5	IP	ID	송신자와 수신자간의 식별자
6	IP	Flag	More Bit Fragmentation Bit Reserved
7	IP	TTL	Time to live, 홉 카운터값 사용
8	IP	PROTOCOL	Encapsulated data의 프로토콜을 식별
9	IP	CRC	헤더 오류 체크
10	UDP	SPORT	소스 포트
11	UDP	DPORT	목적 포트
12	UDP	LENGTH	udp 헤더 및 데이터 길이
13	UDP	CRC	UDP checksum
14	DNS	QR	0:query 1:response
15	DNS	OPCODE	0:standard query/1:inverse query/2:server status/3:reserved
16	DNS	AA (Authoritative Answer)	도메인 주소에 대해 인증된 서버로부터 응답임을 표시
17	DNS	TC(truncation)	전송채널에서 데이터 길이가 길어 절단되어 전송되었음을 지칭
18	DNS	RD(recursion desired)	재귀 요구 플래그

19	DNS	RA(recursion available)	재귀 유효 플래그
20	DNS	Z	reserved
21	DNS	RCODE	응답 오류 값
22	DNS	ID	DNS헤더 송수신자 식별자
23	DNS	QDCOUNT	질문의 수
24	DNS	ANCOUNT	응답 수
25	DNS	NSCOUNT	authority record에서 네임서버의 응답 수
26	DNS	ARCOUNT	additional record에서 응답 수
27	DNS	QTYPE	쿼리 타입
28	DNS	QCLASS	쿼리 클래스

3) 실험 결과

Table 3의 결과에서와 같이 UDP 패킷의 소스포트, CRC 값, 패킷 길이와 IP 패킷의 플래그값, 식별자, CRC 값, 패킷길이 및 TTL값, DNS 패킷에서는 식별자 속성들이 파라미터로 통계적 유의성이 보장되는 분류 기준이 정리되며, 비정상 도메인을 분류하기 위한 효과적인 파라미터 구성으로 참조된다. 고유값이 높은 차순의 성분들이 가지고 있는 데이터가 비정상 도메인의 성향을 나타내는 주성분이 된다.

Table 3. Principal Components Analysis

고유값	번호	속성벡터
63933	10	UDP/SPORT
63933	6	IP/Flag
63933	5	IP/ID
63933	13	UDP/CRC
63933	22	DNS/ID
63751.90431	9	IP/CRC
60171.41127	7	IP/TTL
59375.56591	12	UDP/LENGTH
59375.56591	4	IP/Total_LEN
35581.68459	1	IP/Diff
34406.71387	27	DNS/QTYPE
0	그외 값들	

4. 결론 및 향후 연구

최근의 C&C 서버와 좀비PC 간의 네트워크 접속을 위해 DNS 서비스를 이용한 봇넷은 기존의 선행 기술로는 탐지 및 대응이 어렵다. 도메인네임을 이용한 봇넷을 탐지하기 위해 DNS 트래픽을 기반으로 정상도메인과 비정상 도메인을 분류하기 위한 효과적인 분류 기준으로 작용할 수 있는 분석성분(feature)이 요구된다.

본 논문에서는 주성분분석(Principal Component Analysis)을 적용한 수집된 DNS 트래픽을 기반으로 실험한 결과 UDP 패킷의 소스포트, CRC 값, 패킷 길이와 IP 패킷의 플래그값, 식별자, CRC 값, 패킷길이, TTL값, DNS 패킷에서는 식별자 성분들이 비정상도메인을 분류하기 위한 주요 성분으로 도출되었다. 이러한 연구 결과는 기존의 시그니처 기반의 봇넷 탐지 기술과 다르게 DNS 서비스를 사용하는 봇넷을 탐지하기 위해 DNS 트래픽의 가능한 모든 정보들을 수량화하여 통계적 기법으로 도출된 주성분들을 이용한 탐지가 가능하다.

향후 연구에서는 도출된 분석성분들을 기반으로 분류알고리즘을 통하여 정상 도메인과 비정상 도메인 즉 C&C서버 분류(Classification) 기술에 대해 연구하고자 한다.

참 고 문 헌

[1] *http://en.wikipedia.org/wiki/Botnet*

[2] H. Choi, H. Lee, H. Lee,H. Kim, *"Botnet detection by monitoring group activities in DNS traffic"*, in Computer and Information Technology, 2007. CIT 2007. 7th IEEE International Conference on, 2007, pp.715-720.

[3] G. Gu, J. Zhang,W. Lee, *"BotSniffer: Detecting botnet command and control channels in network traffic"*, in Proceedings of the 15th Annual Network and Distributed System Security Symposium (NDSS'08), 2008.

[4] G. Gu, R. Perdisci, J. Zhang,W. Lee, *"BotMiner: clustering analysis of network traffic for protocol-and structure-independent botnet detection"*, in Proceedings of the 17th conference on Security symposium, 2008, pp.139-154.

[5] J. Liu, Y. Xiao, K. Ghaboosi, H. Deng,J. Zhang, *"Botnet: Classification, attacks, detection, tracing, and preventive measures"*, in EURASIP journal on wireless communications and networking, 2009, Vol.2009, pp.1184-1187.

[6] R. Villamarín-Salomón,J. C. Brustoloni, *"Identifying botnets using anomaly detection techniques applied to DNS traffic"*, in Consumer Communications and Networking Conference, 2008. CCNC 2008. 5th IEEE, 2008, pp.476-481.

[7] J. Dietrich, C. Rossow, F. Freiling, "On Botnets that use DNS for Command and Control".

[8] R. Villamarín-Salomón,J. C. Brustoloni, *"Bayesian bot detection based on DNS traffic similarity"*, in Proceedings of the 2009 ACM symposium on Applied Computing, 2009, pp.2035-2041.

[9] H. Tu, Z. Li,B. Liu, *"Detecting botnets by analyzing DNS traffic"*, Intelligence and Security Informatics, pp.323-324, 2007.

[10] Jiawei Han and Micheline Kamber, "*Data Mining 2nd Edition*", 2007.

[11] L. Bilge, E. Kirda, C. Kruegel,M. Balduzzi, *"EXPOSURE: Finding malicious domains using passive dns analysis"*, Proceedings of the Annual Network and Distributed System Security (NDSS)(February 2011).

임 선 희

e-mail : capsunny@etri.re.kr
1999년 고려대학교 컴퓨터학과(학사)
2005년 고려대학교 정보보호학과(공학석사)
2010년 고려대학교 정보보호학과(공학박사)
1999년~2002년 한화 정보통신 연구원
2010년~현 재 한국전자통신연구원 선임연구원
관심분야 : 무선이동통신보안, 정보보호, 사이버보안, 융합보안기술

김 종 현

e-mail : jhk@etri.re.kr
2000년 오클라호마주립대 컴퓨터학과(공학석사)
2005년 오클라호마주립대 컴퓨터학과(공학박사)
1995년~1997년 삼성전자 연구원
2000년~2001년 삼성SDS 시스템컨설턴트
2005년~현 재 한국전자통신연구원 선임연구원
관심분야 : 정보보호, 사이버보안, 역추적기술

조 재 익

e-mail : chojaeik@korea.ac.kr
2005년 동국대학교 컴퓨터학과(학사)
2008년 고려대학교 정보보호학과(공학석사)
2012년 고려대학교 정보보호학과(공학박사)
2009년~2011년 llinois Institute of technology 선임연구원
2012년~현 재 삼성전자 연구원
관심분야 : 네트워크 모델링, 패턴인식

이 병 길

e-mail : bglee@etri.re.kr
1991년 경북대학교 전자공학과(학사)
1993년 경북대학교 전자공학과(공학석사)
2001년 경북대학교 전자공학과(공학박사)
1993년~2001년 LG데이콤종합연구소(현 유플러스 종합연구소) 선임연구원
2001년~현 재 한국전자통신연구원 융합보안연구팀 팀장(책임)
관심분야 : 융합보안기술, 사이버보안기술, 홈랜드보안기술 등

비정상도메인 분류를 위한 DNS 쿼리 기반의 주성분 분석을 이용한 성분추출

임 선 희[†] · 조 재 익[††] · 김 종 현[†] · 이 병 길[†††]

요 약

최근 봇넷(Botnet)은 탐지 기술을 피하기 위하여 C&C(Command and Control)서버 접속시 DNS(Domain Name System) 서비스를 이용하고 있다. DNS 서비스를 이용한 비정상 행위에 대응하기 위해서 DNS 트래픽 기반의 분석 연구가 필요하다. 본 논문에서는 좀비PC의 C&C서버 도메인주소 질의와 같은 DNS트래픽 기반의 비정상 도메인 분류(Classification)를 위해서 DNS트래픽 수집 및 지도학습(Supervised Learning)에 대해 연구한다. 특히, 본 논문에서는 PCA(Principal Component Analysis) 주성분분석 기술을 통해 DNS 기반의 분류시스템에서의 효과적인 분석 성분들을 구성할 수 있다.

키워드 : DNS, Botnet, DDoS, Malware Detection, C&C Server