DSCI 550 Project 1 - Report

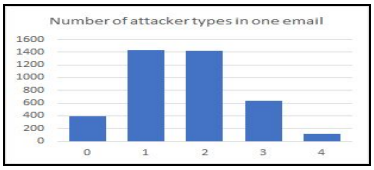
Team Name: MIMECRAFT

Team Member: Joshua Huang, Saumya Shah, Sungho Lee, Xinran Liang, Yue Liu

Analysis of Attack Types and Stylometrics

The first step of our analysis is visual inspection. We went through some emails and realized that all the spam emails have a fixed template, which means the patterns of the spam are repetitive and trackable.

For all the spam emails, the attacker types are categorized in 4 kinds in total. They are reconnaissance, social engineering, malware and credential phishing. Each fraudulent email contains more than 1 attacker type. According to our analysis, almost 1400 emails contain 1 to 2 attacker types in one email. Only a small proportion of the spams (approximately 100 emails) contain all four attacker types.

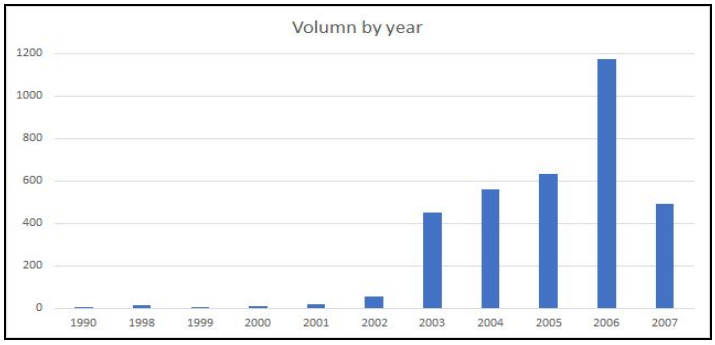


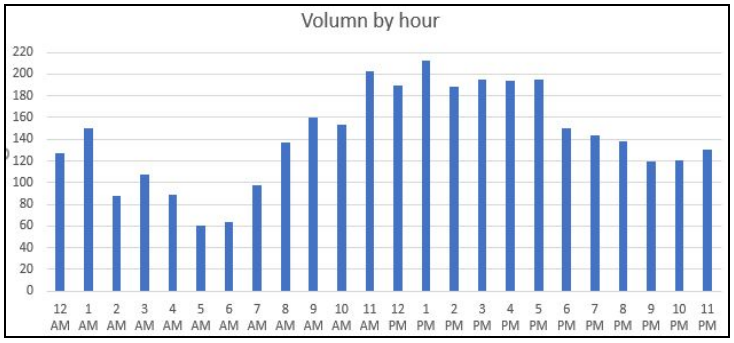
The next step was to figure out attacker stylometrics. The first part is the attacker title. Based on our analysis, the most frequently used attacker titles are "son", "manager", "director", “wife”, and “auditor”. Because most spam emails are trying to get money, attackers always pretend to be someone in charge of money. Hence, they often state they are the son or wife of a rich person who is in trouble. Or, they would state they are the manager, director or auditor who needs help to transfer or cover up money.

The second part is the urgency expressed from the fraudulent emails. We think that the urgency, related to social engineering, can be represented by the frequency of key words and time expressions. For example, "urgent" and "in # minutes”. We have 3 urgency levels in total, represented by 0 to 2, corresponding to the lowest to the highest level. By counting words with their conjugations/plural forms and certain time expressions, we had 436 out of 3992 emails in level 0, 1551 emails in level 1 and 2005 emails in level 2. Besides, the most frequently used level-2-word is "now"(3403 times). For level 1, the word is "need"(3030 times).

Title	Count of email
son	288
manager	285
director	172
wife	140
auditor	104

The third part is the date and time of the email attack. Based on our results, we can see that a dramatic increase in attack volume existed in 2003, almost 4 times of the attack in 2002. Since then, the email attack increased every year and experienced a double increase, also the peak of the email attack, in 2006. Our group then generated the average volume of email attacks per month. April, October and December are the three months with the least email attacks (all less than 250). June has the highest number of attacks, almost close to 350.





In order to further specify the details of the email attack's volume, our group then analyzed the attack on days of week base and hour base. For days of week, weekends have relatively smaller volume compared to work days. However, an obvious difference exists between weekends: Saturday has double the number of attack's volume than Sunday. The volume of attacks is pretty average for work days, but Wednesday has the highest volume of attacks. For attack volume based on hours, the highest attack volume existed from 11am to 5 pm, which is also the working hours for most people. The lowest volume interval existed from 2 am to 7 am, which is approximately the sleeping time for most people.

The next part is for the attack offering. We categorized the attacker offering into 3 categories: money, service and unknown. Based on our analysis, almost 86% (3449 out of 3992) of the offering is about money. The service and unknown categories take comparatively small proportions in offering.

After the offering, our group did analysis for attacker location. The logic behind our analysis is get location either from IP address or from geographic location, like the country name, mentioned in the email content, if IP addresses are not available. We used 500 IP addresses and these IP addresses provided specific city names to us. According to the results, the top 3 locations of the fraudulent emails are Nigeria (489 emails), South Africa (287 emails), and the United States (181 emails). Besides that, the remaining top 10 countries are Senegal, Ghana, Iraq, the United Kingdom, Liberia, and Zimbabwe.

For the attacker relationship part, our group found out that the most frequent relationship among "met online", "friend of a friend", and "met the victim in person before" is "friend of a friend" (286 emails, 7.2%), which is linked to the attacker type "reconnaissance" probably. Besides that, most fraudulent emails are suitable for categorizing them as "having relationships with the victims". 77.4% of them are strangers, and 15.4% of them can not detect any relationship with the victims.

For the sentiment analysis part, 97% of the attacker emails have positive sentiment scores. Among these emails with positive sentiment scores, most of them have polarity scores ranging from 0 to 0.25. For subjectivity score, all the emails are positive, and most of them stayed in the score interval from 0.25 to 0.5.

Count of email	Random Caps			
Miss Spelling	1	2	3	Grand Total
1	1387	393	149	1929
2	891	489	195	1575
3	212	210	66	488
Grand Total	**2490**	**1092**	**410**	**3992**

The next part is for language style. Our group found out that almost 5% of the fraudulent emails have typo errors, including misspelling and random capitalization. For the misspelling part, our group used the ratio of number of typos to the number of words in content to categorize 3 different levels. Level 1, the lowest level, of the misspelling is from 0% to 5%; level 2 is from 5% to 10%; and level 3, the highest level, is bigger than 10%. Based on our results, 12.2% of the emails are from level 3, which means they have very serious misspelling problems. 39.5% of emails are from level 2, and 48.3% of them are from level1.

We used the same methodology on analyzing random caps. The ratio we used is the number of capital letters to the number of letters in content. The 3 levels, from lowest to highest, are level 1 from 0% to 10%, level 2 from 10% to 50%, and level 3 bigger than 50%. We also found that 10.3% of the emails are from level3, which means 10.3% of them have serious random cap problems. 27.4% of the emails are from level2, and 62.4% of the emails are level1.

The next part is the age predictor. After analyzing, we found that the average age of attackers is approximately equal to 33 years old (=132415.59499992165 / 3992). The range of attacker age is about 20 (19.67) years old. What's more, 69.8% of the attackers are approximately from 33 to 34 years old.

The last part we analyzed for the original features of fraudulent emails is IP phisher. According to our results, we used 558 IP addresses during the analysis, which are extracted from the metadata. Among all the 558 IP addresses, only 7.5% of emails were labeled as “very high” risk, and 5.7% of emails were labeled as “high risk”. Surprisingly, 80.47% of them (449 out of 3992) were labeled as low risk.

IP risk	Count of email
high	32
low	449
medium	35
unknown	3434
very high	42
Grand Total	**3992**

IP risk exclude unknown	Count of email
high	5.73%
low	80.47%
medium	6.27%
very high	7.53%
Grand Total	**100.00%**

Analysis of Additional Datasets

Datasets Selected	MIME Type	Features Extracted
Disaster (Earthquake & Tsunami)	TEXT/TSV	Total Deaths and Injuries, Total Damage, Disaster Count
World Bank GDP per capita	APPLICATION/ XML	GDP per capita, Internet Use, Age Dependency
Malicious URL classification	TEXT / CSV	URL Label, URL Length, URL Classification

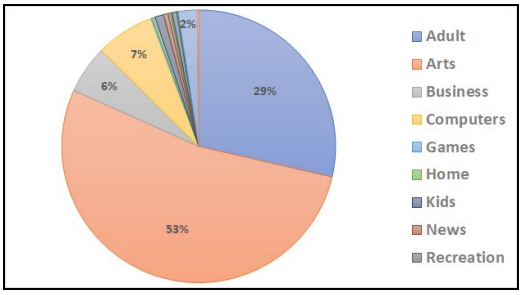
The above three datasets were identified to contribute in our journey of analyzing these spam email datasets. The disaster dataset was considered as most of the spam emails used disaster as an excuse for help and to analyze the trends of how these natural events affect spam email counts across the years 1992-2007. We found the “Total Damage” and “Total Deaths & Injuries” to be maximum in the years 1999 and 2004, but the number of spam emails sent out in 2004 is 100 times more than the ones sent out in 1999. This suggests that email spamming was not that common earlier as people just started getting familiar with this concept and obviously attackers did not have that many targets to attack. We observed an inverse trend in the years 2003-2007 where the number of spam emails sent were high during the year with least “Total Damage”. For “Total Deaths” we observe a direct correlation with “Urgency of the attack” parameter which was “high” (level 2) for most of the spam emails indicating desperation of the attackers.

On the other hand, for “Number of Disaster”, we didn’t observe any correlation with “Urgency of the attack”. We can assume that the depth of the disaster is more important than the number of disasters. For example, the number of disasters in 2005 was less than any other years, but the total death and injuries were second highest among 2003 to 2007.

Year	Total Deaths and Injuries	Total Damage ($Mil)	Number of Disaster
2003	96,784	11,282	82
2004	464,473	48,757	89
2005	237,562	6,680	68
2006	48,992	3,372	72
2007	7,891	12,812	86

Our second dataset helped us in analyzing if the spam emails have any correlation with average GDP per capita of a country. Considering Nigeria specifically, we can observe as the GDP per capita increased from 2003 to 2007 the count of spam emails in these years decreased. We can also observe the Average Urgency of the email to also gradually decrease in this year span. It's contrasting to note that despite the average “Internet Use” increased in this duration however the number of spam emails reduced, indicating that people were more aware of these attacks. The average “Age Dependency” indicator implies the dependency burden that the working-age population bears in relation to children and the elderly and this parameter follows a trend similar to GDP.

Country: Nigeria					
Year	Average Urgency	Count of emails	Average of GDP per capita	Average of Internet Use	Average of Age dependency
2003	1.6	109	1,673.3	0.6	86.8
2004	1.5	78	1,781.9	1.3	86.7
2005	1.4	53	1,848.2	3.6	86.6
2006	1.4	72	1,909.7	5.6	87.1
2007	1.4	42	1,982.7	6.8	87.4



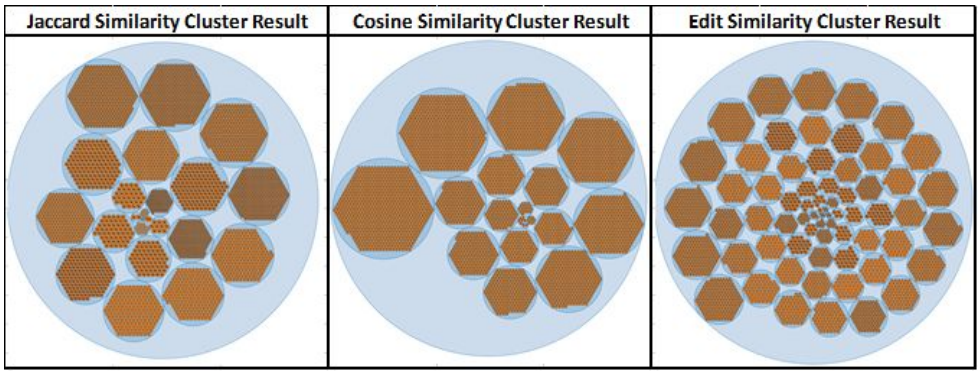
In our third dataset we have classified the URLs from the email content body into good or bad labels as provided in our dataset. We have observed that 80% of URLs were classified as good URL while rest were classified as bad. The URLs were also classified across their type and 52.99% of them were classified under “Arts” and 29% were classified under “Adult”. These were the two dominant classifications. 83.4% of emails having URLs had a “short” URL which corresponds to being shorter than the average length of URLs. The image on the left depicts the URL Classification list found in our Spam emails.

Similarity Inferences

Edit Similarity has the most number of clusters(66), and Jaccard and Cosine all have 23 clusters. Corresponding to that, Jaccard similarity and Cosine similarity have a bigger average size of clusters(~60.48). Edit’s average cluster size is ~181.45. Different from other two methods, clustering over Cosine similarity yields some clusters which contain few objects(< 10, most of these have only 1 object).

Cluster Result	Number of groups	Average size	Median size	Standard deviation
Jaccard	23	173.6	135	162.09
Cosine	23	173.6	97	200.96
Edit	66	60.5	49	52.14

From the perspective of algorithms, Edit similarity is more sensitive to differences between objects, and Jaccard, Cosine similarity are more dependent on intersection/closeness of data points. Our input data is formed by some string attributes, like "Attacker Offering" or "Attacker Location", many floating point numbers, such as "Total Deaths and Injuries" and few True or False features. This implies that the difference between numbers will have a large influence on clustering over Edit similarity score, so it shows us the most number of clusters. We do “Attacker Location” feature in a very precise way, in which we present the country, region, city and country shown in content, so for those emails which don't contain IP address, region and city columns have to be ‘unknown’. This, combining with True or False in Attack types, makes some objects have many overlaps and therefore clusterings based on Jaccard and Cosine show less but bigger clusters. At last, because unlike strings would have pretty different coordinates, uniquely spelled string attributes take some objects away from others, which causes generation of outlier clusters.



For Jaccard and Cosine similarity we have observed that the “blank” values in the features dominate while clustering. For instance, we have a date column with the “Year” parameter which is missing for (14.2%) emails however our additional datasets are joined to the Spam email feature datasets based on “year” or/and “country” columns itself. So if an email feature has a missing year it percolates to the level of combining our external datasets creating separate clusters with these unknown values. In terms of numbers, our original dataset contained 2 columns having unknown values, and when we combine it with our additional dataset having 6 columns, the weight of these missing values increases by a factor of 6 affecting our clustering result. This also accounts as one of our “unintended consequences” that occurred while adding the additional dataset by matching the year.

In our opinion, we believe the cluster result based on the Jaccard similarity algorithm is the most accurate measurement for our final datasets after joining with three other external datasets. The Edit similarity result divides the emails into too many clusters and most of the clusters rely heavily on numerical features. The results are too specific, which generates too much noise to the dataset. And in between Jaccard and Cosine similarity cluster results, there are eight groups with smaller than ten emails in Cosine method while only three in Jaccard method. Moreover, the standard deviation value for the Jaccard method is also smaller than the Cosine method. Hence, we like Jaccard similarity cluster results better.

Additional Observations

If questions from the assignment instruction are not here, they have been already addressed in previous analysis sections.

1. Are there clusters of attackers with similar features that tend to attack victims the same way?
   No, there is no significant difference in any of the four attacker styles for clusters in Jaccard, Cosine and Edit similarity results.

2. Does the time of day of attack matter?
   Yes, it matters. But, the success rate for the reconnaissance attack type doesn’t follow a linear trend line. The lowest 43% is at 7am while the highest 71% is at 9pm. In addition, from 6am to 7am, the attacks are more likely to fail. Our hypothesis is most people will wake up and go to work at 6am, with the habit of checking emails. So most email attacks at 6am are labeled as successful, and, after that, people need to commute or start to work, not paying much attention to attack emails.

3. Is there a set of frequently co-occurring features that induce the email to be read?
   Yes, there is. The success rate for the reconnaissance attack type is higher with the following features.
   - Attacker type is not social engineering
   - Attacker relationship is friend of friend
   - Attacker offering is service
   - Urgency level is 0(lowest)
   - Sentiment-polarity is equal to or higher than 0.19
   - Sentiment-subjectivity is equal to or higher than 0.5
   - Style-random capital is 0(lowest)
   - IP known as phisher is high or very high

4. Also include your thoughts about Apache Tika – what was easy about using it? What wasn’t?
   - Parser: It is very straightforward to use for extracting text from PDF or HTML files. However, for tables, like XML files, it's hard to use. When we tried to extract things from the fraudulent email file in this assignment, Tika generated redundant metadata in the email content part. In other words, the metadata was generated

again in the content part, and the real email content we needed was generated after the repeated metadata part.

- Similarity: Tika similarity is a good tool for parsing into files' metadata and calculating similarity between them. For the source code, it provides clear functions for calculations. Besides, it exports file name pairs and similarity scores in .csv, which increases efficiency in future data processing. Although, in our case, directory of json files is not a proper format as input object in source code, thanks to contributor, Jiarui Ou, we can then read into csv files, compare features and make calculations. Therefore, one possible improvement on Tika similarity should be adding more interfaces for different input file formats.
- Age Predictor: USC Age Predictor needs easy configuration and simple operations, and its results can precisely show authors' age for texts containing some key indicators (e.g. sentences in sample testing set).
- Platform compatibility: Tika is Java based and also has a Python version package. It works well in Windows, MacOS and Linux systems.
- Offline vs Online Environment: When we tried to establish an offline local environment of Tika, the installation did not work. The computer always experienced file path errors. However, for an online environment, Tika worked very well.

References

- Tika Similarity: Contributor: chrismattmann; link: https://github.com/chrismattmann/tika-similarity
- Tika-similarity modified codes: jaccard.py, edit.py, cosine.py and circle-packing-for-all.py, Contributor: Jiarui Ou
  https://uscdatascience.slack.com/archives/G01J5KX0L8L/p1615576178359800
  https://uscdatascience.slack.com/archives/G01J5KX0L8L/p1615657423404100
- Tika AgePredictor: Contributor: chrismattmann; link: https://github.com/USCDataScience/AgePredictor
  @article{hong2017ensemble, title={Ensemble Maximum Entropy Classification and Linear Regression for Author Age Prediction}, author={Hong, Joey and Mattmann, Chris and Ramirez, Paul}, booktitle={Information Reuse and Integration (IRI), 2017 IEEE 18th International Conference on}, organization={IEEE}, year={2017}}
- spaCY: en_core_web_lg: https://github.com/explosion/spacy-models/releases//tag/en_core_web_lg-3.0.0
- blog.text.csv: Contributor: Rachael Tatman; link: https://www.kaggle.com/rtatman/blog-authorship-corpus

Contribution:

- Joshua Huang: Tika installation and setup, convert original text file into JSON, attacker title, date/time, external datasets(disaster and world bank), Tika-similarity(tried to setup but failed, reported error to editdistance package owner on git), report graphic, charts and analysis, readme file
- Saumya Shah: Tika installation and setup, attacker relationship, attacker offering, social engineering, External dataset3(URL), External dataset1(disaster), Report analysis - clustering and content, readme file, Attacker estimate age (Tried the set up with Java, but did not succeed on Windows)
- Sungho Lee: Tika installation and setup, parsing JSON with Tika, malware, credential phishing, attacker offering, attacker location, attacker language style, attacker IP known as phisher, external dataset1(disaster), external dataset2(worldbank), external dataset3(URL), combine all features into TSV file, readme file
- Xinran Liang:Tika installation and setup, convert original text file into JSON, reconnaissance, urgency of the attack email, external datasets 1(disaster) and 2(worldbank), Attacker estimate age with USC Age Predictor(training + testing), Tika-similarity: similarity score calculating, clustering and visualization, report compiling, readme file
- Yue Liu: Tika installation and setup, convert original text file into JSON, attacker sentiment analysis, attacker IP known as Phisher, social engineering, external dataset2(worldbank), external dataset 3(URL), report analysis