Analyzing Tweets for Cybersecurity

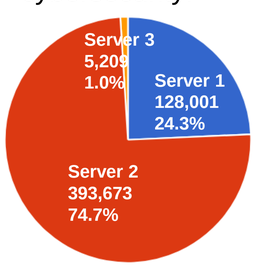
Kyle Murfitt, Eunbi Kim, Devon Smart, Ching-yu Huang
School of Computer Science, Kean University, Union, NJ, chuang@kean.edu

Abstract

We aim to perform sentiment analysis on Twitter tweets related to cybersecurity. This project focuses on developing models and approaches to identify cybersecurity terminology and identify patterns of cyber threats in tweets. Tweets related to cybersecurity comprise the data we utilize as the training set, which will help us study related terminologies and patterns concerning cybersecurity in social media.

Background

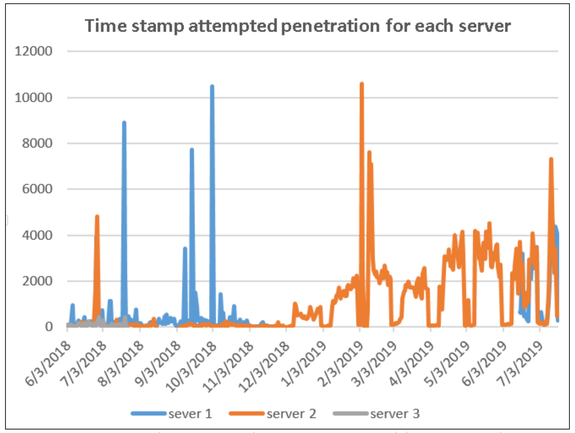
This project tried to find patterns of cyber threats by developing a database of cybersecurity terminology using Twitter tweets. It provides an opportunity to apply data mining technology and research on social media and cybersecurity. Outcomes will be very helpful to establish and support the Center for Academic Excellence Cybersecurity mentorship program at Kean. Also have the opportunity to be involved with the research on social media and cybersecurity.



The attempted penetration for each server

Methods

1. Develop programs to download the tweets automatically use with cronjob.
2. Develop programs that can extract information from the tweets, and transform the data into the database.
3. Design database schema and storage architecture that can store tweets and cybersecurity related information.
4. Develop integrated programs to perform sentiment analysis on the text, summarize the patterns of cyber threats, and identify the connections between users and cyber hackers.



Country	Count
China	117,109
United States	94,169
France	44,290
South Korea	33,678
India	19,230
Brazil	16,830
Germany	16,707
Singapore	15,296
Canada	13,242
Taiwan	12,908

Top 10 country which tried to attack

IP	Count	Country
211.50.130.9	10,852	South Korea
162.243.131.116	10,502	United States
40.143.1.60	10,460	United States
27.105.92.6	8,328	Taiwan
117.22.228.210	8,097	China
198.148.118.199	5,386	United States
119.10.114.9	4,631	China
202.125.157.67	3,590	Pakistan
211.234.110.126	3,382	South Korea
59.46.36.114	2,293	China

Top 10 IP which tried to attack

user	Count
root	119,300
admin	17,129
test	10,412
user	5,397
ubuntu	4,077
oracle	3,755
postgres	3,485
ftpuser	3,211
mysql	2,734
nagios	2,698

Top 10 login which tried to attack

Results

Total number of data

Log File	IP Address	Twitter Tweets
526,883	25,110	36,722,960

Example of log file data

server	log_date	log_time	ip	port	type	user	rawinput
yoda	2018-07-22	05:40:15	121.166.99.177	41438	Invalid User	jira	Jul 22 05:40:15 yoda sshd[4637]: Failed passw…
yoda	2018-07-22	06:01:28	139.59.251.117	52508	Failed password	root	Jul 22 06:01:28 yoda sshd[4665]: Failed passw…
yoda	2018-07-22	06:01:44	35.168.172.225	34218	Invalid User	chloe	Jul 22 06:01:44 yoda sshd[4668]: Failed passw…
yoda	2018-07-22	06:23:42	112.25.214.199	35849	Invalid User	oracle	Jul 22 06:23:42 yoda sshd[4683]: Failed passw…
yoda	2018-07-22	06:25:40	164.77.220.115	39248	Invalid User	pi	Jul 22 06:25:40 yoda sshd[4687]: Failed passw…

Example of Twitter data

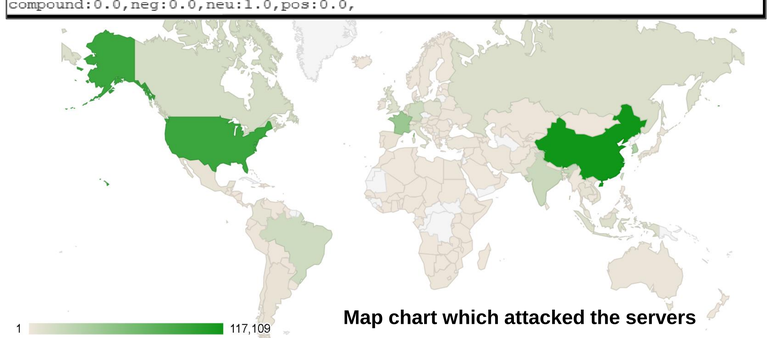
id	twitter_name	user_id	date	follower	friend	text	place_name	country_code	latitude	longitude
1146431649929977857	~Secilygrace~	9069088828…	Wed Jul 0…	119	972	To report sc…	Fl	US	34.83408568	-87.66652946
1146432026683142145	~Secilygrace~	9069088828…	Wed Jul 0…	119	972	To report sc…	Fl	US	34.83408568	-87.66652946
1146432042906923008	???????????…	1316862031	Wed Jul 0…	475	349	D-Link agre…	Wa	PL	52.210569	21.021937
1146432081603358720	TMJ-MEP Sec…	139452990	Wed Jul 0…	271	262	If jobs with…	Do	US	43.1978624	-70.8736698
1146432113652011008	Nassau Security	573867108	Wed Jul 0…	310	281	Have you e…	Va	US	40.6642699	-73.7084645

Example of IP data

ip	city	country	country_code	latitude	longitude
1.100.143.227	Daegu	South Korea	KR	37.51120000	126.97410000
1.100.212.195	Daegu	South Korea	KR	37.51120000	126.97410000
1.109.28.115	Daegu	South Korea	KR	37.51120000	126.97410000
1.109.92.244	Daegu	South Korea	KR	37.51120000	126.97410000
1.119.10.198	Beijing	China	CN	39.92890000	116.38830000

Tweet Polarity Calculation using Vader

```
I, and others, thought it might've been an DDoS. I was wrong. https://t.co/EeBRj
PjLRf
compound:-0.4767,neg:0.22,neu:0.78,pos:0.0,
@lbiredneck @dbroncos78087 @charles_wilkie @Pavox @realDonaldTrump @USNavy They
emailed spyware to the defense. Read about it.
compound:0.128,neg:0.0,neu:0.903,pos:0.097,
Malwarebytes Crack 3.8.3 Premium With Full Keys | Cracksmod https://t.co/JfRn8Hf
yw4
compound:0.0,neg:0.0,neu:1.0,pos:0.0,
Egal ob Ransomware, Verlust, oder Diebstahl - über das Thema Backup machen sich
viele leider erst Gedanken, wenn sc… https://t.co/eu2fVdlawR
compound:0.0,neg:0.0,neu:1.0,pos:0.0,
@TriiiiiKz Je l'ai croisé une fois en ranked sans ddos je l'ai poutre 3z il a ddo
s mdr
compound:0.0,neg:0.0,neu:1.0,pos:0.0,
China's Border Guards Secretly Installing Spyware App on Tourists' Phones https:
//t.co/7dTOe9ZoND via @TheHackersNews
compound:0.0,neg:0.0,neu:1.0,pos:0.0,
#Cyberthreats are not a #Technology issue but rather a Business risk &gt;&gt;&gt
; #PwC via @MikeQuindazzi &gt;&gt;&gt;… https://t.co/wyx7xr3zBs
compound:-0.3919,neg:0.159,neu:0.841,pos:0.0,
Creativo eufemismo de hackeo por malware ????????: "desestabilización de la frec
uencia" de la energía. ???????
compound:0.0,neg:0.0,neu:1.0,pos:0.0,
```



Map chart which attacked the servers

Materials

- **Platform: Linux OS, MySQL Database**
- **Frontend: HTML, CSS, JavaScript**
- **Backend: PHP, Python**
- **API: Tweepy, Geoip-db, Google Charts**
- **IP, country, city, latitude and longitude**
- **Secure log files: July 2018 - July 2019**
- **Twitter data: 7/3 - 9/1, 2019**

Summary

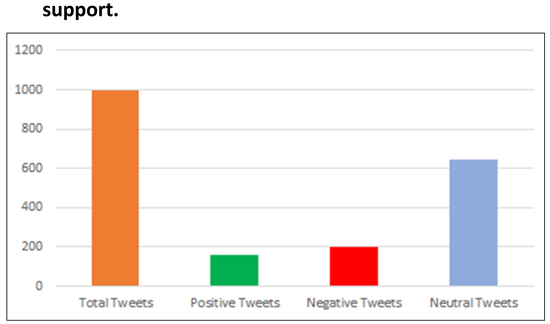
Log files are used to keep track of all the users that have accessed a server. From log files, we collected a total of 25,110 IP addresses with 526,883 data logs who attempted penetrations. We found out China was the highest country and the US was the second highest country who attempted penetrations. From the twitter data files of geo location, the US was the highest country who mentioned about cyber attack.

References

1. **Google Maps API tutorials, Google.**
2. **Twitter API (Tweepy), Twitter.**
3. **Geo-Location, https://geoip-db.com**

Acknowledgement

This research was funded by the Students Partnering with Faculty (SpF) grant supported by the Office of Research and Sponsored Programs (ORSP) at Kean University. We really appreciate Kean University's support.



Number of Tweets and Tweet Polarity