DESIGN OF A PREDICTION GAME IN THE DOMAIN OF COMPUTER SECURITY

An Undergraduate Research Scholars Thesis

by

SIDDHARTH SUNDAR

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Frank Shipman

May 2020

Major: Computer Science

TABLE OF CONTENTS

ABSTRACT

Design of a Prediction Game in the Domain of Computer Security

Siddharth Sundar
Department of Computer Science and Engineering
Texas A&M University

Research Advisor: Dr. Frank Shipman
Department of Computer Science and Engineering
Texas A&M University

Prediction Games are games where players analyze historical data and make predictions about future events. The predictions are scored which gives the players an idea of where they stand. This game is appropriate for domains where data is coming in frequently and where there is a decent quantity of historical data for participants to explore. Knowledge in Computer Security carries high value personally and professionally and statistics in this data domain is collected by many organizations for various reasons. This thesis explores the design of a prediction game in the field of Computer Security. The goals of this project include identifying data sets that could be used for a prediction game, designing a prediction activity which will be helpful to players, and developing a prototype version of the prediction game. A heuristic evaluation of the prototype will provide feedback for improvements to the game mechanics such as user interface and data visualizations. At the end of the project, there will be a greater understanding of the availability of computer security data, how it can be used for developing prediction games, and tradeoffs in the design of computer security prediction games.

ACKNOWLEDGMENTS

I would like to start off by first thanking my research professor, Dr. Shipman, for the guidance and support he provided throughout the course of my research. Without his ideas and suggestions, this thesis would have been difficult to construct.

I also would like to thank my friends at Texas A&M University who gave me some perspectives that I may not have considered.

Finally, I would like to thank my research mates Gabriel Dzodom and Ben D'Antonio for their insights and support.

NOMENCLATURE

PG	Prediction Game (s)
NFL	National Football League
CERT	Computer Emergency Readiness Team
US	United States

CHAPTER I

INTRODUCTION

Is there any other way of gaining data domain knowledge other than the usual textbooks, class lectures, and online tutorial videos? Is there a way to not only gain domain knowledge, but also be motivated in doing so? There is a way of accomplishing both and that "method to generate such motivation is gamification" (Dzodom and Shipman 2016). Gamification is the concept in which a player participates in some form of online or in person activity, individually or in teams, where the player earns points when completing a task successfully. Most people know this concept via video games, but this can be applied in education as well.

Prediction Games

One type of activity that falls under the category of gamification is Prediction Games. A Prediction Game (PG) is a game in which an individual looks at historical data and uses that data to make future predictions. Points are awarded accordingly. Many games can be classified as PG and we will discuss Fantasy Football, an existing PG, in further detail.

Fantasy Football

Imagine you are a coach of a NFL team and need to pick the best players in the upcoming NFL draft to help you win (Sablich). Fantasy Football is basically the scenario I described except it is a virtual experience. Many players (in most cases it is 10) participate in a draft where each participant looks at historical data and some current data to pick the NFL players that they think will be best for their team. In this analogy, the participant acts as the coach and the NFL players are the set of NFL players. Now, that a participant has their team, they compete against other participants head to head each week with a common goal, to win. Many factors are considered by

each participant in terms of who they want in their lineup each week which shows the concept of PG: Data Analysis. Without that, participants could not make the best lineup that would give them the best possible chance of winning.

Computer Security

Today, the entire world lives on technology, with everyone on their phones, laptops, tablets, desktops, and other electronic devices. However, most people do not know about the potential security threats these devices pose. For example, we take our phones out almost every time we receive an email notification, but most people will just click on the email notification and then click whatever link(s) is/are in the email and boom you have been hacked. However, problems like the one I described above would not happen if people had the basic knowledge in Computer Security. This knowledge in Computer Security is extremely valuable and can help the entire world at making better decisions about the data stored on their personal devices.

Goals for Computer Security Prediction Game

My research group believes that individuals can improve their data analysis skills in a particular domain through these PG. The domain of interest is Computer Security and I plan to test this hypothesis through Fantasy Phish Tank, a game where players will find areas that are most and least vulnerable to phishing attacks.

CHAPTER II

METHODS

There are a few methods I used throughout my Prediction Games research in the area of Computer Security. These methods include identifying potential datasets, selecting one dataset to be used as a first prototype, constructing a PG around the dataset, and enhancing the PG to aid players in analyzing the data.

Identifying Potential Datasets

For the first couple weeks, I was looking through the internet for datasets that could potentially be used in creating a prototype prediction game. The major requirements for the dataset included a dataset that gets new data at least once a week, a data set that has numbers (numerical data), and a dataset that is user friendly for importing into our Java framework. I was able to find two datasets that fit these technical specifications: the first was a PhishTank database and the second was the CERT database.

Selecting the Appropriate Dataset

After finding the potential datasets that fulfilled the requirements, my next task was figuring out which dataset should be used. Three factors were considered in the selection of the dataset. The first factor was the organization of the data, the second factor was the difficulty in the understanding of the data, and the third factor was the data feed methods. Although both the PhishTank database and the CERT database had JSON feeds, CERT's JSON feed was not organized in a user-friendly manner. The poor organization of the CERT data ultimately led to the difficulty in the understanding of the data (“National Vulnerability Database”). In addition, PhishTank some extra data feed options including CSV which allowed for easier fetching of the

database (“Statistics about Phishing Activity and PhishTank Usage”). Due to the reasons mentioned above, the PhishTank dataset was selected over the CERT dataset.

Designing the Prediction Game

What question would I ask the players to predict? This question was the main question that needed an answer in order for me to create the PG. When pondering upon this question, I considered Fantasy Climate and Fantasy Precipitation, two games that were developed by Gabriel Dzodom. In Fantasy Climate, a player needed to predict from weather data and a given list of cities, the city that is the warmest relative to its historic norm and the city that is the coolest relative to its historic norm. Similarly, in Fantasy Precipitation, a player needed to predict from precipitation data and a given list of cities, which cities would have precipitation or not (“Prediction Games”). These two games were extremely resourceful because these helped in developing the concept of my Prediction Game.

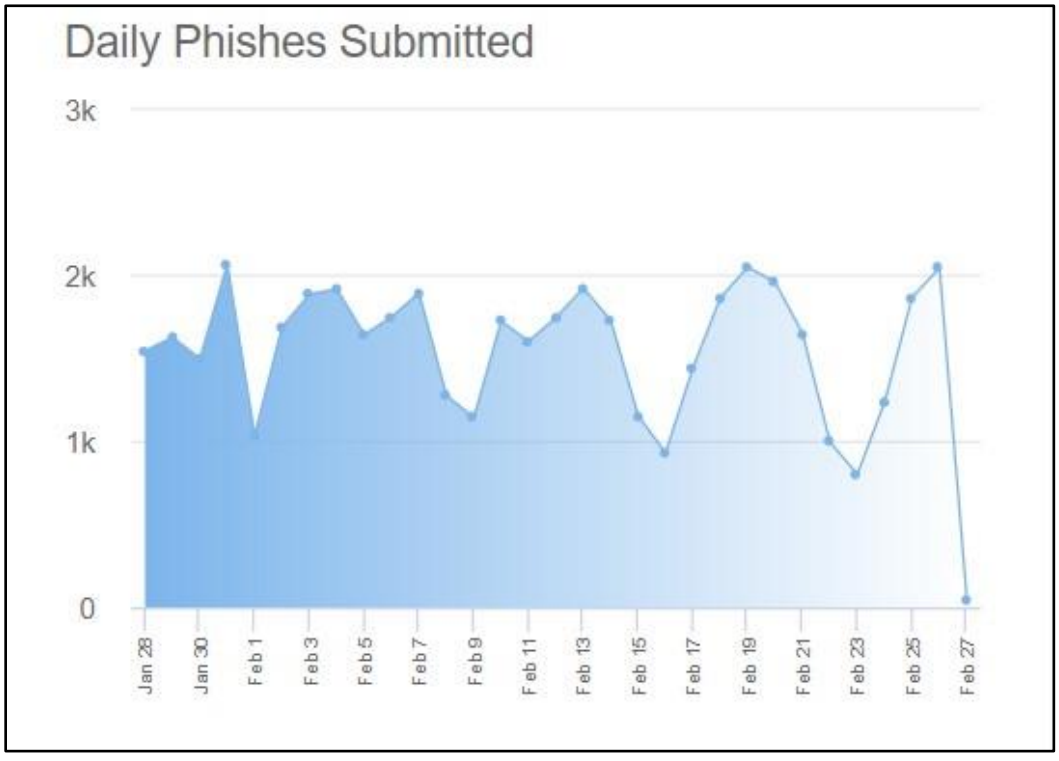

Figure 1. Line Graph of Submitted Phishes on a Daily Basis (PhishTank 2020)

The initial concept of my Prediction Game was to determine how many phishing threats would occur for a given period. Figure 1 depicts the number of Phishes that were submitted on a daily basis from January 28 till February 27, 2020. A player would be given this data visual and interpret this visual accordingly to make their prediction on the initial concept. However, there is a problem with this approach. The problem is that the data is fluctuating, which would make the game extremely difficult from a player's standpoint. The player would not be able to make a prediction solely off of the data and this would result in an extremely difficult and boring prediction game.

Enhancing the Prediction Game

In order to solve the problem mentioned in the previous paragraph, the data needed to be normalized. The new concept used the same idea of predictions based on the quantity of attacks but considered the geographic location of the incidents: predicting countries of increasing activity. This idea partially worked when normalizing data, but was not 100% effective due to the fact that certain parts of the world consistently had high phishing activity such as Europe and the United States. The massive variation in phishing activity amongst the countries can be contributed by cultural and political factors. For example, China might try to avoid reporting any kind of phishing activity due to their communist government. On the other hand, the United States would not place restrictions on reporting phishing activity due to their federal republic government. Another fact to note is that the amount of people with technology varies from country to country. If a country has limited access to technology, then their phishing reports would definitely show less cases than that of a country that has unlimited access. These cultural and political factors would cause the PG to be extremely easy for players because they would not even have to look at the data and can predict with ease and certainty that this country or that

country would have the highest number of phishing attacks. In this case, highest number just means comparing the total amount of attacks in an area with no respect to population and determining which is greater. After a period of time, the game would become boring to the point that players would stop playing. A final decision was made to consider the United States as the region of choice, similar to that of Fantasy Precipitation and Fantasy Climate.

CHAPTER III

RESULTS

Data Visualization

After deciding on the United States, I needed to decide on the data visualizations I would create off of the historical data for players to use to make their predictions. The first idea was a Heat Map of phishing activity in the US. Figure 2 shows a Heat Map of phishing activity in the US where data was accumulated for one week.

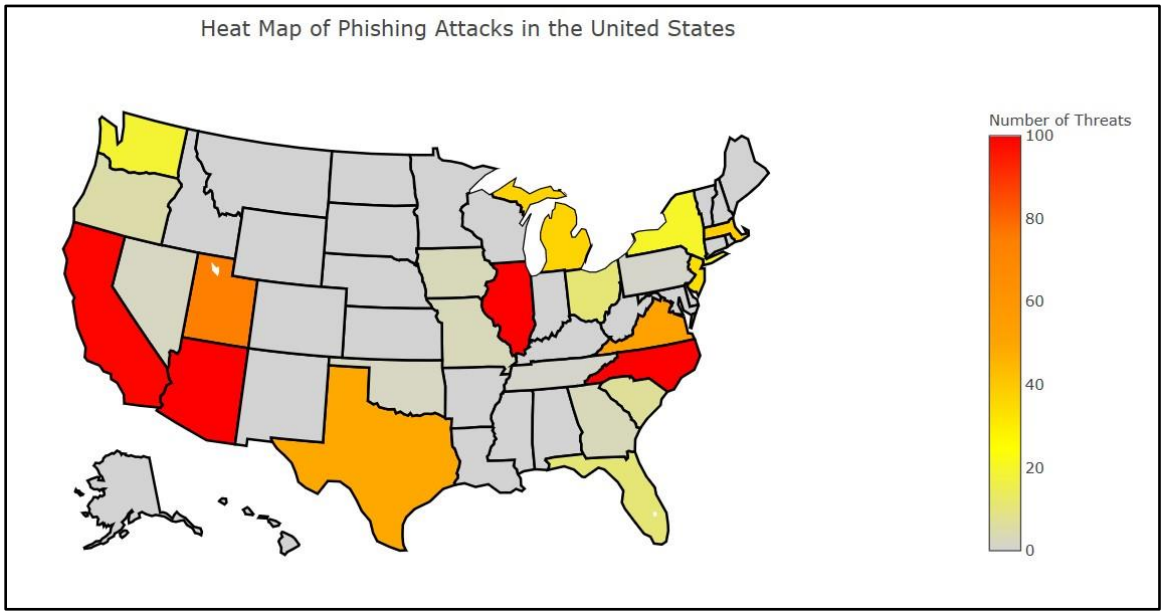

Figure 2. Heat Map of Phishing Attacks in the United States

A player can use this graph to predict which state in the US had higher levels of phishing activity and lower levels of phishing activity with some difficulty. The reason for this difficulty is because the heat map actually has a few places where phishing activity is really high (indicated by the high intensity in color). When there is more than one place where phishing happens frequently, it makes the game slightly difficult and fun for the player because a concept of

strategy is introduced. The second idea was a Line Graph of Phishing Activity over time in the United States. Figure 3 shows a line graph of phishing activity for 5 regions (Northeast, Southeast, Southwest, Midwest, and West) in the United States over the course of last year (2019). In this graph, the x-axis label "Date" refers to each day in the year 2019.

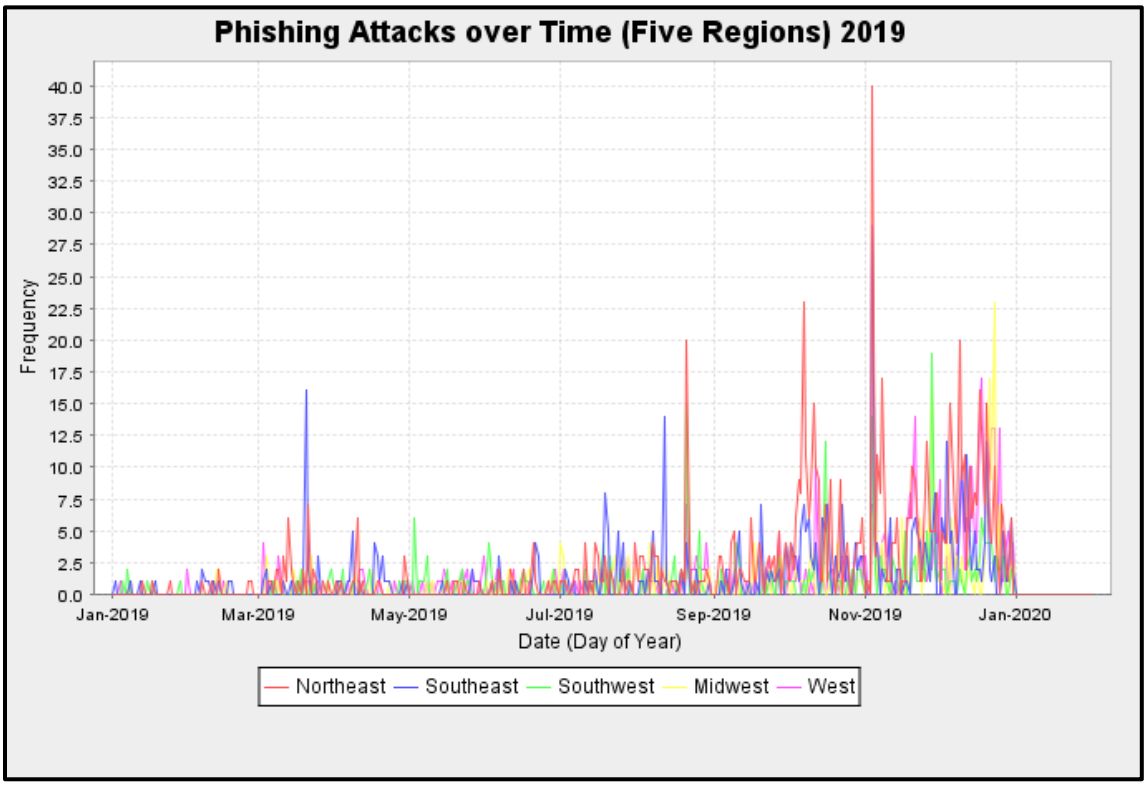

Figure 3. Line Graph of Phishing Attacks over Time on a Daily Basis (Five Regions)

After looking at this graph, I noticed one thing: there were too many fluctuations in the data. The data did not become normalized with this approach and I tried a different time metric. Instead of looking at the number of phishing attacks on a daily basis, I decided to take a look at

the number phishing attacks on a weekly basis (approximately 52 points a year). Figure 4 depicts this scenario.

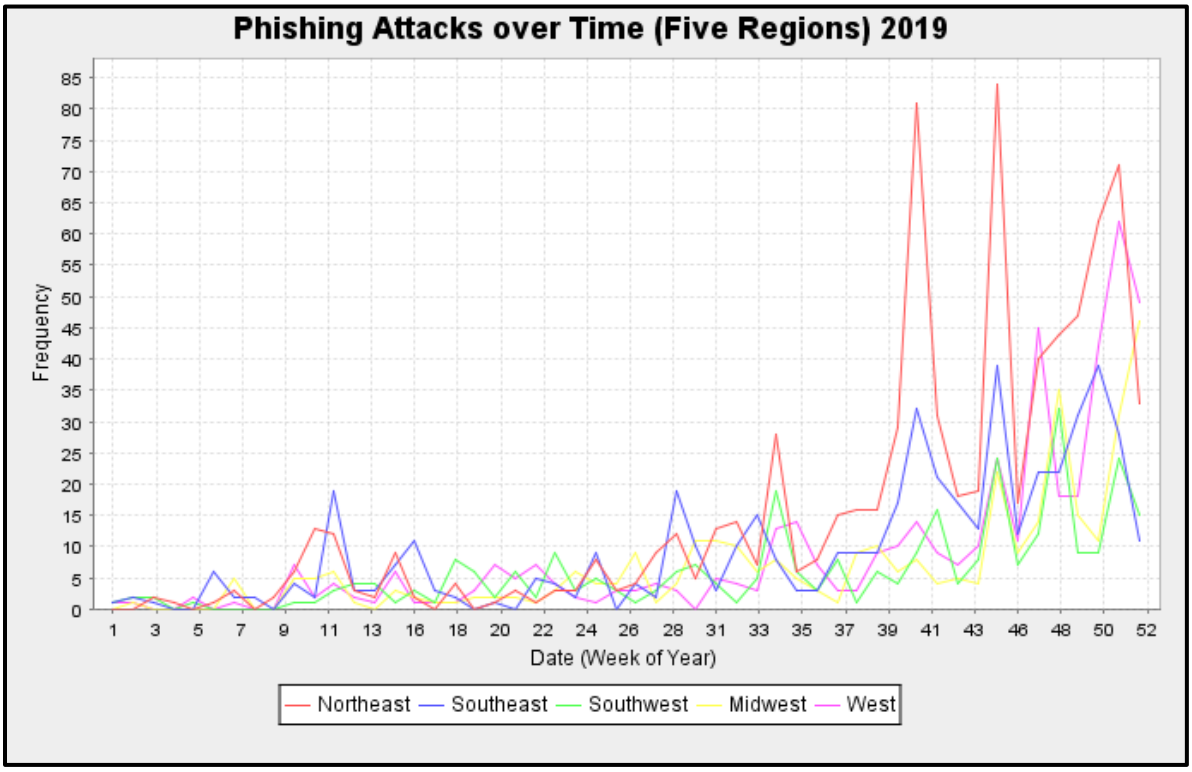

Figure 4. Line Graph of Phishing Attacks over Time on a Weekly Basis (Five Regions)

After looking at Figure 4 above, it seemed like the number of fluctuations went down in comparison to that of Figure 3, but I felt that it could be better. Instead of a 1 week time interval, I decided to look at a 2 week time interval of the same 5 regions for the year 2019. Figure 5 portrays this scenario.

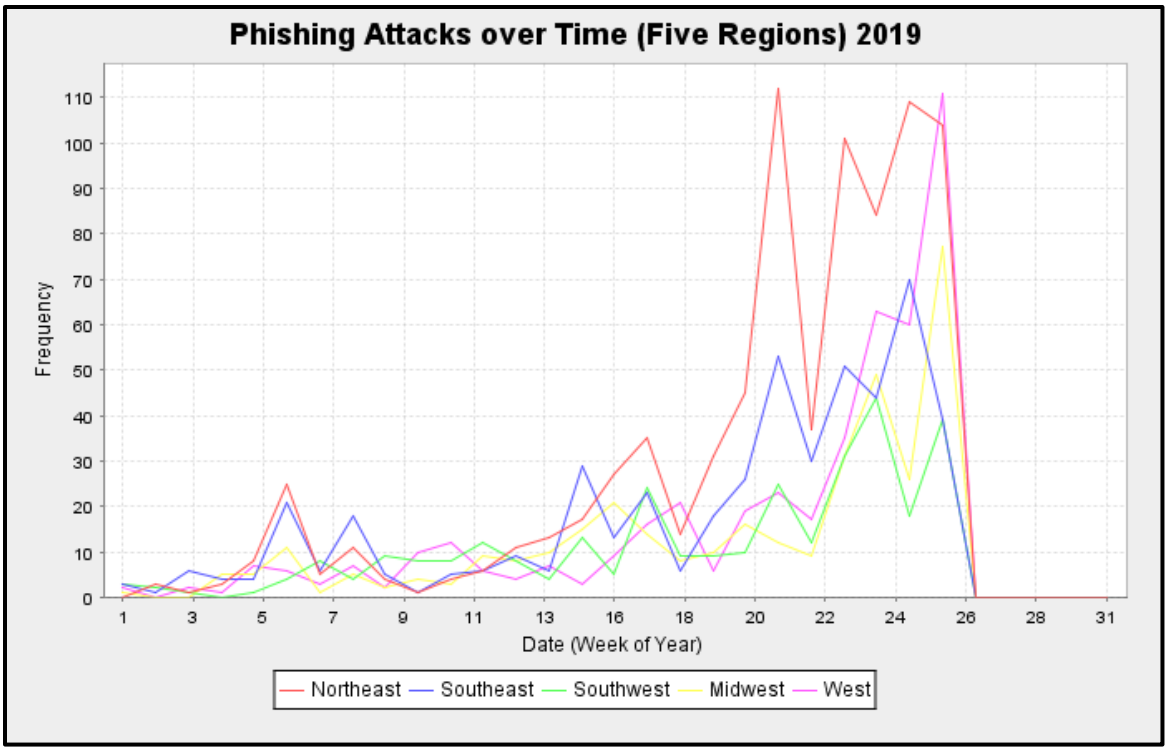

Figure 5. Line Graph of Phishing Attacks over Time on a Biweekly basis (5 regions)

Based on Figure 5, the amount of fluctuation in the data lowered, but not enough to my satisfaction. So far, I only looked to vary the time metric. The time metric was either one day, one week or two weeks, but what if the area metric was changed? I decided to look at the amount of phishing activity for all three time intervals, but with 5 random states. Figure 6 depicts the phishing activity on a daily basis for 5 random states (Texas, California, Virginia, Ohio, North Carolina). Figure 7 depicts the same concept as Figure 6 except for using one week time intervals instead of one day time intervals. Figure 8 shows the same idea as Figures 6 and 7 except for using two week time intervals instead of one day or one week time intervals.

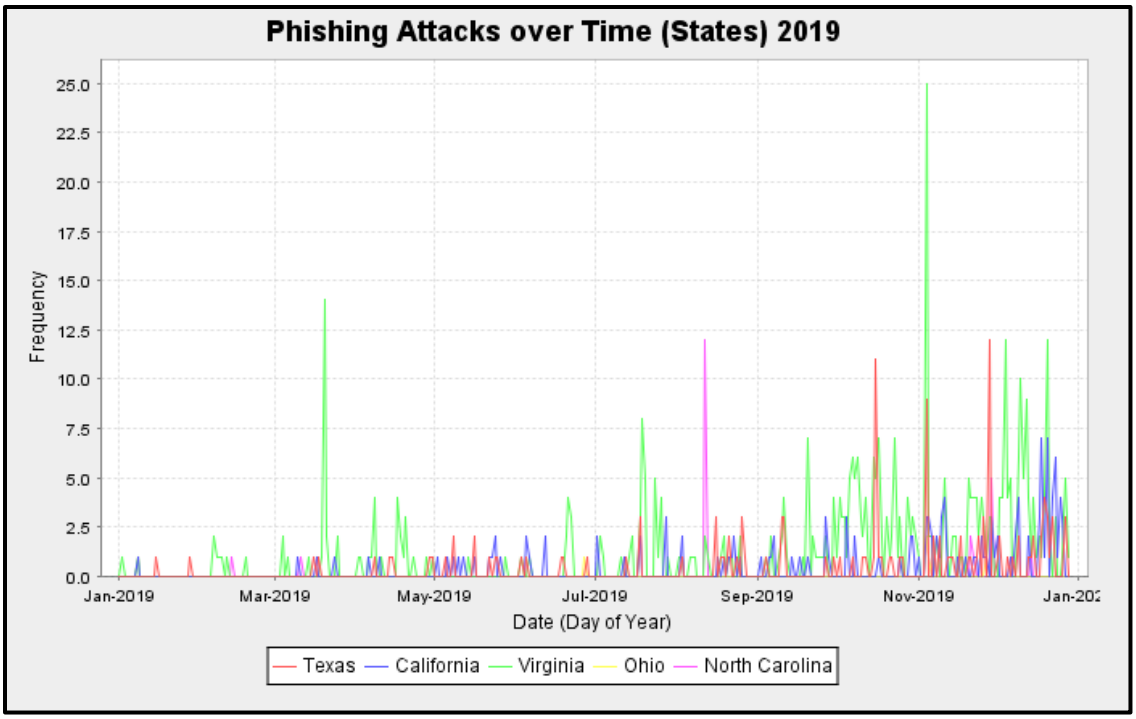

Figure 6. Line Graph of Phishing Attacks over Time on a Daily Basis (Five States)

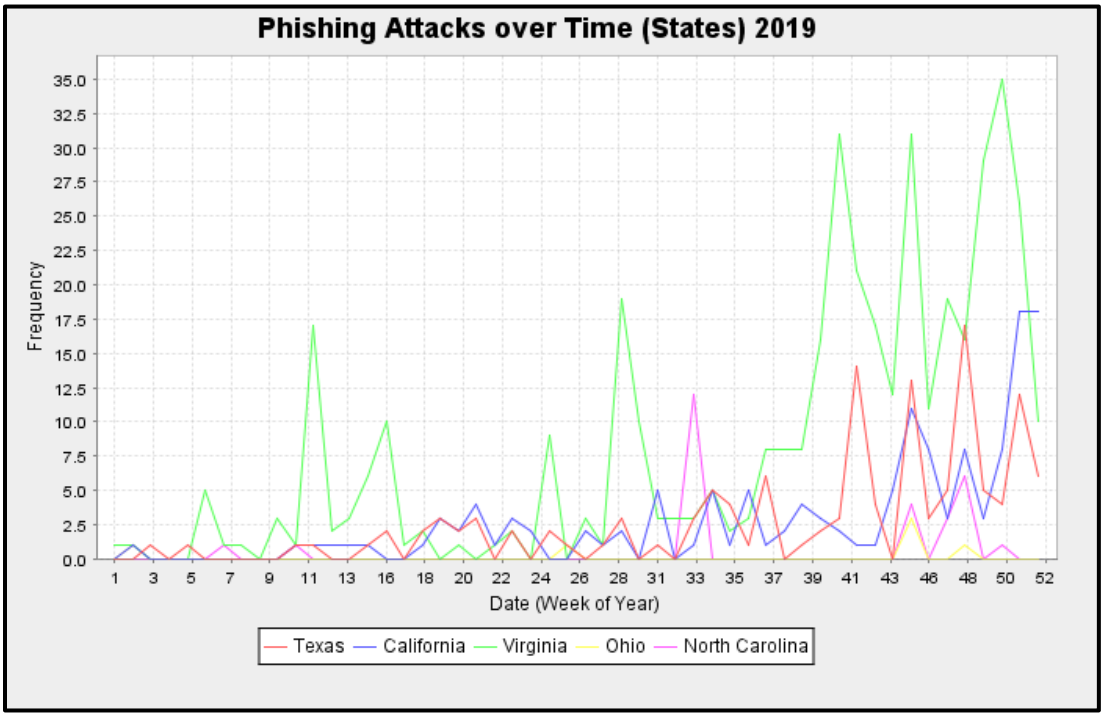

Figure 7. Line Graph of Phishing Attacks over Time on a Weekly Basis (Five States)

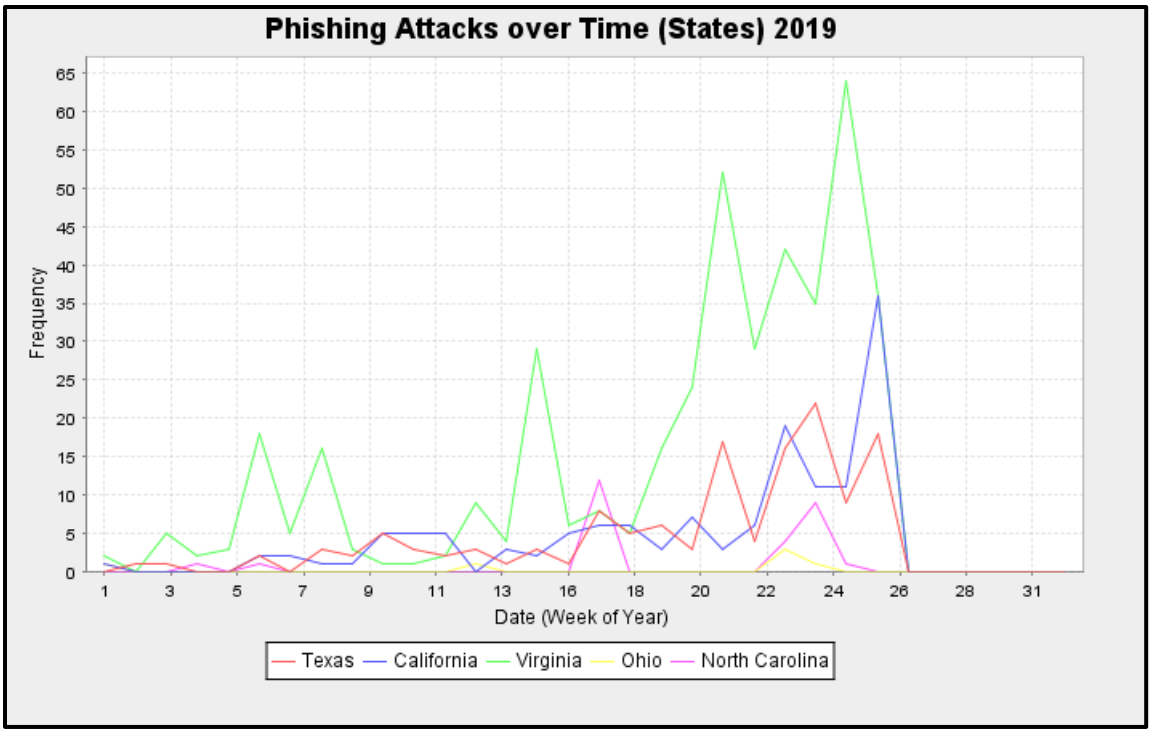

Figure 8. Line Graph of Phishing Attacks over Time on a Biweekly Basis (Five States)

The amount of fluctuation in the data is lower in Figure 8 than that of Figures 6 and 7. If we compare the same 2 week window with varying area type (Figure 5 and Figure 8), we can see that the data looks more normalized when using states as the area type as supposed to regions.

User Interface

The data visualizations have been chosen (Figure 2 and Figure 8), so the final step is to create the actual user interface of the prediction game for players to play. For the interface, I decided that three screens would be needed: a home screen, a how to play screen, and the game screen. Most games in general follow this scheme and I believed it would be appropriate for my game as well. Below are pictures of each of those screens respectively.

Figure 9. Home Screen of Prediction Game

Figure 10. How To play Screen of Prediction Game

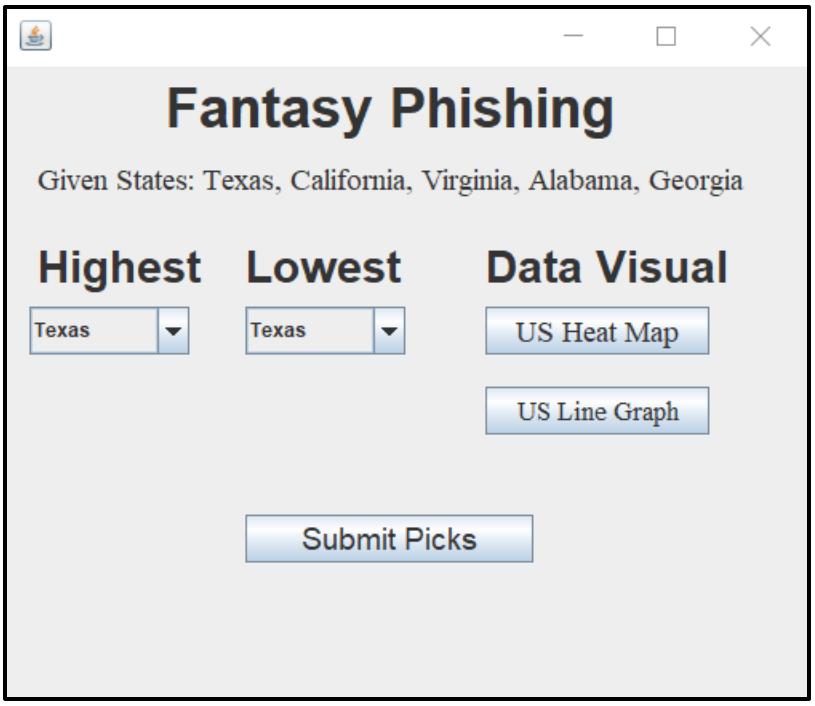

Figure 11. Game Screen of Phishing Game

Figures 9 and 10 are pretty self-explanatory for the player. However, there are a few things in Figure 11 that the player should make a note of. This screen has one drop down for the highest, one drop down for the lowest, two data visualizations, and a submit picks button. The highest drop down selection should be for the state that experienced the greatest amount of phishing activity with no respect to population over the two week time period and the opposite logic is applied for the lowest drop down selection. The visualizations are available if the player chooses to use them, but they can only use those and their critical thinking to make their selections. Once their selections have been made, the submit picks button must be pressed to confirm their submission and they will know at the end of the time interval whether their selections were accurate or not. Now, we have a working prediction game, with a user interface and data

visualizations, that players can play to improve their critical thinking skills in the domain of Computer Security.

CHAPTER IV

CONCLUSION

Summary

For the past two semesters, I have analyzed datasets in Computer Security, with the hope of finding one that could be used to make a prototype of a Prediction Game. I found the dataset that was appropriate for the project which was the PhishTank database and looked into creating the data visualizations and the user interface. The data visualizations I chose for the game include a HeatMap of the United States and a Line Graph of Phishing Activity over Time on a 2 week basis. I decided to make a simple User Interface, with a home screen, a how to play screen, and a game screen. Connecting the data visualizations with the User Interface turned out to be the hardest part, but I was able to complete the task successfully and made a working prediction game named “Fantasy Phishing.”

Future Work

Although I made a prototype prediction game that works, there are some limitations that I would like to address here. First, I was not able to make a chatting channel, a window that most prediction games like Fantasy Sports have, where players can communicate with their fellow opponents on anything they want to talk about. This would make the prediction game interactive and interesting for the players in the league (Junnutula). Second, I was not able to make an aesthetically pleasing User Interface for the players. Most of the amazing games in the world today are successful due to the User Interface that players get attracted to (Petoskey). To further explain, one concept that exists in both Fantasy Climate and Fantasy Precipitation is a news feed which players could use just to update themselves on the latest information regarding the specific

domain. My User Interface would look nicer if there was a feature to display news like in Fantasy Climate and Fantasy Precipitation. Third, although I normalized the data by looking at the number of phishing attacks on a state by state basis, the normalization does not take population into account. The states that have had higher amounts of phishing activity have generally been the ones that have a large population in comparison to the other states. Most players could easily figure out the state that has the highest number of phishing attacks just by knowing the relative populations of each of the states in the prediction activity. However, this can be avoided by further normalizing the data with respect to population. Finally, I only considered Computer Security in terms of phishing activity, but this domain is extremely broad. The game that I created is extremely specific and could get boring for players after a period of time, so if other tasks were incorporated into the game like number of cyberattacks and malware attacks ("Vizsec"), the game would be challenging and fun for the players.

Takeaway Message

Computer Security is a hot topic in the world today. Data is everywhere and people want to always have the feeling that their data is safe. Prediction Games in Computer Security and other related games will help increase awareness of the existence of vulnerabilities and threats around the world and provide an understanding of their spread/patterns both temporally and geographically.

REFERENCES

"Data Sets." *Data Sets*, Vizsec, vizsec.org/data/.

Dzodom, Gabriel and Frank Shipman. "Data-driven Prediction Games." ACM Digital Library. 2016. ACM. 12 Sept. 2019 <https://dl.acm.org/citation.cfm?doid=2851581.2892546>.

Kulkarni, Akshay (2016). Effect of Visualization of News Articles in Data Driven Games. Master's thesis, Texas A & M University. Available electronically from http://hdl.handle.net/1969.1/156977

Junnutula, Meghanath Reddy (2015). Asynchronous and Synchronous Communications' Effect on User Engagement in Prediction Games. Master's thesis, Texas A & M University. Available electronically from http://hdl.handle.net/1969.1/155530.

"National Vulnerability Database." *NVD*, 27 Mar. 2020, nvd.nist.gov/.

Petoskey, Carl. "Importance of User Interface for Online Gaming Sites." *TechGYD.COM*, 19 Dec. 2019, www.techgyd.com/importance-of-user-interface-for-online-gaming-sites/42840/.

"PhishTank > Statistics about Phishing Activity and PhishTank Usage." *PhishTank*, 26 Feb. 2020, phishtank.com/stats.php.

"Prediction Games." Prediction Games, predictiongames.tamu.edu

Sablich, Justin. "A Beginner's Guide to Playing Fantasy Football." *The New York Times*, The New York Times, 24 Aug. 2017, www.nytimes.com/2017/08/24/sports/fantasy-football-draft-guide-beginners.html.