2018 International Conference on Computational Science and Computational Intelligence (CSCI)

Data Visualization in Cybersecurity

Syed Fahad Nadeem and Ching-Yu Huang
School of Computer Science, Kean University, Union, NJ 07083, USA
chuang@kean.edu

Abstract – **Cybersecurity is receiving more attention these years because an increasing number of hackings have been reported. Visualization of data can assist in bringing attention to this issue. Log files are used to keep track of all the users that have accessed a server. Charts and graphs can help get a better understanding of the security log files. The IP addresses extracted from log files can be used to track the location of the machines that are trying to connect to the servers. A third-party geo-locating API is available to match IP addresses to identify where the user resides. Google Maps API can mark where the IP addresses are on the map based on the location given. This research focuses on visualizing the unauthorized access attempts on the servers by extracting the user data from the log files that were collected from 3 servers over a 2-month period.**

Keywords: *Apache web server, cybersecurity, logs, MySQL database, Google Charts, visualization*

1. INTRODUCTION

Security of technology has become a big issue and resulted in costing companies at least $15 million (Griffiths, 2015). Cybercrimes can be expensive if threats are not caught on time or addressed properly. One way to gain an understanding of cyber security is to visualize it. The visualization of data can be used to create a better understanding for individuals on cybersecurity.

Angelini et al. (2017) focused on visualizing how anti-malware utilities on web browsers determine whether a download is safe for the user. Angelini et al. (2017) uses a tool to download web traffic and store it onto a database which is then visualized by using tables on location data using geo-location from the IP address. Angelini et al. (2017) is using IP and location data to create an association of what come from each region. This can allow the security operator to focus on specific IPs and regions (Angelini et al., 2017).

Hao et al. (2017) is using visualizations to show security data. Hao et al. (2017) project involves finding methods to interpret network security alerts and flow traffic as a visualization system. Visualization is created by "using HTML5 and JavaScript for the user interface and visualization components, and MySQL and PHP for server-side data management" (Hao et al., 2017, p. 3).

Yan et al. (2017) project consists of taking equipment status data and geographical data and making the data intuitive and convenient to understand. This equipment would take a route for a task that it needed to complete and create a visualization using a map to show where the equipment went (Yan et al., 2017) Using a map, it gives a better idea of where the work is being done.

Theron et al. (2017) looks at visualizing intrusion detection on a network. Theron et al. (2017) uses a tool called Principal Component Analysis to detect the intrusion. Using this tool, it creates "highly flexible and intuitive interface that allows the user to navigate through the enormous amount of data collected in the network, in order to find anomalous or unexpected behaviors" (Theron et al., 2017, p.1). Allowing the user to look at large amount of data assists in prevent attacks from happening. Theron et al. (2017) discusses the simplicity of showing a visualization of their data: "We propose a workflow for the interactive visual analysis that hides the mathematical complexity of the models behind" (p.7). Taking the mathematical complexity and simplifying it to be visualized makes the data more accessible.

Goodall (2009) analyzes the comparison between a visualized application of looking at network packets versus a traditional interface. Taking 8 students with basic networking knowledge, they looked at network packet data with a visualized application versus a traditional interface to look at network packets (Goodall, 2009). Students were more enthusiastic and were able to get the information easier with the visual tool rather than the Ethereal (Goodall, 2009). It is important for the visualization of cybersecurity community to assure that there are people that can test visualizations (Goodall, 2009). Although there is no testing involved in this project, it is still important to create a visualization of the data that is readable and can be easily understood as seen in all of the literature reviews.

At Kean University, there are many servers that hold information about students, faculty, and sensitive information about them. This research analyzes 3 servers on the kean.edu network to visualize how these servers are affected by cybercrime. These 3 servers are used for educational purposes and used by Kean students and faculty. Taking the log files of these servers, we can take

978-1-7281-1360-9/18/$31.00 ©2018 IEEE
DOI 10.1109/CSCI.2018.00017

data from it to see if we are being affected by unauthorized attempts.

2. DATA SOURCE AND STORAGE

Hardware used for the project are three servers which host all the files programed and the extracted data stored in a database server. These servers are running the operating system Red Hat and are using apache2 to run the webservers for the front end. Tools that were used for programming were HTML and JavaScript to provide front end for this project. For the backend PHP and MySQL were used to process the data and store it. All graphs and maps used for this project were available using Google Maps and Google Charts. IP location data was retrieved from https://extreme-ip-lookup.com/ API.

2.1 Log Files

For this project a PHP script was written to take the date, time, server, type of entry, user, IP address, and port from the log files retrieved from the server. Each line in the log file is read by the script and for the line to be considered for data entry it must contain either a "Failed password" string or a "Invalid user" for these are the warning signs of an attempted unauthorized entry. Once the line is determined to be a valid attempt, each string in the line is broken apart and assigned a variable. Finally, a MySQL query is created, and the data is entered into the MySQL database.

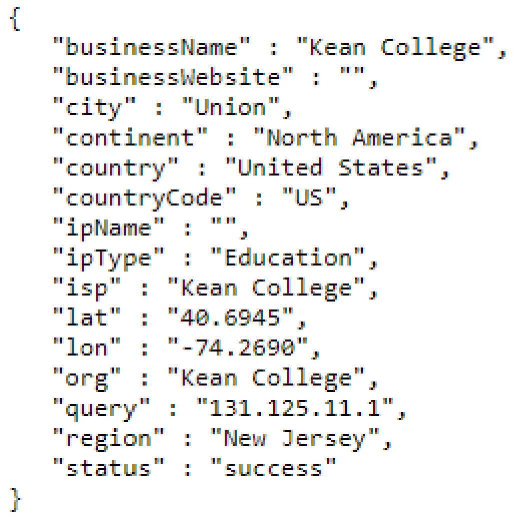
MySQL is being used to store all the data that is retrieved in the log file. Every data point that is taken from the PHP program is being stored into the database along with the actual line itself and a timestamp. With the data in the MySQL database, statistic can be seen using different MySQL queries. This will be used in conjunction with PHP to display all the results.

2.2 Visualization

HTML and JavaScript are being used to display the Graphs, Charts, and Maps provided by Google. On the Google Documentation, it provides us with skeleton code from which we can modify the data that goes into the chart as well as how the chart looks. Using this we can use PHP to bridge MySQL and the frontend aspects of this project. "PHP can generate HTML, and HTML can pass information to PHP" ("PHP: PHP and HTML - Manual," n.d.). This same concept applies to JavaScript except JavaScript cannot pass information on to PHP. With this, we can manipulate the JavaScript code from Google and write it using PHP which we then use to populate the graphs and map with information from the database.

2.3 Location Data

Using the IP addresses provided by the log files, we were able to use a third-party location API to get the origin of the IP address. The API that was used for this project was extreme-ip-lookup.com. When an IP address is given to the API it returns a JSON file which can then be decoded in the PHP program.



```
{
   "businessName" : "Kean College",
   "businessWebsite" : "",
   "city" : "Union",
   "continent" : "North America",
   "country" : "United States",
   "countryCode" : "US",
   "ipName" : "",
   "ipType" : "Education",
   "isp" : "Kean College",
   "lat" : "40.6945",
   "lon" : "-74.2690",
   "org" : "Kean College",
   "query" : "131.125.11.1",
   "region" : "New Jersey",
   "status" : "success"
}
```

Figure 1. When making the request to get location data from the website, it will return the information in JSON format. With this, we can extract whatever data points we need from it as it returns multiple points of data.

The information that was taken for this project was the city, country, country code, latitude, and longitude. This database will be responsible for displaying the information required for the implementation of google maps.

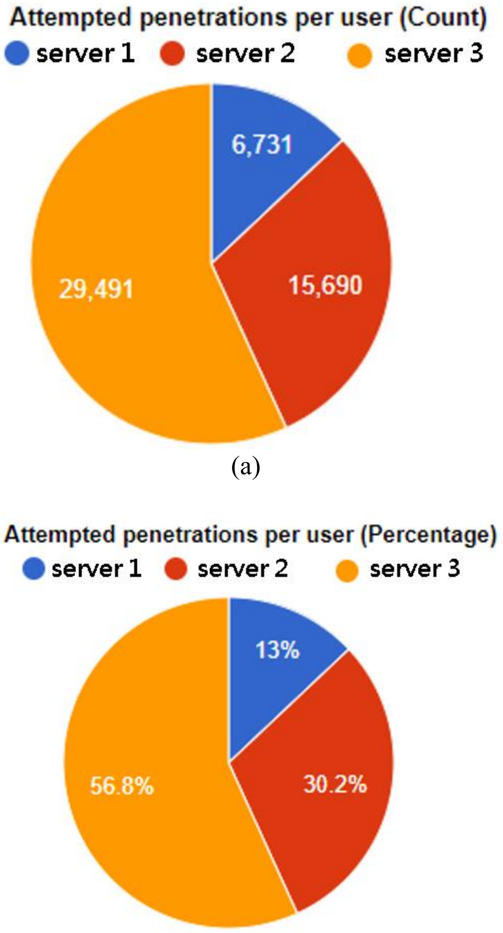
Google maps can be used to display the location given by the location API. We can do this by taking the latitude and longitude that was stored into the database and displaying it on the location marker on the map. Embedding google maps into a webpage requires mostly JavaScript to configure and hold the data for the map and HTML to display it on the webpage. The data in the database then needs to be able to follow the format that google uses for its maps. This can be done by using PHP to extract the data from the database and then give that data to JavaScript for it to display the information. With this method, we can successfully plot out locations with the latitude and longitude data.

3. FINDINGS

3.1 Server Statistics

Data collection were between the dates of June 3rd and July 29th. Between these days, there were 51912 attempts of unauthorized entry between the three servers at Kean

University. A break down per server can be visualized using Google Charts.

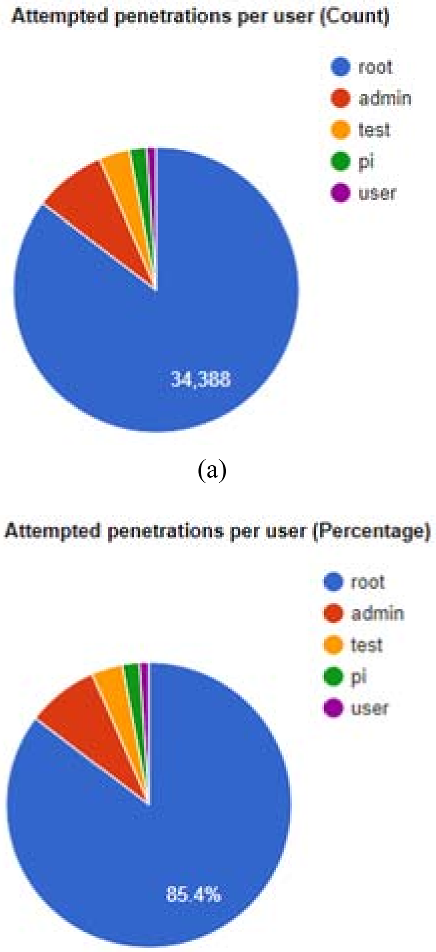


(a)

(b)

Figure 2. Visualization of count from 3 servers. The top chart (a) shows actual count of each server and bottom chart (b) is the percentage per each server. Both charts are screenshotted from the webpage.

Root user is used for system administration ("Superuser," 2018). If this user is accessed by an unauthorized system, the results can be catastrophic. Across all three servers, root was the user that was most attempted to be accessed accounting for over 85.4% of all unauthorized attempts. When examing the raw log files, we could see that 0% of all unauthorized attempts were successful. The attempted entries can be visualized using charts:

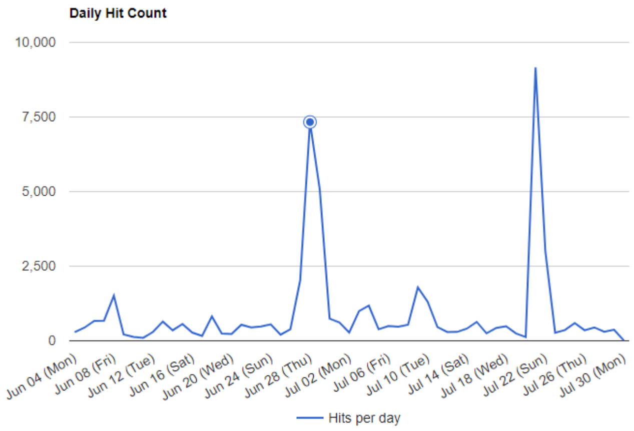


(a)

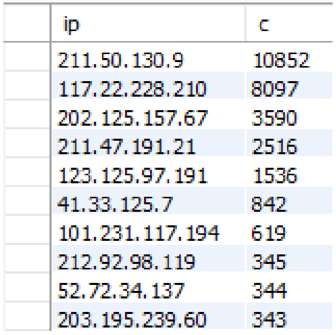
(b)

Figure 3. Pie chart of top 5 users that were attempting to gain unauthorized access. Top chart (a) shows the count of the graph but since the 'root' user takes up so much of the graph, the other text will not display which shows the significance of the data. The bottom chart (b) is similar except with percentage instead of count.

Date data is another statistic we can look at from this project. The two date which saw a spike where the dates of June 28th at 7,330 hits and July 21 at 9159 hits. This was caused by only one or two IP address doing thousands of attempts in one day. Below, a chart of the dates can be seen in a line graph.



(a)



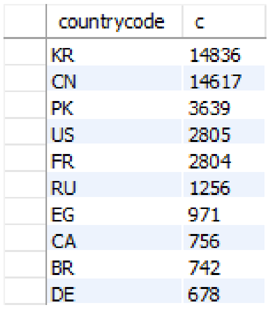
ip	c
211.50.130.9	10852
117.22.228.210	8097
202.125.157.67	3590
211.47.191.21	2516
123.125.97.191	1536
41.33.125.7	842
101.231.117.194	619
212.92.98.119	345
52.72.34.137	344
203.195.239.60	343

(b)

Figure 4. Chart (a) shows data for dates between June 3rd and July 29th. Spikes can be seen for the two days where one IP does thousands of hits per day as seen on (b) with the top ten IP address with the amount of times they attempted entry.

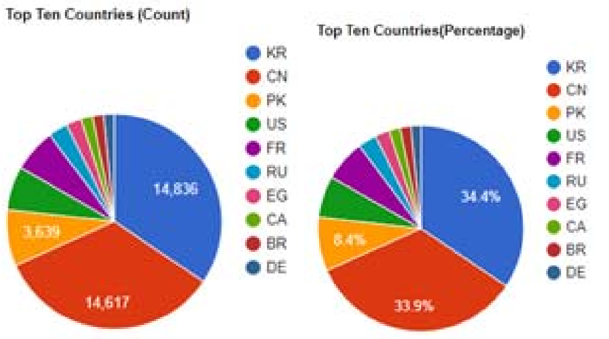
4.2 Location Data

There were 133 distinct countries from where unauthorized attempts had come from. The top ten countries in order where Korea, China, Pakistan, United States, France, Russia, Egypt, Canada, Brazil, and Germany. This data was found by using a MySQL query that looks at both tables to get the IP from the log files and the location from the IP table. Using google charts, we can see how significant each country is to contributing to unauthorized access attempts.



countrycode	c
KR	14836
CN	14617
PK	3639
US	2805
FR	2804
RU	1256
EG	971
CA	756
BR	742
DE	678

(a)



(b) (c)

Figure 5. The chart (a) and (b) show the top ten countries that were found in our IP database. Chart (c) shows the significance of two countries which takes over 50 percent of the top ten countries that have tried unauthorized entries. The image on the right is a raw count of the pie charts. This is the data that is populating the pie charts.

This data can also be visualized using google maps. Using the same techniques as seen with the charts, we can populate the data field using the country code's latitude and longitude to give a location.

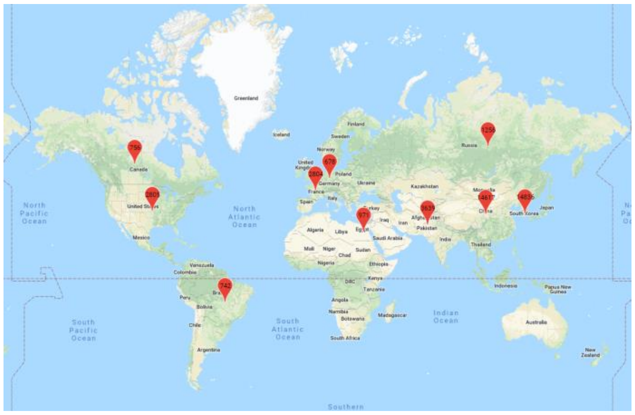


Figure 6. Location data is visualized using an actual map. Each marker has the number of attempts per country.

4. CONCLUSIONS

Overall, we can see that most of the unauthorized attempts come from outside of the United States. The data gained from this project allows the server administrator to examine exactly how much activity the server is really getting. With result as big as this in a span of two months does bring up some questions about the security of the network that these servers are on. With unique server names only known by specific students how are people able to locate these servers and attempt to hack them? Could it be a vulnerability in the network? With the data, it allows us to ask these questions and move onto the next step to better secure our servers.

In this day and age, the threat for cyber-attacks is as high as they are for environmental disasters (Johansson, 2018). With this it is important that we at least stay informed on what is going on in the field of cyber security. This research demonstrates that data that was once not understandable to someone who is not in the field, can be visualized and distributed to allow people to be more informed.

In the future, we hope to refine the research and to create a dynamic website which can show statistics for all the data that we have in the database. Another way to expand on this work would be to collect log files from a span of several months or several years to look at certain trends to see whether unauthorized attempts are provoked by time of year, time of day, or time of week.

5. REFERENCES

Angelini, M., Aniello, L., Lenti, S., Santucci, G., & Ucci, D. (2017). The goods, the bads and the uglies: Supporting decisions in malware detection through visual analytics (pp. 1–8). IEEE. https://doi.org/10.1109/VIZSEC.2017.8062199

Cybersecurity operations and the role of visualization, design, and usability. (2018, January 26). Retrieved August 14, 2018, from https://sdtimes.com/data/cybersecurity-operations-and-the-role-of-visualization-design-and-usability/

Goodall, J. R. (2009). Visualization is better! A comparative evaluation. In *2009 6th International Workshop on Visualization for Cyber Security* (pp. 57–68). Atlantic City, NJ, USA: IEEE. https://doi.org/10.1109/VIZSEC.2009.5375543

Griffiths, J. (2015, October 8). Cybercrime costs the average U.S. firm $15 million a year. Retrieved August 14, 2018, from https://money.cnn.com/2015/10/08/technology/cybercrime-cost-business/index.html

Hao, L., Healey, C. G., & Hutchinson, S. E. (2015). Ensemble visualization for cyber situation awareness of network security data (pp. 1–8). IEEE. https://doi.org/10.1109/VIZSEC.2015.7312766

Johansson, G. (2018, January 18). Cyber-attacks one of the biggest threats to the world in 2018 says WEF. Retrieved August 15, 2018, from https://beta.scmagazineuk.com/article/1473450

Kwan-Liu Ma. (2006). Guest Editor's Introduction: Visualization for Cybersecurity. *IEEE Computer Graphics and Applications*, *26*(2), 26–27. https://doi.org/10.1109/MCG.2006.33

McKenna, S., Staheli, D., Fulcher, C., & Meyer, M. (2016). BubbleNet: A Cyber Security Dashboard for Visualizing Patterns. *Computer Graphics Forum*, *35*(3), 281–290. https://doi.org/10.1111/cgf.12904

PHP: PHP and HTML - Manual. (n.d.). Retrieved August 14, 2018, from http://php.net/manual/en/faq.html.php

Superuser. (2018). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Superuser&oldid=854234549

Theron, R., Magan-Carrion, R., Camacho, J., & Fernndez, G. M. (2017). Network-wide intrusion detection supported by multivariate analysis and interactive visualization (pp. 1–8). IEEE. https://doi.org/10.1109/VIZSEC.2017.8062198

Yan, H., Wang, J., & Xia, C. (2017). Research and Application of the Test Data Visualization (pp. 661–665). IEEE. https://doi.org/10.1109/DSC.2017.110