Testing Effectiveness of AI-Enabled Phishing Attacks based on Public Information

By

JACKY LIN
THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Houman Homayoun, Chair

Matt Bishop

Chen-Nee Chuah

Committee in Charge

2023

Abstract

As modern technology advances and more users are utilizing the internet, people's information has become accessible to the public. Although sensitive information such as ID numbers, bank accounts, or passwords might not be publicly discoverable, an individual's name, interests, education, and social connections may be discoverable.

This research aims to assess the effectiveness of spear phishing emails built upon publicly available information, focusing on individuals within academia as the subjects. The social network of the subjects would be constructed using an information scraper built with Python and machine learning algorithms. The content of the emails would be generated by a Large Language Model (LLM). The experiment aims to evaluate the efficacy of AI-driven target selection through the response rates of email opening and engagement of the embedded link.

Our hypothesis posits that publicly disclosed personal information potentially threatens an individual's online security and privacy. This vulnerability is manifested through a greater susceptibility to spear phishing attacks and inferred private information using machine learning techniques.

Additionally, we would discuss some mitigation strategies from the perspectives of email service providers, organizations, and users. Our goal is to warn the public regarding the potential threats of publicly available information being accessible to attackers. Being aware of phishing attacks that rely on personal social networks could reduce the success rate of such attack angle.

Table of Contents

Chapter 1 - Introduction

1.1 Misuse of LLM

Over the past few decades, the internet has a growing impact on daily communication and information exchange. While it is an effective tool for exploring content and expanding knowledge online, unsafe browsing habits or sensitive information exchange may pose a vulnerability to individuals. Viruses, worms, trojans, spyware, ransomware, or any other malware could penetrate into the system's firewall, potentially leaking user's information and his social connections. The leaked information may be utilized to perform social engineering attacks on various communication platforms, specifically emails. In recent years, there has been a growing concern about the practice of spear-phishing attacks that are designed to deceive victims into disclosing their personal information, transferring money, or even installing malware. Such attacks are often accomplished by luring the victims into clicking malicious links [3]. This study specifically focused on spear-phishing attacks in malicious email scenarios. We also work closely with the Institutional Review Board (IRB) and the university's IT department to ensure the safety of this research.

Although the advancement of LLM has brought many great applications to the world, there are also some downsides. In fact, it has already created many fake creations, such as forged emails, image generation, and even videos. These may be deceiving and sometimes difficult to determine with human perception. Even if it may be illegal, numerous fake content is still being created constantly [18]. Since LLM is still a relatively new trend, legal policies for the use of LLM are still in the process of being certified and still require modification. As a result, possible legal gaps exist to escape this technology's misuse. It can be menacing for people without knowledge of LLM and its uses.

1.2 Phishing events over the past couple of years

Phishing, in general, has always been a critical and sensitive topic on the internet. Especially with the recent advancement in LLM, spam and phishing emails may be fine-tuned to have more variations and in different tones depending on the targeted individuals. It may be possible to massively forward thousands of phishing emails to the victims within a short amount of time. Over 48% of the emails sent in 2022 were detected as spam [4]. Although spam emails are different than phishing emails, there were still 40 million emails that were discovered to include malicious content in 2022 [5]. With the targeted use of LLM, the generated emails may require more work for individuals to determine their legitimacy. Although there might be some deviation from the actual data, it was discovered that there were 300,497 phishing victims, with a total loss of $52,089,159 in the U.S. alone in 2022 [6]. Moreover, between 2021 and 2022, there was a 29% increase in malicious files in emails, and the number of unknown malware increased by 46%. As of today, about 3.4 million spam emails are sent daily [4]. At this rate, it would not be difficult to believe that the amount of spam emails would rise even further this year. These spam emails may be phishing emails that could include malware. The damage these phishing events could cause may be immeasurable if it continues to increase over the years. The increase of phishing emails and their success chance may also be affected by the advancement of LLM, knowing its capability to generate human-like messages.

1.3 Spear-fishing

Although it is true that there has been a large amount of mass email spamming going on recently, most phishing attempts performed through emails are presented as fake invoice scams, likely pretending as some well-known company [7]. It is still a relatively undiscovered field regarding whether or not people are likely to respond to individuals within their social network.

Spear-phishing lies within the category of phishing attacks, characterized by generating tailored messages to specific individuals. Such attacks are often designed to present information that appears trustworthy enough, convincing the victims to fall for the phishing bait. The attackers often choose to manually construct detailed messages to ensure a high probability of success.

A phishing message may be generated by anyone with enough proficiency in a language. However, the success chance would only be high if enough information about the victims were discovered to construct a convincing message as the phishing bait. This may be challenging since the attackers would have to primarily rely on publicly available information, assuming without penetrating the victim's system.

An individual may post his information and personal interests on social media or any publicly accessible platform. This information may often include the individual's email or workspace, which is actually relevant enough to perform the phishing attack. As more platforms are becoming publicly accessible, it is also becoming easier for an attacker to gather more information regarding a person, even information that people did not publish themselves. For instance, an individual's friends might post pictures of that individual on social media, possibly revealing their names, connections, and even locations.

1.4 Research overview

This research explores the possibility of utilizing machine learning and existing publicly available information to infer unpublished information regarding the victims, specifically focusing on the social network in the field of academia. When a researcher publishes an article online, it proves the social connections he has with the other co-authors of the paper. We may also utilize this information to infer whether the authors are professors or students. In other words, a researcher’s published works and profiles that are publicly available may be exploited to construct a social network based on his professional history. However, our intention focuses on alarming the public with such phishing attacks, we do not plan on performing any malicious act that might be fatal for the subjects.

Recent research has found that 34 out of 35 well-known email providers can be penetrated with forged emails, even with authentication checks on the user end. Additionally, only 9 out of those providers implemented security indicators [1]. It may be alarming for the public to start paying attention to incoming emails in their inboxes. This research will also discuss some defense strategies that may be applied to enhance phishing protection from the perspectives of email service providers, organizations, and users.

This study aims to explore the vulnerability of publicly available information and perform phishing experiments under safely regulated circumstances. These phishing events would be done by sending forged emails to the subjects without prior knowledge about the experiment (with IRB approval). An embedded link would be contained in those emails, and the subjects' click-through events would be recorded. Once the experiment was concluded, we obtained their consent to continue using their data while keeping their personal information anonymous. None of the data presented in this paper would reveal the subjects' identification or association with the data.

Chapter 2 - Background

2.1 The 3 phases of spear-phishing

A phishing attack is a type of social engineering attack in which the attacker disguises as a credible party to send deceptive messages to the victims, with the goal of manipulating the victims to perform fatal actions. This type of attack is designed to convince the victims to perform actions that they normally would not undertake with an unverified source, such as disclosing sensitive personal information, transacting money, or installing malware.

In this research, we will focus on the spear-phishing attack analysis, which is a targeted phishing attack that consists of three fundamental phases: a profiling phase, a spoofing phase, and a payload phase. In the profiling phase, the attackers would gather information about the targets and construct dedicated profiles to generate phishing messages tailored toward those victims. Information like email, workspace, and social connections are all viable sources for constructing a convincing phishing message. In the spoofing phase, the forged messages would be sent to the victim, disguised as a trustworthy party using the previously constructed profiles. The messages would be tailored to align with the victims' interests, providing a false sense of security and luring them to click on the malicious link. In the payload phase, the malicious link would often offload malicious files to the victim's device as soon as they click on the link. Another possible scenario is that the link would be directed to a bogus website that asks for the victim's critical information.

2.2 Growing concerns of phishing

As phishing attacks constitute a significant issue in cybersecurity for their widespread use and effectiveness, there is also growing interest in the mitigation side of things. More people are concerned about their sensitive credentials, making them also becoming more interested in the defensive strategies against phishing attacks [20]. In recent years, more attention has been put on evaluating how public information is handled, especially regarding digital footprints [19]. An individual's digital footprint consists of traceable online activities, such as social media activities, publications, or browsing habits. Attackers might be able to exploit these traces to perform malicious behaviors.

With the growing ubiquity of internet utilization in our daily lives, an individual may inadvertently accumulate a substantial digital footprint, possibly without realization as well. Even if the individual is aware of leaving their footprints online, he might not understand the potential threats that lie within those traces. While digital footprints may be beneficial for companies to understand a user's browsing interests, we argue that this information may pose privacy concerns and security threats to the users that they are not aware of. To solidify our argument, we propose a semi-automated, AI-enabled spear phishing attack framework that exploits the publicly available information of academics and is capable of performing phishing attacks on a large scale.

2.3 Attack Intention

While AI and machine learning-driven phishing in cybersecurity has increased in attention within academic discourse over the past few years, most existing studies only explore the problem on a theoretical basis [21]. As of the date this research is published, we are currently unaware of studies or experiments that attempt to measure the impact of LLM in phishing attacks in real-world scenarios, especially focusing the subject analysis in academia. This study aims to fill this knowledge gap, offering insights into AI-powered attacks while also suggesting potential defense mechanisms.

The subject pool for the attack is restricted to those in academia. Most researchers have their information available online, including their social network, which makes it easier for attackers to gather their information. Proving the viability of the proposed attack framework may illustrate the potential effects if it were to be performed on the general public. Generalization of the proposed attack framework would mostly depend on how the targets' information is scraped online, possibly from social media rather than the research databases.

Previous social engineering and information security research have discussed the effects of deception on experimental results. Specifically for our research, if a participant is aware of the fact that they will be receiving a phishing email, the subject may pay extra attention to the emails that appear in their inboxes. In other words, the disclosure of the research may cause the subjects to behave differently than usual [2]. In real-world scenarios, there would be no warning as to when the victims will be receiving a phishing email, making them more prone to phishing baits if they normally do not practice cyber hygiene. Therefore, we have acquired approval from the IRB to perform the experiment on subjects without providing them with any prior information regarding the research. However, we would debrief the subjects regarding the research and acquire their consent after the experiment has been conducted. The experiment is expected to reveal the most honest behavior in the final results.

Chapter 3 - Methodology

3.1 Methodology Framework

As mentioned in the background section, there would be 3 phases that consist of a typical phishing attack: the profiling phase, the spoofing phase, and the payload phase, as shown in Figure 3.1. In the profiling phase, we would focus on scraping two groups of authors from research databases: one for training and testing the machine learning model, and another for performing the experiment. After scraping, the data would be cleaned and used to construct the social networks with the scraped authors. Then, the data would be passed into machine learning classifiers for training and testing. The pre-trained model would be saved for later when we need to predict the labels for the experiment group of authors. In the spoofing phase, LLM would be utilized for forged email generation, while disguising as professors. Gmail accounts would be created for sending the generated emails to the subjects. Finally, in the payload phase, we would track the click-through of the emails and the embedded links. We do not intend to perform any actual malicious act when the subjects fall for the phishing bait, strictly following the IRB guidelines.

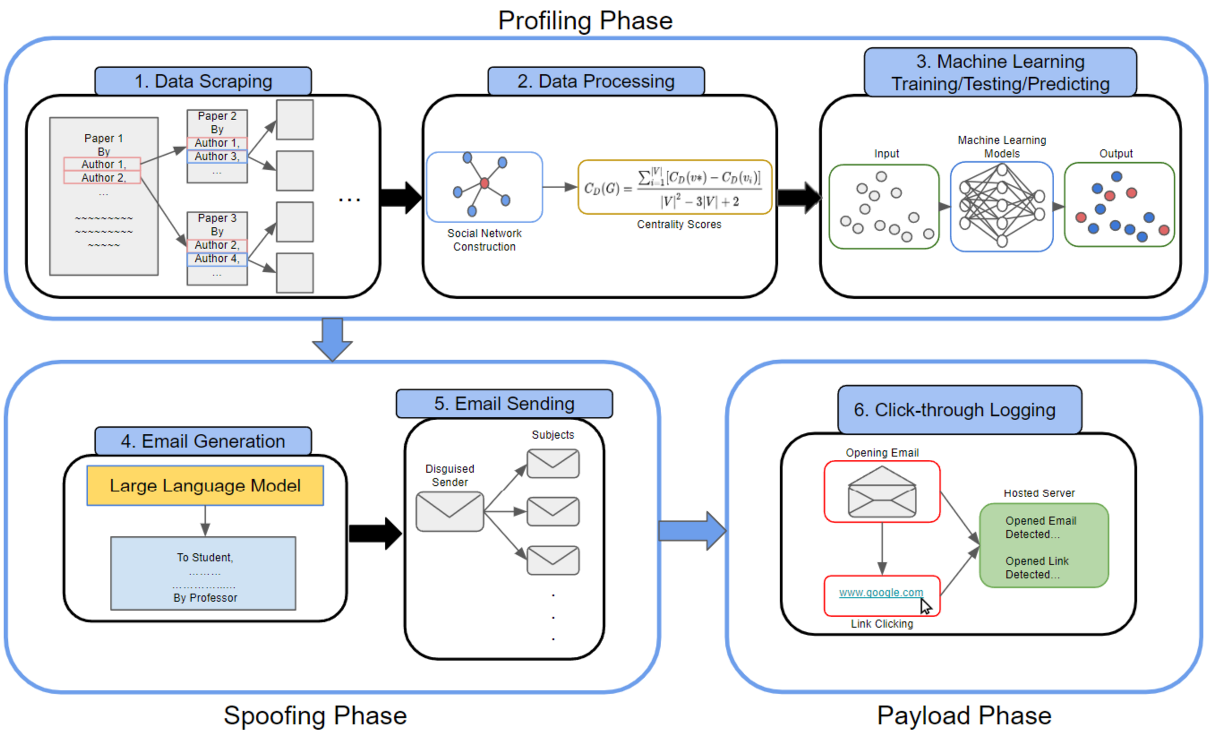

Profiling Phase
1. Data Scraping
Paper 1
By
Author 1,
Author 2,
...
~~~~~~~~~
~~~~~~~~~
~~~~~
Paper 2
By
Author 1,
Author 3,
...
Paper 3
By
Author 2,
Author 4,
...
...
2. Data Processing
Social Network
Construction
$C_D(G) = \frac{\sum_{i=1}^{|V|}[C_D(v*) - C_D(v_i)]}{|V|^2 - 3|V| + 2}$
Centrality Scores
3. Machine Learning
Training/Testing/Predicting
Input
Machine Learning
Models
Output
4. Email Generation
Large Language Model
To Student,
.........
..................
By Professor
5. Email Sending
Subjects
Disguised
Sender
6. Click-through Logging
Opening Email
Hosted Server
Opened Email
Detected...
Opened Link
Detected...
www.google.com
Link Clicking
Spoofing Phase
Payload Phase

Figure 3.1: Methodology framework - the three phishing phases

## 3.2 Data collection for training and testing with machine learning classifiers

Before experimenting, it is necessary to train the classifier to distinguish the author as either a professor or a student for building the social network. The profiling phase of phishing is to scrape data from the AMiner dataset to train the classifiers. The AMiner data set from Open Academic Graph (OAG) was selected due to its coverage of authors within different fields and varied collaborations. After training the machine learning models, the next step would be scraping data from online research databases such as IEEE or ACM and then using the pre-trained models to predict the role of the author.
~~~~~

When performing the experiment, we acquired consent from 6 professors to disguise as them to send the emails. Therefore, 46 authors would be extracted from the scraped social network to experiment, prioritizing the connections of the collaborated professors. The social network would illustrate the professor-student relationship, indicating that the professor could be supervising a group of students in the network. We would also verify their roles through manual searching to ensure the precision of the predicted result from the pre-trained classifier. Figure 3.2 illustrates the general concept of the data collection process. Further details will be explained in the following sections.

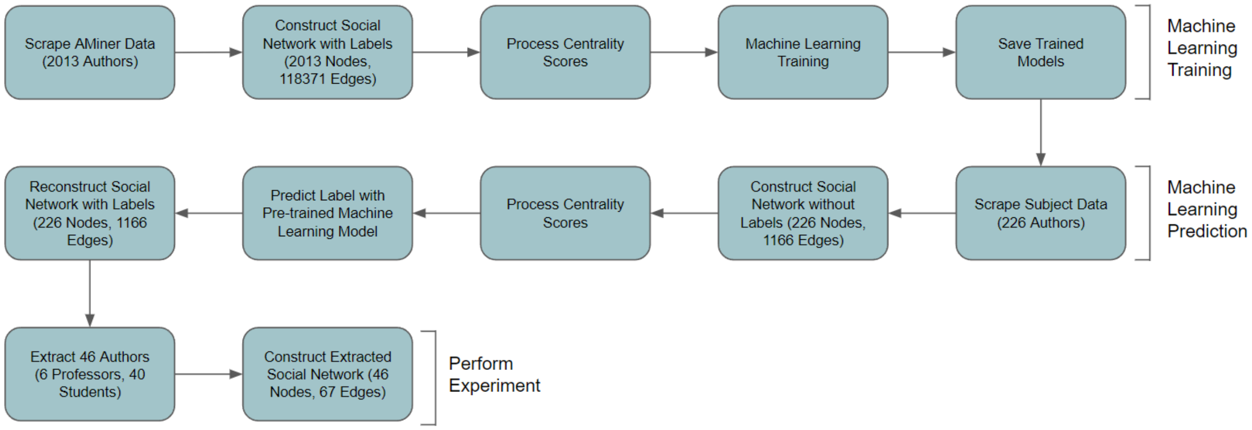

Scrape AMiner Data (2013 Authors)
Construct Social Network with Labels (2013 Nodes, 118371 Edges)
Process Centrality Scores
Machine Learning Training
Save Trained Models
Machine Learning Training
Scrape Subject Data (226 Authors)
Construct Social Network without Labels (226 Nodes, 1166 Edges)
Process Centrality Scores
Predict Label with Pre-trained Machine Learning Model
Reconstruct Social Network with Labels (226 Nodes, 1166 Edges)
Machine Learning Prediction
Extract 46 Authors (6 Professors, 40 Students)
Construct Extracted Social Network (46 Nodes, 67 Edges)
Perform Experiment

Figure 3.2: Data collection flow chart

3.2.1 Processing the AMiner Dataset

Initially, the plan is to train the machine learning classifier with both AMiner and Microsoft Academic Graph (MAG) datasets, both provided by the OAG. However, the AMiner data includes pre-labeled positions for the authors, whereas MAG does not. Therefore, we have dedicated the training process solely to using the AMiner data.

The AMiner dataset serves as the foundation for generating the structured graph and node files. In the node file, each entry consists of the author ID, name, last authorship, number of publications, number of citations, degree centrality, closeness centrality, betweenness centrality, and author's label. Throughout the paper, ‘last authorship’ indicates whether or not an author appears to be the last author in any paper at least once. Each entry in the node file corresponds to an author's information. In the graph file, each entry consists of two connected edges represented by the authors’ ID, signifying that two authors collaborated on a paper.

The centrality scores found in the node file are also computed based on the connections established in the graph file. Degree centrality represents the ratio of the number of edges attached to a node over the entire graph. A higher degree simply means the node is connected to more edges compared to the overall graph, while a lower degree means the opposite. In our context, degree centrality represents the number of associations (popularity) for the authors, which should be higher for faculty than students. Looking at the equation in Figure 3.3, the degree centrality of the node (v) is equal to its connected edges (e) over the total number of edges (E) in the graph. To calculate the degree centrality for the entire graph, accounting for all the nodes, the equation in Figure 3.4 is used. This equation would account for any graph G:= (V, E), while V means vertices and E means edges of the graph.

$$C_D(v) = \frac{e}{E}$$

Figure 3.3: Degree centrality equation for a single node

$$C_D(G) = \frac{\sum_{i=1}^{|V|}[C_D(v*) - C_D(v_i)]}{|V|^2 - 3|V| + 2}$$

Figure 3.4: Degree centrality equation for the entire graph

Closeness Centrality measures the shortest path between a node and all other nodes. In other words, it represents how close a node is when reaching other nodes in the graph. Higher closeness centrality indicates that the node is able to interact with other nodes more efficiently and is likely more centrally located. The equations for closeness centrality are shown in Figure 3.5, N means the total number of nodes, and d(u,v) represents the distance between u and v.

$$C(v) = \frac{N-1}{\sum_u d(u,v)}$$

Figure 3.5: Closeness centrality equation for a single node

Betweenness Centrality indicates the frequency of a node serving as a connection between two other nodes along the shortest path. In our case, it represents the middleman that acts as a bridge of connection between two authors. Figure 3.6 shows the equation of betweenness centrality for a single node. The sigma in the equation represents each pair of vertices (s,t) and the shortest path between them.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Figure 3.6: Betweenness centrality equation for a single node

3.2.2 Finding author connection

The AMiner dataset includes two files: the author file that contains the list of all authors and their bibliography, and the paper file that contains the list of all papers with authors' names included. Both files are parsed in JSON format. To search for authors' connections in the dataset, an algorithm that uses breadth-first search (BFS) is created. To begin with the process, a few authors were randomly selected from the AMiner author file. Those authors would be stored as elements in a list. Each element would be treated as the root node of BFS in the paper file with a queue and a visited array.

In the first iteration, an author would be popped from the list. Then, BFS would be performed with that popped author. In other words, if any paper contains the popped author, all the other authors in that paper are appended to the queue, and the popped author would be moved to the visited array. Going down the process, authors in the queue would be continuously popped and appended to the visited array using BFS to recursively access the AMiner paper file to discover more authors until there are no more author connections to be found. All the discovered author connections would be stored as graph edges in a TSV-formatted file. We will be referring to it as the graph file in this paper. Then, the second iteration would begin with another author popped from the list of randomly selected authors. A new visited array would also be created for the second iteration and compared against the visited array from the first iteration for length. A simple illustration in Figure 3.7 shows the structure of the algorithm as described in the paragraph.

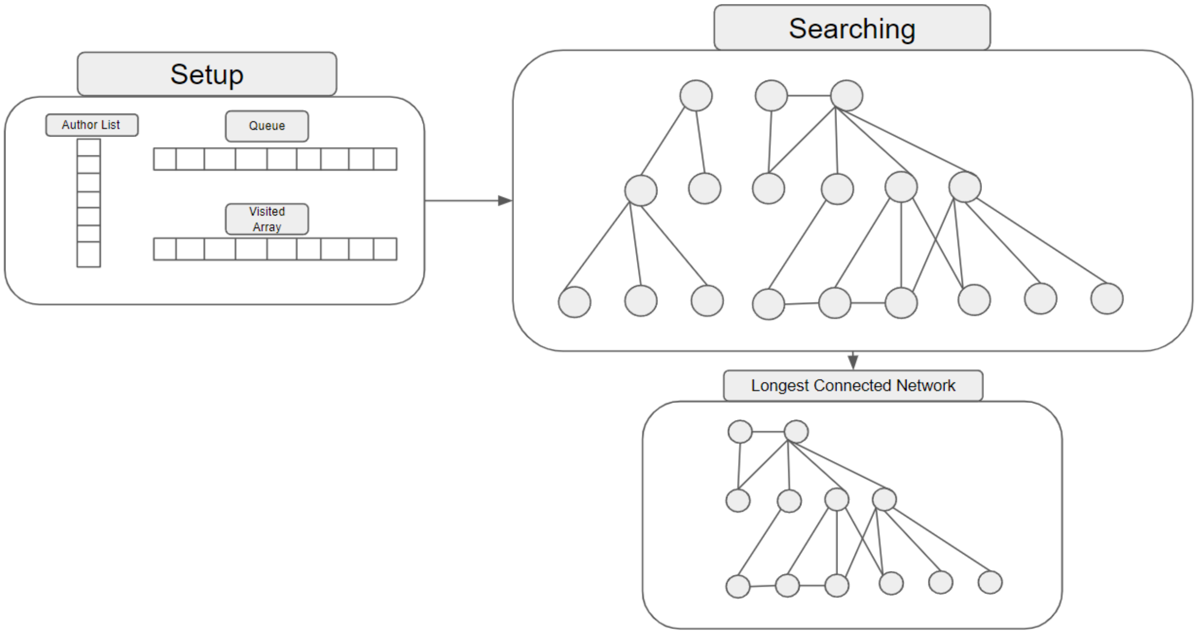


Figure 3.7: Searching algorithm framework

After all the iterations, the longest visited array will be kept with all the edges in the graph file. A TSV-formatted node file would be generated based on the existing authors in the graph file. The author information may be found in the AMiner author file. Each entry in the node file will document an author's ID, name, last authorship, number of publications, number of citations, and position.

3.2.3 Labeling last authorship

Not all authors in the AMiner author file have positions clearly listed. Many would only label the authors as a researcher or indicate the organization they worked in. Therefore, some rules have to be applied to label the authors correctly. The last author of a research paper is often the project investigator or supervisor of the research [8 - 9]. Therefore, we are including the last authorship as an intuitive feature for each author, serving as a possible indication that the author might hold a faculty title.

Furthermore, with the help of natural language processing (NLP), if an author has a bibliography or title that can describe him as a professor, such as "professor" or "prof" in his bibliography, or that he is the last author in a paper, or that he has a significantly high amount of publications and citations, he will also be assigned with the role of a faculty. Similarly, the student position is labeled by using NLP to detect key information in the author's bibliography, checking the last authorship, number of publications, and number of citations that the author has.

3.2.4 Processed Social Network

For the final fully connected graph, there are 2013 nodes and 118371 edges. There is an abundance of edges due to the connections between the authors in a paper. For instance, in a paper with 5 listed authors, each author would establish a connection with every other author in the paper. Therefore, in this example, each person would have 4 direct connections, resulting in 10 edges in total, as shown in Figure 3.8 below. Our goal is to find out who has direct connections with one another, strictly specifying the connections among the colleagues. If there is a middleman who connects with everyone in the paper, it would be almost impossible to distinguish that specific group of authors when one of them is connected with other groups, resulting in a strand of authors rather than groups of authors. Since the professors will likely have more connections with other nodes, if one of the authors in a group connects with many other groups, there is a high chance that the author may be a professor. With this in mind, we may disguise as that professor to send spoofing emails to all other authors in that example group, who are likely the professor's students.

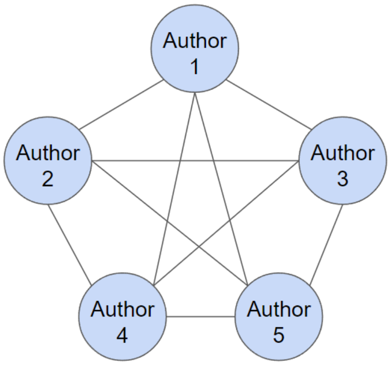


Figure 3.8: Example author connections in a paper

The social network graph is illustrated in Figure 3.9. The diagram might look a little clustered due to its excessive connections. The circles in blue represent the professors, and the circles in red represent the students. However, it is distinguishable that most of the professors are located in the inner circle, while the students mostly appear in the outer region. It would be more apparent when the number of subjects is reduced in the experiment to provide a clearer view of the relationship network. This is just to show that the professors should be in the inner area, whereas the students should be near the outer circle.

Additionally, once the number of authors reduces, it may be easier to distinguish which students are likely working under which professors for us to identify the mentor-mentee relationship among the authors. Knowing this information, we may be disguised as specific professors for sending out benign phishing emails to their students.

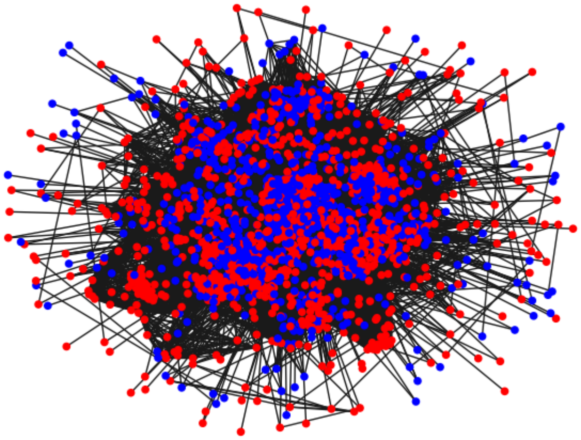


Figure 3.9: Social network from the AMiner dataset

Knowing that the social network diagram may be challenging to distinguish, The features of the collected dataset are represented in Figure 3.10 and Figure 3.11. Especially for Fig. 3.10, we can observe that the professors have a higher degree of centrality, more publications, and more citations. At the same time, the students are statistically lower in these three fields, which intuitively makes sense. Note that although a few students have higher closeness centrality, meaning they are closer to all other nodes, the averages of the professors' closeness centrality are still higher. In Figure 3.11, we can observe a close relationship between the number of publications and citations, indicating that more published works tend to yield more citations.

While the betweenness centrality does not seem to contribute much to the overall classifier results, the degree centrality and closeness centrality have shown a significant correlation, illustrating how a node with more connection would result in higher centrality scores. The centralities also present the relationship with several citations and publications, showing a trendline indicating that more publications and citations often represent the higher value of centrality scores, indicating that professors with more published work also tend to have more connections with others. Overall, it is clear that although some students perform well in terms of publications and citations, most of the trendlines are represented by the faculty members. The few exceptions of the authors labeled as students may even be caused by mislabeling during the data cleaning process.

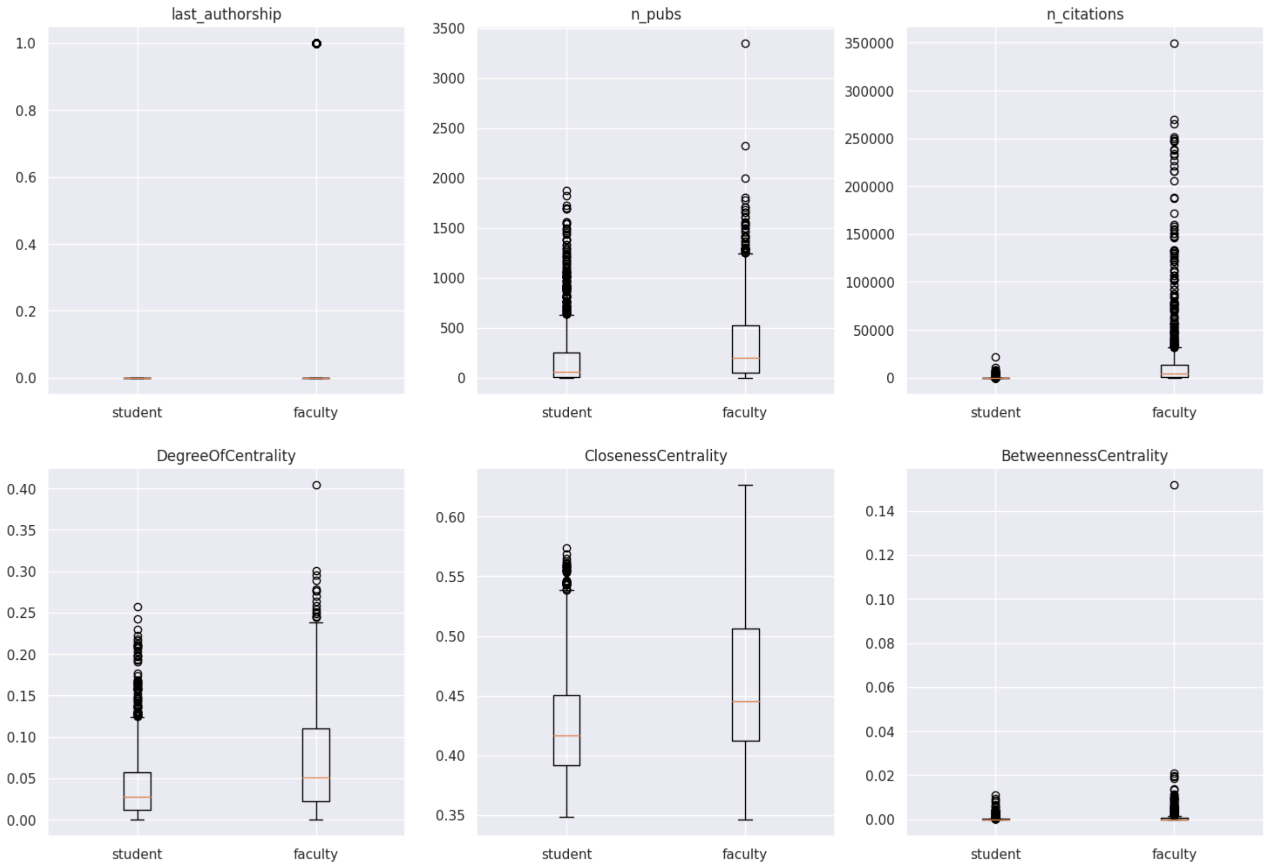


Figure 3.10: Aminer data - features representation

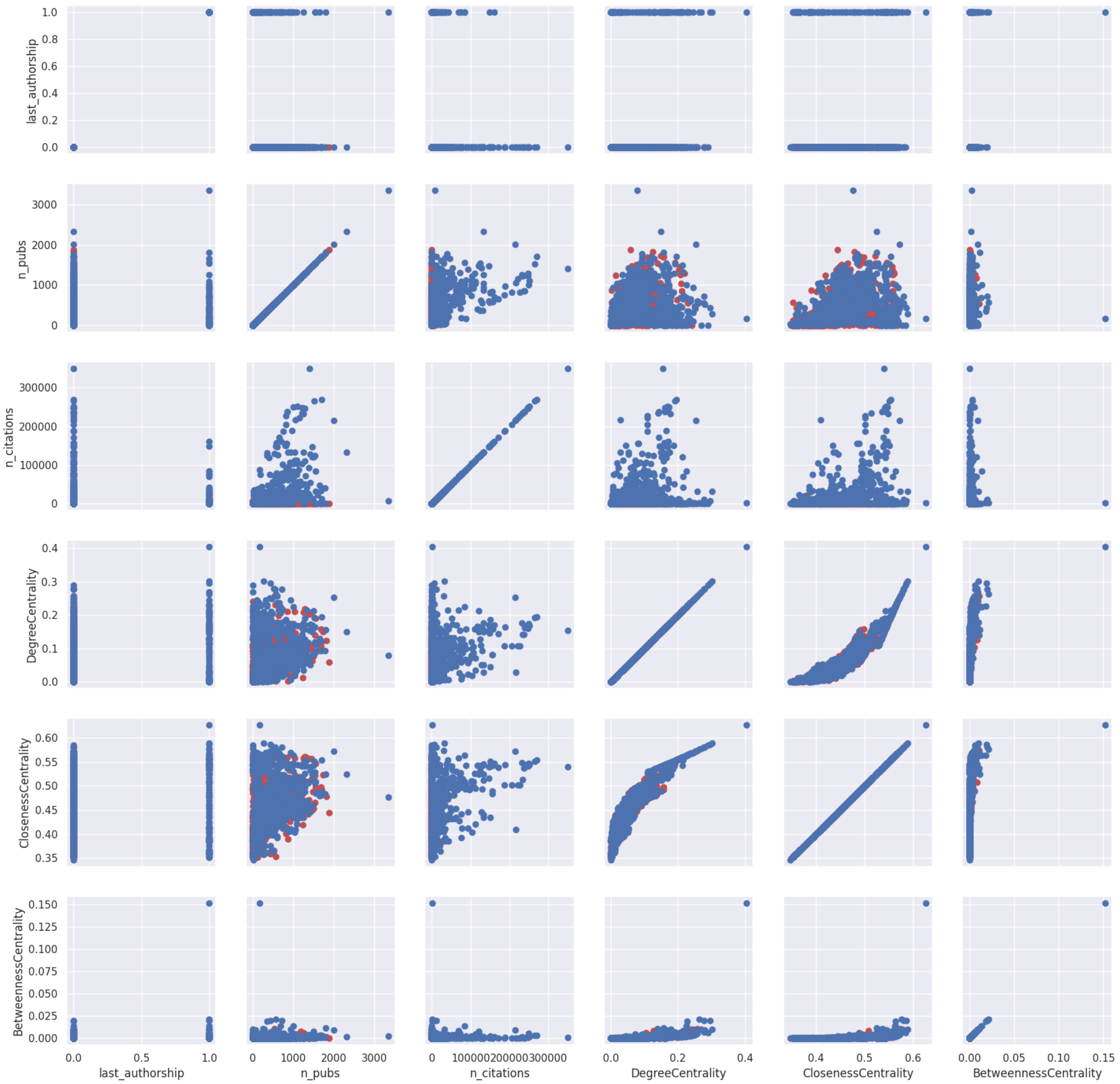


Figure 3.11: AMiner data - cross-features representation

3.2.5 Machine learning classification

The collected AMiner data is used to train and test the machine learning model that classifies the author as a professor or a student. The classifiers used for the experiments are: Random Forest, Logistic Regression, scaled RBF SVC, auto RBF SVC, Linear SVC, Non-Linear

SVC with auto gamma, Non-Linear SVC with scaled gamma, K-Nearest Neighbors, Gaussian Naive Bayes, and Bernoulli Naive Bayes. The features used for training are the last authorship, number of publications, number of citations, degree centrality, closeness centrality, and betweenness centrality. The label is the author's position, indicating the author as either a professor or a student.

With the aid of centrality scores, the accuracy would be increased due to the centralized connections of the professors, raising roughly 5-6% accuracy on average. In other words, authors connected to more authors and closer to the center of the graph are more likely to have a higher degree of centrality scores, indicating more likeliness as faculties. Furthermore, with the help of the last authorship, intuitive features also helped to increase the accuracy of the classifiers, raising about 20% accuracy on average.

Figure 3.12 presents the final accuracy, precision, recall, and f1-score with all features included for all 9 classifiers. The highest accuracy is 98.2%, achieved by the Random Forest classifier. Its precision, recall, and f1-score are 97.52%, 98.75%, and 98.13%, respectively, demonstrating its effectiveness in distinguishing the authors' positions. Even the worst classifier has an accuracy of 62.71% produced by the Non-Linear SVC with scaled gamma. As illustrated in Figure 3.12, most classifiers can achieve an accuracy of above 80%. Overall, with the implementations of centrality scores and the last authorship, the average accuracies rose from 60-70% to 90-95%.

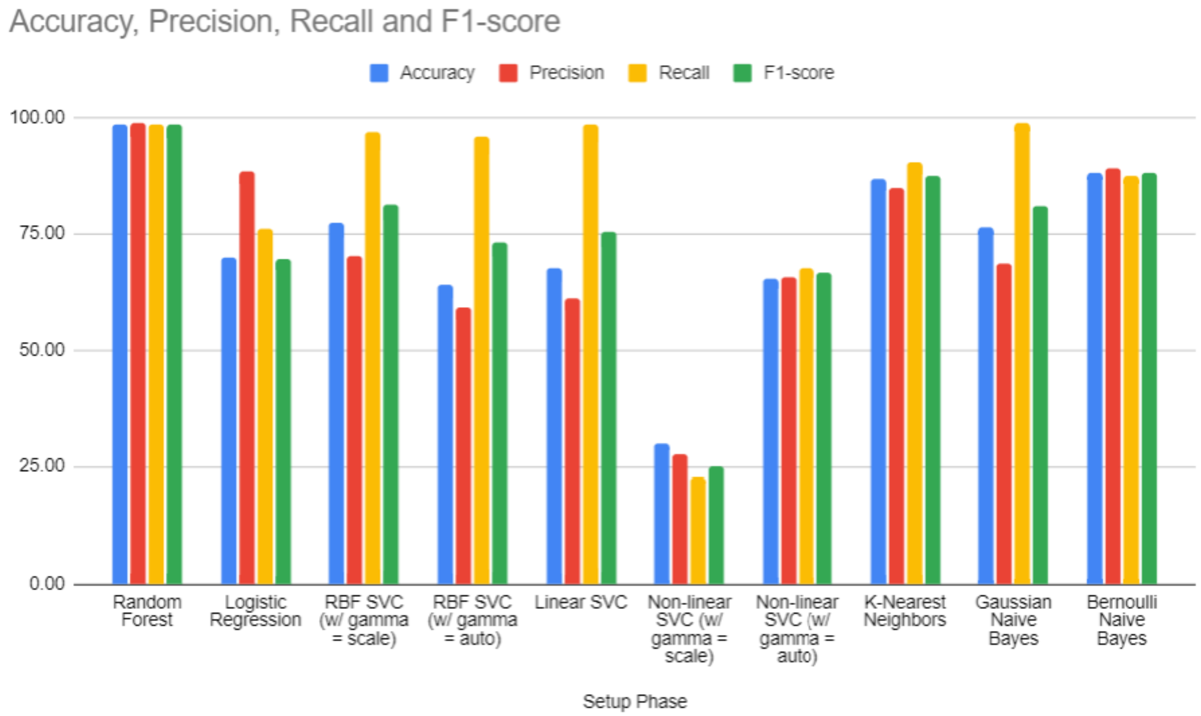


Figure 3.12: Machine learning training result

For training the data, the results from the machine learning models demonstrate that with the features of publication count, citation count, etc., the authors may be correctly classified as either a professor or a student from the AMiner data set. Once the IEEE and ACM data are scraped and cleaned, they can be passed into the pre-trained machine-learning models for classification. Those data would then be used to construct the social network graph of the authors for implementing the attack in real-world experiments.

3.3 Data collection for prediction with pre-trained machine learning model

3.3.1 Scraping author information from online research databases

As mentioned earlier, the experimental data would be scraped from the online research databases that would be used to create a social network among the authors. The data collection process would be similar to when collecting the AMiner data set. We have gotten 6 professors who agreed to participate in our experiment and kindly permitted us to use their names to email others. Therefore, we are using those 6 professors as the root authors in the algorithm discussed in Section 3.2.2 for creating the social network graph. Additionally, the 6 professors all have collaborated with one another before, so the final graph would be fully connected. Another thing to note is that the professors we collaborated with are mostly in the computer science (CS) or engineering departments. Therefore, authors scraped from the research databases would also be more engineering-focused.

After the collection process, we ended up with 226 nodes and 1166 edges to construct the social network. The authors' information includes: names, affiliated institutions, years of active publication, personal biography, attended conferences, citation counts, number of published papers, published papers, and email addresses. However, since we cannot accurately acquire the active years for the AMiner dataset to train the models, we have decided to omit that feature when labeling the authors of research databases. Therefore, to match the entry with the trained classifier, only the name, last authorship, publication count, citation count, and the calculated centrality scores would be stored in the TSV file as features for each author.

Following similar procedures, each collected author would have their corresponding ID. The node and graph lists would be created based on the authors' connections in the scraped data. A network of authors who are directly or indirectly associated with each other would be constructed. This network may be utilized to determine the disguised senders and receivers for the phishing emails. Only authors within this network are to be considered as the subjects of the study.

3.3.2 Prediction with pre-trained machine learning model

The next step is to predict the author labels with the pre-trained machine learning model. Precision is crucial for impersonating professors and not students when sending emails. Therefore, we selected the classifier with the highest accuracy during training and testing in the previous step, which is the Random Forest classifier. Its confusion matrix after testing with the AMiner dataset is shown in Figure 3.13, having an average of 98% accuracy. Using this pre-trained model, the labels for the scraped data are predicted. We then verified the result through manual searching of the authors to confirm the prediction result, and only one author was labeled incorrectly. The results are shown in Table 3.1, illustrating an accuracy of 45/46 = 98%. In real-world attacks, attackers do not even have to verify the author's roles. Even if some authors were mislabeled, as long as the majority of the authors are predicted correctly, the mistakes would be negligible and the attack would still be successful.

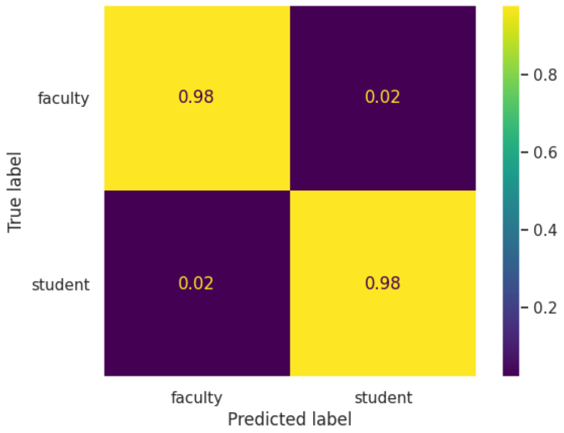


Figure 3.13: Confusion matrix from testing (AMiner data)

	Predicted faculty	Predicted Student
True faculty	6	0
True student	1	39

Table 3.1: Confusion matrix from prediction (scraped data)

3.3.3 Constructing the social network with the scraped data

After successfully labeling the authors, it is now possible to construct the social network graph. Figure 3.14 offers a significantly clearer perspective than Figure 3.9, which has a greater abundance of nodes and edges. Same as Figure 3.9, the circles in blue represent the professors, and the circles in red represent the students. Figure 3.14 also illustrates that the professors are more centralized in the graph, whereas the students mostly appear in the outer region. It is also interesting to note that the professors tend to collaborate with each other and sometimes other professor’s students as well.

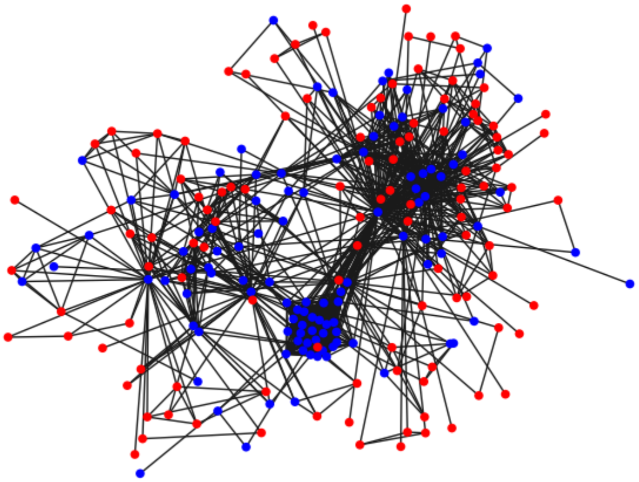


Figure 3.14: Social network for authors scraped from research databases online

As mentioned previously, we have acquired 226 authors by scraping the research databases online. The features of the data are illustrated in Figures 3.15 and 3.16. Similarly to the training data, the scraped data features show that professors tend to have a higher number of publications, citations, and centrality scores. They also tend to be the last authors of a paper. As shown in Figure 3.16, there appears to be an apparent trendline between the number of citations and the number of publications. Although it is not always true, it does indicate that more published work would often result in more recognition and therefore more citations. As expected, the trendlines can also be observed between the centrality scores, indicating how centralized the authors are in the graph. Some other interesting trends are the relationship between closeness centrality, number of citations, and number of publications. As the closeness centrality increases, the number of citations and publications responds with a slightly linear pattern. This could indicate that those well-known authors who publish much work would likely have more connections with other authors. Overall, it is evident that the trendlines mostly occur for the faculties, while the students' data mostly remain in the lower left corner, which makes sense as most students do not have much published work or reputation for creating too many connections with other authors.

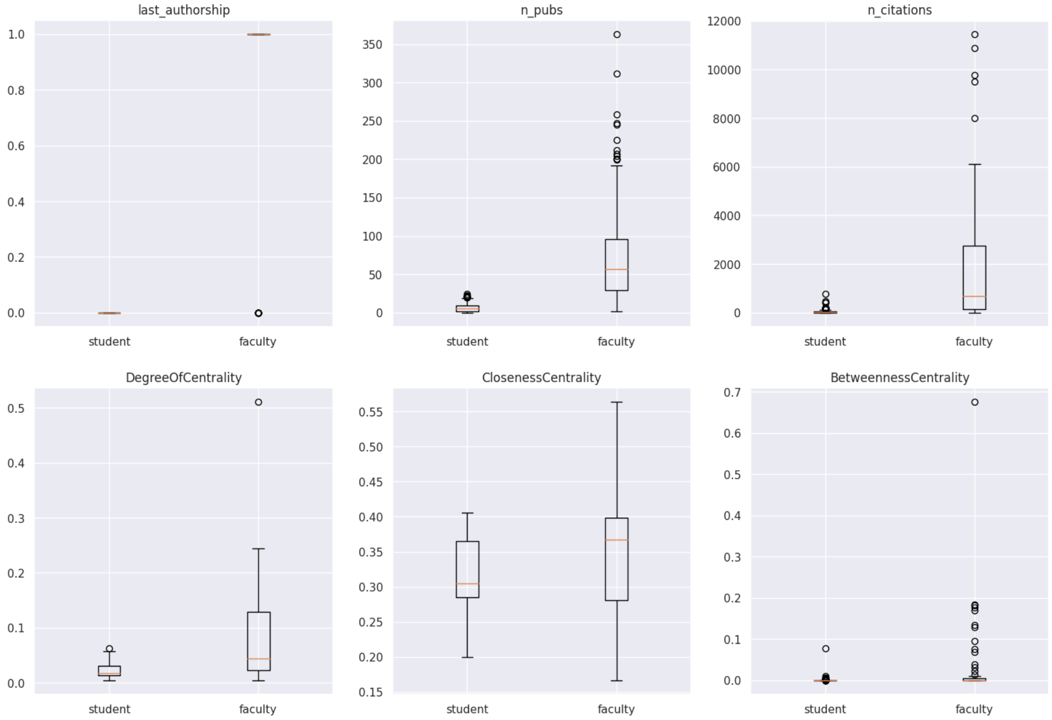


Figure 3.15: Scraped data - features representation

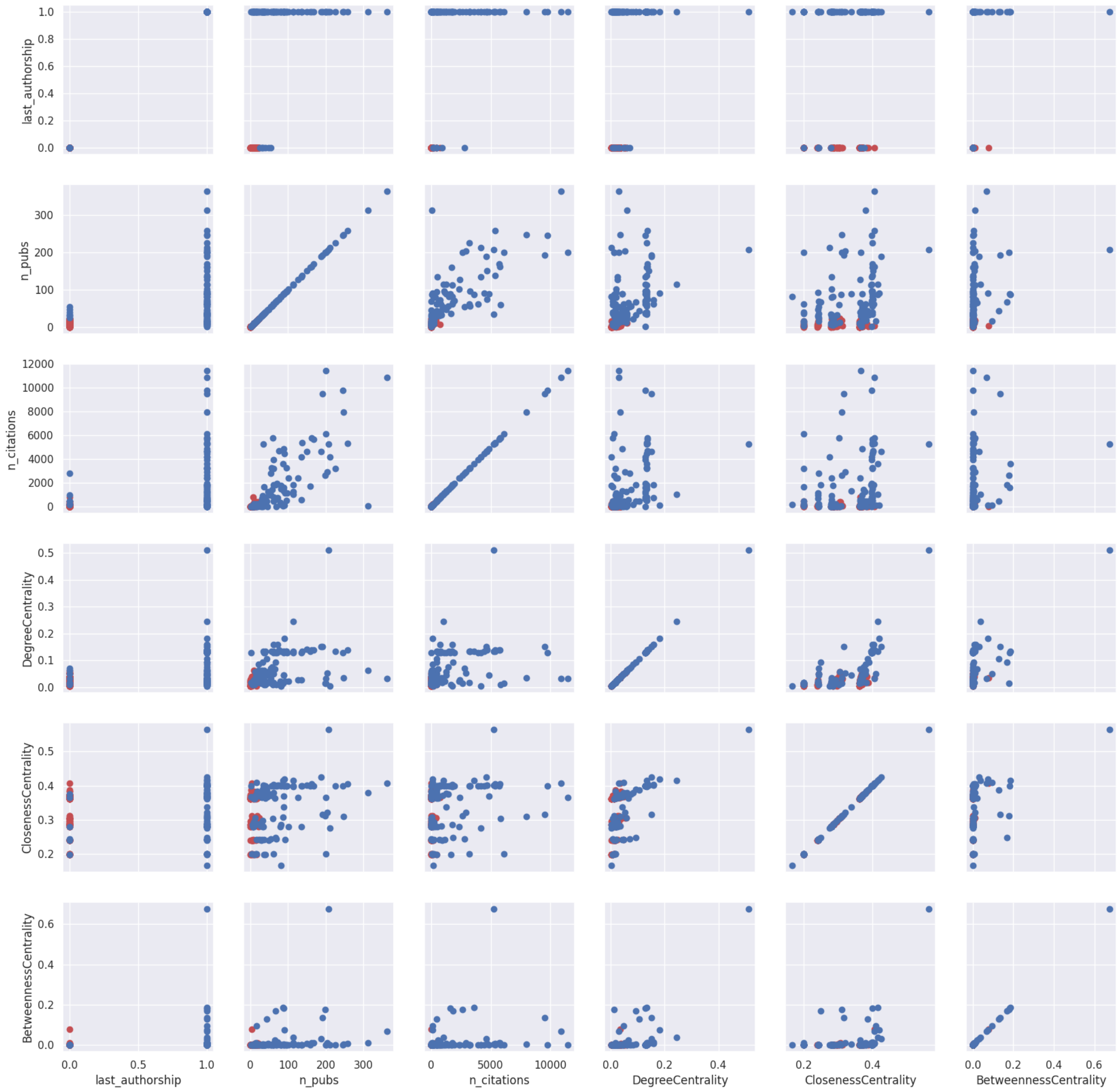


Figure 3.16: Scraped data - cross-features representation

As we want to ensure the professor-student relationship is accurate, we are only performing the experiments with whom we have close contact as referenced in the IRB. All the professors we have disguised agreed to allow us to use their names in the research. This would ensure professional courtesy of using their names. Additionally, since we are mostly limited to working with the professors that our research group is close with, we ended up with a subset of the scraped data, which has precisely 40 subjects.

A sub-network from the original social network was extracted to better demonstrate the differences, resulting in the graph in Figure 3.17. Similar to Figure 3.14, the professors are represented in blues and the students are represented in reds. The diagram illustrates that either the professor and student wrote a paper together at some point or that the students appeared on their respective professor's homepage. It is also interesting to note that some students may have multiple connections with different professors, indicating past collaborations. The students also tend to work with one another under the supervision of the same professor.

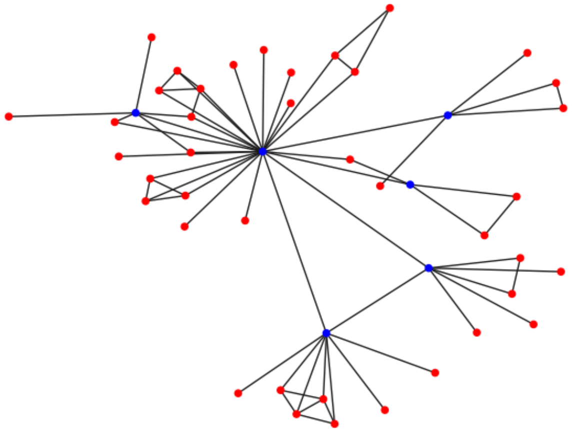


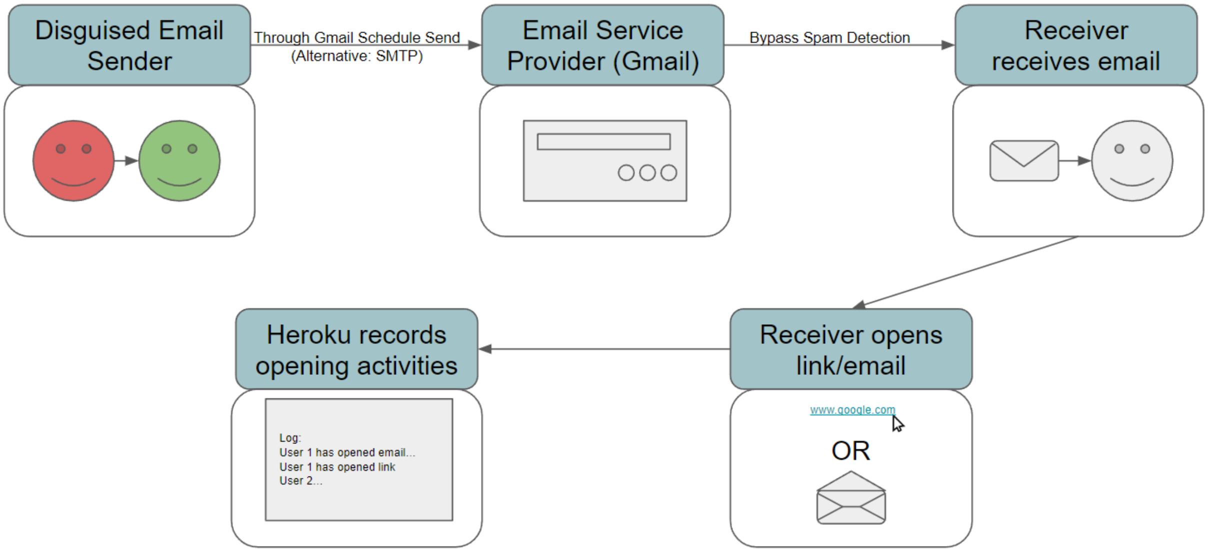
Figure 3.17: Extracted social network from scraped data

The relationship between professors and students can also be determined based on the occurrences of the authorship in the collaborated papers as well as the connections in the graph. If a student tends to have a specific professor as the last author in his papers, then it is very likely that the professor is that student's major professor. Before performing the experiment, the subjects' relationships would be verified by checking the professors' homepage. This avoids the incident where we impersonate a student instead of a professor, keeping the experimental process consistent and in the same dimension, maintaining the qualitative factors within control.

Chapter 4 - Experiment

4.1 Experiment overview:

To perform the experiment, we would use AutoGPT to construct the email scripts based on the subjects' research interests. AutoGPT is an experimental open-source AI-assistant driven by gpt-3.5-turbo [24]. Since we only have a limited number of subjects, and knowing that the emails might be flooded in receivers' emails over the day, we used Gmail's schedule send to send off the emails at a set time, depending on the subjects' time zones. According to the studies, people tend to open their email inboxes around 9 - 11 a.m. in their local time, peaking at 10 a.m. [10-11]. Therefore, we would send out the emails at around 10:00 a.m. in the subject’s respective timezone. This creates the fairness of the received time while ensuring the highest possibility for people to notice the emails in the inboxes. We would only target 40 subjects for the experiment, therefore should not raise any flags of spam detection on the email provider side, accounting that sending the emails from the same IP is part of the concern. Figure 4.1 illustrates the general process of the experiment.


Disguised Email Sender
Through Gmail Schedule Send
(Alternative: SMTP)
Email Service Provider (Gmail)
Bypass Spam Detection
Receiver receives email
Receiver opens link/email
www.google.com
OR
Heroku records opening activities

```
Log:
User 1 has opened email...
User 1 has opened link
User 2...
```

Figure 4.1: General process of the experiment

This real-world experiment would primarily use Gmail accounts to perform the sending process since it accounts for 36.5% of email opens globally in 2021, ranking as the most common email service provider people use [23]. Since there are 6 professors who agreed to participate in the research, we created 6 Gmail accounts that are similar to those professors' legitimate institutional accounts, differing by a few letters. The goal for the forged email addresses is to make them look as promising as if the professors were using their personal Gmail accounts when sending the spoofed emails.

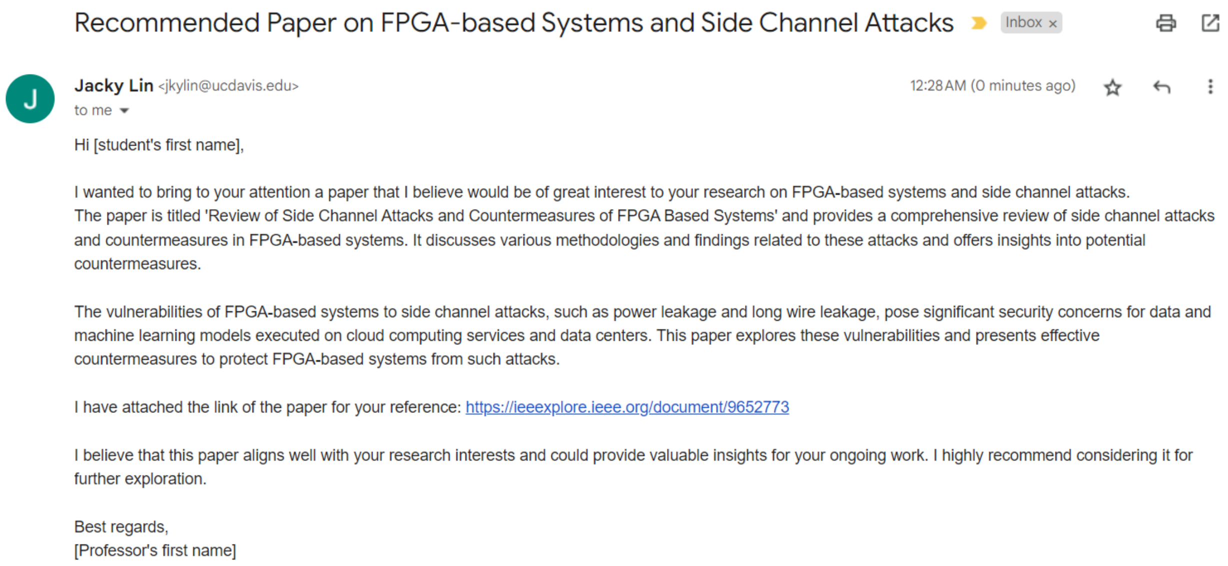
For the dimensionality of the experiment, we only plan to send emails from professors to their students instead of professor-to-professor or student-to-professor. This is because we do not have enough collaborated professors to support a legitimate comparison. The data might appear to be too biased with only 6 professors in the same engineering department, whereas the quantity of students is 40.

4.2 Email generation

The creation of email bodies relies on the utilization of AutoGPT, a prompt-based AI assistant driven by gpt-3.5-turbo as its fundamental LLM. We may pass it with a prompt that asks it to generate emails based on a specific student's interest coming from his professor. It could perform a Google search for the subjects and analyze the subjects' research interests by browsing research databases such as ResearchGate, IEEE, and Google Scholar regarding the subjects. Those search results would then be analyzed to find the subjects' recent research interests and use those interests to search for recent papers related to them. After finding the related paper, AutoGPT will analyze and summarize the paper, then use that information to write an email draft disguised as the subjects' advisor or major professor, intending to send the email to the students, specifically recommending the students to read the paper. Additionally, the email title is also automatically generated by LLM based on the topic. The tone of the email was tuned to be brief and casual. However, the LLM still retains some politeness knowing that it is assigned to construct an email. Since we are uncertain how other professors communicate with their students, we decided it might be better to leave it as is. Phishing emails will be sent out for a duration of 2 weeks, then the effectiveness will be assessed afterward.

Overall, the email bodies should be designed to fit each subject's interests and written on behalf of their major professors, introducing them to a paper and guiding them to click on the embedded link. The link is a redirection link that will take them to the actual paper after logging their click-through activity. Creating such phishing emails can prevent the email provider from flagging them as spam, ensuring the delivery of the emails. All the papers we suggested for students to read should be free for institutional access to avoid the student having to pay to read them.

Although we utilized the LLM model to design unique emails for each subject, a noticeable pattern in the generated emails still persists. As shown in Figure 4.2, it usually starts with a greeting, introducing the paper's title to the student. The remaining content would be providing a brief summary of the contents, persuading the students that it might be a good paper for them to read, and luring them to click on the embedded link that is supposedly going to redirect them to the source of the paper. In reality, the link would first redirect to a website that we host for tracking their click-throughs, then redirect them to the actual paper. As a result, it may be possible to trick the students into thinking that the professors may just be using their personal email accounts to recommend a paper, especially since the embedded link eventually leads to an actual paper.



Recommended Paper on FPGA-based Systems and Side Channel Attacks

Jacky Lin <jkylin@ucdavis.edu>

12:28 AM (0 minutes ago)

Hi [student's first name],

I wanted to bring to your attention a paper that I believe would be of great interest to your research on FPGA-based systems and side channel attacks. The paper is titled 'Review of Side Channel Attacks and Countermeasures of FPGA Based Systems' and provides a comprehensive review of side channel attacks and countermeasures in FPGA-based systems. It discusses various methodologies and findings related to these attacks and offers insights into potential countermeasures.

The vulnerabilities of FPGA-based systems to side channel attacks, such as power leakage and long wire leakage, pose significant security concerns for data and machine learning models executed on cloud computing services and data centers. This paper explores these vulnerabilities and presents effective countermeasures to protect FPGA-based systems from such attacks.

I have attached the link of the paper for your reference: https://ieeexplore.ieee.org/document/9652773

I believe that this paper aligns well with your research interests and could provide valuable insights for your ongoing work. I highly recommend considering it for further exploration.

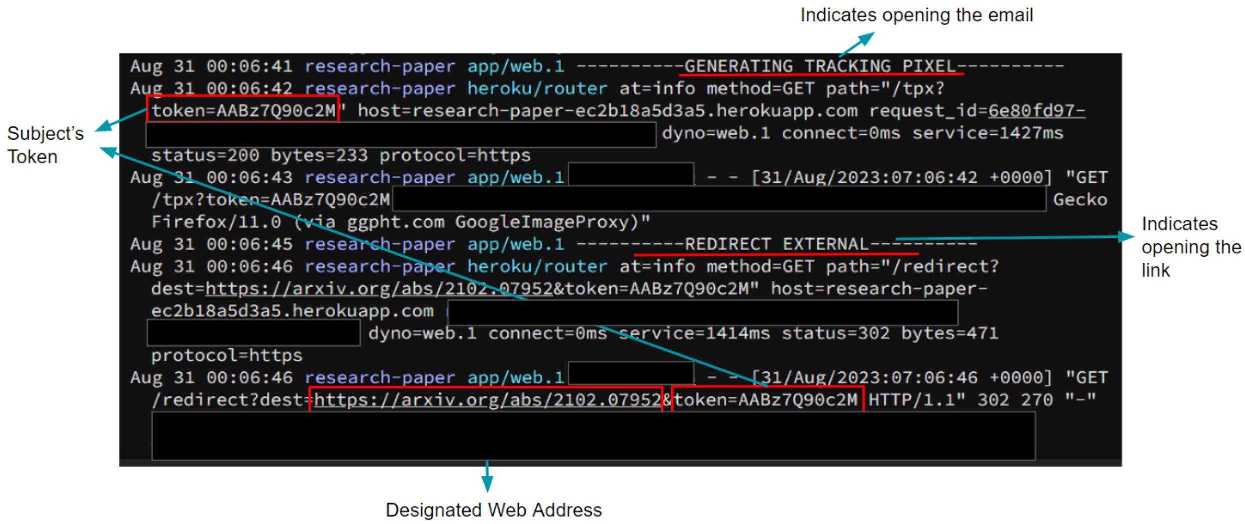
Best regards,
[Professor's first name]

Figure 4.2: LLM-generated sample email

4.3 Email tracking with Heroku

The email bodies will contain a one-pixel image block and a benign phishing link that we control using Heroku. Heroku is a platform that helps to operate applications in the cloud, allowing us to host a website that can log the subjects' click-throughs [25]. When the subject opens the email, the one-pixel image will be detected by Heroku, logging that the user has opened the email. When the subject clicks on the embedded link within the email, they would be directed to our hosted website on Heroku, log the click-through, and then be redirected to the legitimate webpage mentioned in the email, which is often a paper or a conference page depending on the subjects' research interests. Essentially, we would be logging two activities: whether the subject has opened the email and whether the subject has clicked on the embedded link in the email. Whether the user opens the email or links on a laptop, a personal computer, or a mobile device, Heroku would be able to track the accesses of the email and the link.

Each subject would be assigned a token related to him. As shown in Figure 4.3, Heroku's Papertrail add-on presents the access log that indicates which subject opens the email and clicks on the link at what time. The IP addresses are hidden to enforce security and user privacy. The only data that we keep would be the subjects' click-through events and the access time, omitting any other information shown in the log. Additionally, Heroku is also connected to Google Drive. A Python script has been set so that whenever Heroku detects activities of opening an email or the phishing link, not only will it log the response through its Papertrail add-ons but also clean the log, save the access time and token, and indicate whether it is an email or link that has been opened. Afterward, all the information would be stored in a Python dictionary, then converted to a pickle file and saved in Google Drive for easier reference and readability.



```
Aug 31 00:06:41 research-paper app/web.1 ----------GENERATING TRACKING PIXEL----------
Aug 31 00:06:42 research-paper heroku/router at=info method=GET path="/tpx?
  token=AABz7Q90c2M" host=research-paper-ec2b18a5d3a5.herokuapp.com request_id=6e80fd97-
  [redacted] dyno=web.1 connect=0ms service=1427ms
  status=200 bytes=233 protocol=https
Aug 31 00:06:43 research-paper app/web.1 [redacted] - - [31/Aug/2023:07:06:42 +0000] "GET
  /tpx?token=AABz7Q90c2M [redacted] Gecko
  Firefox/11.0 (via ggpht.com GoogleImageProxy)"
Aug 31 00:06:45 research-paper app/web.1 ----------REDIRECT EXTERNAL----------
Aug 31 00:06:46 research-paper heroku/router at=info method=GET path="/redirect?
  dest=https://arxiv.org/abs/2102.07952&token=AABz7Q90c2M" host=research-paper-
  ec2b18a5d3a5.herokuapp.com [redacted]
  [redacted] dyno=web.1 connect=0ms service=1414ms status=302 bytes=471
  protocol=https
Aug 31 00:06:46 research-paper app/web.1 [redacted] - - [31/Aug/2023:07:06:46 +0000] "GET
  /redirect?dest=https://arxiv.org/abs/2102.07952&token=AABz7Q90c2M HTTP/1.1" 302 270 "-"
  [redacted]
```

Indicates opening the email

Subject's Token

Indicates opening the link

Designated Web Address

Figure 4.3: Heroku tracking logs

4.4 Email Spoofing

As mentioned in Chapter 3, there would be 40 subjects in the experiment. The phishing emails would be sent out to these people without prior consent to observe the subject's natural behavior. However, we debriefed the subjects and acquired their consent for the continued use of their data after the experiments. We have gotten approval from the IRB for every step taken in the experiment, ensuring ethical practices are followed. Since the professors that we collaborated with are mostly faculties from either CS or engineering departments, their students, which are also the subjects, are mostly CS and engineering students as well. It is under the assumption that these groups of people should be more aware of the phishing attack compared to the general public. Therefore, it is expected that their biting chance for the bait may not be exceptionally high. Additionally, since we collected the authors solely based on publicly available information, the students should have published at least a paper or showed up on their professors' homepages for us to verify their roles and social connections.

The subjects are selected based on the constructed social networks in Figure 3.17. The disguised sender and the receiver would have a distance of 1-2 nodes in the network graph, representing the sender's close connection with the receiver. In this case, the sender would be a professor who advises a group of students and has assisted on at least one paper. In most cases, the receiver should express some research interest that causes them to fall for the phishing attack sent by the disguised author. Since the sender we impersonate holds a professor title while the receiver is a student, we do expect an increase in the likelihood for the receiver to fall for the bait compared to completely random targets.

The email-sending process was initially set up to use a Python script that enables the Simple Mail Transfer Protocol (SMTP) library for mass-sending emails. However, recent changes with the Gmail policies and SMTP library for Python require authentication with a mobile device. Furthermore, a limited use of authentication under the same mobile number is enforced, so the attacker would need to verify the emails with different mobile devices if he wants to send them using the SMTP library. Additionally, since we want to send out the emails at the desired time to accommodate the timezone difference for different email recipients and SMTP does not seem to support this process without leaving the computer hanging to run the script, we decided to simply use Gmail's schedule send function [26]. The process would simply be copying and pasting the generated email bodies to Gmail to set up a schedule send that would first send the emails to Google's email server, and then have the emails sent off to the recipients at the designated time.

4.5 Wrapping up the experiment

The experiment lasted two weeks after the emails were sent out. At the conclusion of the experiment, subjects who opened the benign phishing email and subsequently clicked on the embedded link will receive a debrief form, revealing the details of the study. Their inclusion without prior consent would also be explained in the debrief form. Following the debrief, the subjects' consent would be acquired for continued utilization of their experimental data. Subjects are also allowed to withdraw from the experiment at any point before the research is concluded. After collecting all the necessary consents, the results will be analyzed and reported.

Consent forms were sent out to the subjects via emails. If we do not receive an affirmative from the subjects, the inclusion of their data will not contribute to the final statistics. The experimental procedures would be transparent upon debriefing, without giving the technical details.

As for the payload phase of phishing, since we do not intend to perform any dangerous or malicious acts, we would only be tracking whether the subjects fall for the phishing bait or not using Heroku. Once the participants' responses are tracked through our server over the course of 2 weeks, the data will be finalized and the research may be concluded.

During the research process, all the data related to the experiment will be kept securely on drives that are only accessible by the members who directly work on this research. Additionally, none of the personal information or identification that could reveal the subjects' association with the data will be shown in this paper. At the end of the research, all subjects' information and their data would be deleted to enforce the protection of subjects' privacy. The entire research and experiment process is followed strictly through the IRB guidelines.

Chapter 5 - Results

5.1 Experiment results

The results of the experiment turned out to be quite decent. In Table 5.1, we can see that almost everyone we sent the emails has opened the email, reaching a 30/31 = 96.8% rate. Furthermore, about 15/31 = 48.4% of people have clicked on the embedded link. The results have shown quite a decent number of people falling for the phishing attack. This means that for every 31 people, there would be 15 people who fell for the bait. Assuming there are 2 million student researchers in the world, almost 970,000 people could possibly be phished, which is an enormous amount, proving the necessity for the public to be aware of such an attack method that is built upon publicly accessible information and social networks.

	Subject Count
Total Subjects	40
Forfeited (No Subject Approval - Does Not Contribute to Overall Statistics)	9
Remaining Subjects	31
Opened Email	30
Opened Email & Clicked the Link	15
Replied to Spoofed Email	11

Table 5.1: Experiment results

Furthermore, it is interesting to note that 11 subjects actually responded to the phishing email that we sent, stating that they would read the paper and get back to the disguised sender. This proves that the emails generated by LLM were trustworthy enough at first sight for the subjects to click on the link and reply to us, indicating a sincere belief from the students that the spoofed emails are legitimate.

Following our IRB guideline, we would debrief the experiment to the subjects if they clicked on the link. Although it is likely that they never found out about the experiment because the link we sent is merely a redirection link to a legitimate paper source, we still debriefed them about the research so that they can be more aware of such phishing attacks. Upon debriefing, we also attempted to ask for their consent for the use of their results. Since not everyone is willing to allow us to use their results, we had to omit a few subjects' results in the end.

In case there are people who discovered that the phishing emai is an experiment, we would reveal the research to them by showing them the debrief notice, explaining to them that this experiment is solely intended for research purposes. This is done by having a Heroku homepage in which if they only enter the Heroku address without the token, they may access that debrief notice on the homepage.

Figure 4.4 represents the result of the experiment in the form of a social network. In the figure, blue represents the professors, yellow represents those who have opened the email but did not click on the link, red indicates subjects who opened the email and subsequently clicked on the link, green represents those who did not open the email nor click on the link, and gray represents those who did not want to share their results. We can observe that there would be at least a couple of people under each professor who fell for the bait. Note that the subjects are students across different institutions in the United States, ruling out the possibility that an area in the U.S. is more likely to fall for the bait. The social network proves the high possibility for an individual to be phished, regardless of demographic location. However, it might not be true if the subject pool is increased to include thousands of people.

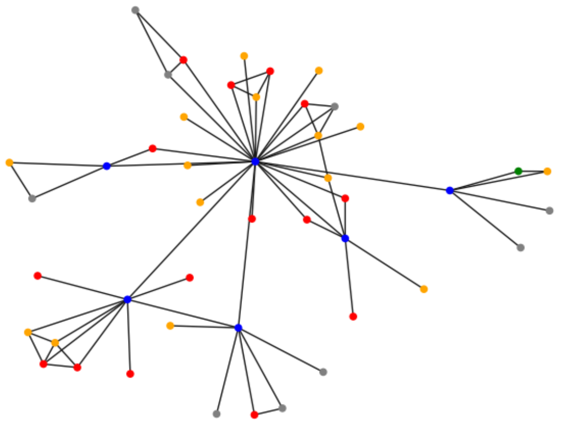


Figure 4.4: Resulted social network

Additionally, it is important to note that most of the subjects are students who major in computer science or engineering-related fields, which is expected that they would be more sensitive to phishing events. However, even these talented individuals would fall for the phishing bait, further proving the power and potential threats of publicly available information and social networks.

The results proved the feasibility of using LLM or GPT to generate human-like messages. As it is also possible to adjust the tone and style of the language that LLM uses, it may be even more human-like if the attackers know the victims well enough. Fortunately, noticing this danger in advance may be useful for discovering solutions early on as well before it is too late, especially with the current bloom of machine learning algorithms and LLM technology.

5.2 Access time of the emails and links

On top of the result, we also extracted the access time of email and link opening from Heroku's log. Our results indicate that the majority of the subjects opened the email and the link almost immediately or within one day upon receiving them, while some others opened them after a couple of days. Most subjects who would click on the link would do it within a day after opening the email, indicating the trustworthiness of the generated emails. However, it should also be noted that even though we include the access times below, all that really matters is that the subject will eventually click on the link, assuming the attacker would perform malicious behavior as soon as the link is accessed.

	# of subject opened email when received	# of subject opened link when received email
t < 5 mins	10	5
5 ≤ t < 30 mins	1	2
30 mins ≤ t < 2 hrs	0	4
2 hrs ≤ t < 1 day	9	1
1 day ≤ t < 2 days	4	0
2 ≤ t < 5 days	2	3
t ≥ 5 days	4	0
Total	30	15

Table 5.2 Email access time

Chapter 6 - Mitigations

6.1 Difficulties

Specifically for our experiments, it may be difficult to flag the phishing email with the use of LLM involved in the attack. Unless there is a consistent method that checks whether the text is generated by LLM or not, the attacker may fine-tune the LLM with custom data to target specific subjects, designing the email personally based on his relationship network. Further, since every generated email has different contents, it may be difficult for the email providers to mark it as a spam message.

It is possible that Microsoft may have encrypted tokens within the generated text for defensive mechanisms to realize that it is an AI-generated message. Even then, having a simple Python script that replaces a few words or punctuations using NLP may still be possible to break the encryption like how it would be done similarly to the machine learning adversarial attacks (e.g. one-pixel attack).

6.2 Defensive strategies

In this section, we will discuss the possible mitigation strategies from 3 perspectives: the email provider, the organization, and the user. For this research, we have mainly targeted researchers who study in the institutions, so the majority of their emails used to communicate should end with "<institution>.edu". Therefore, verifying that the email is from an educational institution and not a random Gmail account may be a valid solution. In which the victims can then verify the legitimacy of the emails by looking at the sender's email address. This may also be applied to any organization since they usually have their own email addresses, making it possible to create a filter to eliminate any external emails. Additionally, adoptions of Sender Policy Framework (SPF), DomainKeys Identified Mail (DKIM), and Domain-based Message Authentication Reporting & Conformance (DMARC) are methods to consider for filtering out unwanted domains or content for organization or institutional uses [27-31]. Furthermore, although there are already some institutions and organizations that flag external emails, the UI design might not be obvious enough. Some might only have a small flair that marks the external emails rather than clearly displaying a banner above the email.

As for the general public, identifying phishing emails may be challenging if the contents are skillfully tailored to mimic the typical communication emails people regularly use. This mimicry blurs the distinction between a legitimate email and a phishing attempt, especially when the email address exhibits only a minor variation, such as a single letter or number difference. However, a few strategies may still be relevant to the user end. One practical approach is to cross-reference the usual email address for communication if a similar but different email address is encountered. Further confirming the email address with the usual sender may also be necessary. Another approach is to copy the embedded link address and paste it somewhere without directly accessing it to see if it appears to be a legitimate website. If it cannot be recognized by simply looking at the web address, pasting it to some malware-checking website to verify its legitimacy may also be a good option.

For both email providers and organizations, it may be possible to implement a filter that could be used to classify AI-generated emails, explicitly highlighting the tone of the email body. AI-generated messages tend to show politeness and longevity in generated emails by default. Furthermore, assuming the attackers can only access the social network of the subjects but not the email bodies that withhold the communications between the victims, the emails may generally appear to follow a pattern or a tone even if it's targeted toward a selected group of audience (researchers, home sellers, casual conversation, etc.). In general, if the attacker does not modify the tone of the LLM, most generated emails appear in a formal and polite tone. Therefore, it may be possible to set up a system that filters out emails that follow a similar pattern or tone. One additional method may be hosting tutorials or seminars regarding email security.

Additionally, Hu and Wang have shown that security indicators may positively reduce effective phishing events [1]. Security indicators for unauthenticated/unverified emails may be an important factor to take note of. Especially having a decent UI design that clearly warns the users that the external emails may be phishing intent, asking the user to verify the email addresses or confirming it with the sender through a known email address may be a valid action to reduce the biting chance of the phishing emails.

Whenever an email is entered on Google applications such as Google Sheets, Google Docs, etc., the profile information regarding that email pops up. Disabling the auto-recognition of emails and making the profiles not publicly accessible may be a potential act performed by the email providers. It may be possible to show profile pictures and information only if communication between the sender and receiver is established. In other words, both the sender and receiver should have sent at least one email to one another to confirm that they have a valid connection. This can ensure that the attacker will not be able to verify the legitimacy of the email discovered online.

Chapter 7 - Future Directions

To extend our experiment further, we could consider other qualitative factors. For instance, we could have another phase of the experiment that asks for people's consent prior to the experiment to observe their click-through rates of the phishing link. We may then compare the results with our current experiment to observe the differences. This may indicate the effectiveness of the phishing attack under the assumption that the subject has sufficient knowledge about the attack to be more aware of the phishing attack. This also opens up the possibility of inviting more people, even if we are unfamiliar with them, to participate in our research. This was not possible due to the restriction of the IRB, enforcing us to only collaborate with professors whom we are close with, limiting the subjects mostly to the same departments or a somewhat more connected social circle.

Currently, most of our subjects are students who major in the engineering field. If there is a chance for us to increase the subject pool size, it might also be possible to consider analyzing demographic factors, educational levels, associated departments, age, and so on in our analysis.

Defensive mechanisms may be considered as a follow-up research topic against the studied spear phishing technique. Since we have all the models required for the attack, it may be easier to interfere with the attack process. As we mentioned in the defensive strategy section, things could be done on the email provider, organization, and user sides to prevent possible phishing attacks. Performing our experiment on email providers with and without methods like SPF, DKIM, or DMARC may be a valid test for the email service provider side of testing. Organizations may usually have their specific email addresses, making it possible to create different filters for incoming emails. Performing the attack with and without a clear UI flag of external emails may test the trustworthiness of our content on the organizational side as well.

Purposely training a group of subjects with enough phishing knowledge and another group without may test the effectiveness of user prevention. However, negotiating with other organizations to collaborate on the research may be challenging.

For the attack improvements, we may fine-tune the tone and content for the email bodies generated by LLM using custom conversation data. Additionally, we may also test out different LLMs to see which fits best for generating emails. Currently, we are using AutoGPT, which is driven by gpt-3.5-turbo, while numerous other LLMs out there might perform better, such as BERT, T5, or any other trained models on Hugging Face. Hugging Face is an AI company that offers machine learning tools for building applications, while also having a community for people to share trained machine learning models [32]. In fact, with sufficient data, we may even fine-tune the LLM models ourselves to make the emails more believable.

Chapter 8 - Related Work

As this project involves many different areas of research, a brief literature survey was conducted on phishing, privacy, and social engineering attacks, all relevant to profiling a target based on available online information.

Researchers have discussed the impact of public information that may be misused for phishing attacks, explicitly stating that information published online may be a dangerous cue for people to be targeted as phishing victims [13, 16]. In our research, we are exploring the possibility of using publicly available information to construct a social network for our implementation of the attack.

Some papers introduced the concept of social attacks in various forms, including spear phishing, and we are extending their concepts into real-world practice. In fact, they have even performed phishing attacks using a social network constructed based on the information they gathered online [12-13]. However, their relationship network is built upon social media groups and students in the same university. Their senders and targets hold the same position, and their disguised targets are usually friends of one another. On the other hand, our relationship network is built upon professors and their students, specifically focusing on the aspect of disguising themselves as professors to send emails to students. This creates an educational hierarchy difference between the sender and the receiver, which we value in our research.

Heartfield performed social engineering attacks in situations where the subjects were aware of the experiment, being more sensitive than they would have normally been [17]. On the other hand, our attack is performed with the impression that the subject only knows about the experiment once we debrief the research to them. We have worked closely with our IRB to ensure the safety and ethical rules of the experiment are accounted.

There are also research that use a recurrent neural network that learns to tweet phishing posts on Twitter, targeting specific users depending on their life interests [14-15]. We have developed a similar pipeline but rather focused on the subjects' research interests since our subjects are mostly students instead of a random group of people on Twitter. Other research relies on the use of NLP for generating email bodies, which appear to look like spam messages [14]. However, they do prove a point that even though the click rate of the phishing link may not be high, the few successes may yield a high return on the investment of the setup [15]. With the advancement in LLM, we implemented LLM as the primary tool for generating emails. While NLP is decent at processing the immediate context of text, LLM is trained on massive amounts of data, allowing it to analyze information and fuse it with the generated text while having the flexibility to generate languages in various contexts.

Hu and Wang have extensively examined the spoofing aspect of phishing attacks by conducting detailed end-to-end measurements of spoofed messages and examining user reactions through real-world experiments [1]. Although their work involves experimenting with different measures such as authentication, security indicators, email content, and UI, they did not have any intention to consider impersonation in the metrics. In other words, while they focused on how different security measures can impact the phishing rate, we focus on the impact of social networks involved in the phishing process.

Chapter 9 - Conclusion

Although the experiment only lasted two weeks, the scraping scripts and machine learning models took over a year to set up. Furthermore, the preparation process took another half a year due to the incredible effort to acquire approval from the IRB and the university's IT. Persuading subjects to provide consent for the research took a couple of months as well. However, we did end up getting enough approvals and consent to finish this paper.

When collecting emails during the data collection process, we noticed that some email accounts may be automatically identified by Gmail, showing people's profile pictures. This could be a way for the attacker to confirm that the email they discovered online is indeed legitimate, increasing the risk of a successful attack.

Additionally, it is important to note that the experiments took place in the summer so we also have to account for the likelihood of students opening and responding to emails over the summer. As there is usually not much happening for a student who does not take summer courses or work on an internship, it may be very likely for students to ignore the incoming emails, especially considering the fact that we targeted solely institutional emails rather than personal emails for the experiment. It is likely that the subjects never checked their institutional emails or had a long period of checking them.

The subjects included in this research are primarily graduate students (Master's and Ph.D.) who major in CS or engineering departments. It should be a factor to consider, knowing that the subjects may be more familiar with the concept of phishing attacks. With this in mind, if the experiment is applied to the general public, the phishing chance may be even higher than the result shown in this research. This further proves the effectiveness of our experiment as the attention on social media increases.

In fact, it is interesting to note that when asking the subjects to fill out the consent form, people are so paranoid about phishing that they even think the link for the consent form is bait. When our research group contacted a professor to ask for his students to fill out the consent form for our research, he was under the impression that the consent form itself could possibly be a phishing link and we are performing the experiment on him.

The research is meant to discover more angles for possible incoming attacks with the future of LLM vigorously evolving. It is not a paper meant to perform any actual malicious acts, therefore we did not perform any action asking for any subject's personal data besides what is already publicly available. Overall, this research demonstrates the effectiveness of phishing emails constructed based on people's publicly available information and their social networks.

REFERENCES

[1] Hu, H., & Wang, G. (1970, January 1). *{end-to-end} measurements of email spoofing attacks*. USENIX. https://www.usenix.org/conference/usenixsecurity18/presentation/hu

[2] Finn, P., & Jakobsson, M. (2007). Designing ethical phishing experiments. *IEEE Technology and Society Magazine*, *26*(1), 46–58. https://doi.org/10.1109/mtas.2007.335565

[3] *Phoneypot: Data-driven understanding of telephony threats*. NDSS Symposium. (2023, July 25). https://www.ndss-symposium.org/ndss2015/ndss-2015-programme/phoneypot-data-driven-understanding-telephony-threats/

[4] Griffiths, C. (2023, August 18). *The latest phishing statistics (updated August 2023): Aag it support*. AAG IT Services. https://aag-it.com/the-latest-phishing-statistics/

[5] Trend Micro. (2023, May 31). *Worldwide 2022 email phishing statistics and examples*. Trend Micro. https://www.trendmicro.com/en_us/ciso/23/e/worldwide-email-phishing-stats-examples-2023.html

[6] Main, K. (2023, July 17). *Phishing statistics by state in 2023*. Forbes. https://www.forbes.com/advisor/business/phishing-statistics/#:~:text=Phishing%20statistics%20show%20that%20in,widely%20varying%20amounts%20of%20losses.

[7] Woods, E. (2022, November 4). *The most common examples of phishing emails*. usecure Blog. https://blog.usecure.io/the-most-common-examples-of-a-phishing-email#:~:text= and%20avoid...-,1.,never%20even%20ordered%20or%20received.

[8] Bouziane, A. (2020, June 3). *Who should be the last author on a research paper?*. Editage Insights. https://www.editage.com/insights/who-should-be-the-last-author-on-a-research-paper#:~:text=The%20last%20author%20is%20usually,also%20often%20the%20corresponding%20author.

[9] Pain, E. (2021, May 6). How to navigate authorship of Scientific Manuscripts - Science | AAAS. https://www.science.org/content/article/how-navigate-authorship-scientific-manuscripts

[10] Ellering, N. (2023, February 21). *What 14 studies say about the best time to send email*. CoSchedule Blog. https://coschedule.com/blog/best-time-to-send-email

[11] Huang, K. (2023, July 17). *When is the best time to send an email? [Data & charts]*. Litmus. https://www.litmus.com/blog/whats-the-best-time-to-send-email-we-analyzed-billions-of-email-opens-to-find-out

[12] Salahdine, F., & Kaabouch, N. (2019). Social Engineering Attacks: A survey. *Future Internet*, *11*(4), 89. https://doi.org/10.3390/fi11040089

[13] Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94–100. https://doi.org/10.1145/1290958.1290968

[14] Seymour, J., & Tully, P. (n.d.). Weaponizing Data Science for Social Engineering: Automated E2E spear phishing on Twitter. https://www.blackhat.com/docs/us-16/materials/

us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf

[15] Seymour, J., & Tully, P. (2018, February 14). *Generative models for spear phishing posts on social media*. arXiv.org. https://arxiv.org/abs/1802.05196

[16] Moore, T. W., & Clayton, R. (2012, March 15). *The impact of public information on phishing attack and Defense*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2020325

[17] Heartfield, R., Loukas, G., & Gan, D. (2016). You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks. *IEEE Access*, *4*, 6910–6928. https://doi.org/10.1109/access.2016.2616285

[18] Cor&icirc;ci, M. (2023, March 22). *The good, the bad, and the ugliness of llms*. LinkedIn. https://www.linkedin.com/pulse/good-bad-ugliness-llms-marius-cor%C3%AEci/

[19] Madden, M., Fox, S., Smith, A., & Vitak, J. (2020, August 17). *Digital footprints*. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2007/12/16/digital-footprints/

[20] Violino, B. (2023, January 10). *Phishing attacks are increasing and getting more sophisticated. here's how to avoid them*. CNBC. https://www.cnbc.com/2023/01/07/phishing-attacks-are-increasing-and-getting-more-sophisticated.html

[21] Schneier, B. (2023, April 10). *LLMs and Phishing*. Schneier on security. https://www.schneier.com/blog/archives/2023/04/llms-and-phishing.html

[22] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

[23] *Email service providers: 8 most popular email providers*. Mailchimp. (n.d.). https://mailchimp.com/resources/most-used-email-service-providers/#:~:text=Sign%20up-,Gmail,email%20opens%20globally%20in%202021.

[24] *Significant-gravitas/auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous.* GitHub - Significant-Gravitas. (n.d.). https://github.com/Significant-Gravitas/Auto-GPT

[25] Heroku. (n.d.). *Heroku*. Cloud Application Platform. https://www.heroku.com/

[26] Google. (n.d.). *Schedule emails to send - computer - gmail help*. Google. https://support.google.com/mail/answer/9214606?hl=en&co=GENIE.Platform%3DDesktop

[27] Google. (n.d.). *Help prevent spoofing and spam with SPF*. Google Workspace Admin Help. https://support.google.com/a/answer/33786?hl=en#:~:text=SPF%20is%20a%20standard%20email,send%20email%20for%20your%20domain.

[28] *What is the meaning of the SPF email standard and how does it work?*. Advanced Email Security Solutions for Enterprises. (n.d.). https://www.agari.com/blog/what-is-spf#:~:text= What%20is%20SPF%20for%20Email,fraudsters%20to%20spoof%20sender%20information.

[29] Google. (n.d.). *Help prevent spoofing and spam with dkim*. Google Workspace Admin Help. https://support.google.com/a/answer/174124?hl=en#zippy=%2Chelps-prevent-spoofing

[30] *What is DKIM? - how it works, Definition & More: Proofpoint us*. Proofpoint. (2023, April 7). https://www.proofpoint.com/us/threat-reference/dkim

[31] *What is DMARC? how does dmarc work?*. Fortinet. (n.d.). https://www.fortinet.com/resources/cyberglossary/dmarc#:~:text=Domain%2Dbased%20Message%20Authentication%20Reporting,Policy%20Framework%20(SPF)%20protocols.

[32] *Hugging face – the AI community building the future.* Hugging Face. (n.d.). https://huggingface.co/