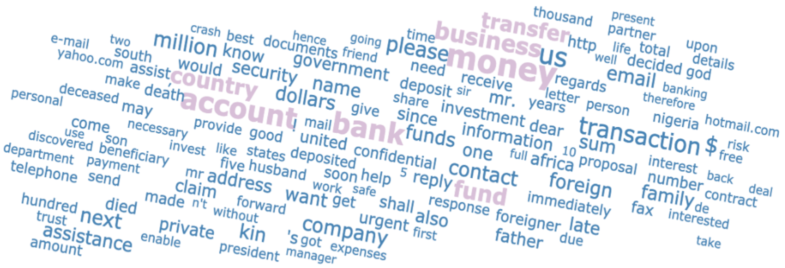
Assignment 3 Project Report

Team members: Zixi Jiang, Peizhen Li, Xiaoyu Wang, Yuchen Zhang, Xiuwen Zhang, Nat Zheng
GitHub Repo: https://github.com/Anthonyive/DSCI-550-Assignment-3

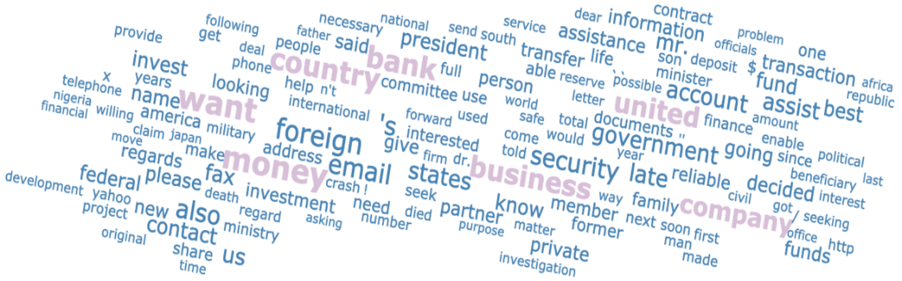
Why did you select your 5 D3 visualizations? How are they answering and showing off your features from assignments 1 and 2 and the work you did?

Visualization 1 - Email Content Word Clod

4000 Emails Word Cloud

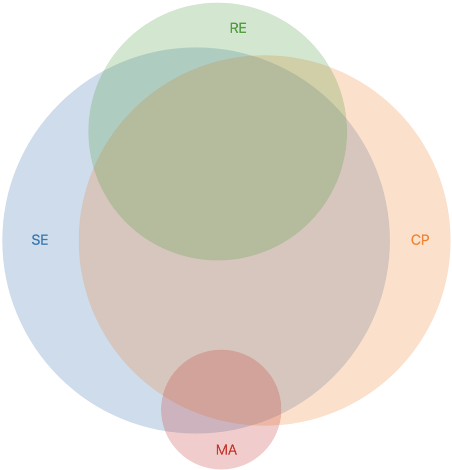


GTP-2 Generated Emails Word Cloud



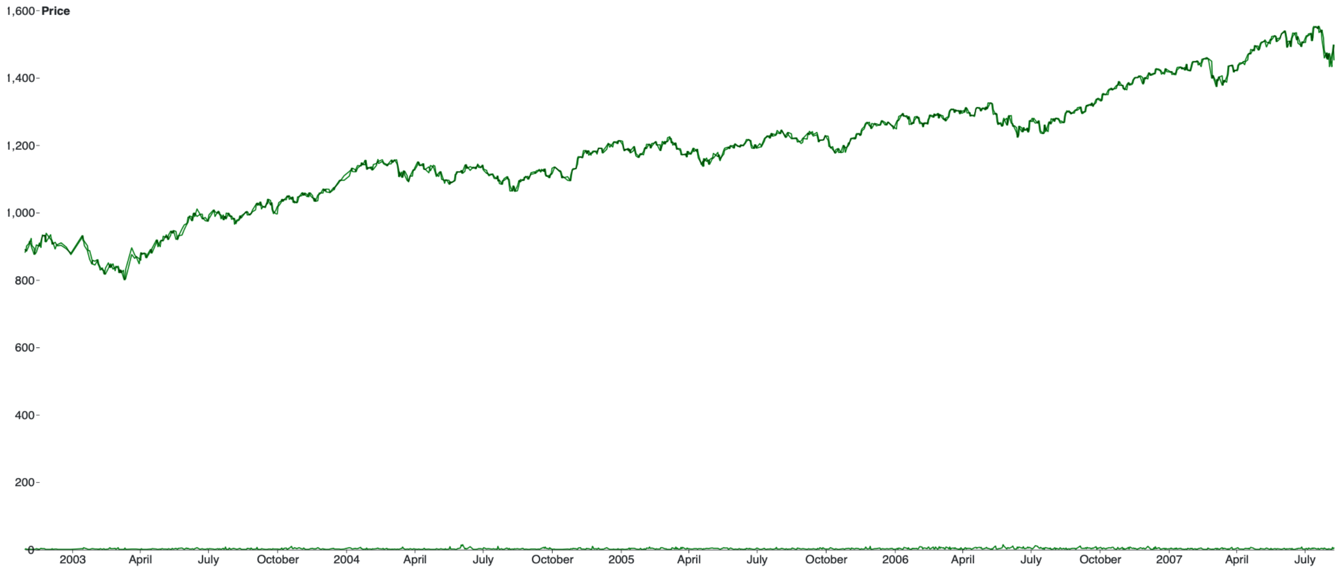
We decided to make clickable word cloud images for visualizing the emails' text content from assignment 1 and 2. The intent is to compare the original 4000 email corpus with the machine generated email text and understand how attackers phrased. The top 150 words with the most frequency are displayed in the cloud proportional to their appearance in the emails, with punctuations and stop words removed. We can see that in a lot of words appeared often in both corpus, such as ‘ bank’, ‘money’, ‘fund’, and ‘business’. It is intuitive to understand that phishing emails mention words relating to money and fund a lot and this is closely connected to their attack types.

Visualization 2 - Attack type popularity and connectivity



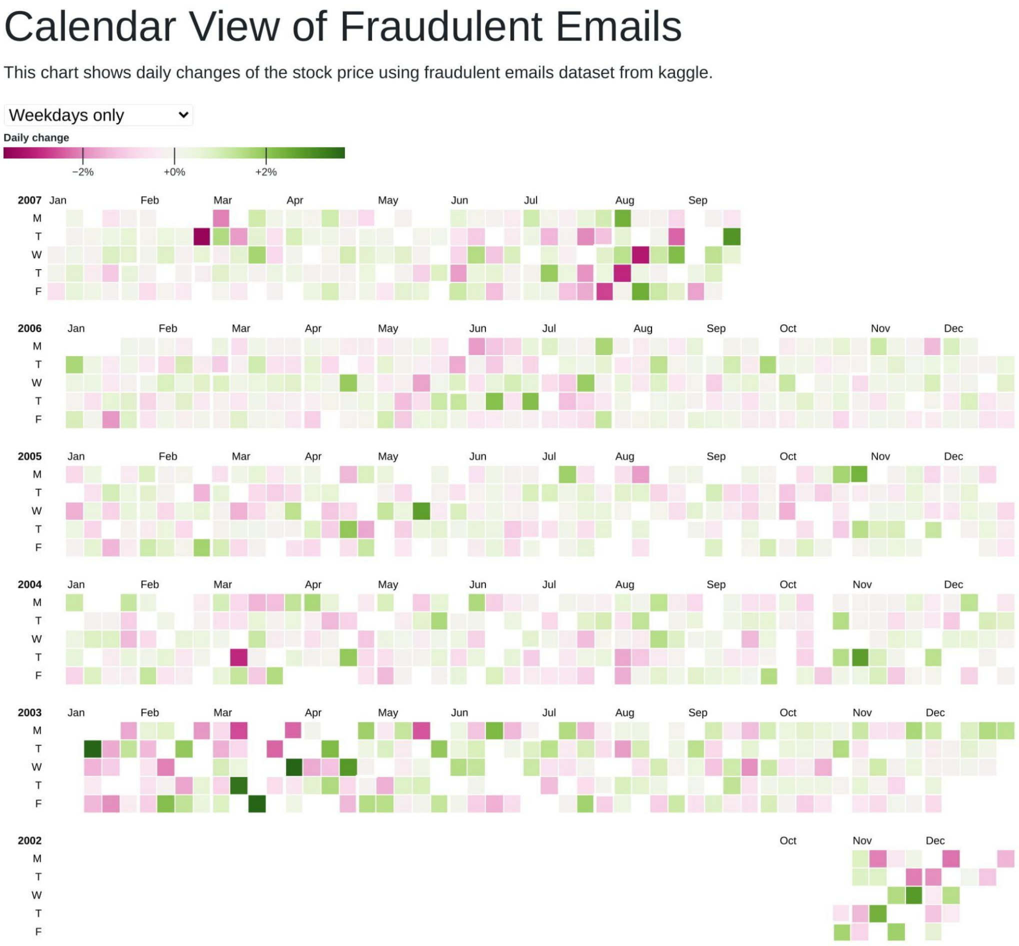
To the very beginning, we are required to seek the correlation among attack types and other fraudulent emails' features. However, we believe that the inner relationships among these four attack types might be ignored. So we decided to see the popularity and connectivity of each attack type by using Venn Diagram. The "RE","SE","CP","MA" each stands for "Reconnaissance", "Social Engineering", "Credential Phishing", "Malware". And according to the graph, we can find that social engineering is very likely to appear together with credential phishing and malware is more likely to be an independent attract type.

Visualization 3 - Multi-Line Chart of Date & Stock



We would like to find out whether the number of fraud emails for each day from 2003 to 2007 is related to the daily opening and closing stock prices. We chose a line chart for this because the line chart can clearly show the trends of how data vary. There is only little difference between opening and closing prices, but it's clear that the prices are continually increasing over time. From the line at the bottom (which can be hardly seen), we can see that the number of fraudulent emails doesn’t vary too much. Thus, the stock prices seem to have little influence on the fraud emails.

Visualization 4 - Calendar View of Fraudulent Emails



Calendar View of Fraudulent Emails

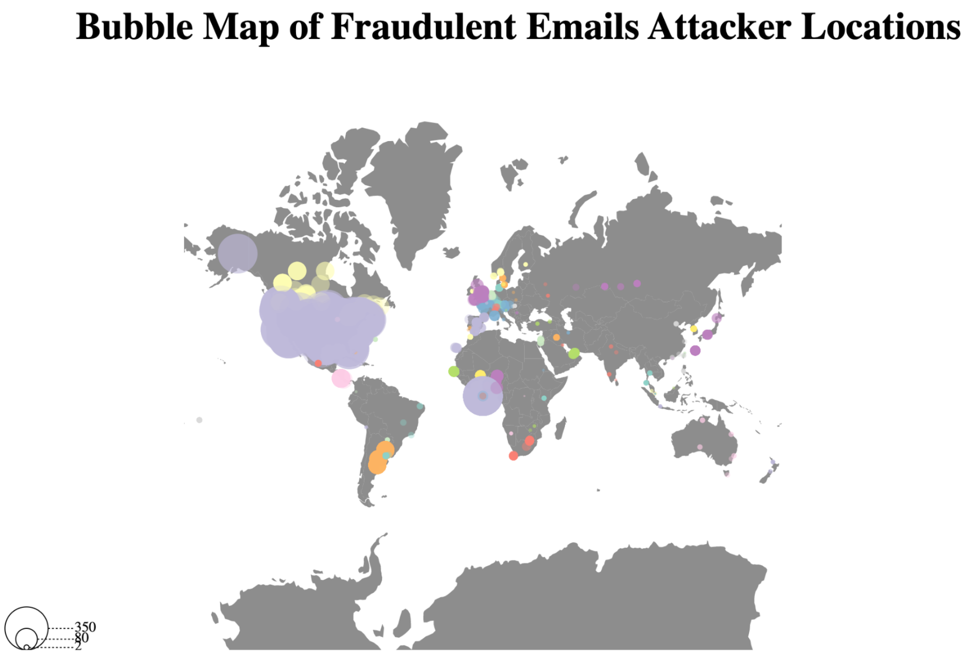
This chart shows daily changes of the stock price using fraudulent emails dataset from kaggle.

Weekdays only

Credit: Calendar View of Fraudulent Emails by Yuchen Zhang

We wanted to find out how the stock market changed over the time period of these fraudulent emails. Besides the multi-line chart that shows the overall stock market price changes during 2003 to 2007, we also wanted to show how the daily opening and closing stock price related to the previous day. Calendar view plot is a great way to show off our intent. The greener the blocks are, the more daily changes it has. We also have some white blocks since we don’t have all the prices of these dates. However, sometimes we may have multiple stock prices on the same day, then we will take the average. We can see that the changes most happened around the winter/spring of 2003 and the summer of 2007.

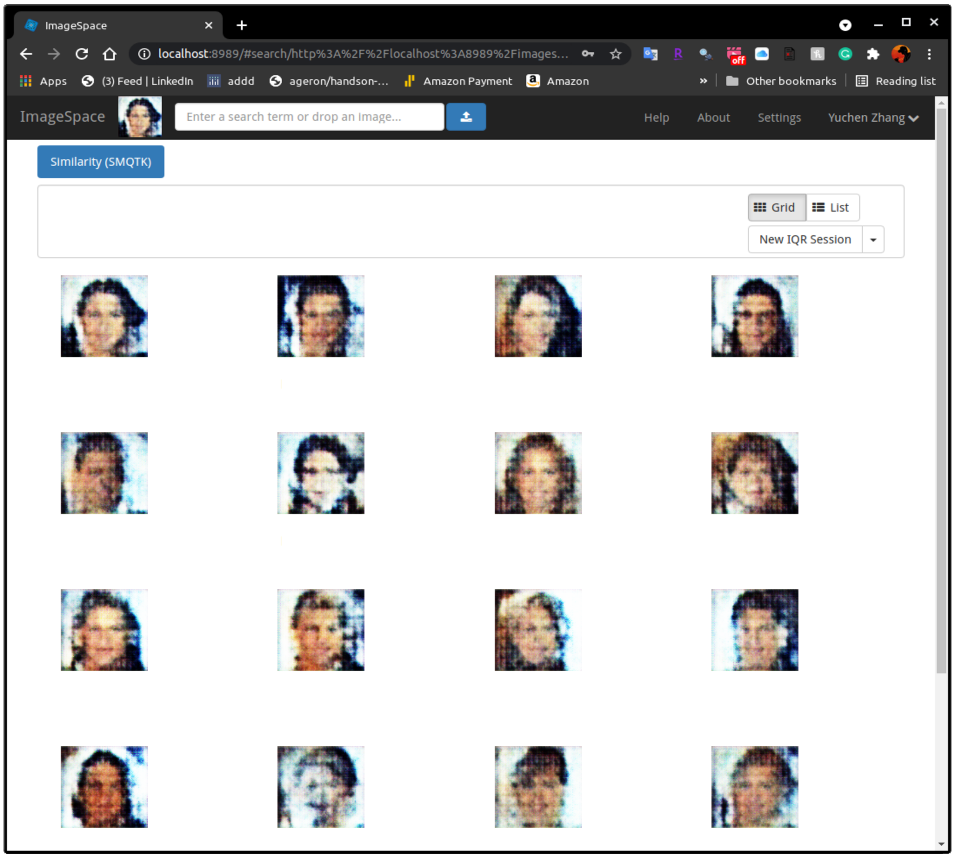
Visualization 5 - Bubble Map of Fraudulent Emails Attacker Locations



This is a bubble map for fraudulent email attackers' location or the countries that they mentioned in the email contents. Size of the bubble indicates the number of spam emails sent from a certain location. The bubble map informed us that the majority of the attackers are located in North America. Some spam email attackers clustered in Africa and South America. Only a few attackers highlighted Asian countries in their emails or sent out emails from Asian countries. We would not be able to easily conclude where the majority of the spam emails came from just by extracting ip address or mentioned locations from email. Generating graphs are visually more straight-forward and convincible.

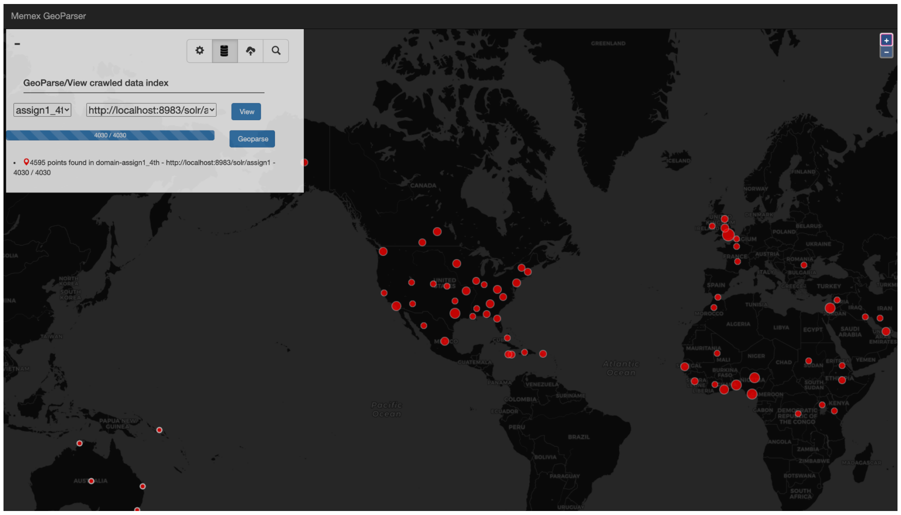
Did Image Space allow you to find any similarity between the fake attacker images that previously was not easily discernible?

Yes, Image Space allows us to see similar fake attacker images.We also included the tarred Solr index from deploy_imagespace-solr_1. Interestingly, the first few pictures do seem very promisingly similar to my eyes. If we don't have imagespace, it would be very hard to find similar images by going through 800 fake face images.



What type of location data showed up in your data? Any correlations not previously seen, e.g., from assignment 1?

The location data we have are coordinates, but we did not have such a clear view of location until we used GeoParser. After parsing our location data through Apache Solr, we can see that the distribution of points is mainly across the North American continent, Europe and Africa, shedding a light on the location of the attacker, or the location of attacks. The points are evenly distributed across the United States, and our assumption is that the attacker might be using fake IPs that were randomly selected across locations.


Memex GeoParser
GeoParse/View crawled data index
assign1_4t
http://localhost:8983/solr/a
View
4030 / 4030
Geoparse

4595 points found in domain-assign1_4th - http://localhost:8983/solr/assign1 - 4030 / 4030

Your thoughts about Image Space and ImageCat – what was easy about using them? What wasn’t?

We had some problems with image space dockers where the similarity container, SMQTK, doesn’t exist any more. However, Professor Chris helped us fix this problem, so everything works as expected.

Me, Yuchen, who worked on the imagespace and imagecat, haven't had any experience with docker, so I learned some basics for the imagespace to work. The good thing about imagespace is that once it finishes its docker, we can see if it works by going to the localhost using our browser. Dealing with “GUI”s and buttons are much easier than command lines.

However, exposing the port of solr needs to edit the docker-compose.yml file. This is not a hard thing to do, but it could be tricky if we didn’t get any help from Slack.

Contributions

Yuchen Zhang ● Convert tsv to json using python3 compatible etllib ○ Pull request: https://github.com/Anthonyive/etllib.git ● Set up Flask for team member’s visualizations ● Visualization 4 - Calendar View of Fraudulent Emails ● Run Image Space using face generator output from assignment 2 ● Help on Task 5 data conversions	Xiuwen Zhang ● Visualization 3 - Multi-Line Chart of Date & Stock ● Run and test GeoParser
Nat Zheng ● Cleaned and reorganized the attacker locations from assignment 1 and stored stored them to a valid input csv file for visualization ● Visualization 5 - Bubble map of fraudulent attackers’ location	Zixi Jiang ● Made Visualization 1 - Clickable Word Cloud ● Ran GeoParser and got the location graphs
Peizhen Li ● Task 3- ingest Task1 and Task 2 data into ElasticSearch ● Task 5- Using Elasticdump to generate new json files with Index for Task 3 ● Upload the folder and zip	Xiaoyu Wang ● Visualization 2 - Attack type popularity and connectivity ● Modify scripts for GeoParser, prepare datasets for geological data in assignment1&2, and test GeoParser