Homework: Analysis of Cyber Phishing EMails: Fraudulent Emails and Social Engineering
Due: Friday, March 12, 2021 12pm PT

---

1. Overview

```
From r  Thu Oct 31 17:27:16 2002
Return-Path: <obong_715@epatra.com>
X-Sieve: cmu-sieve 2.0
Return-Path: <obong_715@epatra.com>
Message-Id: <200210312227.g9VMQvDj017948@bluewhale.cs.CU>
From: "PRINCE OBONG ELEME" <obong_715@epatra.com>
Reply-To: obong_715@epatra.com
To: webmaster@aclweb.org
Date: Thu, 31 Oct 2002 22:17:55 +0100
Subject: GOOD DAY TO YOU
X-Mailer: Microsoft Outlook Express 5.00.2919.6900DM
MIME-Version: 1.0
Content-Type: text/plain; charset="us-ascii"
Content-Transfer-Encoding: 8bit
X-MIME-Autoconverted: from quoted-printable to 8bit by sideshowmel.si.UM id g9VMRBW20642
Status: RO

FROM HIS ROYAL MAJESTY (HRM) CROWN RULER OF ELEME KINGDOM
CHIEF DANIEL ELEME, PHD, EZE 1 OF ELEME.E-MAIL
ADDRESS:obong_715@epatra.com

ATTENTION:PRESIDENT,CEO Sir/ Madam.

This letter might surprise you because we have met
neither in person nor by correspondence. But I believe
it is one day that you got to know somebody either in
physical or through correspondence.

I got your contact through discreet inquiry from the
chambers of commerce and industry of your country on
the net, you and your organization were revealed as
being quite astute in private entrepreneurship, one
has no doubt in your ability to handle a financialbusiness transaction.

However, I am the first son of His Royal
majesty,Obong.D. Eleme , and the traditional Ruler of
Eleme Province in the oil producing area of River
State of Nigeria. I am making this contact to you in
respect of US$60,000,000.00 (Sixty Million United
State Dollars), which I inherited, from my latefather.

This money was accumulated from royalties paid to my
father as compensation by the oil firms located in our
area as a result of oil presence on our land, which
hamper agriculture, which is our major source oflivelihood.

Unfortunately my father died from protracted
diabetes.But before his death he called my attention
```

Figure 1: Phishing Email Example from the Fraudulent Email Corpus on Kaggle at https://www.kaggle.com/rtatman/fraudulent-email-corpus

In this assignment we will explore several of the topics discussed in the early portion of class – Big Data – MIME types and their taxonomy – Data Similarity – and so forth. To do this, we will leverage the dataset highlighted in Figure 1 – a set of more than 2,500 "Nigerian" Fraud Letters, dating from 1998 to 2007. The-emails are valid RFC 822 emails, with headers including the examples below:

- Return-Path: address the email was sent from
- X-Sieve: the X-Sieve host (always cmu-sieve 2.0)
- Message-Id: a unique identifier for each message
- From: the message sender (sometimes blank)
- Reply-To: the email address to which replies will be sent
- To: the email address to which the e-mail was originally set (some are truncated for anonymity)
- Date: Date e-mail was sent
- Subject: Subject line of e-mail
- X-Mailer: The platform the e-mail was sent from
- MIME-Version: The Multipurpose Internet Mail Extension version
- Content-Type: type of content & character encoding
- Content-Transfer-Encoding: encoding in bits
- X-MIME-Autoconverted: the type of autoconversion done
- Status: r (read) and o (opened)

These attacks represent a form of *social engineering*. In these types of attacks, there are attackers, and victims as shown in Figure 2. The attackers typically have an "ask" of the victims. They are: 1) **reconnaissance**, trying to see if you actually read the email, or will reply back or not, and the signals associated with it; 2) **social engineering** in general, trying to identify themselves as "your friend" or referencing some key life event, using words like "urgent" and so on; 3) **malware,** providing a link to perhaps an application/exe file and trying to get you to click a link to it; or finally 4) **credential phishing**, trying to get you to provide an SSN, Date of Birth (DOB), Account number, or some personally identifiable information (PII) that will allow them to perform identity theft and digitally act as you. Some of these 'asks' may be overlapping. In practice, the goal would be for 'asks' to automatically be identified using some type of perhaps machine learning approach. Coincidentally there was a large Defense Advanced Research Projects Agency (DARPA) project in this area trying to create automatic defender bots that would thwart the attackers in various ways. We'll get back to these ways in later assignments.

These four types of attack vectors are all present in the corpus as described above. As you have learned thus far, RFC822 is the standard for emails originating from ARPANet and that project. Emails are rich Media Types that may reference other MIME types including file attachments, and Emails are the reason that we have such a rich definition of the MIME taxonomy for file types. This particular corpus doesn't include any multipart or message composite types, and instead you are given a single file containing thousands of RFC 822 emails.

In this dataset, however, you are given a tranche of Emails that have signals in them that you need to extract out. You are going to play the role of cyber investigator in this project, scanning through the emails as if they were just sent to your entire IT department and company and the CIO has asked you to perform some investigations on them.

The collective investigation that you are about to undertake will benefit from combining data from other types. Perhaps for example, there is a location referenced in the Email that identifies or attributes the author and that relates to some other type of attack. Or as another example, perhaps the Emails indicate some form of progeny, and they found the victim from a friend either at the company or elsewhere, or perhaps the attackers claim to know the victims from meeting online. Identifying the origination source will help you trace the social path of attack. Finally, maybe you could – based on the text provided – get an estimate of the attackers age, by examining the writing style and text used.

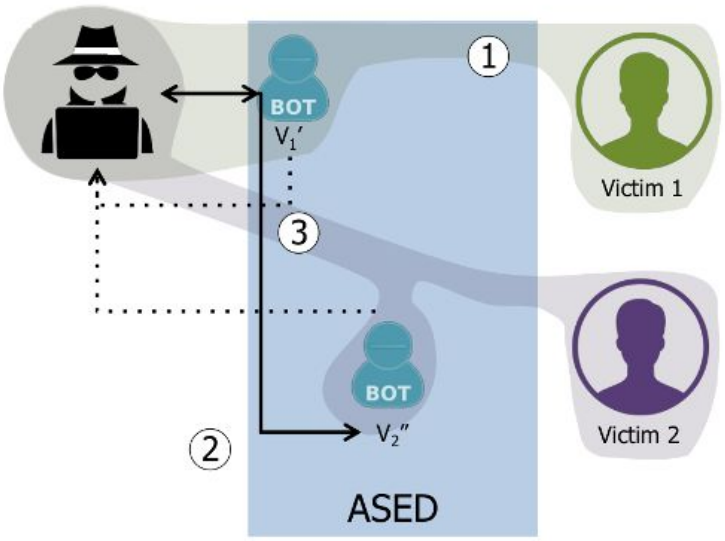

Figure 2: Traditional Social Engineering Attack Vectors: Bad actors (upper) left attack Victims (right). Active Social Engineering Defense involves making "defender bots" to protect against various attack vectors.

Alternatively, instead of the specific email content perhaps the RFC822 headers matter too. As an example, the RFC 822 headers as referenced before include a Subject, a Date, and also whether the email was opened by the victim (receiver) or not. In this fashion, you could state that everyone that opened an Email in fact was a successful victim of reconnaissance as an example. Or, for example, if the Emails sent late at night were all ignored, but those in the morning were all opened, you could discern further useful signals and patterns based on this information.

We call these properties **stylometrics** and these signals are used to **attribute** malicious attackers and used for further investigation or to play defense against them. You will engage in a cyber response investigation using your knowledge gained so far about Big

Data, the five Vs, the MIME taxonomy, Data Similarity and information that you will be learning about content and text extraction.

2. Objective

Looking at the Fraudulent Emails dataset, a couple things may jump out at you. First, it's in an inconvenient format, a single text file representing many sequential RFC822 emails concatenated together. You'll deal with this as one of the tasks for the assignment, but consider that somehow, we could easily make the data formattable as a set of independent easy to process and join, JSON records, or perhaps a big TSV table of them.

What would you join the data to? What features could you derive simply from the existing data? The first thing you could likely derive are the specific "asks" of the attackers. For example, you could classify the asks (repeated again for your benefit as):

1. Reconnaissance
2. Social Engineering
3. Malware
4. Credential Phishing

Getting these classes could be done with your existing dataset. You could for example write some software that did keyword scanning in each email content for words like "Reply back", "Click here", "Keep Confidential". This would indicate that the attacker is looking for the victim to have seen the email, and to engage in conversation (the recon worked in short). Or alternatively you could use the read/not read email header field too. Or some combination of them.

For Social Engineering, you could scan the content for words like "Your friend", indication of "some life event", "Urgent", "Children need help", or "there was a threat to my life". You don't have to do exact matches, you could actually do text processing as we have discussed already in class, and find these patterns.

For Malware, you could look for links to application/exe files, or phrases like "Click here". Finally, for Credential Phishing you may look for "asks" related to SSN, DOB (date of birth) or "Account number". These are just suggestions and there are clearly many ways of deriving all of this information. You will explore them in this assignment.

One difficulty is that likely many of these classes are overlapping and there are likely emails that are both performing Credential Phishing, and asking you to click Malware. There are likely emails that are both Reconnaissance gathering, and Social Engineering. So you will need to account for this.

What other features can you think of, and how would you get them? As an example, you may want to scan the content of the emails and pull-out information like the attacker's "official title" (a prince, a military person, a nonprofit worker?) because it may attribute them somehow, even if faked. Additionally, did the attacker use words to reflect urgency

in response? Perhaps they are trying to create a conduit for the victim to reply right away. You could easily flag this. Are they trying to get money? What are they offering to the victim? Are they offering money to the victim? Services? What are they offering?

Or alternatively where are the attackers located? Can you use a geoNames database lookup to more precisely place the location of the attackers, and join that ancillary information about that location including its population? Can you do a reverse lookup or a search of the email attacker's IP against existing phishing databases to see if this is a known scammer? You could also perform some data joins to generate a subsequent profile of the potential victims and gather properties about what was going on in the world at the time that the emails were sent? Were there a lot of commercials about giving? Was it during a time in which there was some catastrophic event or atrocity in the Attacker's area that may make the potential victims more sympathetic?

You will choose at least three publicly accessible datasets along these lines to join the Fraudulent email data to, and you must add at least three new features per dataset that you join. The datasets you select may not all belong to the same MIME top level type – that is – you must pick a different MIME top level type for each of the three datasets you are joining to this Fraudulent Emails dataset

Once the data is joined properly, you will explore the combined dataset using Apache Tika and an associated Python library called Tika-Similarity. Using Tika Similarity, you can evaluate data *similarity* (as discussed during the Deduplication lecture in class; and also during data forensics discussions). Tika similarity will allow you to explore and test different distance metrics (Edit-Distance; Jaccard similarity; Cosine similarity, etc.). And it will give you an idea of how to cluster data, and finally it will let you visualize the differences between different clusters in your new combined dataset. So, you can figure out how similar attackers are within the data given your stylometrics, and ask questions of your new augmented Fraudulent Emails dataset. For example, you may ask; how many Credential Phishing attacks came from Lagos, Nigeria around 8am to your company in which the victim clicked on the email, the attacker worked at a nonprofit and only asked for a transfer of less than 10,000 having referenced knowing the victim from previously meeting online.

The assignment specific tasks will be specified in the following section.

3. Tasks

1. Download and install Apache Tika
    a. Chapter 2 in your book covers some of the basics of building the code, and additionally, see http://tika.apache.org/1.25/gettingstarted.html
    b. Install Tika-Python, you can pip install tika to get started.
        i. Read up on Tika Python here: http://github.com/chrismattmann/tika-python
2. Download and install D3.js
    a. Visit http://d3js.org/
    b. Review Mike Bostock's Visual Gallery Wiki

        i. https://github.com/mbostock/d3/wiki/Tutorials
3. Download the Fraudulent emails dataset from Kaggle
    a. https://www.kaggle.com/rtatman/fraudulent-email-corpus
    b. Make a copy of the original dataset (because you are going to modify/add to it in this assignment)
4. Begin by converting the original dataset format into JSON
    a. Use Tika
    b. After converting the email dataset into JSON, ensure that each email is a separate JSON that will ensure easy processing and aggregation later
5. Add and expand the dataset with the following features
    a. Attack Type (aka "Ask") – may contain more than one
        i. Reconnaissance
        ii. Social Engineering
        iii. Malware
        iv. Credential Phishing
    b. Attacker Stylometrics
        i. Attacker title (e.g., Prince, Colonel, Nonprofit worker)
        ii. Urgency of the attack email (word strength, "urgent", "now")
        iii. Date/time of the email attack
        iv. Attacker offering (Money? Services?)
        v. Attacker location
            1. Resolve the location using the geoNames.txt dataset and compare the location referenced of the attacker either in the text, or via the Attacker's IP
            2. Add relevant geoNames.txt features including information about the locality
        vi. Attacker relationship (how do they claim to know the victim, 1) met online? 2) friend of a friend?, 3) met the victim in person before)
        vii. Attacker email sentiment
            1. You will use the USC Data Science Sentiment analyzer capability to perform this, see here: https://github.com/USCDataScience/SentimentAnalysisParser
        viii. Attacker language style
            1. Many misspellings (test different thresholds)
            2. Random capitalization (test different thresholds)
        ix. Attacker estimated age
            1. You will use the USC Data Science AgePredictor, here: https://github.com/USCDataScience/AgePredictor
        x. Attacker IP known as Phisher?
            1. See: https://scamalytics.com/ip as an example of how to look this up
6. Identify at least three other datasets, each of different top level MIME type (can't all be e.g., text/*)
    a. Check out places including: https://catalog.data.gov/dataset (Data.gov)

    b. For each dataset, develop a Python program to join the data to your new Fraudulent Emails dataset
        i. For each non text/* dataset, be prepared to describe how you featurized the dataset
    c. Each dataset that you join must contribute at least three features (in addition to the features you are adding described in part 5)
    d. For each feature you add, be prepared to discuss what types of queries it will allow you to answer and also how you computed the feature
7. Download and install Tika-Similarity
    a. Read the documentation
    b. You can find Tika Similarity here (http://github.com/chrismattmann/tika-similarity)
    c. Compare Jaccard similarity, edit-distance, and cosine similarity
        i. Compare and contrast clusters from Jaccard, Cosine Distance, and Edit Similarity – do you see any differences? Why?
    d. How do the resultant clusters generated highlight the features you extracted? Be prepared to identify this in your report
8. Package your data up by combining all of your new JSONs with additional features into a single TSV (tab separated values) file where the columns represent the features and the rows are the instances of email attack.
9. (**EXTRA CREDIT**) Add some new D3.js visualizations to Tika Similarity
    a. Currently Tika Similarity only supports Dendrogram, Circle Packing, and combinations of those to view clusters, and relative similarities between datasets
    b. Consider adding
        i. Feature related visualizations, e.g., time series, bar charts, plots
        ii. Add functionality in a generic way that is not specific to your dataset
        iii. See gallery here: https://github.com/d3/d3/wiki/Gallery
        iv. Contributions will be reviewed as Pull Requests in a first come, first serve basis (check existing PRs and make sure you aren't duplicating what some other group has done)

4. Assignment Setup

4.1 Group Formation

You can work on this assignment in groups sized at **minimum 2, and maximum 5**. You may reuse your existing groups from discussion in class. Please fill out the group details in the Google Form after class on Thursday, February 18. Only one form submission per team. If you have any questions, contact Keerti via her email address with the subject: DSCI 550: Team Details.

4.2 Fraudulent Emails dataset

Access to the data is provided by Kaggle. The dataset itself is 5.8Mb zipped and 17.8Mb unzipped. You may want to distribute the data between your team-mates since the data is fairly small (for now).

4.3 Downloading and Installing Apache Tika

The quickest and best way to get Apache Tika up and running on your machine is to grab the `tika-app.jar` from: http://tika.apache.org/download.html. You should obtain a jar file called `tika-app-1.25.jar`. This jar contains all of the necessary dependencies to get up and running with Tika by calling it your Java program.

Documentation is available on the Apache Tika webpage at http://tika.apache.org/. API documentation can be found at http://tika.apache.org/1.25/api.

Since you will be using Tika Python, you will want to read up on the Tika REST API, here: https://cwiki.apache.org/confluence/display/TIKA/TikaServer. The Tika Python library is a robust REST client to the Java-side REST API.

You can also get more information about Tika by checking out the book written by Professor Mattmann called "Tika in Action", available from: http://manning.com/mattmann/.

5. Report

Write a short 4 page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. What questions did your new joined datasets allow you to answer about the Fraudulent Emails and the attackers previously unanswered? What clusters were revealed? What similarity metrics produced more (in your opinion) accurate measurements? Why? What did the additional datasets suggest about "unintended consequences" related to the Cyber Attacks? You should also clearly explain which datasets you used to join the Fraudulent Emails and how you extracted the new features from each dataset.

Thinking more broadly, do you have enough information to answer the following:

1. Are there clusters of attackers with similar features that tend to attack victims the same way?
2. Does the time of day of attack matter?
3. Are specific types of asks more prevalent?
4. Is there a set of frequently co-occurring features that induce the email to be read?
5. What insights do the "indirect" features you extracted tell us about the data?
6. What clusters of Attacks and Attacker stylometrics made the most sense? Why?

Also include your thoughts about Apache Tika – what was easy about using it? What wasn't?

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail **dsci550spring2021@gmail.com.** Use the subject line: DSCI 550: Mattmann: Spring 2021: BIGDATA Homework: Team XX. So if your team was team 15, you would submit an email to dsci550spring2021@gmail.com with the subject "DSCI 550: Mattmann: Spring 2021: BIGDATA Homework: Team 15" (no quotes). **Please note only one submission per team**.

- All source code is expected to be commented, to compile, and to run. You should have at least a few Python scripts that you used to join three other datasets, and what you used to extract additional features.
- Include your updated dataset TSV. We will provide a Dropbox or Google Drive location for you to upload to.
- Also prepare a readme.txt containing any notes you'd like to submit.
- If you used external libraries other than Tika Python and Tika Similarity, you should include those jar files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.
- Save your report as a PDF file (TEAM_XX_BIGDATA.pdf) and include it in your submission.
- Compress all of the above into a single zip archive and name it according to the following filename convention:

  TEAM_XX_DSCI550_HW_BIGDATA.zip

  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.
- If your homework submission exceeds Gmail's 25MB limit, upload the zip file to Google drive and share it with dsci550spring2021@gmail.com.

Important Note:

- Make sure that you have attached the file when submitting. Failure to do so will be treated as non-submission.
- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.
- Again, please note, only **one submission per team**. Designate someone to submit.

6.1 Late Assignment Policy

- -10% if submitted within the first 24 hours
- -15% for each additional 24 hours or part thereof