MapReduce 환경에서 네트워크 공격 탐지를 위한 실시간 로그 분석 시스템 개발*

장진수, 신재환, 장재우†
전북대학교 IT정보공학과
{lklin1749, djtm99, jwchang}@jbnu.ac.kr
†Corresponding author

Real-time log analysis system for detecting network attacks in a MapReduce environment

Jae-Hwan Shin, Jin-Su Chang, Jae-Woo Chang†
Dept of Information Technology and Engineering, Chon-Buk University

요 약

네트워크 기술의 발전으로 인터넷의 보급률이 증가함에 따라, 네트워크 사용량 또한 증가하고 있다. 그러나 네트워크 사용량이 증가함에 따라 악의적인 네트워크 접근 또한 증가하고 있다. 이러한 악의적인 접근은 네트워크에서 발생하는 보안 로그를 분석함으로써 탐지가 가능하다. 그러나 대규모의 네트워크 트래픽이 발생함에 따라, 보안 로그의 처리 및 분석에 많은 시간이 소요된다. 본 논문에서는 MapReduce 환경에서 네트워크 공격 탐지를 위한 실시간 로그 분석 시스템을 개발한다. 이를 위해, Hadoop의 MapReduce를 통해 보안 로그의 속성을 추출하고 대용량의 보안 로그를 분산 처리한다. 아울러 처리된 보안 로그를 분석함으로써 실시간으로 발생하는 네트워크 공격 패턴을 탐지하고, 이를 시각적으로 표현함으로써 사용자가 네트워크 상태를 보다 쉽게 파악할 수 있도록 한다.

1. 서론

최근 네트워크 기술의 발전으로 인터넷의 보급률이 증가함에 따라, 네트워크 사용량이 증가하고 있다. 그러나 네트워크 사용량이 증가함에 따라, 악의적인 네트워크 접근 또한 증가하고 있다. 이를 방지하기 위해 네트워크 공격 탐지를 위한 연구가 수행되었다.[1], [2]

이러한 연구들은 시각화를 통해 악의적인 네트워크 접근을 확인할 수 있지만, 공격의 종류에 대해서는 사용자가 직접 판단해야 하는 한계점이 존재한다. 또한, 해당 연구는 네트워크 공격 탐지 및 네트워크 공격 패턴 분석을 수행할 수 없다. 또한, D.S.Choi의 연구[2]는 방대한 양의 보안 로그를 Hadoop[3] 기반의 MapReduce[4]를 통해 빠른 분석을 지원한다. 그러나 해당 연구는 실시간으로 발생하는 보안로그에 대한 처리 및 분석을 지원하지 못하는 한계점이 존재한다.

본 논문에서는 MapReduce 환경에서 네트워크 공격 탐지를 위한 실시간 로그 분석 시스템을 개발한다. 이를 위해, 침입 탐지 시스템인 Snort[5]를 이용하여 보안 로그를 수집하고 분석 시간을 단축시키기 위하여 MapReduce를 이용하여 수집된 보안 로그로부터 분석에 필요한 요소만을 미리 추출함으로써, 네트워크 공격 패턴을 빠르게 분석한다. 마지막으로 이를 시각화하여 사용자가 네트워크 상태를 쉽게 파악할 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 보안 로그에 대한 효율적인 처리 및 분석에 관련된 연구를 기술하고 3장에서는 제안하는 시스템의 전체 구조를 제시하고, 보안 로그 전처리 및 분석 과정을 설명한다. 4장에서는 실시간 로그 데이터 및 VAST 2012 데이터를 이용하여 제안하는 시스템을 테스트한 결과 및 분석 내용을 설명한다. 마지막으로 5장에서는 제안하는 시스템의 결론 및 향후 연구를 기술한다.

2. 관련연구

네트워크 사용량이 증가함에 따라 악의적인 네트워크 접근이 증가하고 있다. 이를 해결하기 위해, 네트워크 공격을 탐지하는 연구가 활발히 진행되고 있다.

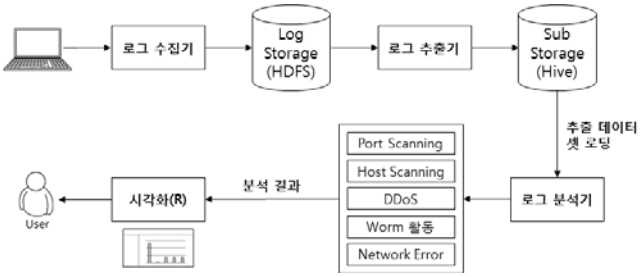
D.H.Lee의 연구[1]는 보안 위협을 탐지하기 위해 3가지 요소인 Source IP, Destination IP, Destination Port를 추출하여 각각의 비율에 대한 RGB색을 이용하여 시각화를 제공한다. 그러나 RGB색의 변화를 통해 네트워크 로그의 상태만을 시각화하여 보여주기 때문에 네트워크 공격을 사용자가 직접 판단해야 한다는 한계점이 존재한다. 한편, 실시간으로 발생되는 방대한 양의 로그 데이터를 처리하기 위해서는 많은 시간이 소요된다. 이를 해결하기 위해, D.S.Choi의 연구[2]는 대용량의 보안 로그 데이터를 빅데이터를 효율적으로 처리 및 분석하기 용이한 플랫폼인 Hadoop[3] 기반의 MapReduce[4]를 이용한다. 그러나 해당 연구는 실시간으로 발생하는 보안로그에 대한 처리 및 분석을 지원하지 못하는 한계점이 존재한다.

3. 제안하는 시스템 설계

본 논문에서는 MapReduce 환경에서 네트워크 공격 탐지를 위한 실시간 로그 분석 시스템을 제안한다. 이를 위한 전체 시스템 구조는 <그림 1>과 같다. 첫째, 로그 수

*이 논문은 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음(IITP-2017-R0113-15-0005, 대규모 트랜잭션 처리와 실시간 복합 분석을 통합한 일체형 엔지니어링 기술 개발). 또한 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1A3B03935298)

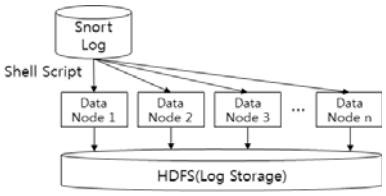
집기는 다양한 속성의 네트워크 정보를 포함하고 있는 네트워크 로그를 실시간으로 수집하고 Log Storage인 HDFS에 삽입한다. 둘째, 로그 추출기는 HDFS에 저장되어 있는 로그에서 필요 요소만을 추출한 후 Hive[6]에 저장한다. 마지막으로, 로그 분석기는 Hive에 저장되어 있는 데이터를 로드(Load)한 후 주요 공격 패턴 분석을 수행한다. 아울러, 분석 결과를 시각화하여 사용자에게 웹 인터페이스 형태로 제공한다.



<그림 1> 실시간 네트워크 공격 패턴 분석 시스템의 전체 구조

3.1 로그 수집기

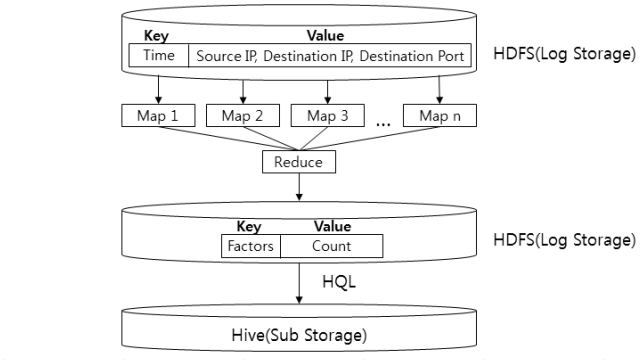
로그 수집기는 침입 탐지 시스템인 Snort를 이용하여 해당 서버에 실시간으로 들어오는 네트워크 로그를 수집하며, 과정은 <그림 2> 와 같다. 첫째, Snort를 이용하여 네트워크 로그를 수집하며 로그 정보로는 Time, type, Source IP, Port, Destination IP, Port, ID, Len 등이 포함된다. 둘째, 수집된 네트워크 로그는 Shell Script를 통해 n개의 데이터 노드에 분산되어 처리된다. 마지막으로 n개의 노드에서 처리된 결과는 HDFS에 저장된다.



<그림 2> 네트워크 로그 수집 과정

3.2 로그 추출기

로그 추출기는 HDFS에 저장되어 있는 네트워크 로그의 여러 요소 중 필요한 요소인 Time, Source IP, Destination IP, Destination Port 정보만을 추출한다. 추출과정은 네트워크 로그 필수 요소 추출 단계, 네트워크 로그 Count 요소 추출 단계, 추출 데이터 Hive 저장단계로 구성되며, 이는 <그림 3> 과 같다. 첫째, 네트워크 로그 필수 요소 추출 단계에서는 Hadoop 기반의 MapReduce를 사용하여 추출요소 중 Time을 Key로 하고, 나머지 요소들을 Value로 하여 추출한다. 아울러, 추출된 각 요소인 Source IP, Destination IP, Destination Port값을 Count 요소 추출단계의 입력 데이터로 사용한다. 둘째, 네트워크 로그 Count 요소 추출 단계에서는 MapReduce를 이용하여 Source IP, Destination IP, Destination Port에 대한 Count 값을 추출 및 전체 로그의 양을 추출한다. 아울러, 추출된 요소 및 전체 로그의 양을 Key로 하고 각각의 Count 값을 Value로 하여 HDFS에 저장한다. 마지막으로, 추출 데이터 Hive 저장 단계에서 HDFS에 저장된 데이터 및 Count 값을 Hive[6]에 테이블 형식으로 저장하며 이는 다양한 질의를 통해 보안 로그 분석에 사용된다.



<그림 3> 네트워크 로그 요소 추출 과정

3.3 로그 분석기

로그 분석기는 네트워크 로그를 분석하여 네트워크 공격 패턴을 인지할 수 있도록 한다. 본 연구에서는 VAST Challenge 2012[8]에서 제공한 IDS 및 방화벽 로그 데이터 셋과 Snort를 사용하여 실시간으로 수집한 로그 데이터를 이용하여 공격 탐지 조건을 설정한다. 본 논문에서 설정한 공격 탐지 조건은 Intel(R) Core(TM) u7-6700k 4.00GHz, 32GB RAM 서버를 기준으로 각각의 threshold 값을 설정하였으며, 이는 네트워크 환경에 따라 변할 수 있다.

3.3.1 Port Scanning

Port Scanning 공격은 악의적인 사용자가 특정 목적지에서 사용 가능한 Port에 대해 전체적인 탐색을 시도하는 공격으로써, 공격 발생 시 보안 로그 상에서 Destination Port의 수가 상대적으로 급증한다. 공격 탐지 기준에 대한 수식은 다음과 같다.

$$\text{Port Scanning} = \frac{D_{Port}}{avg(D_{Port})} > \alpha_1 \text{ AND } D_{Port} > \beta_1 \quad \text{--- (식 1)}$$

(식 1)과 같이 보안 로그에서 Destination Port의 평균 발생량을 나타내는 $avg(D_{Port})$에 대해 단위시간 Destination Port의 발생량을 나타내는 D_{Port}의 비율이 threshold α_1를 초과하고 D_{Port}의 값이 설정된 threshold β_1를 초과하는 경우 해당 시점에서 Port Scanning 공격이 발생하였다고 분석할 수 있다.

3.3.2 Host Scanning

Host Scanning 공격은 다수의 목적지의 IP를 스캔하는 공격이며 평상시보다 접근하는 목적지가 다양하다는 특징이 있다. 공격 발생 시 보안 로그 상에서 Destination IP의 수가 상대적으로 급증하는 공격이다. 공격 탐지 기준에 대한 수식은 다음과 같다.

$$\text{Host Scanning} = \left(\frac{D_{IP}}{avg(D_{IP})} > \alpha_2\right) \text{ AND } \left(D_{IP} > \beta_2\right) \quad \text{--- (식 2)}$$

(식 2)와 같이 보안 로그에서 Destination IP의 평균 발생량을 나타내는 $avg(D_{IP})$에 대한 단위시간 Destination IP의 발생량을 나타내는 D_{IP}의 비율이 threshold α_2를 초과하고, D_{IP}가 설정된 threshold β_2를 초과하는지 검사한다. 위 조건을 만족하는 경우 평상시에 비해 Destination

IP의 발생량이 급증한 것이므로 Host Scanning 공격으로 분석할 수 있다.

3.3.3 DDoS

DDoS 공격은 악의적인 사용자로부터 감염된 다수의 PC로부터 특정 목적지에 대하여 비정상적인 네트워크 트래픽을 발생시켜 해당 목적지의 서버를 다운시키는 공격이다. 공격발생 시 보안 로그 상에서 전체 로그의 양이 증가하고, Source IP의 수가 급증하지만, 상대적으로 Destination IP, Port는 낮은 증가율을 보인다. 공격 탐지 기준에 대한 수식은 다음과 같다.

$$\text{DDoS} = \{(\frac{Total_{Log}}{avg(Total_{Log})} > \alpha_3) \text{ OR } (\frac{S_{IP}}{avg(S_{IP})} > \beta_3)\}$$
$$\text{AND } (\frac{S_{IP}}{D_{IP}} > \gamma_3) \text{ ---------- (식 3)}$$

(식 3)과 같이 다수의 PC로부터 비정상적으로 많은 로그가 발생하였는지 확인하기 위해 보안 로그의 평균 발생량을 나타내는 $avg(Total_{Log})$에 대한 단위시간 보안 로그 발생량 $Total_{Log}$의 비율이 설정된 threshold α_3를 초과하거나 Source IP의 평균 발생량을 나타내는 $avg(S_{IP})$에 대한 단위시간 Source IP의 발생량을 나타내는 S_{IP}의 비율이 설정된 threshold β_3를 초과하는지 검사한다. DDoS 공격은 다수의 근원지로부터 특정 목적지를 대상으로 수행되므로 S_{IP}에 대한 D_{IP}의 비율이 설정된 threshold γ_3을 초과하는지 검사하여 조건을 만족하면 DDoS 공격으로 분석한다.

3.3.4 Worm 활동

웜 활동이 발생되면 해당 서버에서 자기 자신을 복제하여 서버를 마비시키고, 네트워크를 통하여 다수의 목적지에 웜을 전파하여 해당 네트워크를 손상시키고 파일 등을 악의적으로 암호화 한다. 또한 웜 활동은 일반적인 네트워크 트래픽이라고 보기 어려운 규칙적인 활동 양상을 보이기도 한다. 웜 활동 발생 시, Destination IP, Port의 양이 증가하며, 규칙적인 주기로 Source IP의 발생량이 증가한다. 공격 탐지 기준에 대한 수식은 다음과 같다.

$$\text{Worm} = (\frac{D_{IP}}{avg(D_{IP})} > \alpha_4 \text{ AND } \frac{D_{Port}}{avg(D_{Port})} > \beta_4)$$
$$\text{OR } (\frac{S_{IP}}{avg(S_{IP})} > \gamma_4) \text{ ---------- (식 4)}$$

(식 4)와 같이 보안 로그에서 $avg(D_{IP})$, $avg(D_{Port})$에 대한 D_{IP}, D_{Port}의 비율이 각각 threshold α_4, β_4을 초과하거나, $avg(S_{IP})$에 대한 S_{IP}의 비율이 threshold γ_4를 초과하는 경우 설정된 시점에 대한 Worm 활동으로 분석할 수 있다.

3.3.5 Network Error

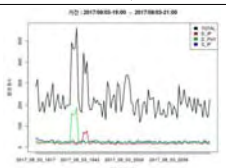
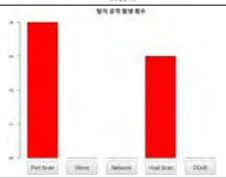
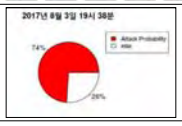
보안 로그를 수집하면 방대한 양의 로그가 축적되며 이는 해당 서버에서 수용 불가능한 네트워크 트래픽이 발생되어 서버가 다운될 수 있다. 또한, 해당 서버의 네트워크 또는 방화벽 스위치 장애가 발생하는 경우에도 네트워크 장애가 발생할 수 있다. 네트워크 장애가 발생하면 전체 로그의 양이 급증하고 몇 분 뒤, 로그 발생량이 급감하는 특징이 있다. 네트워크 장애를 탐지하기 위한 기준은 해당 네트워크의 특정 지점의 로그 발생량이 일정 Threshold를 초과하고, 특정 지점의 단위시간 로그 양과 다음 시간의 로그 양을 비교하는 것이다. 공격 탐지를 위한 수식은 다음과 같다.

$$\text{Network Error} = (Total_{Log(Q)} * \beta_5 > Total_{Log(Q+1)})$$
$$\text{AND } (Total_{Log(Q)} > \alpha_5) \text{ -------- (식 5)}$$

(식 5)와 같이 전체 로그 수가 해당 시스템에서 처리 가능한 범위 이내인지 확인하기 위해, 현재 시간의 보안 로그의 수를 나타내는 $Total_{Log(Q)}$이 threshold α_5를 초과하는지 검사한다. 또한 시스템 다운으로 인해 수분 뒤 전체 로그 수가 급감하는지 확인하기 위해 $Total_{Log(Q)}$에 threshold β_5를 곱한 값이 다음 시간의 보안 로그의 수를 나타내는 $Total_{Log(Q+1)}$를 초과하는지 검사한다. 해당 조건이 만족되면 Network Error로 분석한다.

3.4 시각화 인터페이스

시각화 인터페이스는 보안 로그를 각 공격에 대하여 분석을 수행한 후, 통계시각화 도구 R[7]을 사용하여 사용자에게 웹 인터페이스 형식으로 제공한다. 시각화 인터페이스는 꺾은선 그래프, 막대 그래프, 원형 그래프가 존재하며, 각 그래프의 특징 및 형태는 <표 1>과 같다.

종 류	특 징	형 태
꺾은선 그래프	로그 발생 시간 대비 로그 발생량을 나타내며, 보안 로그의 각 속성을 4가지 색으로 구분하여 각 속성을 시각화하여 제공한다.	
막대 그래프	검색한 시간에 5가지 네트워크 공격 패턴이 탐지된 횟수를 나타내는데 사용된다.	
원형 그래프	탐지된 공격이 발생한 시간 및 해당 시간에 공격이 발생한 확률을 보여준다.	

<표 1> 시각화 인터페이스의 특징 및 형태

4. 네트워크 공격 패턴 분석

4.1 Port Scanning

공격 분석에 사용된 데이터는 Snort를 통하여 실시간으로 수집된 보안 로그이며, 네트워크 검색 및 Scanning이 가능한 툴인 Nmap을 사용하여 해당 서버에 Port Scanning 공격을 시도하여 테스트를 진행하였다. 분석 결과는 <표 2>와 같다.

Time(시:분)	D_{Port}	$avg(D_{Port})$	$rate(D_{Port})$
19:35	25	25.94	0.963
19:36	21		0.809
19:37	24		0.925
19:38	**160**		**6.168**
19:39	**154**		**5.936**
19:40	**156**		**6.013**

<표 2> Port Scanning 공격에 대한 분석

(식 1)에 적용한 결과로 $rate(D_{Port})$가 threshold α_1(예를 들면, 1.8)을 초과하고, D_{Port}가 threshold β_1(예를 들면, 100)을 초과하는 것을 확인할 수 있다.

4.2 Host Scanning

사용 데이터는 실시간으로 수집된 보안 로그 및 Nmap을

사용하여 테스트를 진행하였다. 분석 결과는 <표 2>과 같다.

Time(시:분)	D_{IP}	$avg(D_{IP})$	$rate(D_{IP})$
19:43	18	18.22	0.987
19:44	18		0.987
19:45	**72**		**3.951**
19:46	**64**		**3.512**
19:47	**78**		**4.281**
19:48	19		1.042

<표 3> Host Scanning 공격에 대한 분석

(식 2)에 적용한 결과로 $rate(D_{IP})$가 threshold α_2=1.8 을 초과하고 D_{IP}가 threshold β_2=40 을 초과하는 것을 확인할 수 있다.

4.3 DDoS

공격 분석에 사용된 데이터는 VAST Challenge 2012에서 제공하는 방화벽 로그를 사용하였으며, 분석 결과는 <표 4>와 같다.

Time(시:분)	D_{IP}	S_{IP}	$avg(S_{IP})$	$Total$	$avg(Total)$
18:19	0	0	198.79	0	734.12
18:20	6	1		6	
18:21	0	0		0	
18:22	**32**	**1240**		**2611**	
18:23	**32**	**1248**		**2696**	
18:24	**23**	**2652**		**8459**	
18:27	**30**	**2650**		**8439**	

<표 4> DDoS 공격에 대한 분석

(식 3)에 적용한 결과로 $avg(Total)$에 대한 $Total$의 비율이 threshold α_3=10 을 초과하며 $avg(S_{IP})$에 대한 S_{IP}의 비율이 threshold β_3=5 을 초과하는 것을 확인할 수 있다. 그리고 D_{IP}에 대한 S_{IP}의 비율이 threshold γ_3=50 을 초과하는 것을 확인할 수 있다.

4.4 Worm 활동

분석에 사용된 데이터는 VAST Challenge 2012에서 제공하는 IDS 로그를 사용하였으며, 분석 결과는 <표 5>와 같다.

Time(시:분)	D_{IP}	$avg(D_{IP})$	D_{Port}	$avg(D_{Port})$	S_{IP}	$avg(S_{IP})$
19:30	**1**	3.59	**2**	3.88	**52**	10.29
19:31	1		1		4	
19:32	0		0		0	
19:33	1		1		4	
19:34	1		1		4	
19:35	1		1		6	
19:36	**1**		**1**		**32**	

<표 5> Worm 활동에 대한 분석

(식 4)에 적용한 결과로 $avg(S_{IP})$에 대한 S_{IP}의 비율이 threshold γ_4=2 를 초과하는 것을 확인할 수 있다. 또한 threshold α_4=3, β_4=3 는 Worm의 확산 시도를 탐지하기 위한 가중치이다.

4.5 Network Error

분석에 사용된 데이터는 Snort를 이용하여 실시간으로 보안 로그를 수집하여 사용하였으며, 이를 (식 5)에 적용하여 분석하였다. 분석 결과는 <표 6>과 같다.

Time(시:분)	$Total$	$avg(Total)$
07:46	118	115.37
07:47	123	
07:48	154	
07:49	**4910**	
07:50	57	

<표 6> Network Error에 대한 분석

(식 5)에 적용한 결과로 $Total_{(Q)}$(AM 7시 49분)가 threshold α_5=1000 을 초과하고, $Total_{(Q+1)}$(AM 7시 50분)에 대한 $Total_{(Q)}$ 의 비율이 threshold β_5=10 을 초과하는 것을 확인할 수 있다.

5. 결론

본 논문에서는 MapReduce 환경에서 네트워크 공격 탐지를 위한 실시간 로그 분석 시스템을 제안하였다. 개발된 시스템은 첫째, 5가지 네트워크 공격 패턴의 특징을 파악하고 공격 탐지 기준을 설정하여 실시간으로 네트워크 공격 패턴에 대한 분석을 수행하였다. 둘째, 보안 로그의 분석 결과를 시각화하여 사용자가 네트워크 상태를 쉽게 파악함으로써 네트워크 공격에 정확하고 빠르게 대응할 수 있도록 하였다. 마지막으로 분석을 통해 본 논문에서 제시한 공격 탐지 조건의 타당성을 제시하였다.

향후 연구로는 본 논문에서 지원하는 5가지 공격 이외에 보안 로그의 이상 패턴에 대한 탐지 조건을 설정하여 다양한 네트워크 공격에 대한 탐지를 수행하는 것이다.

참고문헌

[1] 이동건, et al. "RGB Palette 를 이용한 보안 로그 시각화 및 보안 위협 인식." 정보보호학회논문지 25.1 (2015): 61-73.

[2] 최대수, et al. "MapReduce 를 이용한 대용량 보안 로그 분석." 한국정보기술학회논문지 9.8 (2011): 125-132.

[3] White, Tom. Hadoop: The definitive guide. "O'Reilly Media, Inc.", 2012.

[4] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.

[5] Roesch, Martin. "Snort: Lightweight intrusion detection for networks." Lisa. Vol. 99. No. 1. 1999.

[6] Thusoo, Ashish, et al. "Hive: a warehousing solution over a map-reduce framework." Proceedings of the VLDB Endowment 2.2 (2009): 1626-1629.

[7] Team, R. Core. "R language definition." Vienna, Austria: R foundation for statistical computing (2000).

[8] Cook, Kristin, et al. "VAST Challenge 2012: Visual analytics for big data." Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. IEEE, 2012.

[9] Lyon, Gordon Fyodor. Nmap network scanning: The official Nmap project guide to network discovery and security scanning. Insecure, 2009.