Analyzing the Targeted Nature of Phishing Attacks

Jesse Truong, Abigail Thomas, Jerry Teng, Egan Johnson, Eric Zhang

The University of Texas at Dallas

ACM RESEARCH

Introduction

Students at UT Dallas receive enormous amounts of phishing emails on a regular basis. Our project, Analyzing the Targeted Nature of Phishing Attacks, aims to address the issue of phishing attacks that target college students. Our goal is to research the factors that cause college students to be targeted by phishing emails. We approach this issue from two fronts. A survey to research students' responses to phishing emails and a parser to better understand how an automated system might be able to combat phishing emails.

Survey Component

Upon addressing the main focus of our project, we decided to conduct a survey to gauge student responses to phishing and non-phishing emails. In order to remove as much bias as possible from the survey, we took a variety of measures:

- no backtracking
- 15 second time limit to view each email
- randomized question order

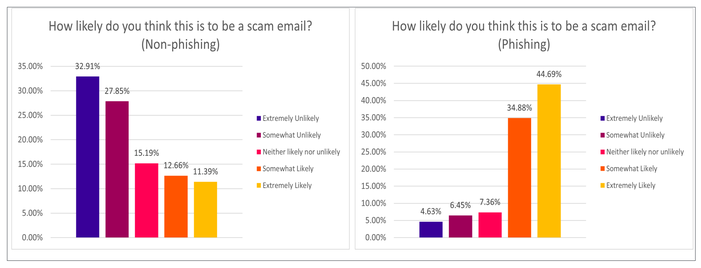

Fig. 1: Graphs of survey responses documented by percentage.

We asked survey participants to rate the likelihood of various emails of being phishing emails, as well as what factors made each one suspicious and trustworthy. We came to the conclusion that a larger amount of survey participants were able to recognize the phishing emails throughout the survey with more certainty, but the non-phishing emails were more divided among participants in terms of whether they were phishing or non-phishing.

Data Extraction

A brief foray into existing APIs like mailparser.io revealed some flaws including pricing and lack of customizability.

Features of Our Custom Parser

- filters out HTML text
- reads text from images and gifs
- parses PDFs and other text attachments
- generates language analysis scores
- performs sentiment analysis

Challenges

- relatively small dataset
- duplication in dataset
- unstandardized email formatting

Data and Analysis

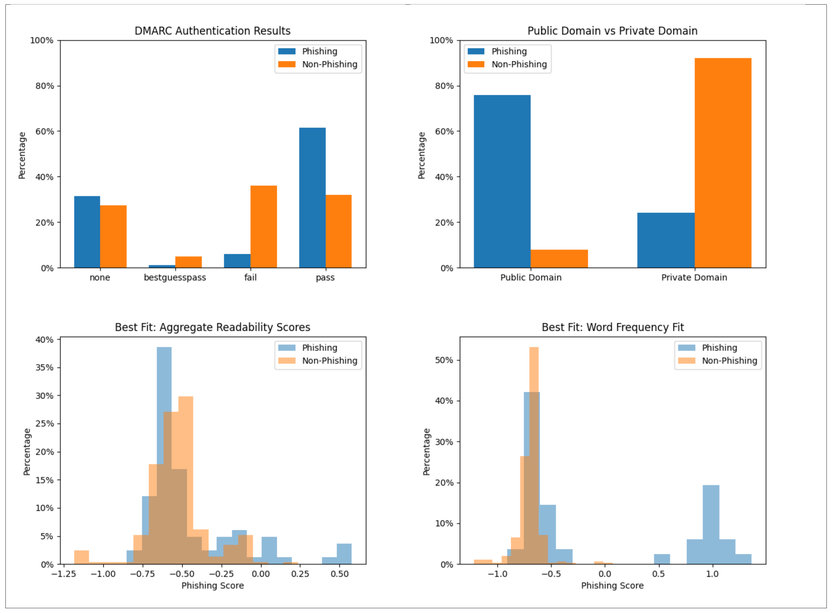

Fig. 2: Various graphs that represent data extracted using the email parser.

We selected our metrics based on their ability to separate the phishing and non-phishing emails. We tried a variety of metrics, some of which we decided were a poor predictor. We also had the additional issue of some metrics over performing due to matching a subset of the data that was largely identical.

Metric	Status	Status Justification
Sender Domain	✓	High degree of separation component of determining if sender spoofed
SPF Auth.	X	Inaccurate predictions against normal value
DKIM Auth.	X	Inaccurate predictions against normal value
DMARC Auth.	X	Inaccurate predictions against normal value
Readability Scores	X	Low separation, possible over-fitting to data
Reply to sender	✓	Useful predictor, component of determining spoof
Attachment count	✓	Moderate predictor, can white-list a small subset
Word frequency	✓	Accurate separator. Careful not to over fit
Subject Line intensity (caps, ! chars)	X	Not widely applicable

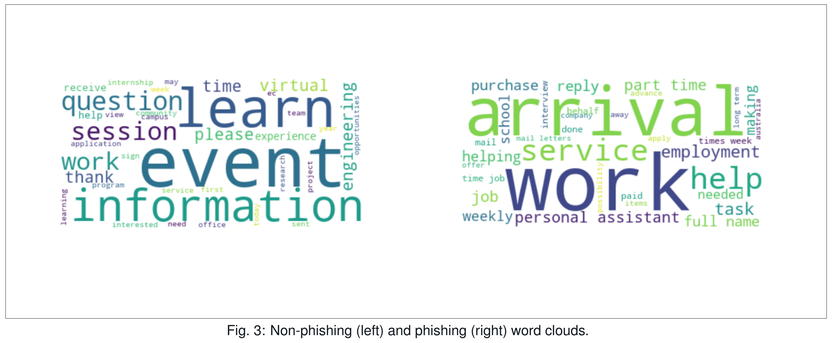

Fig. 3: Non-phishing (left) and phishing (right) word clouds.

Detection

In order to visualize the data and derive useful insights, we calculated three scores as defined below.

- Word Frequency Score: This score is calculated from the occurrences of various words frequently found in phishing and non phishing emails. Positive weights are assigned to each occurrence of a "phishing word," while negative weights are assigned to occurrences of "non-phishing words."
- Source Score: This score is determined by the source of the email. More specifically, email addresses were split into 'utdallas,' personal, and unknown domains.
- Reply-to Score: The reply-to score is another discrete variable which represents whether or not the reply-to field matches the sender. Again, three categories were derived from this measurement: reply-to not specified, reply-to matches from, and reply-to does not match from.

The three scores have been mapped to three coordinates (x, y, and z, respectively) and graphed in a 3-dimensional scatter plot for visualization:

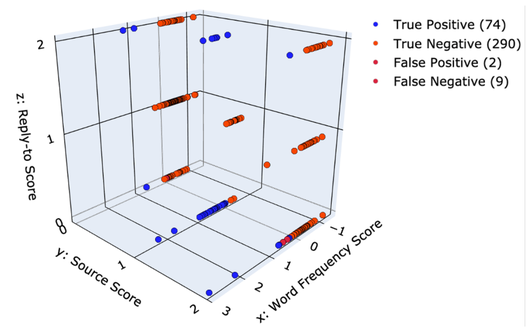

Fig. 4: 3-dimensional scatter plot visual

An interactive version can be generated by classifier.py

Our final detection script classifies emails into phishing and non-phishing categories. It was able to detect 89.16% of phishing emails, while incorrectly labeling 0.68% of non-phishing emails as phishing. Both of these rates were better than the rates achieved by the current system, which is Outlook's general-purpose spam filter.

Conclusion

From our survey results, parsing tool, and detection tool, we have deepened our own understanding of the underlying motives and characteristics of phishing emails—specifically those targeting college campuses. Our analysis demonstrates the prevalence and danger of these targeted phishing attacks. Through the analysis of various features in emails, we have created a detection script that captures 40% more phishing emails than the current solution, Outlook's built in general-purpose spam filter.

Outside of data and analysis, we have also raised awareness about these targeted phishing attacks through our survey. We hope that our findings help colleges to identify and combat phishing emails to better protect students.

References

[1] 61% of emails being read for more than 8 seconds
https://www.litmus.com/press/litmus-report-email-read-time-increased-by-21-percent/