시뮬레이션 환경 구축을 위한 패킷 분포에 따른 네트워크 데이터 셋 구성 방안

조재익, 구본현, 이민수, 문종섭
고려대학교 정보보호대학원
{chojaeik, koo191, leesle, jsmoon}@korea.ac.kr

A Study on the Dataset Construction for Network Simulation base on Packet Distribution

Jaeik Cho, Bonhyun Koo, Minsoo Lee, Jongsub Moon
GSIS/CIST, Korea University

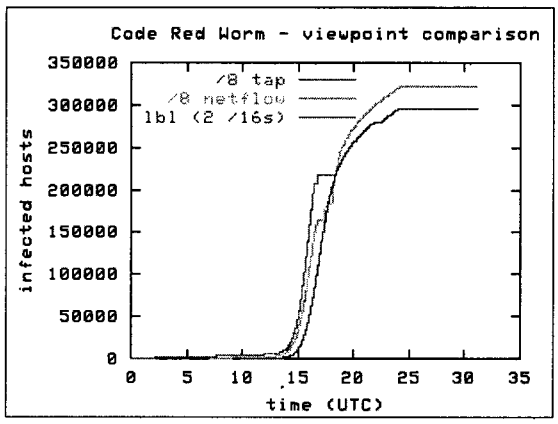
요 약

국내에는 많은 수의 네트워크 망 제공 업체로부터 고속 네트워크가 제공되고 있다. 이러한 네트워크 망에서 또한 많은 침입 탐지 시스템을 필요로 하고 있고 또한 많은 새로운 웜 바이러스의 출현에 따른 연구도 필요로 하고 있다. 그러나 현재 실정에 맞는 네트워크 데이터 셋이 구성되어 있지 않고 이러한 문제점으로 하여 정확한 침입 탐지 혹은 웜 바이러스의 효과적인 탐지와 차단에서 어려움이 있다. 이러한 문제를 해결하기 위해 본 논문에서는 실제 환경과 흡사한 데이터 셋 구성을 위한 방안에 대해서 제안 한다.

1. 서 론

현재 우리나라는 대부분의 가정에서 고속 네트워크를 사용하고 있으며 또한 회사, 관공서에서도 마찬가지로 사용하고 있다. 고속 네트워크로 국가 전체가 하나의 통합 망 전체로 구성되어 있으며 이는 많은 부가가치를 창출하고 있다. 그러나 국가 전체가 네트워크로 연결 되었을 때의 문제점들 또한 많이 있다. 불과 몇 해 전까지만 해도 악의적인 해커가 그 대표적인 문제점이었으나 근래에 들어 네트워크의 단절, 개인 정보의 불법 수집 등을 목적으로 확산되는 웜 바이러스가 출현하고 있다. 이러한 악의적인 목적으로 사용되는 네트워크의 침입을 막기 위해 대부분의 네트워크 망에서는 침입 탐지 시스템을 사용하고 있다. 이러한 침입 탐지 시스템은 여러 종류가 있으나 그 기반에는 네트워크에 대한 충분한 실험과 연구가 뒷받침 되어야 한다.

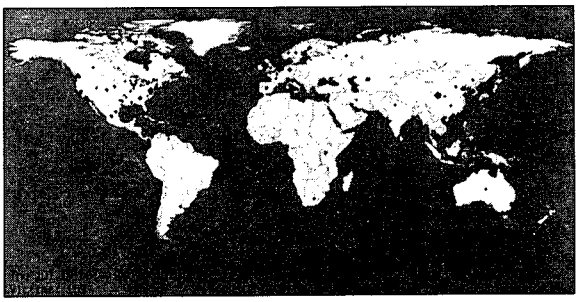
침입 탐지 시스템에서는 대부분 미리 정의된 셋을 이용하여 해당 셋과 동일하거나 흡사한 경우 탐지하는 방법을 사용하고 있다. 이러한 침입 탐지에서도 마찬가지 어떠한 악의적인 코드나 패킷을 발견 후 해당 패킷의 시그너쳐를 입력하는 방법을 사용한다. 웜 바이러스의 경우는 그 정의된 셋을 구성하기가 힘들다. 웜 바이러스의 실행 파일이나 혹은 전달되는 시그너쳐를 구성하여 셋을 구성하였다 하더라도 새로운 웜의 경우 그 시그너쳐를 생성하여 배포하는 시간 보다가 확산 속도가 월등히 빨라 그 대처를 하기가 힘든 실정이다. 이는 1.25대란에서도 그 심각성이 증명되었다. 과거의 대부분의 바이러스와 침입은 분석과 탐색 후 침입 탐지 시스템에 적용하는 시간이 확산 속도 보다가 빠를 수 있었으나 근래에 와서는 그 속도를 수용할 수 가 없는 속도가 되었다.



<그림 1. Code-red 웜 확산 1 [1]>

또한 우리나라는 세계적으로도 거의 제일 많은 수의 인터넷 보급으로 인하여 그 사용량도 엄청나다. 이러한 이유로서 많은 바이러스의 공격과 많은 해커들의 공격에 취약하며 이러한 망의 발달에 미치지 못하는 침입 탐지와 그에 해당 하는 연구의 부족으로 많은 피해를 입고 있다. 이러한 점은 Code-red 바이러스의 전세계적인 피해에서

확산 초기에 우리나라가 감염되는 것만 보아도 알 수 있다.

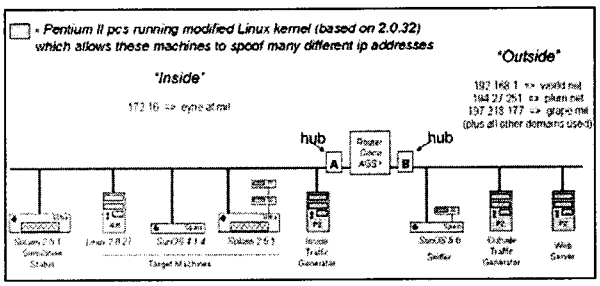


<그림 2. Code-red 의 확산 2 [2]>

이러한 문제점을 해결하기 위하여 본 논문에서는 침입 탐지 연구의 기반이 될 수 있는 데이터 셋 구성 방안을 한국 네트워크 환경에 맞게 제안 하고자 한다.

2. 관련 연구

기존 미국 MIT 의 Lincoln 연구실에서 침입 탐지 시스템 구성을 위한 DATA SET 연구를 많이 진행하고 진행 되었다. 대다수의 데이터 셋에 관련된 연구는 이 연구실의 자료를 많이 이용하고 있다. 그러나 데이터 셋의 직접적 이용에는 문제가 있다. 미국의 환경에서 일반적인 네트워크 환경을 대상으로 한 것이 아니라 일부분의 네트워크 트래픽을 이용하며 그것의 자료를 전체 트래픽이라 가정 하고 인공의 트래픽을 생성, 데이터 셋을 구성 하기도 하였기 때문이다. Lincoln 연구실의 실험용 시뮬레이션 환경은 다음과 같다.



<그림 3. Lincoln 연구실 시뮬레이션 환경 [3]>

일반적인 연구실이나 기업의 전산실인 경우 위와 같은 네트워크의 구성 방식으로 실험 된 데이터 셋을 사용하여도 차이가 없으나 웜이나 해킹의 경우 일반 사용자들에게 더 많은 영향을 끼치고 더 많이 확산 된다. 그래서 현재의 네트워크 상황을 반영한 시뮬레이션 환경을 구축하였다. 네트워크 시뮬레이션 환경을 구축하기 위해 반드시 필요한 것은 현재의 네트워크 상황을 분석하는 것이다. 현재의 네트워크 상황 분석은 위치와 시간에 따라 다르다. 그래서 일반적으로 테스트가 용이한 학교에서 보편적으로 사용하고 있는 네트워크 망에서 실시하였다.

2.1 네트워크 망 분석

네트워크 망 분석에서 네트워크 망에서 출현한 단계별 값을 비교해 보았다. 첫째, 학교 망에서의 일반적인 컴퓨터 사용자의 실제 사용에서의 패킷을 각 단계 (TCP, UDP, ICMP, ARP, NetBIOS, IPX)로 구분하여 패킷 수를 확인하고 그 비율을 비교 하였다. 둘째, 학교 망에서의 일반적인 컴퓨터 사용자의 실제 사용에서의 패킷을 각 단계로 구분하여 역시 확인하였다. 셋째, 사용자가 사용하지 않고 서버용 서비스를 하지 않는 일반적인 개인용 컴퓨터에서의 패킷을 확인하였다. 위와 같이 실험을 한 이유는 첫째, 개인의 사용 량에 따른 차이가 있을 수 있어서 사용하지 않는 윈도우와 사용자가 사용하는 윈도우를 구분하였다. 둘째, 윈도우에서는 다른 네트워크에서 유입되는 웜으로 인하여 그로 인한 트래픽이 발생 할 소지가 있었다. 그것을 배제 하기 위하여 또한 리눅스와 유닉스를 구분하였다. 물론 리눅스용 웜 바이러스가 있으나 학교 네트워크 망에서는 확인 되지 않는 상태였다. 그러나 테스트 해 본 결과 윈도우나 리눅스 양 운영체제 모두에 네트워크 트래픽은 영향을 끼치고 있었으며 이는 전체 테스트 대상이 된 컴퓨터에 비슷한 패킷 종류별 분포를 보이고 있었다. 이는 새로운 웜의 유입 시 기존 유입된 문제 패킷 또한 기반 대상으로 하여 시뮬레이션 하여야 해당 네트워크의 확산을 보다 흡사히 구성할 수 있기 때문이다.

[단위 : 천]

종류	TCP	UDP	ICMP	ARP	Net BIOS	IPX
1	156	202	2	1470	32	299
2	12	175	0.8	1255	18	250
3	17	183	0.8	1407	31	279

종류 1 : 교내 망에서의 사용중인 Windows OS
종류 2 : 교내 망에서의 미사용 중인 Windows OS
종류 3 : 교내 망에서의 사용중인 Linux OS

<표 1. 실제 네트워크 트래픽 비교>

2.2 데이터 셋 구성

데이터 셋을 구성하기 위해서는 2.1 에서 수집한 자료와 같은 자료의 확장 수집으로 가능하다. 수집된 자료는 국내에 반영되어 있는 네트워크의 성능과 그 성능에 영향을 끼

칠 수 있고 현재 활동 중인 악성 코드 (웜 바이러스를 포함)의 활동 트래픽도 역시 포함되어 있기 때문이다. 이러한 이유는 앞서 설명한 바와 같이 실제로 새로운 웜이나 새로운 네트워크 불안을 가져올 수 있는 요소 또한 이러한 환경 기반에서 영향을 끼치기 때문이다. 해당 네트워크 망을 가상의 환경으로 하기 위해 다음과 같이 위의 결과를 이용하여 네트워크 패킷의 구성을 재 구성 하였다.

수집된 종류의 패킷 개체 수 X (해당 환경의 OS / 전체 수)

의미로 보면 해당 네트워크 환경을 대변 할 수 있으며 실험을 하기 위해 실험 환경 자체의 환경에 적합성을 최대한 유지하기 위함이고 또한 이는 데이터 셋을 이용하여 일정 부분의 내부 망 만을 위하여 사용하려 할 때에 해당 망의 특성에 맞는 데이터 셋을 구성하기 위함이다.

실험은 30일을 기준으로 패킷 분포를 확인하였으며 그 중 일반적인 샘플 패킷 수를 수집하여 계산에 이용하였다. 교내 망은 현재의 분석으로 비추어 볼 때 웜 등으로 인한 패킷이 계속적인 네트워크의 소통을 방해하고 있으며 이는 새로운 웜의 출현 시 오히려 외부 망에서 보다가 확산 속도가 오히려 느릴 수 있음을 예상 할 수 있는 결과였다.

3. 결론 및 향후 과제

초기 수집된 데이터 패킷 2.1 에서 수집된 데이터를 전체 분석을 하면 비율에서는 특별한 차이는 보이지 않는다. 다만 일반 네트워크 보다가 비정상 적으로 ARP 의 수치가 높은 것을 보이고 있다. 이 평균 값을 2.3의 수식에 대입하면 다음과 같은 해당 실험 환경에 적합한 데이터 셋 구성에서의 패킷 비율을 구성 가능 하다. 처음 실험에서의 테스트는 1차, 사용하는 윈도우와 2차, 사용하는 리눅스의 패킷을 이용하여 계산 하였으며 이는 윈도우나 리눅스 양 운영체제 모두 실제 사용자가 사용하고 있다는 가정 하에 패킷의 비율을 조정하여야 시뮬레이션 구성에서도 실제 상황에 흡사한 구성을 할 수 있기 때문이다.

[Windows 평균, 전체 개체 수 : 98%] [단위 : 천]

종류	기존 값	변환 값	해당비율
TCP	156	152.88	7.2%
UDP	202	197.96	9.3%
ICMP	2	1.96	-
ARP	1470	1440.6	68%
NetBIOS	32	31.36	1.4%
IPX	299	293.02	13.8%

[리눅스 평균, 전체 개체 수 : 2%] [단위 : 천]

종류	기존 값	변환 값	해당비율
TCP	17	0.34	0.8%
UDP	183	3.66	9.5%
ICMP	0.8	0.016	-
ARP	1407	28.14	73.3%
NetBIOS	31	0.62	1.6%
IPX	279	5.58	14.5%

<표 2. 변환된 패킷 구성 분포>

결과적으로 시뮬레이션 환경을 구축하기 위한 네트워크 패킷의 구성은 다음과 같다.

[단위 : 천]

종류	Windows	리눅스	전체 합	비율
TCP	152.88	0.34	153.22	7.106%
UDP	197.96	3.66	201.62	9.350%
ICMP	1.96	0.016	1.976	0.091%
ARP	1440.6	28.14	1468.74	68.11%
NetBIOS	31.36	0.62	31.98	1.483%
IPX	293.02	5.58	298.6	13.84%

<표 3. 시뮬레이션 환경의 패킷 분포>

교내 망에서의 네트워크 패킷 분포를 반영하여 시뮬레이션을 구성하기 위해서의 결과는 위와 같았다. ARP 가 많은 것이 특징이었으며 이는 기존의 Lincoln 연구실의 일반 데이터 셋 구성을 이용 하는 것 보다가 특정 환경을 위한 구성에서는 더 정확할 수 있다. 그러나 향후 보다 정확한 연구를 위하여 일반 사용자가 사용하지 않는 특정 망에서의 데이터 셋 구성을 하여 실제 사용자와 웜 등의 사용자가 아닌 트래픽을 구분하여 시뮬레이션 환경을 구성하는 것 또한 필요하다. 또한 장기간 패킷의 흐름을 분석하여 장기간에 대한 단위 기간 흐름 분석이 보다 정확한 분포를 확인 가능 하다.

참 고 문 헌

[1] Moore, D., " The Spread of the Code-Red Worm (crv2)," 2001
http://www.caida.org/analysis/security/code-red

[2] Jeff Brown., " Animation of geographic spread of Code-Red worm" , 2001

[3] Lincoln Lab, " Summary and Plan for the 1999 DARPA Evaluation" , 1999