마르코프 체인 모델을 이용한 네트워크 포트 스캐닝의 탐지

한상준[0], 조성배
연세대학교 컴퓨터과학과
sjhan@cs.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Detecting Network Port Scanning Using Markov Chain Model

Sang-Jun Han[0], Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

일반적으로 해킹이 이루어지기 위해서는 공격의 대상이 되는 시스템과 네트워크의 정보를 수집하는 사전단계가 필수적이다. 네트워크 포트 스캐닝은 이 시스템 정보 수집단계에서 중요한 역할을 하는 방법으로 주로 통신 프로토콜의 취약점을 이용하여 비정상적인 패킷을 보낸 후 시스템의 반응을 살피는 방법으로 수행된다. 본 논문에서는 마르코프 체인 모델을 이용한 비정상행위기법 기반의 포트 스캐닝을 탐지방법을 제안하고 여러 가지 은닉/비은닉 포트 스캐닝 방법에 대하여 좋은 성능을 나타냄을 보인다.

1. 서론

해킹을 시도하기 위해서는 공격 목표인 시스템과 네트워크에 대한 정보를 수집하는 사전단계가 필수적이다. 이 사전단계는 네트워크 주소, 도메인 네임, 전화번호 등의 공격에 필요한 기본적인 정보를 알아내는 풋프린팅 단계, 활성화된 시스템과 서비스를 알아내는 스캐닝 단계, 공격의 대상이 될 시스템의 자세한 정보를 수집하는 목록화 단계의 과정으로 이루어진다[1]. 이 과정에서 네트워크 포트 스캐닝이 수행되는데 포트 스캐닝은 공격 목표가 제공하고 있는 서비스, 방화벽에 의해 차단되는 포트 등을 식별하기 위한 수단으로 주로 통신 프로토콜의 취약점을 이용하여 이루어진다. 포트 스캐닝의 결과는 곧바로 시스템의 취약점을 드러내고 후에 직접적인 공격행위로 이어질 가능성이 크기 때문에 시스템에 시도되는 스캐닝을 감시하고 이에 적절히 대응하여 보안사고를 사전에 예방하는 것은 필수적이라고 할 수 있다. 실제로 한국정보보호진흥원의 조사에 의하면 2002년 10월까지 16,782건의 스캔이 탐지 되었고 점점 더 증가하고 있는 추세이다. 주로 공격이 이루어지는 포트들과 최근 추세는 다음의 표1과 같다[2].

표에서와 같이 주로 21번(FTP), 22번(SSH), 25번(SMTP), 1433번(MS-SQL)등의 TCP(Transmission Control Protocol)프로토콜을 사용하는 서비스에 주로 스캔이 이루어지고 있음을 알 수 있다. 이는 대부분의 중요한 서비스들이 TCP를 사용하고 있기 때문인데 따라서 본 논문에서는 TCP 포트 스캐닝의 탐지를 중심으로 실험을 하였다.

본 논문에서는 포트 스캐닝이 주로 프로토콜의 취약점을 이용하는 것에 착안하여 정상적인 TCP연결의 패킷을 마르코프 체인 모델을 이용하여 모델링 한 후 이와 비교해 비정상적인 상태변화를 보이는 연결을 침입으로 판단하는 방법을 제안하였다. 논문의 나머지 내용은 다음과 같다. 2장에서는 이벤트 시퀀스의 모델링을 이용한 다른 침입탐지 방법을 알아보고 3장에서는 마르코프 체인 모델과 논문에서 제안한 탐지방법을 소개하고 4장에서 실험 결과를 분석한 후 마지막으로 5장에서는 결론 및 향후 연구에 대하여 언급하도록 하겠다.

표 1. 포트 스캐닝 현황

포트 번호	2001 합계	2002 8월	2002 9월	2002 10월	2002 합계
21	1,398	152	197	202	1,877
22	57	44	61	49	641
23	88	6	2	3	60
53	6,118	16	16	10	359
80	894	69	70	184	585
110	300	29	27	22	227
111	3,510	32	44	16	730
기타	8,138	868	1,184	1,997	14,178
합계	20,503	1,216	1,601	2,483	16,782

2. 관련연구

시간순서의 이벤트 시퀀스를 이용한 방법은 비정상행위기반 침입탐지 방법으로 매우 효과적이다. 이벤트 시퀀스를 모델링 하는데는 주로 Markov 가정을 이용하는데 복잡성을 줄이면서도 좋은 성능을 얻을 수 있게 하기 때문이다. Markov 가정을 활용한 좋은 통계적인 모델링 방법으로 Markov chain model과, 실제로 어떻게 내부적인 상태 변화 과정을 알 수 없고 다만 출력되는 신호만을 알 수 있는 경우에 사용하는 HMM(hidden Markov model)이 있다. C. Warrender 등은 시스템 호출 감사자료를 모델링하는데 HMM이 다른 모델링 방법에 비해 좋은 성능을 보여준다고 하였고 박혁장 등은 HMM모델의 긴 모델링 시간을 극복하기 위하여 감사자료를 축약하는 방법을 제안하였다[3][4]. T. Lane은 UNIX쉘데이터에 HMM을 적용하는 방법을 제안하였으며 Nong Ye 등은 숨겨진 상태를 사용하지 않는 Markov chain model을 시스템 호출 감사자료에 적용시켜 좋은 결과를 얻었다[5].

이제까지 이벤트 시퀀스 모델링을 이용한 방법들은 주로 호스트기반 침입탐지에 사용되었지만 본 논문에서는 네트워크기반 침입탐지에 많이 사용되는 네트워크 패킷 감사자료에 적용하는 침입탐지방법을 제안하였다.

3. 마르코프 체인 모델

마르코프 체인 모델은 초기 상태 분포 Q와 상태 전이 확률 분포 행렬 P로 이루어진다. 시간t의 상태 i에서 시간 $t+1$의 상태 j로 전이 될 확률을 P_{ij}라고 할 때 n개의 상태를 가지는 마르코프 체인의 상태 전이 확률 행렬 P는 식1과 같이 정의 되며 식 2에서와 같이 한 상태에서 다른 상태로 전이될 확률을 모두 합하면 1이 된다.

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \quad (1)$$

$$\sum_{j=1}^{n} p_{ij} = 1 \quad (2)$$

그리고 시간 0에서 상태i가 될 확률을 q_i라고 할 때 초기 확률 분포 Q는 다음과 같다.

$$Q = [q_1 \quad q_2 \quad \cdots \quad q_n] \quad (3)$$

마르코프 체인 모델의 학습은 데이터로부터 상태 전이 확률 분포 행렬과 초기 확률 분포를 얻어내는 과정이다. 시간 0부터 N-1까지의 상태변화 시퀀스 $X_0, X_1, X_2, \ldots, X_{n-1}$가 학습 데이터로 주어졌다고 할 때 다음과 같은 식에 의해 P와 Q를 얻어 낼 수 있다.

$$p_{ij} = \frac{N_{ij}}{N_i} \quad (4)$$

N_{ij} : 시퀀스중상태i에서j로의전이회수

N_i : 시퀀스중상태i가 나타난 회수

$$q_i = \frac{N_i}{N} \quad (5)$$

N_i : 시퀀스중상태j가나타난회수

N : 모든 관찰 시퀀스의 수

주어진 모델로부터 관찰된 상태변화 시퀀스 $X_0, X_1, X_2, \ldots, X_T$가 나올 확률 $P(X_0 \ldots, X_T)$는 식 6에 의해서 얻어진다.

$$P(X_1, \ldots, X_T) = q_{x_1} \prod_{t=2}^{T} P_{X_{t-1}X_t} \quad (6)$$

4. 제안하는 방법

일반적인 비정상행위기반 침입탐지방법은 정상행위 감사자료 수집, 정상행위 모델 구축, 침입이 들어있는 감사자료의 테스트 과정으로 이루어진다. 본 논문에서는 정상연결들의 기록이 담겨있는 네트워크 패킷 감사자료를 마르코프 체인 모델을 이용하여 모델링하여 정상행위모델을 구축한 후 침입이 담겨있는 테스트 자료로 성능을 측정하였다.

네트워크 감사자료 중 탐지에 사용되는 부분은 TCP 헤더와 IP헤더의 네트워크 주소 부분과 TCP 헤더의 flag부분이다. 네트워크 주소부분은 각 연결별로 구분하여 변화를 관찰하기 위하여 사용되었고 헤더의 flag부분은 마르코프 체인 모델의 관찰 시퀀스로 사용되었다. 실제 연결이 성립되면 많은 패킷들이 오가게 되는데, 이는 정해진 관찰 시퀀스를 정해진 길이의 윈도우 단위로 잘라서 사용하였다. 예를 들어 윈도우 크기가 N이고 현재 시간이 t일 경우 한번에 처리되는 감사자료는 $X_{t-(N-1)}, X_{t-(N-2)}, \ldots, X_t$가 된다. 이 관찰 시퀀스의 평가 값은 식6을 이용하여 다음과 같이 계산된다.

$$P(X_{t-(N-1)}, \ldots, X_t) = q_{t-(N-1)} \prod_{i=t-(N-2)}^{t} P_{X_{i-1}X_i} \quad (7)$$

평가값이 높을수록 주어진 모델에서 일어날 확률이 높은 것이므로 정상적인 연결에 가까운 것이고 평가값이 낮을수록 침입일 확률이 높은 것이 되는데 이때 적절한 임계값을 설정하여 침입여부를 판단한다.

제안하는 탐지 방법의 전체적인 구조는 그림 1과 같다. 전처리 모듈에서는 감사자료 중 데이터 분류에 사용될 네트워크 주소와 TCP헤더의 flag부분만을 걸러 준다. 모델링 모듈에서는 전처리 과정을 거친 정상행위 자료를 마르코프 체인 모델을 사용하여 정상행위 모델을 구축한다. 마지막으로 탐지 모듈에서는 구축된 정상행위 모델과 입력된 테스트 시퀀스를 비교하여 침입여부를 판단한다.

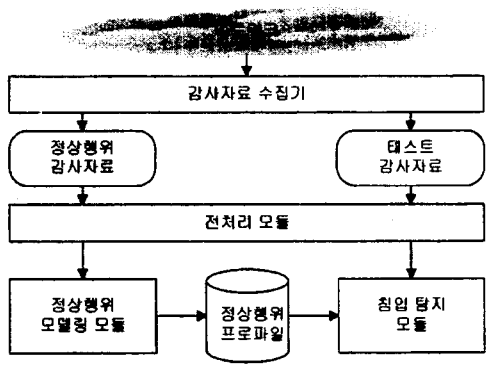

그림 1. 제안하는 탐지방법의 구조

5. 실험 및 결과

실험에 사용된 정상행위 감사자료는 MIT Lincoln Lab에서 1999년에 DARPA프로젝트의 일환으로 수행한 침입탐지시스템 평가프로젝트에서 사용된 자료를 이용하였다[6]. 그 중 침입이 수행되지 않은 첫째 주의 첫날의 내부 네트워크 감사자료를 이용하였으며 크기는 약 340메가바이트이다. 테스트 감사자료는 유닉스 호스트에 직접 포트 스캐닝을 수행하여 얻어 내었다. Solaris 8운영체제 환경에서 tcpdump를 사용하여 네트워크 패킷을 감시하였고 스캐닝 도구는 가장 많이 쓰이는 보안 스캐너 중 하나인 nmap을 사용하였다. 본 실험에는 총 5가지 유형의 TCP 포트 스캐닝 공격이 사용되었는데 각 유형별 내용은 표2와 같다.

표 2. 실험에 사용된 스캐닝의 유형

유형	이름	설명
Open	connect scan	정상적인 연결을 사용
Half-open	SYN scan	Half-open상태에서 종료
Stealth	FIN scan	FIN만 설정후 연결 시도
	Xmas scan	모든 flag를 세팅
	Null scan	모든 flag를 세팅하지 않음

먼저 정상행위 데이터와 침입 데이터에 대해서 각각 평가값을 계산한 후 윈도우 번호에 따른 평가값 변화의 추이를 그래프로 그려 보았다. 정상행위 테스트 데이터는 Lincoln Lab의 정상행위 데이터 중 학습데이터와 다른 하나를 사용하였고 비정상행위 테스트 데이터는 5가지 유형의 스캐닝 공격만을 담은 자료를 사용하였다. 그림 2와 3에서와 같이 정상행위 데이터와 침입 데이터는 평가값에 있어서 큰 차이를 보였다. 각 자료의 테스트 결과를 비교해보면 표3과 같다.

다음은 정상행위와 포트 스캐닝 공격이 같이 존재하는 데이터를 가지고 실험을 해보았다. 원격로그인, 파일 전송 등

의 일반적인 작업을 수행하는 중에 포트스캐닝을 하는 방법으로 실험 데이터를 생성하였다. 그 결과 임계값을 2.044E-08으로 설정하였을 때 100%탐지율에서 5.7%의 false-positive 오류율을 얻었다.

표 3. 평가값의 차이

	정상	비정상
평균	6.037E-04	4.629E-03
표준편차	8.792E-03	6.773E-02
최대	1.191.E-04	2.044E-08
최소	5.041E-21	1.429E-23

6. 결론 및 향후 연구

본 논문에서는 네트워크 침입탐지를 위한 비정상행위 탐지 기법으로 마르코프 체인 모델을 이용한 방법을 제안하였다. TCP 프로토콜의 패킷 시퀀스를 이용하여 정상행위를 모델링하고 여러 가지 유형의 포트 스캐닝 공격을 탐지해보았고 그 결과 정상행위와 비정상행위가 평가값에 있어서 큰 차이를 보인다는 것을 확인하였으며, 낮은 false-positive 오류율에서 높은 침입탐지율을 얻을 수 있었다.

제안한 방법은 비정상행위기반이므로 대부분의 오용탐지 기반 네트워크 탐지 방법에 비하여 예상하지 못한 프로토콜의 약점을 이용한 공격도 탐지가능 하다는 장점을 가진다. 논문의 실험에서는 TCP 포트 스캐닝만을 대상으로 하였지만 UDP등의 통신 프로토콜을 이용한 네트워크 스캐닝도 많이 행해지고 있으므로 다른 프로토콜을 이용한 스캐닝 공격과 최근 문제가 되고 있는 네트워크 DoS공격도 탐지할 수 있도록 하는 연구가 필요하다.

감사의글

본 연구는 대학 IT 연구센터 육성/지원사업의 연구결과로 수행되었음.

참고문헌

[1] G. Kurtz, J. Scambray and McClure, *Hacking Exposed*, McGraw-Hill, 2nd edition, 2000.

[2] 해킹바이러스 통계 및 분석 월보, 한국정보보호 진흥원, 10, 2002.

[3] C. Warrender, S. Forrest and B. Pearlmutter, "Detecting intrusion using calls: Alternative data models," *In Proc. of Symposium on Security and Privacy*, pp. 133-145, May 1999.

[4] H.-J. Park and S.-B. Cho, "Efficient anomaly detection by modeling privilege flows with hidden Markov model," *Computers & Security*, 2003. (To appear)

[5] N. Ye, "A markov chain model of temporal behavior for anomaly detection," In *Proc. of IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, 2000.

[6] MIT Lincoln Labs. 1999 DARPA intrusion detection evaluation. In *http::/www.ll.mit.edu/IST/ideval/index.html*

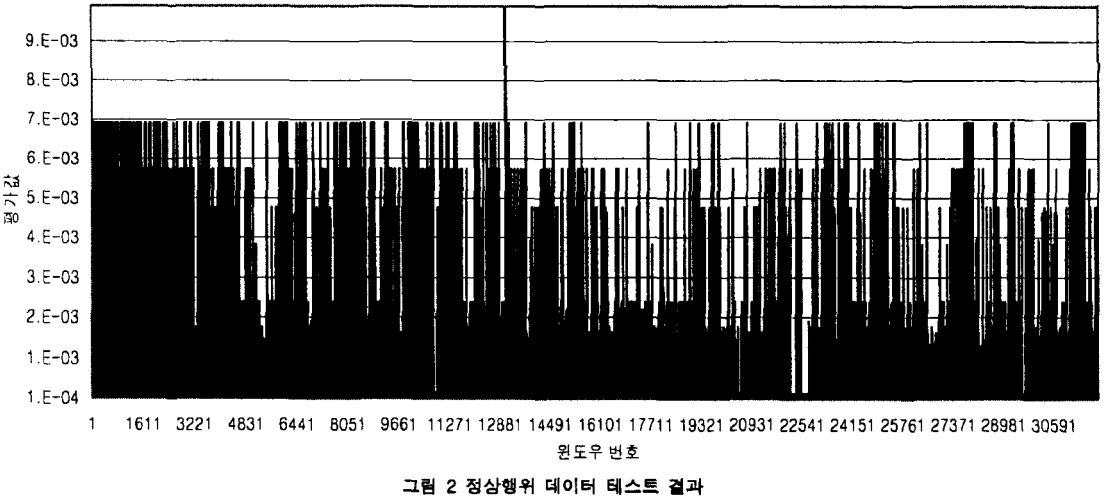

그림 2 정상행위 데이터 테스트 결과

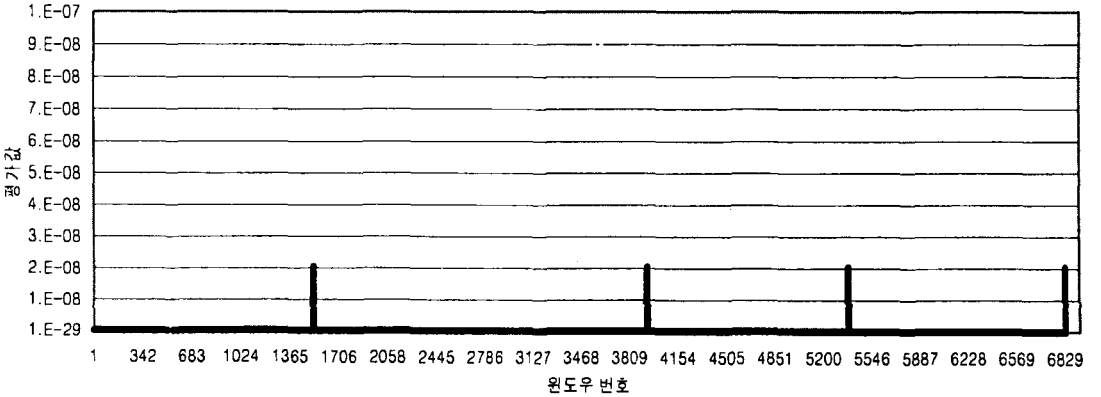

그림 3 비정상행위 데이터 테스트 결과