악성 코드 유포 사이트 탐지에 관한 연구

서동원*, Arindam Khan**, 이희조*
*고려대학교 정보통신대학 컴퓨터·전파통신학과
**Dept. of Computer Science and Engineering, Indian Institute of Technology
e-mail : {aerosmiz, heejo}@korea.ac.kr

A Study on Detecting Malcodes Distribution Sites

Dongwon Seo*, Arindam Khan**, Heejo Lee*
*Dept. of Computer and Radio Communications Engineering, Korea University
** Dept. of Computer Science and Engineering, Indian Institute of Technology

요 약

최근 웹사이트를 통해 악성 코드의 유포가 성행하면서 많은 웹 서비스 사용자들이 위험에 노출되어 있다. 특히, 특정 웹페이지에 접속하는 것만으로도 사용자가 알지 못하는 사이에 악성 코드를 자동으로 다운로드 받도록 함으로써 그 위협은 더욱 커지고 있다. 본 논문에서는 이러한 악성 코드 유포 사이트를 탐지하기 위해 사용하였던 Website relationship graph, Parallel coordination, Amazon Web Service system 을 차례로 소개하고, 각 기법의 장단점과 결과적으로 도출해낸 악성 코드 유포 사이트들의 특징과 그것을 이용한 알려지지 않은 악성 코드 유포 사이트 탐지 기법을 제안한다.

1. 서론

전통적으로 인터넷을 통해 악성 코드를 유포시키는 방법에는 여러 가지가 있어왔다. 이메일의 첨부파일, 인스턴스 메신저를 통한 파일 전송, P2P 서비스의 악용이 그런 예라고 할 수 있다. 하지만, 최근 들어 더욱 정교해진 악성 코드 유포 방법이 유행하고 있는데 웹사이트를 통한 감염이 그것이다. 악성 코드 유포자는 웹서버의 취약성을 이용하여 특정 웹페이지를 조작하여 이 페이지에 접속하는 일반 사용자의 컴퓨터에 악성 코드가 자동으로 다운로드 되도록 하는 것이다. 이 같은 악성 코드 유포의 주 목적은 key logger 나 bot 들을 설치하여 금융 사이트등의 개인 정보를 탈취하거나 DDoS 공격을 위한 좀비 머신으로 활용하기 위함이다. 공격자는 최대한 많은 사용자들을 감염시키기 위하여, 이미 많이 알려진 사이트의 메인 페이지를 조작하려는 시도도 하고 있다. [1] [2]

이렇게 조작된 악성 코드 유포 사이트들을 탐지하기 위해서 우선 감염된 웹사이트와 웹페이지를 분석하는 작업이 선행되어야 한다. 감염된 웹사이트와 정상적인 웹사이트를 비교함으로써 악성 코드 유포 사이트의 특징을 찾아내고, 효과적인 탐지 기법을 도출해 낼 수 있다. 악성 코드 유포 사이트들의 특징을 찾아내기 위해 크롤러를 사용하여 웹페이지를 수집하였고, 이 수집된 웹페이지들을 바탕으로 Website relationship graph 와 parallel coordination [3][4], Amazon Web Service [5]를 이용 정상 사이트와 악성 사이트를 비교하는 실험을 하였다. 그 결과 악성 코드 유포 사이트들은 주로 web 사이트 랭크 ing 와 페이지 뷰가 정상 사이트들에 비해 낮다는 것을 알 수 있었으며, 이 특징을 바탕으로 알려지지 않은 악성 코드 유포 사이트를 효과적으로 찾아 낼 수 있는 기법을 제안한다.

이 후 논문의 구성을 다음과 같다. 먼저, 2 장에서는 관련 연구에 대해 소개하고, 3 장에서는 본 연구의 동기와 목적을, 4 장에서는 웹페이지 수집 방법과 Website relationship graph, parallel coordination, Alexa.com database system 을 이용한 웹페이지 분석 기법에 대해 다루도록 한다. 5 장에서는 실험의 결과를 바탕으로 한 효과적인 악성 코드 탐지 기법을 제안하고, 마지막으로 6 장에서는 논문의 요약과 결론, 향후 과제에 대해 언급한다.

2. 관련 연구

Ying Pan 등은 phishing 페이지를 탐지하기 위한 방법 [6]을 제시하였다. phishing 페이지를 구분 하기 위해서 phishing 페이지에서 보여지는 비정상 상황을 8 가지로 정의한 후, 웹페이지를 정의된 기준을 바탕으로 8 차원 벡터로 나타낸다. 그 후 수집된 웹페이지를 이용하여 미리 학습된 SVM(Support Vector Machine)에 대상 웹페이지를 적용하여 1(phishing site) 혹은 -1(authentic site)로 구분한다. 이 논문은 악성 코드 유포 사이트 탐지를 위한 방법을 제시한 것은 아니지만, 악성 사이트의 특징을 이용해서 정의한 여러 기준을 이용하여 다른 사이트들의 악성 여부를 판단하는 anomaly detection 을 사용하여 알려지지 않은 phishing site 를 판단하는데 효과적이다.

Niels Provos 등이 발표한 Google technical report [7]는 Google 에서 10 개월 동안 수집한 수십 억개의 URL 을 분석한 결과 약 3 백만개(약 0.003%)가 drive-by download 방식의 악성 코드 유포 사이트였으며, 전체

질의의 약 1.3%는 악성 사이트를 결과값으로 보여준다는 통계를 제시했다. 이런 악성 사이트들 중에 약 40%는 중간에 다른 경로 없이, 즉 landing site 를 거치지 않고 바로 악성 코드 유포 사이트 URL 로 접근이 되었으며, 나머지 60%는 많은 수의 landing site 들을 거쳐서 악성 코드 유포 사이트에 도달하는 것으로 나타났다. 이는 공격자가 자신의 위치를 숨기고, 탐지를 어렵게 하기 위함이다.

Neils Provos 등이 발표한 다른 연구 [8]에서는 실제로 웹 기반의 악성 코드를 탐지하기 위한 방법이 제시되어있다. 이 연구는 [7]의 연구를 확장한 것으로 수집된 웹페이지를 바탕으로 URL 을 추출한 후에 그 URL 들을 가상 머신의 인터넷 익스플로러 (보안 패치 미적용 버전)를 이용해 방문을 한다. 이 때 가상 머신에 설치된 여러 개의 Anti-virus 프로그램의 반응을 점수화 하여 악성 코드 유포 사이트의 여부를 판단하게 된다.

3. 연구의 동기와 목적

2 장에서 언급한 연구 [8]은 악성 코드 유포 사이트를 탐지해내기 위한 기법으로서 그 목적은 본 연구와 같다. 하지만, 수집된 웹페이지를 바탕으로 추출된 URL 을 일일이 방문하여 탐지하는 것은 효과적일 수 없다. 인터넷은 하루가 다르게 그 크기가 커지고 있는데 반해 무작정 모든 웹사이트를 방문해 보겠다는 것은 비효율적이기 때문이다. 그래서 우리는 정상 웹사이트와 악성 코드 유포 웹사이트의 특징을 관찰하여 악성 사이트로 판단되는 사이트를 우선적으로 방문하여 검사하는 기법을 구상하였다. 정상과 악성 사이트를 구분하기 위하여 website relationship graph, parallel coordination, AWS 의 세 가지 다른 기법을 응용하였고, 그 결과 악성 사이트에서 대표적으로 발견되는 특징들을 발견할 수 있었다.

다음에 나올 4 장과 5 장에서는 제시된 세 가지의 기법을 어떻게 활용하였고, 그 결과 고안된 기법이 어떤 것인지를 살펴보도록 한다.

4. 웹페이지 수집과 분석

본 장에서는 악성 사이트를 판별하기 위한 웹페이지의 수집과 분석 기법에 대해 설명한다.

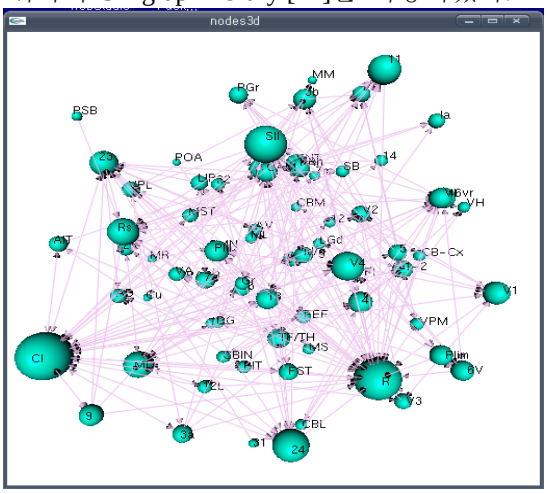
3.1 웹페이지 수집

인터넷에 산재해 있는 수많은 웹페이지를 수집하기 위해 공개 Crawler 인 NetCrawler [9]를 사용하였다. 웹페이지를 수집하기 위한 정책은 아래와 같다.

- 2008 년 6 월 30 일부터 동년 동월 16 일까지, depth 5 (seed 로부터 5 단계 링크)까지의 웹페이지를 수집하였다.
- Alexa.com [10]에서 제공하는 traffic ranking Top 10 을 Crawler 의 seed 로 선정하여 약 248,000 개의 웹페이지를 수집하였다.
- Stopbadware.org [11]에서 제공하는 알려진 악성 사이트 10 개를 seed 로 선정하여 약 119,730 개의 웹페이지를 수집하였다.

3.2 Website relationship graph

정상 사이트와 악성 사이트를 구분하고 특징을 밝혀내기 위한 첫 번째 기법으로 Website relationship graph 를 구현해 보았다. 수집된 웹페이지를 도메인 레벨, 사이트 레벨, 페이지 레벨의 3 단계로 구분하여, 어떤 도메인이 어떤 도메인을 얼마나 많이 링크하고 있는지를 살펴보았다. 이를 위해 웹페이지의 html 코드를 분석하여 <a href>, <object>, <iframe>등의 하이퍼링크가 올 수 있는 태그만을 추출하여 그래프를 작성하였다. 보다 효율적으로 각 웹사이트의 관계를 표시하기 위하여 3D graph library [12]를 사용하였다.



(그림 1) Website relationship graph

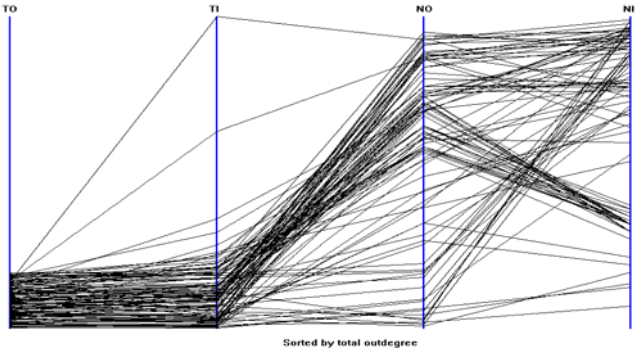
악성 코드 유포 사이트는 외부에서 자신을 가르키는 indegree 값이 outdegree 값보다 상대적으로 높을 수 밖에 없기 때문에 그림 1 과 같은 그래프에서 indegree – outdegree 값이 높은 사이트를 의심 사이트로 판단할 수 있다. 하지만 이 그래프는 무한한 인터넷의 웹사이트들 중에서 극히 일부만을 분석한 결과이기 때문에 이 결과를 있는 그대로 적용하기에는 무리가 있다. 또한 노드의 수가 많아질수록 그래프가 복잡해져서 결국에는 알아보기 힘든 경우가 생길 수 있다. 이를 극복하기 위해서 parallel coordination 기법을 사용하게 되었다.

3.3 Parallel coordination

Total Indegree(외부에서 가리키는 총 링크 수), Total Outdegree(외부로 나가는 총 링크 수), Neighbor Indegree(중복 링크 수를 제외한 Indegree 값), Neighbor Outdegree (중복 링크 수를 제외한 Outdegree 값)의 4 가지 값을 4 개의 평행한 Y 축으로 하여 각 웹사이트들의 특징을 표현하였다.

그림 2 는 수집된 웹페이지의 분석 결과를 parallel coordination 기법을 이용하여 나타낸 것이다. 대부분의 정상 사이트들은 높은 TO, TI 와 낮은 NO, NI 값을 가지는 특징을 보였다. 이 기법은 Website relationship graph 기법보다는 직관적인 결과를 보여주지만, 역시 짧은 기간 동안에 수집된 웹페이지 분석 결과를 바탕으로 하고 있기 때문에 객관적인 신뢰도가 떨어질 수

밖에 없다.

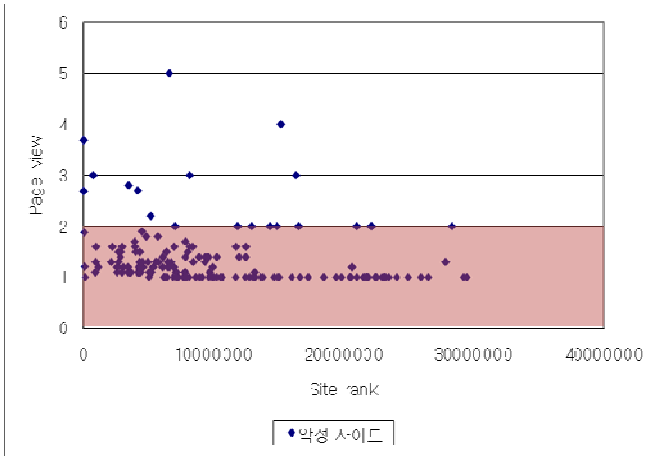


(그림 2) Parallel coordination

3.4 Amazon Web Service

인터넷 트래픽 수집 및 분석사이트인 Alexa.com 에서는 전 세계의 수많은 웹사이트의 트래픽을 분석하고 그 결과를 xml 파일의 형태로 전달해 주는 Amazon Web Service (AWS) 시스템을 제공해주고 있다. 여기에서 수집되고 분석된 결과는 오랜 시간 동안 전 세계의 호스트에서 전송된 트래픽을 기반으로 하기 때문에 객관적인 신뢰도가 높다.

본 연구에서는 이 AWS 를 사용하여 정상 사이트와 악성 사이트를 구분 지을 수 있는 특징을 찾아 낼 수 있었다. AWS 에서 제공해 주는 정보 중에 사용자 당 페이지 뷰 와 사이트 랭크 값의 상관 관계를 조사한 결과 대부분의 악성 사이트들은 낮은 페이지 뷰 값과 높은 사이트 랭크 값을 보이고 있었다. 즉 사용자들은 정상 사이트에 접속했을 때는 여러 개의 웹페이지를 방문해 보지만, 악성 사이트에 접속했을 경우에는 한 두 개의 페이지를 방문해보고 접속을 종료한다는 것을 뜻한다. 또한 이런 악성 사이트들은 대부분 숨겨져 있는 경우가 많기 때문에 사이트 랭크 값이 높게 나타난다. 여기서 사이트 랭크의 값은 낮을수록 방문자 수가 많고, 높을 수록 방문자 수가 낮음을 뜻한다. 그림 3 은 Stopbadware.org 와 malwaredomainlist.com [13]에서 제공해주는 알려진 악성 사이트 167 개를 대상으로 페이지 뷰와 사이트 랭크값의 상관 관계를 나타낸 그래프이다.



(그림 3) 악성 사이트의 페이지 뷰와 사이트 랭크 분포도

이처럼 페이지 뷰는 2 이하, 사이트 랭크는 50,000 이상(적색 영역)을 악성 코드 유포 사이트를 구분 짓는 임계 값으로 설정한 경우 167 개의 알려진 악성 사이트들 중에 157 개의 악성 사이트들을 탐지할 수 있었다.

하지만, 잘 알려지지 않은 개인 홈페이지나 특정 그룹만이 사용하고 있는 웹사이트의 경우에도 낮은 페이지 뷰와 높은 사이트 랭크의 값이 나타날 수 있기 때문에, 이와 같은 오탐을 줄일 수 있는 방법 또한 강구되어야 한다.

5. 효과적인 악성 코드 유포 사이트 탐지 기법

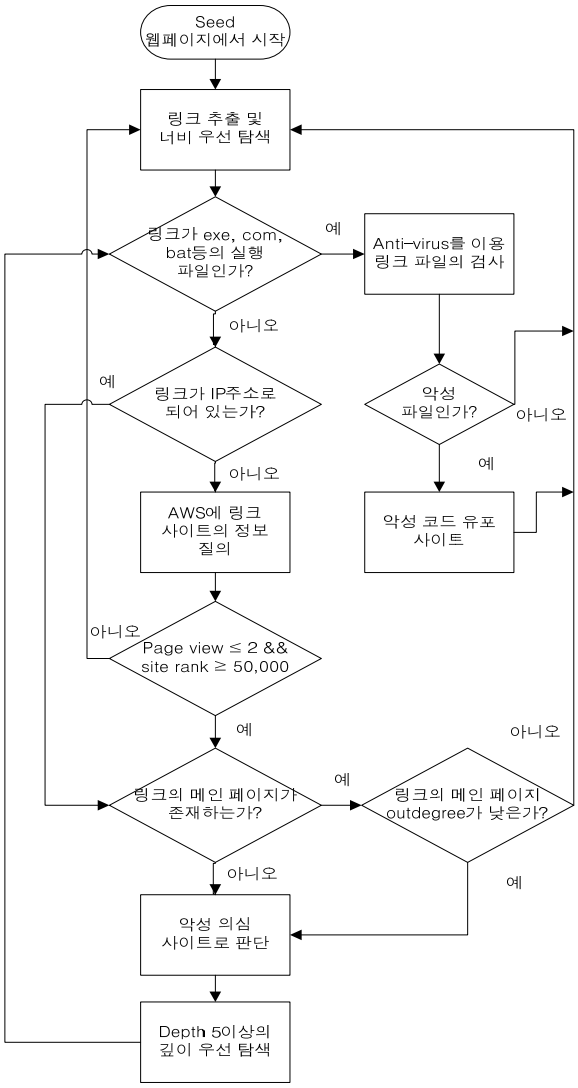
3 장에서 설명한 기법 및 실험들을 바탕으로 악성 코드 유포 사이트의 특징을 살펴 보면 다음과 같다.

- 페이지 뷰가 2 보다 낮고, 사이트 랭크는 50,000 보다 높다.
- Indegree 가 outdegree 에 비해서 현저히 높다.
- 링크가 도메인 네임보다는 IP 주소인 경우가 많다.
- 링크의 최상위 메인 페이지는 존재하지 않거나 매우 간단하게 꾸며진 경우가 많다.

위의 특징들을 바탕으로 기존의 공개된 crawler 를 수정하여 알려지지 않은 악성 코드 유포 사이트 탐지기를 제작 할 수 있다.

제안된 기법은 웹페이지를 단순히 수집하기 위한 기존의 crawler 를 악성 코드 유포 사이트를 빠르고 효과적으로 판단할 수 있도록 개선한 것이다. 우선 링크가 IP 주소로 되어 있으면 링크의 메인 페이지 존재여부와 메인 페이지 구성의 단순성을 검사하게 된다. 만약 링크가 도메인 네임으로 되어있다면 AWS 의 트래픽 정보를 이용하여 페이지 뷰가 낮고, 사이트 랭크가 높은 사이트만을 선택하여 메인 페이지를 검사하게 된다. 링크가 악성 의심 사이트로 판단이 되면, 해당 링크를 새로운 seed 로 하여 깊이 우선 탐색을 수행한다. 이 것은 보통 악성 코드 유포 사이트가 수 단계의 landing site 를 거친 후에 존재하기 때문이다. 깊이 우선 탐색 중에 발견된 실행 파일들은 anti-virus 프로그램을 통해 검사를 하게 된다. 여기서 악성 코드가 발견되면 해당 링크 사이트는 악성 코드 유포 사이트로 판별할 수 있다.

현재의 악성 코드 유포 사이트들은 하루에서 수없이 생겨나고 사라지며, 도메인 네임과 IP 주소를 변경한다. 이러한 변화에 따라 갈 수 있는 탐지 기법 중에 가장 필요한 항목은 어떻게 하면 방문하려고 하는 사이트가 정상인지 악성 의심 사이트인지를 구분하는 것이다. 정상 사이트로 보이지만 악성 코드를 유포할 수 있다. 하지만 빠른 시간 내에 악성 의심 사이트를 많이 방문하여 그 근원지를 차단 할 수 있다면, 많은 정상 사이트에 존재하는 몇몇 악성 링크들도 무력화 시킬 수 있는 것이다. 본 기법은 그러한 빠른 탐지를 하기 위해 고안되었다.



(그림 4) 알려지지 않은 악성 코드 유포 사이트 탐지 기법 순서도

6. 결론 및 향후 과제

본 논문에서는 최근 그 위험성이 증대되고 있는 웹 사이트를 통한 악성 코드 유포를 탐지할 수 있는 기법을 제안하였다. 특히 crawler 를 통해 수집된 웹페이지를 분석한 결과를 website relationship graph 로 표현하고, 각 사이트의 링크 indegree, outdegree 관계를 직관적으로 표현하기 위해 parallel coordination 기법을 도입하였다. 하지만, 전체 인터넷의 극히 일부분만을 분석한 결과라는 단점을 극복하기 위해 전문 인터넷 트래픽 분석 사이트인 Alexa.com 의 AWS 를 이용하여 알려진 악성 코드 유포 사이트는 낮은 페이지 뷰와 높은 사이트 랭크 값을 보인다는 것을 밝혀 냈다. 그 외에도 수집된 웹페이지의 분석 결과 보여진 다른 특징들을 이용하여 알려지지 않은 악성 코드 유포 사이트를 효율적으로 탐지할 수 있는 기법을 제안하였다.

이 기법은 새로운 악성 코드 유포 사이트를 빠르게 찾을 수 있는 가능성을 보였고, 향후 과제로서 실제적인 구현과 실험을 통해 확장 연구를 할 것이다. 이와 더불어 정상적인 사이트이지만 낮은 페이지 뷰 값과 높은 사이트 랭크 값을 보이는 사이트들을 구별해 내기 위한 방법, 높은 indegree 에 비해 상대적으로 낮은 outdegree 를 보이는 악성 사이트의 특징을 회피하기 위해 거짓으로 외부 링크를 많이 설정해 놓은 악성 사이트의 탐지 방법 등이 논의 되어야 하겠다.

참고문헌

[1] 한국정보보호진흥원(KISA). 웹 해킹을 통한 악성 코드 유포 사이트 사고 사례, Jun., 2005

[2] 한국정보보호진흥원(KISA). 취약한 웹서버 공격을 통한 내부망 해킹 및 악성코드 삽입 사례, Dec., 2007

[3] A. Inselberg. The plane with parallel coordinates. The Visual Computer 1, pp 69–91, 1985

[4] Hyunsang Choi, Heejo Lee. PCAV: Internet Attack Visualization on Parallel Coordinates, Int'l Conf. on Information and Communications Security (ICICS), LNCS, Vol. 3783, pp. 454-466, Dec. 2005

[5] Amazon Web Service. http://www.amazon.com/AWS-home-page-Money/b?ie=UTF8&node=3435361, 2008

[6] Ying Pan , Xuhua Ding. Anomaly Based Web Phishing Page Detection, Computer Security Applications Conference, 2006. ACSAC '06. 22nd Annual, Dec., 2006

[7] Niels Provos, Panayiotis, Mavrommatis, Moheeb Abu Rajab, Fabian Monrose. Google technical report provos-2008a : All Your iFRAMEs Point to Us, Feb., 2008

[8] Niels Provos, Dean McNamee, Panayiotis Mavrommatis, Ke Wang and Nagendra Modadugu. The Ghost In The Browser: Analysis of Web-based Malware, Workshop on Hot Topics in Understanding Botnets (HotBots), Apr., 2007

[9] Hatem Mostafa. Netcrawler, http://www.codeproject.com /KB/IP/Crawler.aspx, Mar., 2006

[10] Alexa The Information Company. http://www.alexa.com, 2008

[11] StopBasware.org. http://www.stopbadware.org/home/topsites, 2007

[12] Nodes3D library. http://brainmaps.org/index.php?p=desktop-apps-nodes3d, 2008

[13] Malware Domain List. http://www.malwaredomainlist.com, 2008