ENVIRONMENTAL FACTORS AFFECT SOCIAL ENGINEERING ATTACKS

by

Minglu Li

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer and Information Technology

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Ida Ngambeki, Chair

Department of Computer and Information Technology

Dr. Baijian Yang

Department of Computer and Information Technology

Dr. Yingjie Chen

Department of Computer and Graphics Technology

Approved by:

Dr. John Springer

ACKNOWLEDGMENTS

I received a lot of help from my committee members: Dr. Ida Ngambeki, Dr. Baijian Yang, and Dr. Yingjie Chen. I want to offer my sincere thankfulness for their help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Social engineering attacks can have serious consequences when it comes to information security. A social engineering attack aims at sensitive personal information by using personality weaknesses and using manipulation techniques. [1] Because the user is often seen as the weakest link, techniques like phishing, baiting, and vishing, and deception are used to glean important personal information successfully. This article will analyze the relationship between the environment and social engineering attacks. This data consists of 516 people taking a survey. When it comes to discovering the relationship, there are two parts of the analysis. One is a high-dimensional analysis using multiple algorithms to find a connection between the environment and people's behavior. The other uses a text analysis algorithm to study the pattern of survey questions, which can help discover why certain people have the same tendency in the same scenario. After combining these two, we might show how people have different reactions when dealing with social engineering attacks due to environmental factors.

1. INTRODUCTION

1.1 Problem Statement

Social engineering is a technology using humans as the weakest link to get personal information. Social engineers use lots of techniques to manipulate individuals to extract sensitive information from them.[2] Even though people think they can detect the attack, the truth is they often perform poorly when under attack. Often people are bad at recognizing lies. Our paper will design several scenarios to test people's behavior and collect certain personal information to find a connection between them.

There are various definitions and lots of different models of social engineering attacks. [1] There are direct attacks and indirect attacks in social engineering attacks. The direct attack can divide into bidirectional communication attacks and unidirectional communication attacks. Indirect attacks are often using 3rd party mediums. Email, telephone, and face-to-face communication are often used in bidirectional communication. Unidirectional communication often uses SMS, Email, and paper mail. Finally, indirect communication uses pamphlets, flash drives, and web pages to attack users.[3] Figure 1 shows the framework of social engineering attacks.

1.1.1 Research question

- What impact do environmental factors have on user performance under social engineering attacks?
- What might influence people's choices other than ecological factors when under social engineering attacks?
- Which environmental factors can influence participants performance against recognizing social engineering attacks.

To answer these questions above, this thesis collects survey results from 516 test-takers. This survey has 22 questions about what degree people think a particular scenario can be a social engineering attack. Some of the rest questions ask for personal information like age, gender, education level, annual income, etc. The others require users' experience of social engineering attacks, like how many times they ever fell victim to phishing. After collecting the dataset, the main job is to analyze it. This paper will use machine learning algorithms and text analysis algorithms to find the connection between environmental factors and people's behavior. Machine learning algorithms can help us analyze and display high-dimensional data in the fastest way, and text analysis algorithms can distribute to the reason behind the phenomenon.

1.2 Significance

To better understand this research's significance, we need to discuss the situation of social engineering attacks. The problem of information security involves everyone. However, leakages may still happen even without the help of high technology. This research mainly focuses on vishing, phishing, and ad fraud scenario, which use human weakness only. Under the circumstance of nowadays, everyone can be a victim of social engineering attacks. Whoever has a smartphone can get phone calls, emails, and can click on websites. Attackers can use these techniques to trick people for personal information. For example, Mark receives a call from someone claiming to work for his bank. The caller says that they have observed Mark's account's suspicious behavior and lists several purchases totaling over 2000 made at online sex shops. He claims that unless Mark proves his identity, he will be responsible for the charges. The caller asks him to provide his account number, date of birth, and social security number to verify his identity and remove his account bills. In this case, the caller has a high possibility of being an attacker.

In a successful attack, attackers must first gain people's trust, which means they need to gather information early and then plan the attack. Information gathering is essential if attackers want people to believe their identity. Attackers need to blend in with a credible environment. A website can quickly get people's trust in a fraud case using excuses like job openings, coupons, and current events.[4]

Even though people know social engineering attacks can have serious consequences, not many researchers have addressed causes and solutions. There are several types of research about what factors may affect people's reactions to social engineering attacks.

But based on different datasets, the conclusion from other research can be different.[5] and [6] have opposing opinions on whether gender is a significant determinant. But we can say users' characteristics can be a factor that influences the judgment of social engineering attacks. [5] Many factors can affect personal behavior, like gender, education level, internet and computer knowledge level, and security awareness. [5][7][8][9] Besides these factors, this thesis will also discuss whether the geological location, time, income level, and computer operating system influence people's judgment. If we want to get reliable survey results from the participants, it is necessary to use designed behavior to test their true thoughts. Because sometimes, people don't even know if they are under attack, which means the results can be misleading if we only ask for phishing, vishing, and fraud directly.

1.2.1 Social engineering approach and framework

In the paper of [10]Ashish Thapar, social engineering is classified as computer-based and human-based approach.

A social engineering attack is any act that influences a person to do things that may or may not harm his or her interest. It can be divided into multiple vectors e.g. SMiShing, Vishing, Phishing, and impersonation.[11] Phishing, Vishing, Spam email, popup window, and interesting software are technology-based approach. Pretexting/impersonation, dumpster diving, spying, acting as technical expert, support staff, hoaxing, and authoritative voice are non-technology-based approach. [10]

Vishing is voice phishing, which exploits victims' trust by using telephone service. " Vishing exploits the public's trust in landline telephone services, which have traditionally terminated in physical locations which are known to the telephone company, and associated with a bill-payer." [10] Most of the vishing attacks aims at tricking for personal and sensitive financial information. During the pandemic period, according to the security magazine, "The Federal Trade Commission says that since January 2020 until mid-April, they received

18,235 reports related to COVID-19, and people reported losing 13.44 million dollars to fraud." (p. 1) [12] Phishing often uses email to pretend as a legitimate business. "The letter usually contains a link to a fraudulent web page that looks legitimate with company logos and content and has a form that may request username, passwords, card number or pin details." [10]

Spam emails also can be called junk emails, "unsolicited usually commercial emails sent to a large number of recipients or posted in a large number of places". [13] Spam email maybe one of the most frequent social engineering attacks we face every day, since the spam filter cannot work as we expect.

A social engineer often collects information from OSINT, which is open-source intelligence. Through OSINT, engineers can gather personal information like phone number, name, gender, and address. One thing to be clarified is that most of the time, OSINT does not require high technology.

There are specific steps that social engineering attacks can follow. There are six core phases in the social engineering attack framework: namely attack formulation, information gathering, preparation, develop a relationship, exploit relationships, and debrief.[1] Communication, social engineer, target, medium, goal, compliance principles, and techniques are essential features of social engineering attacks. In our survey, we designed several questions to follow the steps of social engineering attacks. Steps are as follows: attack formulation, information gathering, preparation, develop a relationship, exploit relationship, and debrief. [1]

As [14] categorizes the phases of social engineering as research/information gathering, hook/establishing relationship, play/exploitation, and exit/attack execution. The research/information gathering phase will target the eligible victims, design a reasonable plan, and collect personal information from online source including social media.

The purposes of research are to gather victims' background information, find the target, and plan ahead of the identity. The phase of hook "aims to engage with the target, spin the story, build a level of intimacy, and take control of the interaction." (p. 9)[15] The phase of play aims to "extract information and keep things going long enough to do so: maintain charade, strengthen control of relationship, extract information". And the final phase – exit aims to “close interaction, without arousing suspicion: bring charade to natural end, provide target with reason to keep quiet, cover tracks”[15]

1.3 Scope

This study aims to analyze the impact of environmental factors on people’s response and awareness of online fraud. Studies have shown that environmental factors will cause people to have different security awareness behaviors on the Internet. Our research validates their conclusions based on previous studies and tries to find new relationships between behaviors and environmental factors. Regarding the actual influence of environmental factors on people’s online behavior, the public has a preconceived impression. However, after this research, we may be able to find that although some impression is correct, others may need to be violated. I hope this research can arouse more people’s attention to the cultivation of cyber security awareness.

1.4 Research design

According to the research that has examined how a user behaves towards security threats and vulnerabilities, knowledge level, whether the environment is home or work, whether he/she has been an online fraud victim before, and the career, these are all the factors that can influence the human behavior. In consequence, this research includes the following factors: age, sex, income, education level, whether work in IT related area, time spent on computer at home/work, and whether he/she received cyber security education. In this research, we want to discover that if there are environmental factors that influence human behavior. In order to answer the research question, we designed a survey to collect data from more than five hundred people, which have different background. Other than the environmental factors, the questionnaire also asks for their opinion about one likely scam scenario, and the experience about social engineering attack.

For research design, we designed questionnaire, collected data, cleaned data, analyzed the data, and analyzed the text. By using Qualtrics tool and website we create and distribute the questionnaire. After 500 people taking the survey, we get a 517-row and 141-column

CSV file, including the geo-location, duration and status. For basic analysis, we operate SPSS Statistics to find relationship between each column. Then we use machine learning algorithms to find a deeper connection and visualize the results for a better view. Text analysis we concentrate on question 18 to question 41, which you will find in the appendix. The procedure of data analysis will be discussed in Chapter 4.

1.5 Assumptions

- Taking into account the differences in gender, education level, income level and work field, we assume that people in different groups will have different safety awareness and performance.
- Cyber security related courses can raise people's online safety awareness.
- The structure of network fraud is actually traceable, and we can find a unified structure for analysis.
- People who working or studying in Information Technology are better at recognizing social engineering attacks.
- The ability of recognizing phishing, vishing and ad fraud can be trained.

1.5.1 Hypothesis

- Whether an individual ever had a virus on their machine cannotaffect SEA performance.
- Whether an individual ever had a ransomware on their machine cannot affect SEA performance
- Whether an individual's account ever been hacked cannot affect SEA performance
- Whether an individual ever shared credentials for any password protected account cannot affect SEA performance

- Whether an individual ever been a victim of identity theft cannot affect SEA performance
- Whether an individual ever lost money of any online or phone scam cannot affect SEA performance
- Whether an individual ever been taught any computer science, cybersecurity or related topics cannot affect SEA performance

1.6 Limitations

- Those under the age of 35 accounted for 60% of the test subjects, but those over the age of 35 may be more likely to be scammed by the Internet. Therefore, we should expand the amount of data over 35 years old.
- When we test whether everyone thinks a particular scene is a fraud, we can add a procedure to ask people which keyword makes them aware of fraud. This allows a more detailed understanding of the psychology of the tested users and the conditions under which they can be more vigilant.
- Because this questionnaire completely relies on the user's memory, some of the data may not be completely accurate. Specifically, there may be situations where the user was scammed before but did not realize it.

1.7 Delimitations

- This research will not over-test various ways of online fraud, but focus on three SEAs, including vishing, fishing, and ad fraud.
- This research will not provide suggestions to prevent SEA but provide environmental factors that may affect behavior.
- Rather than studying the structure of sea, we are more to provide a new research idea for future research directions, that is, some keywords that may cause different reactions from people.

- Because there are endless ways of online fraud, people's perception of the same kind of fraud may change within a few months. We can conduct this research again in the future, and there may be new discoveries.

2. LITERATURE REVIEW

2.1 Social engineering attacks definition

According to the report of FireEye about the security predictions 2021, spear phishing is still one of the most popular techniques during nation-state threat activity. “An increasing number of nation-state actors are focusing on intrusion techniques that don’t require any victim interaction, such as exploiting web-facing applications and password spraying.” **2021** “Cybercrime as a whole has increased by 600 hundred of percent since the beginning of the global pandemic.” (p. 2) , and it is believed that the amount of cybercrime will continue to grow in 2021, work from home makes the residential area more vulnerable. [16] There is study found that “from 2013 to 2017 the average cost of cyberattacks has increased by 62 percent” and cyberattacks may cost even more than natural disasters. (p. 1) [17] Social engineering is an essential part of the cybersecurity and often at low cost, which can be avoid by training. According to statistics, nearly 98 percent of cyber-attacks rely on social engineering. Any individual or group can use social engineering methods to obtain public or private information, and the legality varies according to actual operating conditions. [18] These numbers show the importance of training awareness of social engineering attacks. By understanding the techniques and method, it is worthy to train everyone about the cruciality of protecting personal information. Although there are endless ways to be frauded, maintaining a vigilant awareness is the easiest and most convenient way to reduce the probability of being defrauded. [19] The concept of the social engineer has been popular since the last century. Even though some people maybe not familiar with social engineering, they must have been victims of social engineering which can also be called pre-texting, blagging, and conning.

Users can use methods based on the principles of psychology, linguistics, behavior and sociology without knowing it, to obtain valuable information from the target at a lower cost without leaving easy Traces of being traced and forensic. Instead of breaking into a system, social engineering involves getting information from a person. No matter how well-protected the computer is, social engineers still can obtain personal information using psychological tricks. [20] Social engineering can have multiple definitions, depending on where you sit. But one thing to be sure of is that human is the weakest link. Social engineering takes advantage of people's trust in protection and authenticity. [21] By interacting with victims, social engineering attacks can achieve their goals without awareness of users. Social engineering combines the computer science and psychology which makes this technology could cause vital consequences with menial efforts. [15] Social engineering attacks make data breaches much more common, since it is more low-technical than other methods of cybercrime. By applying psychology into the technology, anyone can adopt the technique and achieve an attack. Rather than using high-technical methods to breakthrough the computer firewall, social engineering uses human as the weakest link to save lots of money and time. [15] Social engineering attack is a technique with low cost and high reward, which makes it maybe one of the most dangerous cybercrime forms. [22]

Social engineering is an important part of cyber security. Ira Winkler [23] states that social engineering mainly talks about non-technical attacks [23]. In the book of McClure, they define social engineering as "a description of techniques using persuasion and deception to gain access to information system". (p. 623) [24] "social engineering determines the target and helps attackers devise applicable strategies for accessing, exploiting, and exfiltrating data from target information systems." Kevin[25] defines social engineering as "Using influence and persuasion to deceive people by convincing them that the attacker is someone he is not, or by manipulation. As a result, the social engineer is able to take advantage of people to obtain information, or to persuade them to perform an action item, with or without the use of technology." [25]

When social engineering attacks want to influence people, there are mainly five psychological principles applied, "including fear, diffusion of responsibility, the chance of ingratiation, guilt, and overloading" (p.6)[26]. There are several principles that can be applied in social engineering, when attackers want to manipulate victims subconsciously. 1. People tends to offer a favor when receive one. 2. People are willing to help the person they like. 3. The victims like to remain consistency and commitment. 4. Social elements can influence people do things. 5. People would listen to someone with authority. 6. People would be more eager for items that are limited edition [26].

The point of social engineering is that victims may not aware that they are divulge their personal information of property. In Fillipe Breda's article (2017), he states that social engineering attack can be divided into two types: farming and hunting. "Hunting seeks to execute the social engineering attack through minimal interaction with the target."(p.2) [14]

In conclusion, hunting uses as least as possible interaction and lost connection once the attacker achieves the goal. Framing aims at long-term benefits, involves consistent communication. For example, phishing, vishing, and spam email are approaches that can be categorized as hunting. [15]

Social engineering can be divided into human-based social engineering and computer-based social engineering (software-based social engineering). Using social engineering toolkit to perform spear-phishing emails is an example of software-based social engineering attacks. In addition, social engineering can also be divided into three groups based on approaches: physical approaches, social approaches, technical approaches.[27]

2.2 Social engineering approach and framework

In the paper of Ashish Thapar[10], social engineering is classified as computer-based and human-based approach. "The technology-based approach is to deceive the user into believing that he is interacting with a real application or system and get him to provide confidential information." (p. 2) [10]

A social engineering attack is any act that influences a person to do things that may or may not harm his or her interest. It can be divided into multiple vectors e.g. SMiShing, Vishing, Phishing, and impersonation. [11] Phishing, Vishing, Spam email, popup window, and interesting software are technology-based approach. Pretexting/impersonation, dumpster diving, spying, acting as technical expert, support staff, hoaxing, and authoritative voice are non-technology-based approach. [10] Typical social engineering attacks are "phishing, impersonation on help desk calls, shoulder surfing, dumpster diving, stealing important documents, diversion theft, fake software, baiting, quid pro quo, pretexting, tailgating, pop-up windows, robocalls, ransomware, online social engineering, reverse social engineering, and phone social engineering (vishing)" Figure 2.1 is the taxonomy of social engineering attacks. [17]

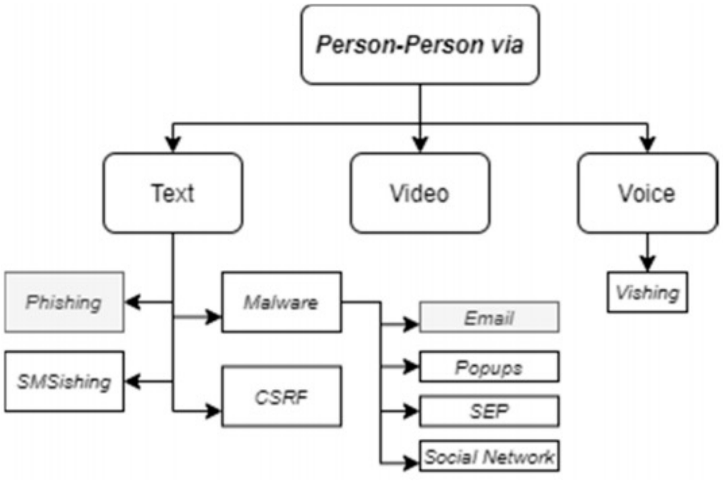


Figure 2.1. Social engineering attack taxonomy [17]

Vishing is voice phishing, which exploits victims' trust by using telephone service. Most of the vishing attacks aims at tricking for personal and sensitive financial information. During the pandemic period, according to the security magazine, "The Federal Trade Commission says that since January 2020 until mid-April, they received 18,235 reports related to COVID-19, and people reported losing 13.44 million dollars to fraud." (p. 1)[12] Phishing often uses email to pretend as a legitimate business. "The letter usually contains a link to a fraudulent web page that looks legitimate with company logos and content and has a form that may request username, passwords, card number or pin details." [10]

Tailgating attacks mainly based on physical techniques. In real life, when the attacker cannot enter restricting area of a building, he or she may lie about their intention and identification to get into the area. For example, radio-frequency identification cards are widely used in lots of companies for security purpose, however, the attack techniques on RFID are common and cheap. "At the physical layer, the RFID devices and the physical interface are targeted to manipulate an RFID communication. At the network layer level, the attacker manipulates the RFID network such as the communication between the RFID entities and data exchange between thesis entities." (p. 6) [17]

SMiShing is short message service phishing. There are two typical types of SMiShing. One is sending a text message that seems sent by authentic cooperation or agent. The other one is sending a text message that contains personal information. For example. it convinces the receiver that account has been stolen, and either texting the password or download malicious software can restore the account. Either way, it highly possible for hackers access sensitive data inside the smartphone [28].

Ransomware attacks have been proved can cause tremendous money loss. Typically, hackers use ransomware attack their system, gather and encrypt important data. By threatening company that hackers may reveal all these data to society, the company needs to pay terrific asset (bitcoin) otherwise data leakage could cause loss of business, customers, and productivity. "A ransomware attack involves six stages: creating the malware, deployment, installation, command and control, destruction, and extortion." The result of ransomware is powerful. Once the attack has been performed, victims only have three options. Either paying the money, restoring system and files, or taking the consequences of losing data. Fake software attacks can also be called fake website. By mimicking the website that victims familiar with, victims may lower they safety consciousness and feel comfortable. There are lots of fake websites pop out when people type keywords within search engine. This approach exploits users' trust of original websites and hack into accounts perform illegal operations [17].

According to Arif Koyun [27], other than the approaches above, dumpster diving and reverse social engineering attacks also are skills that commonly used. Dumpster diving means digging into possible victims' trash for valuable information, in both physical and virtual ways. This skill can be operated by both human and computer software. As for reverse social engineering, rather than starting communication by attackers initially, it makes potential victims contact first [27]. In this case, whoever calls the number of hackers could be more easily fall for the fraud. Reverse social engineering can benefit from the format of online social network. Since the made-up message is posted online, it can bypass the filter of the email application and have a lager target group based on platform. From the attack point of view, reverse engineering attack can be categorized as: direct attack, mediated attack, targeted attack, and un-targeted attack. From the perspective of online social media, reverse engineering attack can be divided into three types: recommendation system-based attack, demographic search-based attack, and visitor tracking based attack [27].

Spam emails also can be called junk emails, "unsolicited usually commercial emails sent to a large number of recipients or posted in a large number of places". [29] Spam email maybe one of the most frequent social engineering attacks we face every day, since the spam filter cannot work as we expect. A social engineer often collects information from OSINT, which is open-source intelligence. Through OSINT, engineers can gather personal information like phone number, name, gender, and address. One thing to be clarified is that most of the time, OSINT does not require high technology.

There are some other approaches not as well-known as others also need to be paid attention, such as shoulder surfing, baiting, watering hole, and advanced persistent threat. For example, shoulder surfing steals information by observing directly – looking at other's screen and keyboard intentionally. Baiting allures people open malware and malicious link by using techniques like pop-up window. Watering hole uses website and technical approaches to gain victims' trust, which requires more high-technical skills. Advanced persistent threat requires long-term relationship using email or website to perform socio-technical approach [27].

There are specific steps that social engineering attacks can follow. There are six core phases in the social engineering attack framework: namely attack formulation, information gathering, preparation, develop a relationship, exploit relationships, and debrief. [1]. Communication, social engineer, target, medium, goal, compliance principles, and techniques are essential features of social engineering attacks. In our survey, we designed several questions to follow the steps of social engineering attacks. Steps are as follows: attack formulation, information gathering, preparation, develop a relationship, exploit relationship, and debrief. [1]

As Filipe Breda (2017) categorizes the phases of social engineering as research/information gathering, hook/establishing relationship, play/exploitation, and exit/attack execution. The research/information gathering phase will target the eligible victims, design a reasonable plan, and collect personal information from online source including social media.

The purposes of research are to gather victims' background information, find the target, and plan ahead of the identity. The phase of hook "aims to engage with the target, spin the story, build a level of intimacy, and take control of the interaction." (p. 9)[15] The phase of play aims to "extract information and keep things going long enough to do so: maintain charade, strengthen control of relationship, extract information". And the final phase – exit aims to "close interaction, without arousing suspicion: bring charade to natural end, provide target with reason to keep quiet, cover tracks" [15]

2.3 Detecting and defending social engineering attacks

2.3.1 Human awareness

Other than using high-technical software to detect social engineering attacks, it is efficient to train the awareness of human. From the perspective of a victim, he or she "is manipulated by social engineering attacks are often made vulnerable by natural tendencies to trust others without requiring much reason". (p. 27) [30] In other words, even though a person questions the credibility of a phone call or email, the tendency of trusting might still affect his or her decision. The success of social engineering attacks based on the unawareness of victims of social engineering strategies and outcome of disclosing personal information [30].

The most efficient way to detect and stop the attack is training security awareness. Research by Arif Koyun [27]demonstrates several suggestions to protect people from social engineering attacks. For example, be aware of people who ask for quick decision very urgently, things too good to be true, and information no one has access to. Also, do not click unknown link without complete trust, do not share personal information with stranger online, and contact IT expert when it comes to work-related security issue. In conclusion, when it comes to future attacks, make strong passwords, do not share too much information online, and verify every contact before sharing [27].

"Defending against social engineering requires the continuous education and training of users."(p. 32) [30] Good cybersecurity behaviors are essential for preventing cybercrime, which means social engineering often exploit behaviors' vulnerability. The figure 2.2 bellow shows eight types of behaviors that can be applied by engineers to gain information. In a report of internet threats trend in 2014, PayPal, Apple, Poste Italian, Barclays Bank, Battle.net, and Sparkasse are the top six most frequently phished properties. [31]

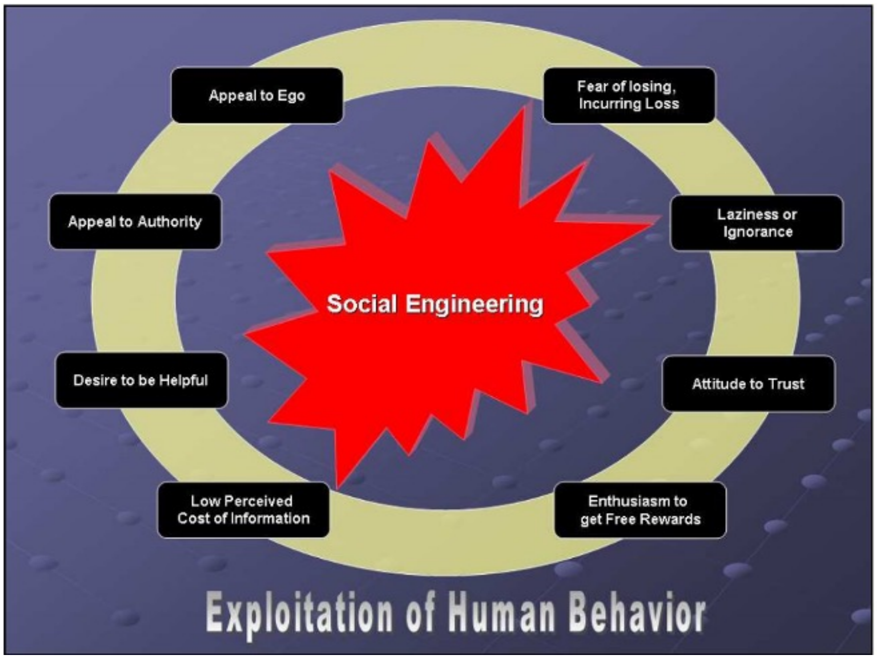

Appeal to Ego
Fear of losing, Incurring Loss
Appeal to Authority
Laziness or Ignorance
Social Engineering
Desire to be Helpful
Attitude to Trust
Low Perceived Cost of Information
Enthusiasm to get Free Rewards
Exploitation of Human Behavior

Figure 2.2. Exploitation of human behavior [10]

”Developing cyber security policies, implementing security awareness training, installing spam filters and anti-malware software, deploying next-generation firewalls, and installing endpoint detection and response are some of the common ways to prevent attacks.” (p. 2) [18] Other than these, companies should also organize opportunities to train and educate about social engineering. For tailgating attacks, employees should be aware of anyone who wants to borrow locks or ID cards. For should surfing attacks is pay attention to the surrounding environment, especially cameras. [17]

2.3.2 Applied technology

Training human’s awareness is the first step of defending against social engineering attacks. The rest should rely on technical approach. Figure 2.3 is the taxonomy of social engineering attacks and their countermeasures. Following are several defending techniques can be applied. 1. Email filtering tools. There are tools that collect email addresses that has

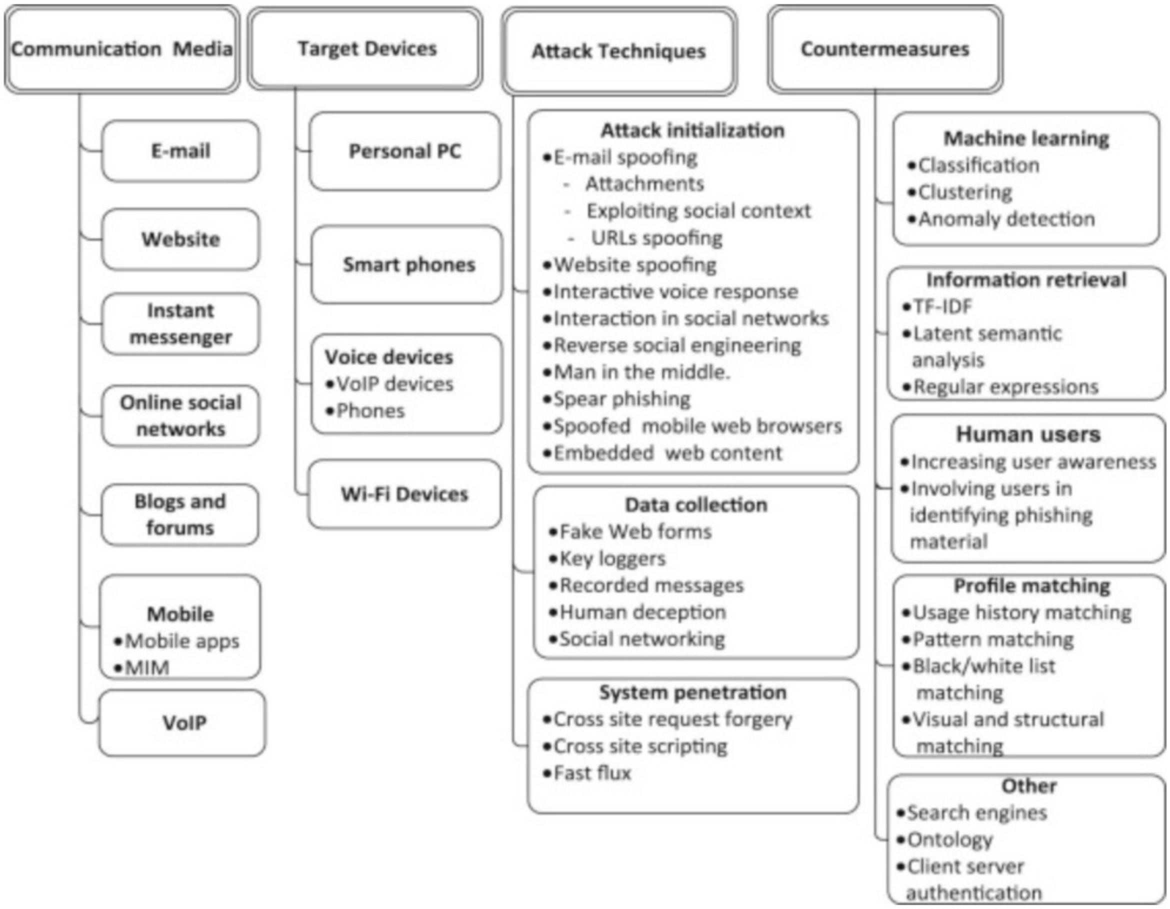


Figure 2.3. Phishing countermeasures taxonomy [32]

been detected with spam history. With these tools, whenever the email is coming from the spam-trap list, they will suspend the process of receiving this email [30]. “Other procedures that can be done include: verifying emails’ sources before clicking on a link or opening an attachment, examining the emails header, calling the known sender if suspicious and discarding emails with quick rich or prize-winning announcements.” (p. 9) [17] 2. Web browsers and application that have built-in alert 3. Blacklisting tools. In a vishing attack scenario, the best way to prevent this attack is ignoring the phone call. Blacklisting tools can help the victim decide whether this phone call is coming from a hacker. In a phishing attack scenario, blacklisting tools can block malicious website, for example, McAfee anti-phishing filter, Microsoft phishing filter, and Web sense.[17] 4. Machine learning algorithms. By using unsupervised learning algorithms, and other six machine learning algorithms, there are tools that can detect phishing attacks. Also anomaly detection algorithms can detect unusual operation and information inside the computer, if the computer is hacked. [17]

2.4 Social cognitive theory

Social cognitive theory is the reason why we choose the environment as a comparing factor. Albert Bandura proposed the social cognitive approach in the 1960s. The idea recognizes that human beings are not mindless automatons responding to stimuli. But mental beings making choices that are influenced by the complexity of their surroundings. [33] This means that the environment can affect people's behavior through control and reinforcement, and this influence can be long-term [34]. For example, if people saw a friend being tricked by a social engineer, they might have a higher possibility of not falling into the same fraud. The influence of friends and other peers is another element of this theory. People's answers to these questions connect to environmental factors. Figure 2.4 shows three types of elements affected by the environment based on Bandura [33]'s theory. And these are the factors we applied to the survey.

The social cognitive theory also has a limitation: we cannot be sure that the environment changes will automatically lead to personal behavior differences. But in this research, we only focus on the overall environment.

2.5 Environmental factors

Prior research has examined how a user behaves towards security threats and vulnerabilities after being provided with some form of formal security course. The results show that knowledge influences the students' awareness of security [35]. Aldossary and Zeki's study sent an online Google Doc survey over email in order to quantitatively analyze a student's behavior to a security threat. The results showed that in terms of gender, males showed safer security procedures. In terms of university, there was no difference in terms of which university was being studied. In terms of major, IT students showed a better awareness for safety procedures. In terms of study level, students at a higher level of a study showed safer

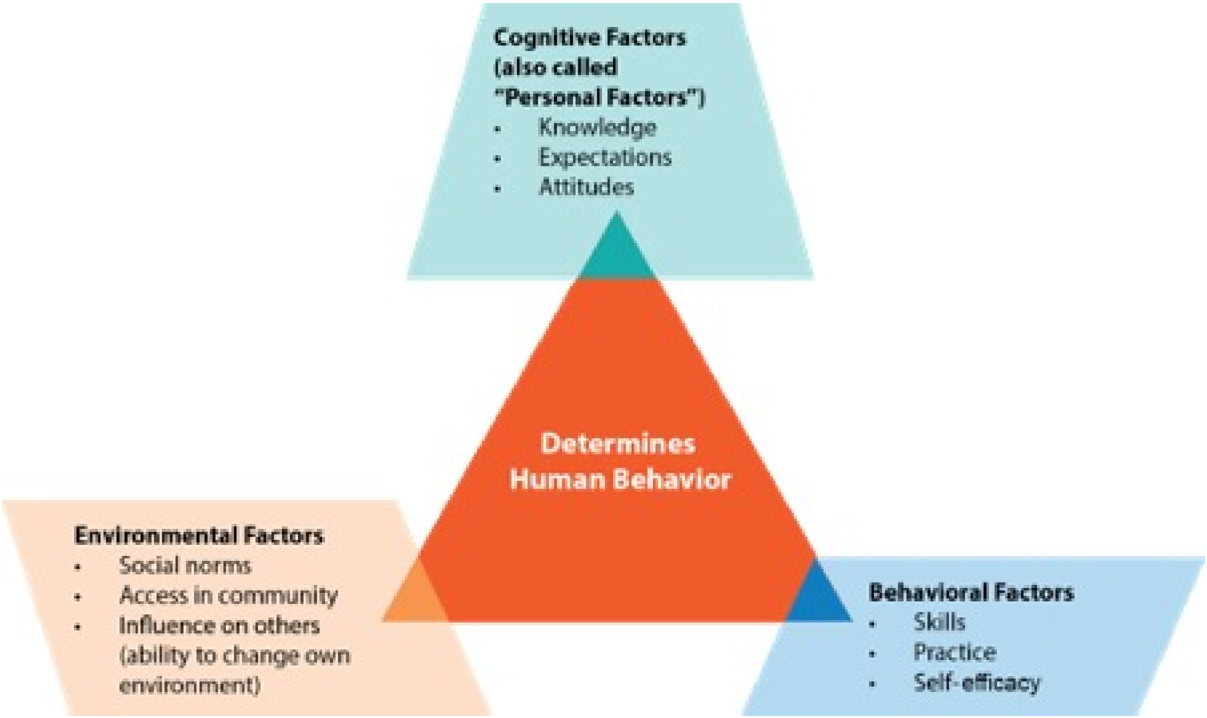


Figure 2.4. Social cognitive theory model
[1]

safety procedures than students at a lower level of study. The researchers concluded that the culture and behavior did overshadow students' knowledge of security awareness to some degree, students with high education did influence their awareness of security [35]. While education on cybersecurity has proven to influence a students' awareness on security, not all students are offered a cybersecurity awareness course in their field of study. These courses are normally aimed to be taught to technical students at a university level. In another study, a three-tiered framework that provided security education inspired students to explore cybersecurity as a career option or at least broaden their knowledge of cybersecurity in order for them to be more security-conscious [36]. The framework for security education includes (1) a formal literacy-based training for students of all backgrounds, (2) inquiry-based learning through security and technically-focused student group and activities, and (3) classical technical-based initiatives. The results proved that course-based education is still needed for students who have a high interest in cybersecurity, but, inquiry and literacy-based learning allows students who have never considered cybersecurity to explore and understand this career track [36]. While the study focuses on getting individuals interested in the field of study in order to pursue a cybersecurity career, it would be interesting to understand if inquiry and literacy-based learning methods could help improve the use of security protocols to those who do not have previous cybersecurity knowledge.

In another research study, there is proof that the biggest misconception that shows up during research in this area is that education probably reduces one's chances of being susceptible to a phishing attack. However, research has proven that even a carefully crafted attack can fool even someone who has all the education about how to avoid an attack. No matter how much training there is, sometimes it is not effective enough to reach all customers [37]. Social engineering makes it easy to drive corporate users to fake websites and corrupt a user's computer. The solution to this problem can be to use more than one prevention method. Organizations should reevaluate the ways they approach targeted attacks [37].

From studies, it has been noticed that individuals show different security behavior at work than at home. Individuals can be more cautious at work and as soon as they bring that sensitive information home, their security behaviors could change as one can feel more 'safe' or 'comfortable' in their own home. In a study conducted by Hazari about factors influencing information security behavior, 200 business students who were enrolled in a state university received an on-line questionnaire where they answered a knowledge quiz which was later quantitatively analyzed [38]. The survey was broken apart into four constructs including: Attitude, Subjective Norm (SN), Perceived Behavioral Control (PBC), and Behavioral Intention (BI). The results of this study showed that the attitudes, subjective norm, perceived behavioral control is related to maintaining information security awareness [38]. The results also proved that experience in using the Internet is not related to knowledge about information security behavior. The results of the study also implied that training is needed and required in order to inform and reinforce the need to maintain information security even when working from home [38]. While this study only studied business students it would be knowledgeable to understand how different majors and people with different backgrounds would respond to a similar questionnaire/survey.

In another study, the research aimed to analyze the level of information security awareness among a group of undergraduate mechanical engineering students [39]. As technology is more apparent and in demand in our world today, we see almost everyone, especially the younger generation, with some type of personal technological device. While this enhancement in technology has made lives more efficient and engaging, individuals may take the security precautions they need too lightly. This study looked at passwords the respondents used to protect their information [39]. This is relevant as we often see in the news about some hacker trying to access unauthorized accounts due to weak security protocols implemented by the user. The study included sending out a personal data questionnaire to the students of Obuda University. The results proved that none of the participants had short passwords ($<$ 6 characters long), none of the participants used only letters in their password, 26.5 percent of the respondents used uppercase letters, numbers and characters as well. However, the results also proved that more than half the respondents (58.8 percent) do not change their password often and 11.8 percent of the respondents stated that they wrote down all their passwords [39]. The study stated that in regards to the sensitive data, it is important to raise the level of information security awareness of the mechanical engineering students in the study [39]. While this study only measured mechanical engineering students and while the study conducted by [38], only measured business students, it will be knowledgeable to understand the difference in security perceptions of individuals from different industries.

Similarly, a study was conducted to understand the knowledge and practice relationship between the workplace and home. The study identified that majority of the learning about information security occurred in a workplace environment compared to at home as there is legislation and regulation in the workplace where employers constantly reiterate the importance of security awareness [40]. In order to understand the relationship between workplace and home, a survey was distributed to a wide range of people regardless of location but with the condition that they were employed and used a computer at home as well as at work. The results proved that if users were given an effective training session at the workplace, that knowledge will transfer home as well [40]. There has been a long history regarding good security and how it cannot be achieved through technical means alone and an understanding of how to protect oneself is imperative when it comes to being security-aware and security-safe.

While research has already proved that the higher the level of education an individual has received, the more information security aware they are, a study was created in order to examine levels of awareness toward information security in terms of perception and behavior. This meant if one had been a victim of cybercrime in the past or not [41]. In order to categorize the preventive actions employed by users and the threats they have been exposed to a scale of four categories were created to conduct this study. These categories includes: Risky Behavior Scale (RBS). Conservative Behavior Scale(CBS), Exposure to Offence Scale(EOS) and Risky Perception Scale(RPS). The results proved that those who have been a victim of cybercrime showed low awareness when it came to reporting the crime and that individuals find it more personal to share their contact information rather than their personal information [41]6). However, it was found that the higher level of education an individual has received, the more their information security awareness is. While this study did not discriminate on age and population, it will be knowledgeable to individuals from different backgrounds react to security protocols depending on their perceptions and behaviors.

Combining all the information learned from these previous studies it led to the research question of this study: Is there a relationship between environmental factors and the susceptibility of a social engineering attack?

2.6 Phishing

Unsolicited commercial e-mail, non-responsive commercial e-mail, list makers and scams are four categories from spam. Scams can be divided into “Nigerian-style” scams, malware, and phishing. In order to identify the group of phishing, the key items are: “bulk-mailing tool identification and features, mailing habits including specific patterns and schedules, types of systems used for sending the spam, types of system used for hosting the phishing server, and layout of the hostile phishing server” (p.8) [42]

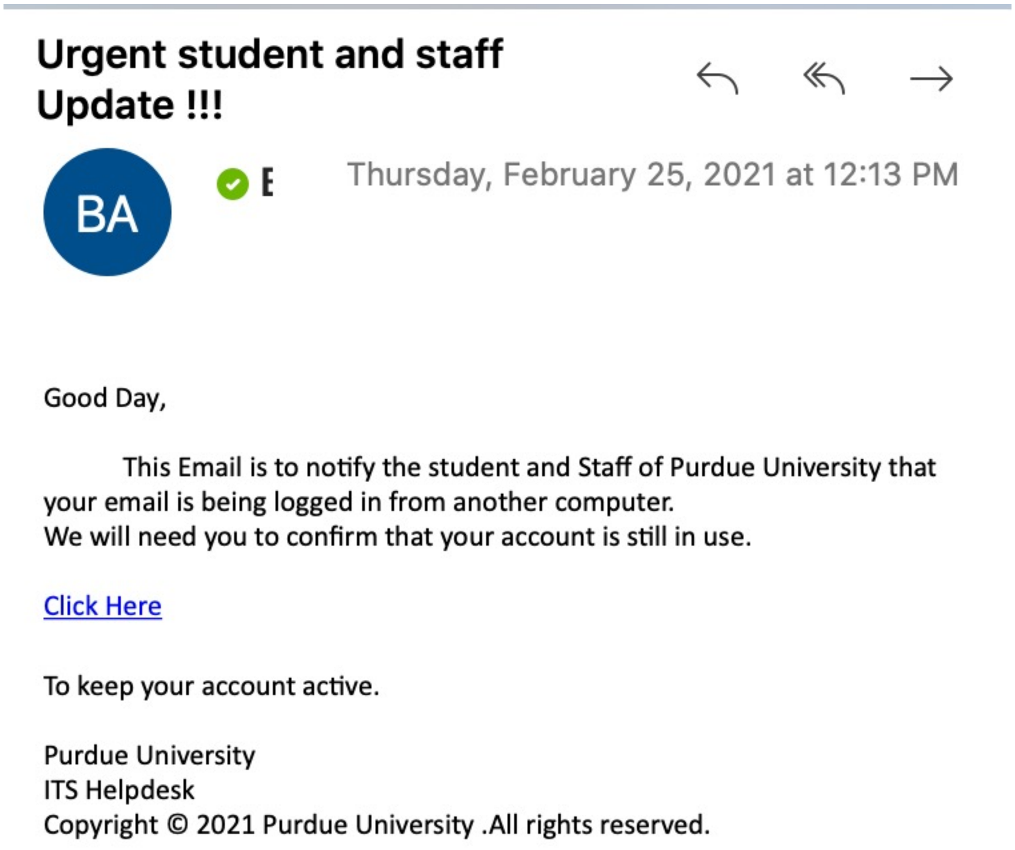
Phishing varies from fishing, which uses baits to get victims swindled. “The term phishing comes from the fact that cyber attackers are fishing for data, the ‘ph’ is derived from the sophisticated techniques they employ, to distinguish their activities from the more simplistic fishing.” (p. 2) [42]. He also mentioned the most common three types of phishing, including impersonation, forwarding, and popups.

One example of impersonation is tricking people to click the link by constructing a fake website. Forwarding is a technology based on real website, but though a third-party server. Popups nowadays are harder to be performed since the computer system are difficult to be cracked. However, popups often are operated by real hackers, which brings a lot of uncertainty. [42] Since the computer system are harder to be cracked, attacking human is a more achievable goal. [43]

Spear phishing aims at individual person and specific group by using the online social media data, it is hard to distinguish the attackers from regular users. Spear phishing appears to have a higher successful rate than other attacks [44]. Whaling is spear phishing when the target is in big cooperation and called big fish. Vishing combines phishing and voice together to perform social engineering via phone. Interactive voice response phishing combines interactive voice system and phishing which instruct targets leaking valuable data step by step. Business email compromise phishing needs to gather information first from company executives by analyzing their online posts. After understanding their working patterns, calendars, clients, and so on, attackers imitate victims' tone and identification to get access to business financial account [45].

Phishing has been emerged since 1995, when there were programs can automatically send e-mails and ask for bank information. According to the research of Bruce Schneier, there are three kinds of network attacks: physical attacks, syntactic attacks, and semantic attacks. Phishing is one approach as the third form of attacks called semantic attack, which uses words to manipulate human and assigns meaning to these words. No matter what the communication medium is, it is possible for a human being accidentally give out personal information. In the research of Bruce, which is 21 years ago, he anticipated that the semantic attacks would be more and more serious than the other two types of attacks. This opinion turns out to be true, phishing has turned into a worldwide problem. Phishing as one way of semantic attacks needs to draw more attention and awareness. [46]

Phishing has become one of the most widespread way for attackers to obtain identity, bank account, address, phone number and so on[47]. Based on statistics, 50 percent of people who tends to fall for phishing attacks usually click on the link with the first 24 hours. [42] Standard email phishing is the most common usage. Techniques that been applied into phishing are: code-based key-logger, in-session phishing, domain name system poison, search engine phishing, mass e-mailing. On one hand, the procedure of phishing can be categorized as: plan attack, setup, execution, fraud, and post attack. (Ahmed Alerouda Lina Zhoub, 2017) On the other hand, Jakobsson divided the procedure into: preparation, send malicious website or email, ask for personal information, compromise, transmit, impersonate, and accomplish. (Jakobsson and mayer, 2006) [47].



Urgent student and staff
Update !!!

Thursday, February 25, 2021 at 12:13 PM

Good Day,

This Email is to notify the student and Staff of Purdue University that your email is being logged in from another computer.
We will need you to confirm that your account is still in use.

Click Here

To keep your account active.

Purdue University
ITS Helpdesk
Copyright © 2021 Purdue University .All rights reserved.

Figure 2.5. Phishing example

Figure 2.5 is an example of standard email phishing. The email address appears to be legit from Purdue University, and it pretends to be a faculty of Purdue University. In the meanwhile, the email illustrates a possible scenario that leads the target to click the link. However, the email address is hacked, as well as the website is fake. Once the victim log into the website, the identification and the password may fall into the hands of lawbreakers.

In the research of [48], they illustrate the phases of an example of how to perform an email-based phishing attack and based on different phases, lots of techniques have been applied into the attack. Nowadays, multiple communication media can be used for phishing attacks, for example, email, website, instant messenger, online social networks, blogs, forums, mobile applications, and voice over internet protocol. Personal laptops, smart phones, and Wi-Fi devices can all be target devices. For example, when it comes to a standard email phishing attack, the steps are as follows:

1. Select email addresses. There are different strategies that can be used when it comes to select the potential victims. One is to use existing email addresses: web crawled email addresses, previous phishing email addresses, and the addresses from other email. The other is generating new email addresses based on domain or other techniques[48].

2. Build the content of the email. Every phishing email needs a background story or a topic. It could be about earning benefit, requesting information from a legal department, getting important information, target's failure or anything else. This phase is the most essential one for gaining the trust from targets, the excuse needs to be authentic, either the engineers generate, edit, or copy the content. The email could be either personalized or not. If it is a personalized email, the attacker needs to collect personal information first or have special background knowledge. In addition, the content could be produced by both human and robot[48].

3. Send out the emails. The strategy of sending emails is based on three factors: the sender's email address, the number of recipients, and the usage of a phishing attack strategy. The receivers of the email can be divided into individual and groups. The senders of the email can use both spoofed email addresses and real email addresses[48].

4. Waiting for response. Since sending emails can be divided into systematic, and non-systematic strategy, non-systematic strategy will only send one email and wait for response. On the contrary, systematic strategy has several follow-up actions, for example, send multiple emails, contact by other channels: calls, social network), involve other people [48].

5. Data gathering. The goal of phishing is obtaining personal information for the receiver. There are two types of data are gathered: association data and secret data. Secret data mostly is more valuable and the ultimate goad in an attack, such as, login passwords,

company data, business data, bank account, social security numbers and so on. Association data might not be as valuable as secret data, but it could be confirmation of the email address and profile data of the receiver [48].

6. Usage of gathered data. How to use the data highly associated with the types of the data. In conclusion, attackers can use these types of data access into financial system, commit fraud, hack company system, and so on [48].

2.6.1 Vishing

Other than email phishing attacks, vishing was invented before the internet, which is more old-fashioned. Vishing and phishing often have the same purpose and target, but according to numbers, vishing has a lower success rate. However, since the smartphone developed really fast, vishing attacks increasing hugely at the same time[49]. In the year of 2019, there are nearly 83 percent of people have experienced vishing. It is hard to track down whoever commits the crime of vishing. In the year of 2020, “over 56 million Americans lost money to phone scams, this is a 30 percent increase compare to last year.” [50] Based on statistics, it appears to 19.7 billion dollars been spent on phone scams for one year, and the average money loss of each victim is 351 dollars. Vishing combines voice over internet protocol with phishing techniques, for example, automatic dialing, manual dialing, and telemarketing calls[51].

During a vishing attack, first is get access of victims’ phone number. “The method of compromise is inconsequential, as long as the attacker is able to install and run software” (p. 33)[52] After controlling victims’ phone, hackers need to install digital private branch exchange. Another thing called caller ID spoofing is often used in vishing. This technique is used for convincing people that phone calls are coming from authentic organization. For example, even though the hackers cannot use the exact phone number of White House, they can make different phone numbers appear to be coming from White House. In order to hiding real identification of hackers, they usually using software or hardware devices so that “digital session initiation protocol can convert to the public switched telephone network”.

(p. 34) [52] Nowadays, there are lots of fake number agent website, like google voice, eVoice and so on to generate fake phone numbers [52].

The differences between vishing and phishing are as follows: vishing more relies on social proofing, while phishing more relies on one-to-one communication. Phishing tends to maintain multiple relationships at the same time, vishing can only communicate with one person at one time. Phishing emails often contain textual and graphical messages. The fundamental reason of the difference between phishing and vishing is that vishing more relies on benefiting victims via social proofing [53].

2.6.2 Ad fraud

Online advertising has become an industry chain with 10 billion assets, and a large number of people become victims of online fraud every day. [54] And Ad fraud has become particularly dangerous when advertisements are everywhere nowadays. As mobile phones become more and more intelligent, ad fraud is not only on computers, all devices that can access the Internet will be under the threat of ad fraud, including but not limited to smartphones, tablets, and smart homes. There has been a report to claim that: "FOR EVERY $3 SPENT ON DIGITAL ADS, FRAUD TAKES $1", "As digital spend continues to reach landmark highs – it hit $27.5 billion for the first half of 2015 – so does ad fraud, which is now estimated to cost the industry about $18.5 billion annually" [55] These are all the proofs that we need to pay more attention and be more careful about ad fraud. Fake websites may be full of fraudulent advertisements. Not only that, but many regular websites are also likely to play advertisements that may leak information after charging high fees. After these advertisements are paid to developers, they attract ordinary people to click, and then may use false publicity messages to attract people to enter private information including bank account passwords for payment. Ad fraud is different from ordinary website advertisements. Rather than attracting more users through publicity websites to attract traffic to achieve the purpose of publicity, ad fraud is more about falsely exaggerating the value of a website or product and deceiving consumers to obtain sensitive information.[56] Although studies have shown that through the analysis of HTTP algorithms, many applications have been able to filter and detect malware and ad fraud, but there are still loopholes in the development of these technologies. Also recording to Hamed Haddadi, the bluff ads are able to prevent the consumers from being an easy target, detect possible fraud timely, and help victims fight back. [57]However, even if it is possible to determine whether certain advertisements are available by analyzing IP addresses dangerous, some providers can still bypass the detection.

Ad fraud mainly uses psychology to attract people to click on malicious websites. The profit model of ad fraud can be divided into: "cost per mille (CPM), where advertisers are charged per thousand impressions; cost per click (CPC), where advertisers are charged per click; and cost per action (CPA), where advertisers are charged per action, such as an online sale." To perform an ad fraud, the basic steps of a consumer are like this: 1) log into a normal search engine, for example google.com 2) visit the target research result which is also isn't a fraud 3) the pop-up window or injected window may have some information to lure victims click 4) the fake website bypass the security and trick victims to leak personal information.[54]

3. METHOD

3.1 Data Collection

The data were collected using a mixed methods survey distributed through Qualtrics (Appendix A). The survey was developed using social cognitive theory as a framework and the three most common social engineering attack vectors as a focal point. The survey was divided into eight sections. The first was a section on consent, the second section asked standard demographic questions, the third section was about knowledge of cybersecurity specifically social engineering literacy. The fourth, fifth and sixth sections were behaviorally based and asked the respondents to identify phishing, vishing, and ad fraud. The seventh section dealt with respondents' past experience with vishing, phishing, and ad fraud, while the eighth section asked about environmental factors. These data were collected using MTurk, a crowdsourcing website validated for on-demand tasks such as survey taking. Data were collected from five-hundred and sixteen participants.

3.2 Data cleaning

In order to filter out the unreliable parts of the data, we set up a validation question. If the answer to the question 14 is wrong, it means the all the data from this participant is unreliable and we need to delete the rows. This is the reason why we have 516 participants, but we only count 495. In addition, if the status of a participant is unfinished, and the values are missing after question 18 can also be deleted.

3.3 Analytical Strategy

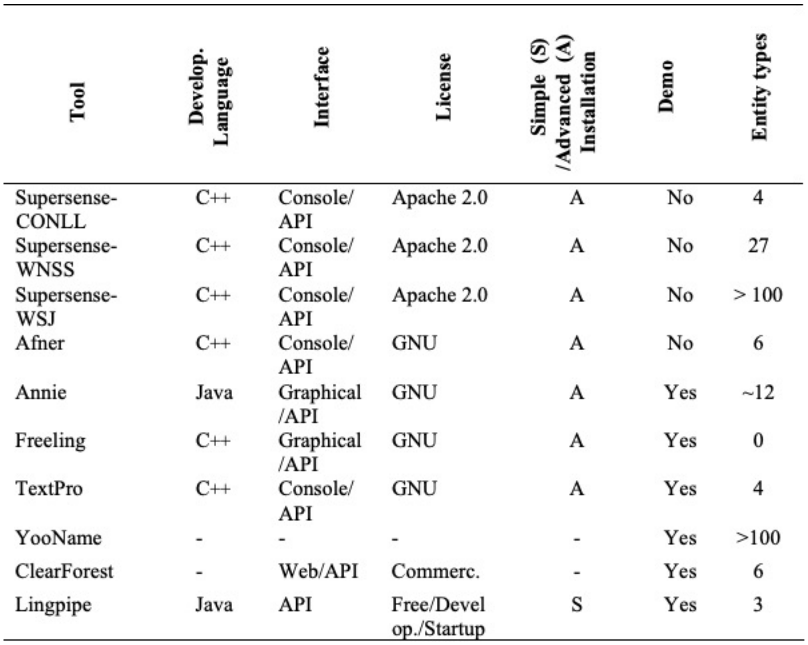
In order to calculate the relationship between user behavior and environmental factors and using SPSS, we calculate the mean value of the how likely question 18 to question 41 is a vishing, phishing or ad fraud scenario. One thing to notice is that if one scenario is unlikely to be a social engineering attack, the question needs to be reverse coded. In this case, we can combine all the questions together as a whole. Table 4.2 shows the descriptions of mean value of phishing, vishing and ad fraud. Standard deviation shows how dispersed the data is in relation to the mean, with higher std. deviation the data is more spread and vice versa. In the next subsections, we will use descriptive statistics including Minimum, Maximum and mean. We also apply the ANOVA analysis and correlation to determine if there is difference between each group. For example, if the ANOVA significance is lower than 0.05, we can support the hypothesis that there is significant difference.

3.4 Data analysis algorithm

The dataset we collected is high-dimensional data. If we want to draw conclusions from this type of data, it is essential to choose the right algorithm, which means we need to apply multiple algorithms and compare them. Before we put data in machine learning algorithms, we need to preprocess the data. For example, because the answers for our survey questions mostly show the levels, if we change the response to number might be easier to get numerical results. In this way, we can get the mean value of a cluster to show the tendency. Besides, we also need to fill out all the NAN values to 0 if we lost data. The clustering algorithm can be useful for this research if we want to plot responses against the test-takers' location. The location consists of longitude and latitude, and they have so many digits. So, it is hard to apply the algorithm directly, which can mislead the conclusion. The clustering algorithm can classify the two location columns into one cluster column. After visualizing the longitude against latitude, I find out that the location data can be divided into four groups. Using a clustering algorithm, we can easily label each row to a new feature. One way to deal with high-dimensional data is by using a dimension reduction algorithm. The next chapter will discuss why this algorithm can get reliable results and how it works. This thesis uses three types of algorithms to compare the products, which are UMAP, t-NSE, and PCA. Each of them has pros and cons; one has better visualization results, one can show more information, and the other is easier to manipulate. If we don't use the dimension reduction algorithm, we could use the clustering matrix directly. The benefit is that we don't lose any information, but the weakness is that we can only analyze the plot intuitively. But still, based on the color of each column, we can divide them into several groups. As for why this happens, we need help from the more individual analysis between test-takers' factors and judgments. Data visualization can aid our understanding of data. With the help of a well-designed visualization, we can save a lot of effort in machine calculating. Time series data can use index charts, stacked graphs, small multiples, and horizon graphs. Statistical distribution help analysts understand the statistical models, like stem-and-leaf plots, Q-Q plots, scatter plot matrix and parallel coordinates. If the data has geological information, maps are worthy of applying, like flow maps, choropleth maps, graduated symbol maps, and cartograms. Survey data can sometimes be organized into a hierarchy. Then we can use node-link diagrams, adjacency diagrams, and enclosure diagrams. Finally, if we want to build a relationship between different questions, we can use networks like force-directed layouts, arc diagrams, matrix views. (Heer et al., 2010) A fundamental goal of most quantitative data analysis in the social sciences is to ascertain whether an observed association between two measured variables can be interpreted as a relationship among the corresponding constructs, which means that the two constructs share a causative connection, specifically that one construct influences the other. For example, one might test whether single parenthood demands account for the lower level of voter turnout among this group compared with married persons as hypothesized. (Aneshensel, 2013) To analyze the high-dimensional data and find the relationship between each question, we can adopt several algorithms. The typical procedures of analyzing are data transformation, visual mapping, view transformation, and make decisions. Transforming data can use dimension reduction, subspace clustering, regression analysis, and topological data analysis. Dimension reduction has a linear projection, non-linear dimension reduction, and precision measures, etc. In our case, the visual mapping mainly consists of axis-based mapping, hierarchy-based mapping, and evaluation, such as scatterplot matrix, radial layout, dimension hierarchy, and scatterplot guideline. (S. Liu et al., 2017) Besides plot-based analysis, we can also use numerical methods like linear regression and correlation value to compute the corresponding value.

3.5 Text analysis

Not only are the answers worthy of analysis, but also the questions themselves can be useful. Our survey has almost 100 items, and 22 of them are pretty long questions. After understanding the relationship between the factors and answers, we can dig deeper into the questions' keywords. After using text analysis, we can collect a couple of keywords and label each item. These labels are like the metadata of survey questions. After this procedure, we can use machine learning algorithms and repeat the analysis one more time. Using these steps, we might be able to find a more straightforward pattern of data behavior. Text analysis can also benefit social engineering attack detection. If we collect a larger dataset of scenarios, we certainly can achieve the goal of predicting the case is social engineering attack or not. Even though our text information is not enough, we can get humans involved in this procedure to get a better conclusion about which words happen more often in a social engineering attack. Our survey has 22 lengthy questions that might have similar keywords. By using visual text analysis, we can better understand the pattern of social engineering attacks. This research may even help future research about identifying social engineering attacks to raise people's awareness. There are two types of information visualization: metadata text visualization and content text visualization. [58] Our research will focus more on content-based text visualization because the question itself does not have metadata. To be more precise, I will use named-entity extraction to analyze the survey file's text information. This method can extract valuable entities from each question; after this, I can build the structure individually in order to find the reason why similar items have different results. On the one hand, the questions in this survey are not overly lengthy. On the other hand, the analysis based on this can contribute to future studies. Coates-Stevens defined entity extraction in 1992, also known as named entity recognition. He proposed the idea of analyzing proper names for robust text understanding [59]. The typical system architecture of entity extraction application consists of text, necessary preprocessor, database lookup, named entity analyzer, categorizer, deep analyzer, and knowledge base [60]. There are different approaches and machine learning algorithms in entity extraction operation. The various types of limitation of named entities recognition system are: 1. Different types and groups of entities like the person, organization, and location 2. Different language 3. Software tools 4. The evaluation algorithms are different in each competition 5. Users' manually labeled dataset 6. Tools in the system may be not available [61]

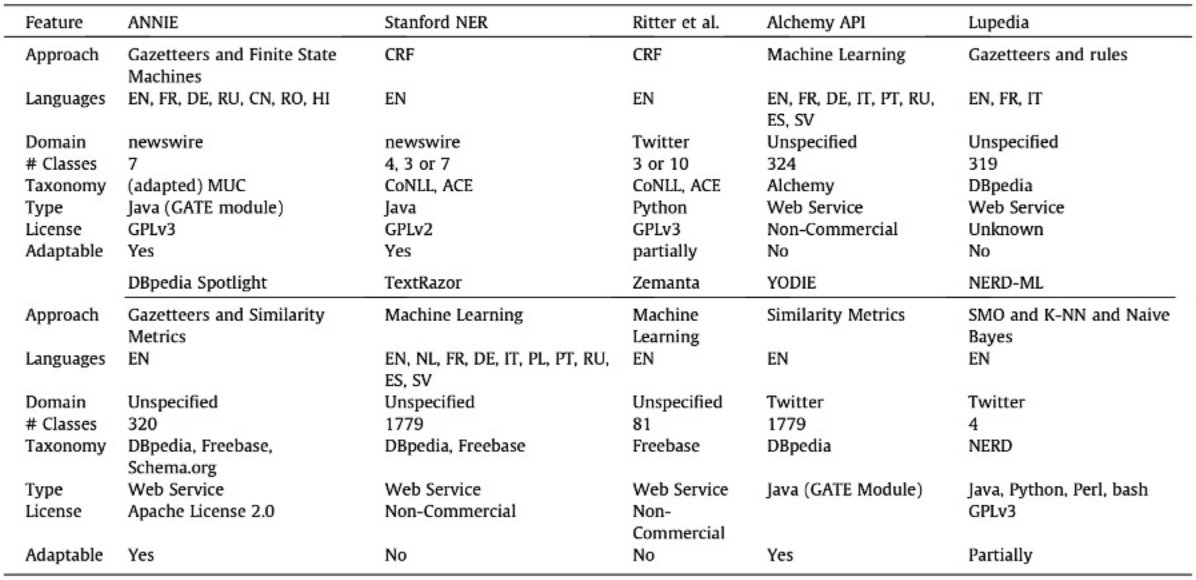


Tool	Develop. Language	Interface	License	Simple (S) /Advanced (A) Installation	Demo	Entity types
Supersense-CONLL	C++	Console/ API	Apache 2.0	A	No	4
Supersense-WNSS	C++	Console/ API	Apache 2.0	A	No	27
Supersense-WSJ	C++	Console/ API	Apache 2.0	A	No	> 100
Afner	C++	Console/ API	GNU	A	No	6
Annie	Java	Graphical /API	GNU	A	Yes	~12
Freeling	C++	Graphical /API	GNU	A	Yes	0
TextPro	C++	Console/ API	GNU	A	Yes	4
YooName	-	-	-	-	Yes	>100
ClearForest	-	Web/API	Commerc.	-	Yes	6
Lingpipe	Java	API	Free/Devel op./Startup	S	Yes	3

Figure 3.1. Tool features
[61]

Figure 3.1 shows the table of features of different tools. There are no most optimal tools when it comes to entity extraction. For example, more entity types do not mean the device is better. I need to apply the system individually so that I can detect a particular module to define questions. According to Marrero's evaluation of these tools, they compute the precision and recall in classification and identification. Among all the tools, the performance is generally over 50 percent, and the After, which has the worst result, can perform better on a specific tag [61]. Derczynski also did similar research comparing systems based on tweets. Figure 3.2 shows the key features of the five strategies. The results of the study indicate that NERD-ML and Stanford have the best performance [62].

Data preprocessing is fundamental when it comes to high-dimensional data. In our case, since the survey has 100 questions, this data is very high-dimensional. So we need to use a dimension reduction algorithm to help with analyzing, like PCA, t-SNE, and UMAP.



Feature	ANNIE	Stanford NER	Ritter et al.	Alchemy API	Lupedia
Approach	Gazetteers and Finite State Machines	CRF	CRF	Machine Learning	Gazetteers and rules
Languages	EN, FR, DE, RU, CN, RO, HI	EN	EN	EN, FR, DE, IT, PT, RU, ES, SV	EN, FR, IT
Domain	newswire	newswire	Twitter	Unspecified	Unspecified
# Classes	7	4, 3 or 7	3 or 10	324	319
Taxonomy	(adapted) MUC	CoNLL, ACE	CoNLL, ACE	Alchemy	DBpedia
Type	Java (GATE module)	Java	Python	Web Service	Web Service
License	GPLv3	GPLv2	GPLv3	Non-Commercial	Unknown
Adaptable	Yes	Yes	partially	No	No
	DBpedia Spotlight	TextRazor	Zemanta	YODIE	NERD-ML
Approach	Gazetteers and Similarity Metrics	Machine Learning	Machine Learning	Similarity Metrics	SMO and K-NN and Naive Bayes
Languages	EN	EN, NL, FR, DE, IT, PL, PT, RU, ES, SV	EN	EN	EN
Domain	Unspecified	Unspecified	Unspecified	Twitter	Twitter
# Classes	320	1779	81	1779	4
Taxonomy	DBpedia, Freebase, Schema.org	DBpedia, Freebase	Freebase	DBpedia	NERD
Type	Web Service	Web Service	Web Service	Java (GATE Module)	Java, Python, Perl, bash
License	Apache License 2.0	Non-Commercial	Non-Commercial		GPLv3
Adaptable	Yes	No	No	Yes	Partially

Figure 3.2. Key features of different NER systems
[62]

Without dimension reduction, data visualization results would be tough to find a connection. Figure 1 shows an example of a raw dataset clustering plot. The lines above and left try to establish the relationship between columns. But it is impossible to tell by intuition. We can choose from the box plot, heat map, cluster map, violin plot, strip plot, and regression plot.

4. RESULTS

4.1 SPSS Analysis

4.1.1 Participant demographics

Table 4.1. Demographics

Demographics	Groups	Frequency	Valid Percent	Cumulative Percent
Age	18-25	61	12.3	12.3
	26-35	250	50.5	62.8
	36-45	111	22.4	85.3
	46-55	41	8.3	93.5
	55+	32	6.5	100.0
Sex	Male	295	59.8	59.8
	Female	197	40.0	99.8
	Choose not to disclose	1	.2	100.0
Education Level	high School Graduate	44	8.9	8.9
	GED	21	4.3	13.2
	Some College	103	20.9	34.1
	Associate Degree: occupational program	33	6.7	40.8
	Associate Degree: academic program	24	4.9	45.6
	Bachelor's Degree	213	43.2	88.8
	Master's Degree	54	11.0	99.8
	Choose not to disclose	1	.2	100.0
Yearly income	Less than $15000	46	9.3	9.3
	$15000 - $24999	64	13.0	22.3
	$25000 - $34999	83	16.8	39.1
	$35000 - $49999	96	19.4	58.5
	$50000 - $74999	118	23.9	82.4
	$75000 - $99999	52	10.5	92.9
	$100000 - $149999	26	5.3	98.2
	$150000 - $199999	2	.4	98.6
	$200000+	3	.6	99.2
	Choose not to disclose	4	.8	100.0

This research collects survey data from 516 participants. After the procedure of data cleaning, the number of effective data is 495. As it shows in the table 4.1, the environmental factors includes age, sex, education level, yearly income, whether the participants work or study in information technology related field, how long do participants spend on a computer at work or at school on an average day, how long do participants spend on a computer at home on an average day, and how long do participants spend on a smartphone or tablet on an average day. In addition, most of the participants are under the age of 35, and the ratio of male to female is 4:6. The proportion of the participants working in IT is 70%, which is higher than we expected. Most participants complete the Bachelor degree, and the second proportion make some college degree. The annual income of most people is mainly distributed in the range of $15,000 to $99,999. Nearly half of people spend time on computer at work more than 5 hours a day. On the other hand, the participants have different distribution when they at home. And 247 people spend 1-3 hours using a smartphone at home.

Table 4.2. Environmental Factors

Environmental factors	Groups	Frequency	Valid Percent	Cumulative Percent
Work/study in IT	No	152	30.9	30.9
	Yes	340	69.1	100.0
computer time at work	less than one hour	25	5.1	5.1
	1-3 hours	68	13.8	18.8
	3-5 hours	88	17.8	36.6
	more than 5 hours	269	54.5	91.1
	never	44	8.9	100.0
computer time at home	less than one hour	21	4.3	4.3
	1-3 hours	165	33.5	37.7
	3-5 hours	143	29.0	66.7
	more than 5 hours	162	32.9	99.6
	never	2	.4	100.0
Smartphone/tablet time	less than one hour	78	15.8	15.8
	1-3 hours	247	49.9	65.7
	3-5 hours	112	22.6	88.3
	more than 5 hours	51	10.3	98.6
	never	7	1.4	100.0

4.1.2 Descriptive Statistics

Table 4.2 is the frequency and percentage results of the environmental factors. There are very few people in some groups, and subsequent analysis may not provide meaningful results. In the meantime, this table will also be used in the next subsection, supporting our analysis.

Table 4.3. Statistics for SEA

	N	minimum	maximum	mean	Std. Deviation
Phish	463	1.00	5.00	3.852	0.643
Vishing	444	2.44	5.00	3.852	0.488
Ad fraud	455	1.60	4.60	3.507	0.455

Table 4.3 is the descriptive analysis of phishing, vishing, and ad fraud, including number, minimum, maximum, mean, and standard deviation. From the table, we can see that the averages of these three groups are all numbers above 3, which is corresponding to our assumption. When we design the questions, the majority are fraud scenario. In consequence the correct average value of how likely these questions are vishing/phishing/ad fraud should be larger than 3. In addition, the standard deviation value shows the extend of deviation of a group. The larger the number, the more scattered this set of data is, and it is also evidence that the data may be more irregular.

The table from 4.4 to 4.11 are results from mean value and standard deviation value of phshing, vishing and ad fraud for each group. The assumption for these 8 environmental factors is that different groups may have variant response, so the mean value and S.D. value should be different. And based on the number, we will try to find relationship between groups. On top of this, in the next subsection we also use ANOVA analysis to support the conclusion.

Table 4.4. Different age groups affect SEA performance

Groups	Phishing Mean	Phishing S.D.	Vishing Mean	Vishing S.D.	ad fraud mean	ad fraud S.D.
18-25	3.7812	0.56777	3.6057	0.57362	3.3930	0.48839
16-35	3.7575	0.69140	3.8150	0.49041	3.5290	0.46455
36-45	3.9648	0.59032	3.9524	0.44957	3.5717	0.43143
46-55	3.9844	0.57956	4.0171	0.30433	3.4300	0.42859
55+	4.1875	0.45833	4.0115	0.43375	3.4429	0.39009

From table 4.4, we found that the three groups over the age of 36 have relatively higher average values, which means that they may be more alert and able to better recognize SEA. At the same time, each group's S.D. value does not perform very abnormally. And according to Table 1, the number of people in each group is not particularly small, so here we can assume that aging will make people more vigilant about SEA.

Table 4.5. Different sex groups affect SEA performance

Groups	Phishing Mean	Phishing S.D.	Vishing Mean	Vishing S.D.	ad fraud mean	ad fraud S.D.
Male	3.8920	0.62071	3.8148	0.51160	3.4879	0.45814
Female	3.8025	0.65868	3.9059	0.44562	3.5408	0.45303
Choose not to disclose	3.2500				3.2000	

According to Table 4.5, we can see that the gap between men and women is very small. Males performed better in phishing, but females performed better in the other two groups. Therefore, we need more evidence to prove that gender has an impact on the performance of people facing SEA. Judging from this result, the impact is not significant.

From Table 4.6, we can see that there is no absolute best performer among all the groups. Generally, people think that the higher the education level, the better at detecting malicious attacks, but from this table we can see that the facts are not exactly like this. Each group

Table 4.6. Different education level groups affect SEA performance

Groups	Phishing Mean	Phishing S.D.	Vishing Mean	Vishing S.D.	ad fraud mean	ad fraud S.D.
High School Graduate	3.9968	0.61871	3.8827	0.48173	3.5385	0.49023
GED	3.7632	0.56962	3.9012	0.52378	3.4444	0.42040
Some College	3.9488	0.52588	3.8822	0.463899	3.6303	0.47669
Associate Degree: occupational program	3.8712	0.61831	3.8459	0.52718	3.5636	0.43719
Associate Degree: academic program	4.0298	0.63234	3.9259	0.41026	3.4455	0.44905
Bachelor's Degree	3.7594	0.70632	3.8002	0.49971	3.4625	0.45193
Master's Degree	3.8673	0.63143	3.9051	0.49848	3.449	0.39637
Choose not to disclose	3.25				3.2000	

performs differently when facing phishing, vishing and ad fraud. So we can only temporarily judge that the impact of education level is not obvious.

Table 4.7. Different yearly income groups affect SEA performance

Groups	Phishing Mean	Phishing S.D.	Vishing Mean	Vishing S.D.	ad fraud mean	ad fraud S.D.
Less than $15000	4.1447	0.32089	3.9571	0.46541	3.5190	0.46130
$15000 - $24999	3.9877	0.60497	3.8249	0.51093	3.5000	0.47657
$25000 - $34999	3.7418	0.77871	3.8164	0.49415	3.5653	0.48141
$35000 - $49999	3.7583	0.65257	3.8458	0.45911	3.5200	0.49519
$50000 - $74999	3.8153	0.65731	3.8660	0.46762	3.5245	0.42265
$75000 - $99999	3.8828	0.55866	3.8420	0.49435	3.4170	0.42901
$100000 - $149999	3.7200	0.65678	3.7169	0.64120	3.3846	0.40762
$150000 - $199999	4.0000	0.0000	3.7222	0.39284	3.5000	0.42426
$200000+	4.4167	0.38188	4.0370	0.32075	3.6667	0.30551
Choose not to disclose	3.8333	0.56366	4.3333	0.15713	3.3333	0.11547

The opposite is the performance of Table 4.7. We found that the two groups with income above $150,000 in this table performed significantly better. But we can't directly judge that income level will have a significant impact. According to Table 4.1, we know that the number of people in these two groups is less than ten. It is difficult for us to confirm this conclusion when the amount of data is not enough.

Table 4.8. Whether work or study in IT affect SEA performance

Groups	Phishing Mean	Phishing S.D.	Vishing Mean	Vishing S.D.	ad fraud mean	ad fraud S.D.
No	3.7066	0.64238	3.6920	0.51939	3.4085	0.44521
Yes	3.9272	0.62563	3.9242	0.45689	3.5524	0.45414

According to table 4.8, we can see that obviously, the performance of group Yes is better than that of group No, and the difference is basically between 0.1 and 0.3. So, we can judge that people who work or study in information technology related area perform better than the people who don't.

Table 4.9. Computer time at work affect SEA performance

Groups	Phishing Mean	Phishing S.D.	Vishing Mean	Vishing S.D.	ad fraud mean	ad fraud S.D.
less than one hour	4.0990	0.45591	4.556	0.39358	3.4957	0.52525
1-3 hours	3.7637	0.53272	3.6508	0.54465	3.3788	0.50000
3-5 hours	3.6540	0.72417	3.7208	0.50982	3.4359	0.39967
more than 5 hours	3.8889	0.66242	3.8916	0.46387	3.5545	0.44359
never	4.0183	0.49494	4.0769	0.36153	3.5805	0.47287

Table 4.10. Computer time at home affect SEA performance

Groups	Phishing Mean	Phishing S.D.	Vishing Mean	Vishing S.D.	ad fraud mean	ad fraud S.D.
less than one hour	3.8988	0.50097	4.0167	0.53987	3.5300	0.56298
1-3 hours	3.8783	0.65526	3.8677	0.45782	3.4848	0.41469
3-5 hours	3.7639	0.67689	3.7760	.54342	3.4594	0.46777
more than 5 hours	3.9030	0.61037	3.8859	0.45011	3.5697	0.46820
never	4.4375	0.26517	4.0556	0.07857	3.7000	0.14142

From Table 4.9, we can see that there is a gap in the performance of each group. However it is difficult for us to conclude that as the time spent on the computer increases, people will be better able to identify fraud, so we need to conduct further analysis To determine whether the grouping has an effect on the results. But we also found that people who spend less than an hour tend to have better performance than other groups.

Table 4.11. Time spent on smartphone or tablet affects SEA performance

Groups	Phishing Mean	Phishing S.D.	Vishing Mean	Vishing S.D.	ad fraud mean	ad fraud S.D.
less than one hour	4.0845	0.49488	4.0432	0.37878	3.5750	0.41103
1-3 hours	3.8977	0.64112	3.9056	0.46430	3.5093	0.45433
3-5 hours	3.6719	0.69109	3.6909	0.52766	3.4796	0.49336
more than 5 hours	3.6436	0.53201	3.6591	0.50912	3.5174	0.42703
never	3.9821	0.46451	3.7302	0.47944	3.0857	0.36253

The same phenomenon also occurs in Table 4.10 and Table 4.11. The two best performance groups are people spend less than one hour and people never spend time on computer and smartphone.

4.1.3 ANOVA analysis and results

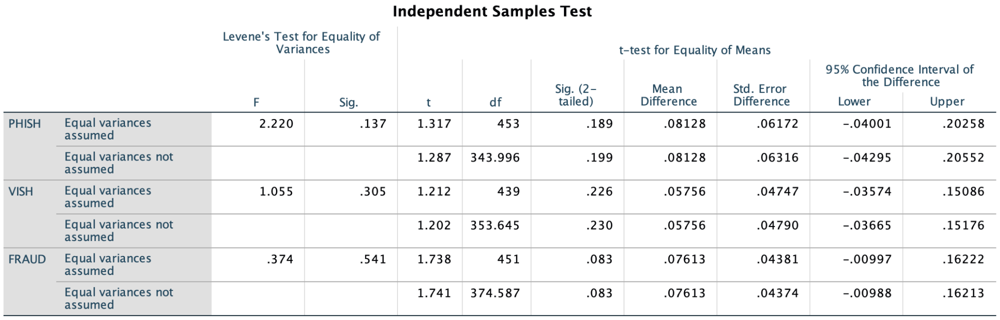
Table 4.12. ANOVA significance value between Phishing, Vishing, and ad fraud against factors

Factors	Phishing	Vishing	ad fraud
AGE	0.001	0.000	0.098
SEX	0.215	0.054	0.386
EDUCATION	0.127	0.697	0.103
YEARLY INCOME	0.024	0.657	0.742
STUDY/WORK IN IT	0.001	0.000	0.002
COMPUTER TIME AT HOME	0.003	0.000	0.027
COMPUTER TIME AR WORK	0.239	0.160	0.283
SMARTPHONE/TABLAT TIME	0.000	0.000	0.091

In this section, we will support our conclusion based on the perspective statistic analysis in the previous section using ANOVA. ANOVA can help us find out whether we can accept or reject the hypothesis we proposed. For our research, the hypothesis is that there is no difference between environmental factors' groups. If the p-value of ANOVA is less than 0.05, we will reject the hypothesis, and assume that there is significant difference between groups.

Table 4.12 is the result table to show significance value of phishing, vishing, and ad fraud based on different groups of age, sex and so on. From all the values, we can conclude that age, whether work in IT related area, the time spend on computer at home, and the time spent on smartphone at home are the factors have significant difference. From section 4.1.2, we also propose that age groups and whether work in IT groups have consistent performance. For the other two factors, even we cannot conclude that with the increasing of the time, the participants are better at recognizing SEA. We still can assume that these two factors can affect people's performance.

4.1.4 T-test analysis and results



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
PHISH	Equal variances assumed	2.220	.137	1.317	453	.189	.08128	.06172	-.04001	.20258
	Equal variances not assumed			1.287	343.996	.199	.08128	.06316	-.04295	.20552
VISH	Equal variances assumed	1.055	.305	1.212	439	.226	.05756	.04747	-.03574	.15086
	Equal variances not assumed			1.202	353.645	.230	.05756	.04790	-.03665	.15176
FRAUD	Equal variances assumed	.374	.541	1.738	451	.083	.07613	.04381	-.00997	.16222
	Equal variances not assumed			1.741	374.587	.083	.07613	.04374	-.00988	.16213

Figure 4.1. An Example of SPSS indepent samples test result

ANOVA and T-test are both used to determine whether there are difference among several groups. ANOVA is often used when the groups number is greater than two, and T-test is often used within two groups. In this section, we will analyze whether there is difference between two groups and propose hypothesis. We set the Confidence Interval value to 0.95, which means if the p-value is less than 0.05 we reject the hypothesis and assume there is difference between groups. In addition, the results of Confidence Interval of the Difference also can complement the significance test results. ”If the CI for the mean difference contains 0, the results are not significant at the chosen significance level.”[63] We use SPSS Statistics tool to perform independent sample test. Figure 4.1 is an example output of SPSS, which has both Levene’s test and t-test at the same time. Levene’s test is the test for equality of variances. If the p-value of this test is larger than 0.05, then we assume there is no variance difference and use the first row for t-test result. If the p-value is less than 0.05, then we assume there is variance difference and use the second row for t-test results. The p-value of t-test measures the equality of means. If the p-value of the test result is less than 0.05, we can reject the hypothesis and assume there is mean difference between groups.

Hypothesis: whether an individual ever had a virus on their machine cannot affect SEA performance

The difference of mean value between yes group and no group is less than 0.1. The p-values of the Levene’s test on phishing(F = 2.220,p = 0.137), vishing(F = 1.055,p = 0.305), and ad fraud(F = 0.374,p = 0.541) are all greater than 0.05. This means that equal variances is assumed, we need to use the first row of the output. In conclusion, we cannot reject the hypothesis, and assume that there is no variance difference between two groups. And then we look at the first row of t-test for equality of means. The two-tailed p-values are also larger than 0.05 as determined by t-test for phishing(t453 = 1.317 ,p = 0.189), vishing(t439 = 1.212,p = 0.226), and ad fraud(t451 = 1.738,p = 0.083). And the confidence intervals of the mean difference also contain 0 for all three of them. In conclusion, we state that there is no significant difference in SEA performance between people who had a virus on their machine and who don’t.

Hypothesis: whether an individual ever had a ransomware on their machine cannot affect SEA performance

The difference of mean value between yes group and no group is between 0.2 to 0.4. The p-values of the Levene’s test on phishing(F = 3.890,p = 0.049) is smaller than 0.05. But the p-values of the Levene’s test on vishing(F = 3.125,p = 0.078), and ad fraud(F = 0.993,p

= 0.320) are both greater than 0.05. In conclusion, we cannot reject the hypothesis, and assume that there is no variance difference between two groups. And then we look at the first row of t-test for equality of means. The two-tailed p-values are smaller than 0.05 as determined by t-test for phishing(t118.025 = -4.348,p < 0.001), vishing(t439 = -5.105,p < 0.001), and ad fraud(t451 = -2.931,p = 0.004). And the confidence intervals of the mean difference does not contain 0 for all three of them. In conclusion, we state that there are significant mean differences in SEA performance between people who ever had a ransomware on their machine and who don't.

Hypothesis: whether an individual's account ever been hacked cannot affect SEA performance

The difference of mean value between yes group and no group is between 0.1 to 0.2. The p-values of the Levene's test on phishing(F = 2.875,p = 0.091), vishing(F = 0.472,p = 0.493), and ad fraud(F = 0.005,p = 0.943) are all greater than 0.05. This means that equal variances is assumed, we need to use the first row of the t-test output. In conclusion, we cannot reject the hypothesis, and assume that there is no variance difference between two groups. And then we look at the first row of t-test for equality of means. The two-tailed p-values of phishing(t453 = 3.078,p = 0.002) and vishing(t439 = 2.782,p = 0.006) are smaller than 0.05. And the confidence intervals of the mean difference does not contain 0 for both of them. The two-tailed p-value of ad fraud(t451 = 1.117,p = 0.265) is larger than 0.05. And the confidence intervals of the mean difference contain 0. In conclusion, we state that there are significant differences in recognizing phishing and vishing but not in recognizing ad fraud between people whose account ever been hacked and who don't.

Hypothesis: whether an individual ever shared credentials for any password protected account cannot affect SEA performance

The difference of mean value between yes group and no group is between 0.1 to 0.3. The p-values of the Levene's test on phishing(F = 0.973,p = 0.324) and ad fraud(F = 0.577,p = 0.448) are both greater than 0.05. This means that equal variances is assumed, we need to use the first row of the output. The p-values of the Levene's test on vishing(F = 9.757,p =

0.002) is less than 0.05. This means that equal variances is not assumed, we need to use the second row of the output. The two-tailed p-values of vishing(t100.130 = -3.795,p < 0.001) and ad fraud(t450 = -2.740,p = 0.006) are smaller than 0.05, but two-tailed p-value of phishing(t452 = -1.473,p = 0.141) is larger than 0.05. And the confidence intervals of the mean difference does not contain 0 for both of them. In conclusion, we state that there are significant differences in recognizing vishing and ad fraud but not phishing between people who shared credentials for any password protected account and who don't.

Hypothesis: whether an individual ever been a victim of identity theft cannot affect SEA performance

The difference of mean value between yes group and no group is between 0.1 to 0.2. The p-values of the Levene's test on phishing(F = 1.891,p = 0.170) and ad fraud(F = 0.013,p = 0.911) are both greater than 0.05. This means that equal variances is assumed, we need to use the first row of the output. The p-values of the Levene's test on vishing(F = 4.499,p = 0.034) is less than 0.05. This means that equal variances is not assumed, we need to use the second row of the output. The two-tailed p-values of phishing(t450 = -2.157,p = 0.032) and vishing(t105.053 = -3.555,p = 0.001) are smaller than 0.05. And the confidence intervals of the mean difference does not contain 0 for both of them. The two-tailed p-value of ad fraud(t448 = -1.181,p = 0.238) is larger than 0.05. In conclusion, we state that there are significant differences in recognizing phishing and vishing but not ad fraud between people who had been a victim of identity theft and who don't.

Hypothesis: whether an individual ever lost money of any online or phone scam cannot affect SEA performance

The difference of mean value between yes group and no group is between 0.2 to 0.5. The p-values of the Levene's test on phishing(F = 3.280,p = 0.071), vishing(F = 0.473,p = 0.492), and ad fraud(F = 2.349 ,p = 0.126) are all greater than 0.05. This means that equal variances is assumed, we need to use the first row of the output. The two-tailed p-values of phishing(t452 = -3.373,p = 0.001), vishing(t438 = -7.350,p = 0.000) and ad fraud(t450 = -2.982,p = 0.003) are all smaller than 0.05. And the confidence intervals of the mean difference does not contain 0 for all of them. In conclusion, we state that there are significant differences in SEA performance between people who lost money of any online or phone scam and who don't.

Hypothesis: whether an individual ever been taught any computer science, cybersecurity or related topics cannot affect SEA performance

The difference of mean value between yes group and no group is between 0.01 to 0.2. The p-values of the Levene's test on vishing(F = 1.135,p = 0.287) and ad fraud(F = 1.126,p = 0.289) are both greater than 0.05. This means that equal variances is assumed, we need to use the first row of the output. The p-values of the Levene's test on phishing(F = 5.599,p = 0.018) is less than 0.05. This means that equal variances is not assumed, we need to use the second row of the output. The two-tailed p-values of phishing(t384.476 = -3.592,p = 0.000) and vishing(t434 = -2.167,p = 0.031) are smaller than 0.05. And the confidence intervals of the mean difference does not contain 0 for both of them. The two-tailed p-value of ad fraud(t446 = -0.382,p = 0.703) is larger than 0.05. In conclusion, we state that there are significant differences in recognizing phishing and vishing but not ad fraud between people who had been taught any CS related topics and who don't.

4.2 PCA and UMAP results

Section 4.1 more concentrate on individual analysis, which conducted analysis between environmental factors and one calculated performance column. In this section, by using data visualization and machine learning algorithms, we can analyze high-dimensional data. The purpose of this section is to find if there is an environmental factor that can influence the performance as a whole. By using PCA algorithm and UMAP algorithm, we can analyze the relationship between each environmental factor and the behavior of Q18 to Q41. Figure 4.2 and Figure 4.3 are results from PCA and UMAP, respectively. By combining two results, we are able to conclude whether this factor can affect user behavior and divide their performance into groups.

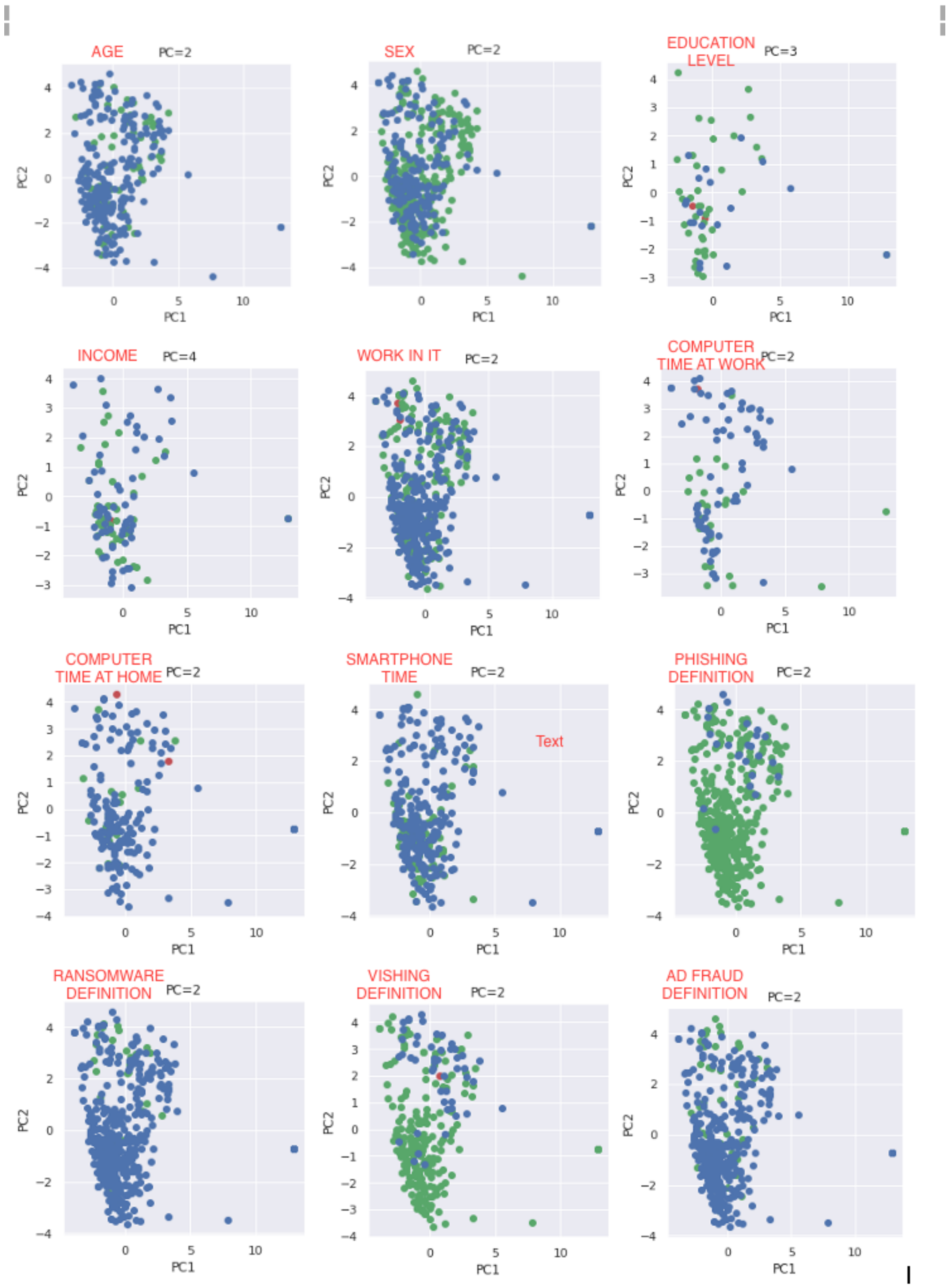


Figure 4.2. PCA result

Through the investigation of factors, we divided all the respondents into several groups. Q18 to Q41 survey people's ability of recognizing SEA. If we can easily group all data points by color from the output image, it means that this factor has an impact on the performance

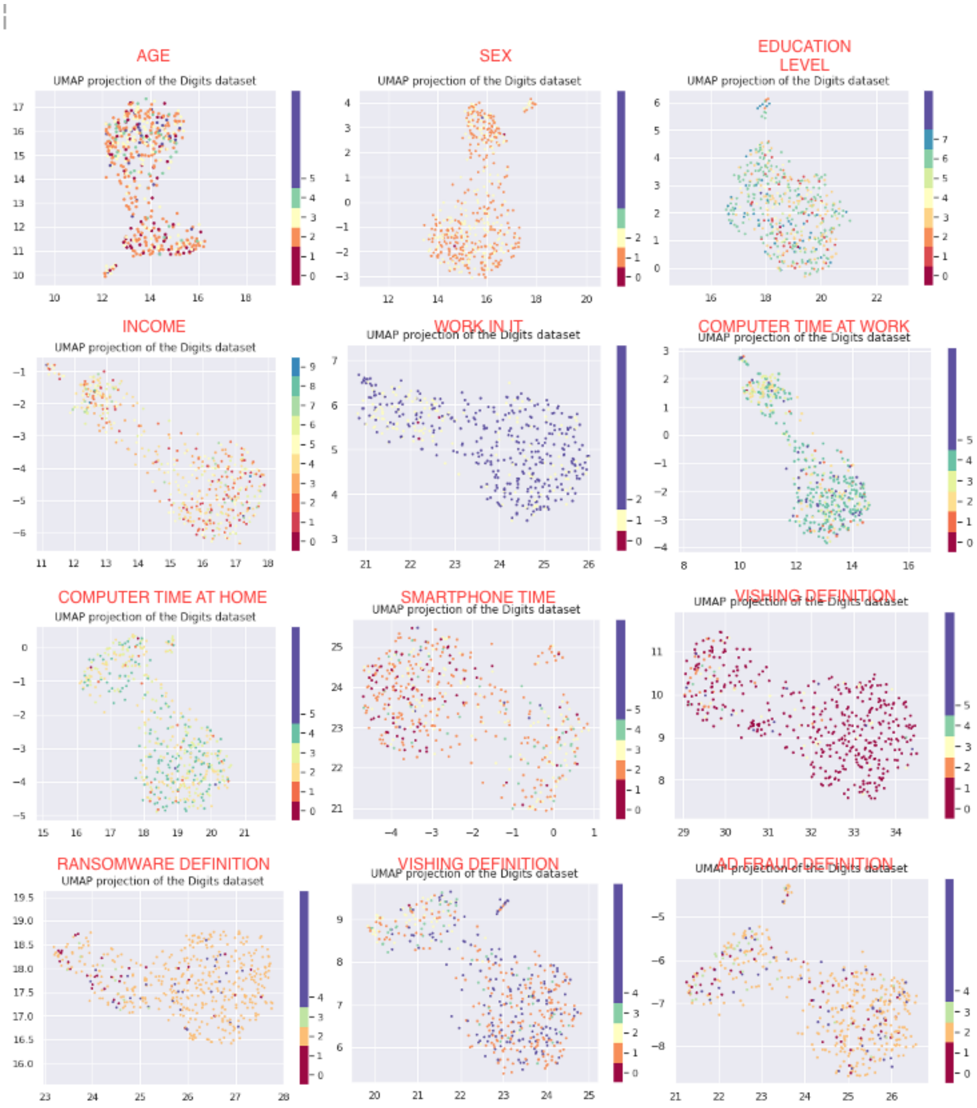


Figure 4.3. UMAP result

of people as a whole. For instance, in the result of age group, all the points of the two colors have no obvious different distribution areas. Their distribution areas basically overlap, so there is no obvious difference in performance in groups based on age. The opposite example is the result of phishing definition group, it is obvious that blue dots and green dots spread in different areas. In between, there are also fuzzy types, such as computer-time-at-work result. Although the distribution of the two colors overlaps a large part, there are some areas that only have blue dots but no green dots. In this case, we also believe that the computer time spent at work will affect people's ability to recognize SEA.

In consequence, from Figure 4.2, we conclude that the different time people spent on computer at work, the definition of phishing they choose, the definition of ransomware, and the definition of vishing are all the factors that can influence an individual's performance towards social engineering attacks.

The main function of Figure 4.3 is to corroborate and supplement our previous conclusions. The advantage of UMAP over PCA is that it can show more details, but the disadvantage is that it interferes with more judgments. Therefore, we will not use UMAP results to refute our previous conclusions, but rather add some details. From the result of work-in-IT group we found that, the yellow dots are more gathered on the left, and the purple dots are more gathered on the right. From the result of smartphone-time group we found that, the red dots mainly distribute on the left. These are all the evidence that whether an individual ever works in IT related area, and the time he/she spent on smartphone can affect an individual's performance towards social engineering attacks.

4.3 Text Analysis

Figure 4.4 is an example after using text analysis tools. Although the extraction results produced by the text analysis well summarize the keywords in the input documents, they are abstract and difficult to understand. We thus developed a visualization for presenting the overall structure and information exchange within the input scenario comprehensively in a single view. Because the input text contains not only the scene of the event itself, but also the background information related to the event including character characteristics and character relations. And there is no doubt that the event background often affects people's judgment of the credibility of the event scene. Thus in order to improve the understand ability of visualization, a two-level structure visualization is adopted to present the event background information and event scene information in the text so that the viewer can make a more comprehensive judgment of the event credibility, based on the analysis of the correlation between the two kinds of information.

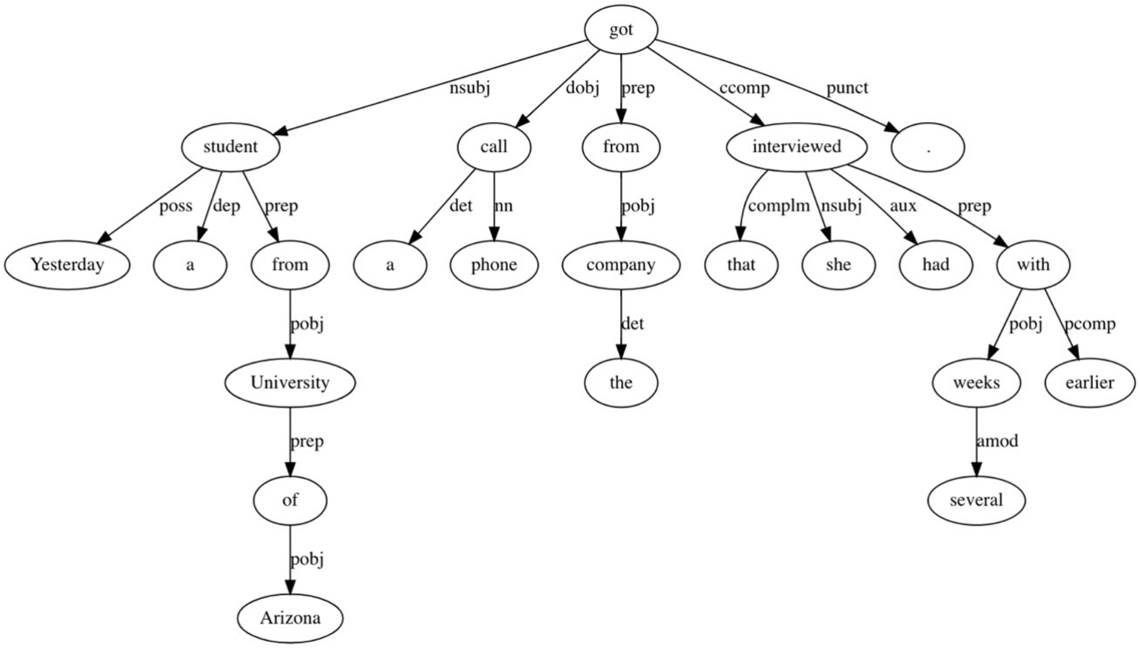


Figure 4.4. Sentence Structure Example

The visualization consists of two parts that correspond to two levels of information in the text. The first is the event background information, which is represented in the visualization by a gray circle in the background. This section corresponds to the text content that describes the basic information about the event. The content is to describe the background information before the event scene, such as the background of the character, the activity of the character, the origin of the event, etc. The second part is the event scene information, which corresponds to the main scene of the information exchange described in the input text and is also the information that the viewer mainly relies on to make the judgment. In visualization, the presentation of event scene information is located in the center part.

In addition, the order of information exchange in an event also affects the credibility of the scene. Thus, the vertical in the visualization represents the time dimension. The one at the top of the visualization represents the event that happened first, or the information that was passed first.

Based on the content of the input text. The person in the event can be divided into the initiator (or the caller) of the scene and the receiver of the scene. The initiator of an event

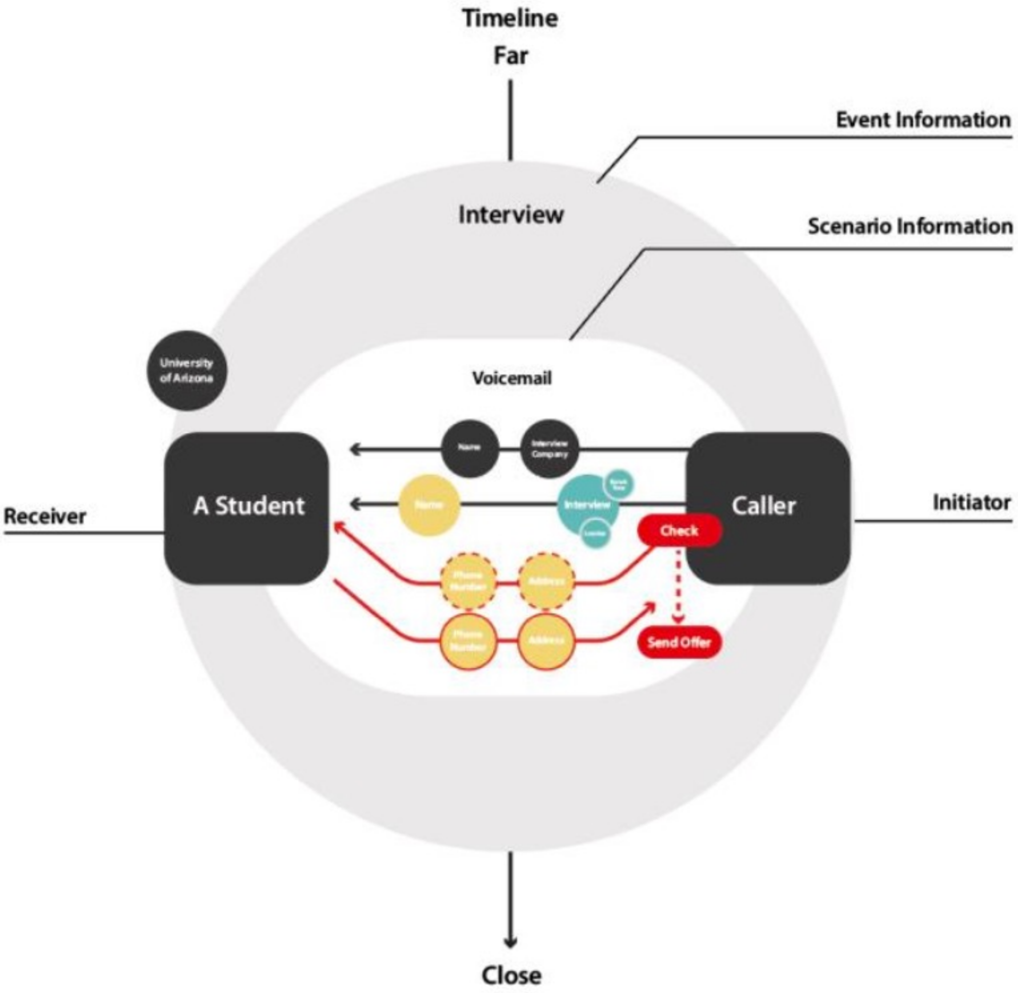
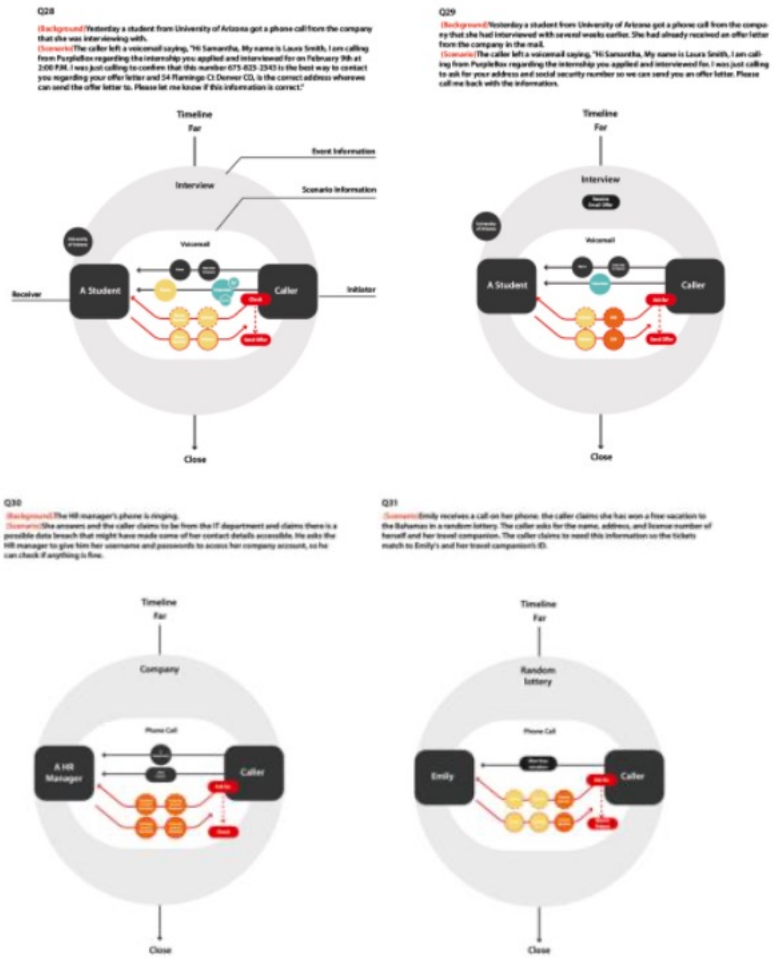


Figure 4.5. Text visualization framework example

scene can be understood as the person that makes the scene happen and the one that the viewer needs to identify its credibility. In the visualization, the two characters are displayed through black rounded squares, and are located on the left and right sides of the event scene information section.

Between the initiator and the receiver, the black arrow and the circle and rounded rectangle on the arrow together show the exchange of information in the event scene. The black arrow indicates that the message is sent only in one direction and does not require any subsequent action by the recipient. The direction of the black arrow indicates the direction in which the information is transmitted. There are two types of information delivered, information about events and information about people. Information about events is displayed in a rounded rectangle. Information about people is displayed in a circle, and the sensitivity of the information is distinguished by color. The colors are green, yellow and orange. Green represents low-sensitive personal information, yellow represents general personal information, such as name and address, etc., and orange represents highly sensitive personal information, such as SSN, bank account, password, etc. By visualizing the direction of information transmission and distinguishing the information sensitivity, the observer can quickly identify the information flow pattern, amount of information and information categories in the event scene.

In addition to the exchange of information, there are often target actions that need to be



Q28

(Background)Yesterday a student from University of Arizona got a phone call from the company that she was interviewing with.
(Scenario)The caller left a voicemail saying, "Hi Samantha, My name is Laura Smith, I am calling from PurpleBox regarding the internship you applied and interviewed for on February 9th at 2:00 P.M. I was just calling to confirm that this number 675-825-2345 is the best way to contact you regarding your offer letter and 54 Flamingo Ct Denver CO, is the correct address where we can send the offer letter to. Please let me know if this information is correct."

Q29

(Background)Yesterday a student from University of Arizona got a phone call from the company that she had interviewed with several weeks earlier. She had already received an offer letter from the company in the mail.
(Scenario)The caller left a voicemail saying, "Hi Samantha, My name is Laura Smith, I am calling from PurpleBox regarding the internship you applied and interviewed for. I was just calling to ask for your address and social security number so we can send you an offer letter. Please call me back with the information.

Q30

(Background)The HR manager's phone is ringing.
(Scenario)She answers and the caller claims to be from the IT department and claims there is a possible data breach that might have made some of her contact details accessible. He asks the HR manager to give him her username and passwords to access her company account, so he can check if anything is fine.

Q31

(Scenario)Emily receives a call on her phone. the caller claims she has won a free vacation to the Bahamas in a random lottery. The caller asks for the name, address, and license number of herself and her travel companion. The caller claims to need this information so the tickets match to Emily's and her travel companion's ID.

Figure 4.6. The rest of text visualization

performed in the event scenarios that are described by input text. For example, the initiator requires the receiver to provide some personal information or asks the receiver to perform certain behaviors. The target action to be performed in these scenarios is also an important reference for judging the credibility of the event scenario. Thus, in visualization, the target action is performed independently of the exchange of information but is presented in conjunction with the exchange of information to assist the observer in joint analysis.

Specifically, the action information in the event scene is presented by red arrows and the circle with red edges. The direction of the arrow represents the direction in which the action information is transmitted. Labels at the beginning of the arrows represent categories of action information, such as “ask for,” “check,” and so on. Unlike before, the information with the dotted edge represents the specific content without the specific content. In the example above, for example, the red arrows that from the caller to the student represent that the caller put forward an action information. And the information on the line is “phone number” and “address” with dotted line edge, which means that the caller only proposed that the two types of information should be checked but did not provide the specific content of these two information, that is, the specific phone number and address information. To demonstrate the usefulness of the developed visualization method, we have applied it to different types of event text: the event with interview background the event with bank affair background, and the other two types of event. In summary, this approach can be applied to the visualization of text describing different background event scenarios. Compared with the text display method, the visualization method has the following advantages. First, it can quickly perceive the structure of the event scene, which is helpful to help the observer form an overall impression of the scene. In addition, in the details of the event scene, different visualization methods, such as layout, color contrast, etc., are used to let the observer check the specific information exchange.

5. DISCCUSION

The purpose of this study is to identify the relationship between environmental factors and participants' ability of recognizing social engineering attacks. To address this question, we conduct basic descriptive statistics analysis, ANOVA analysis, propose and test hypothesis using t-test, and machine learning algorithms analysis. In order to collect data, we used Qualtrics to build and spread questionnaire. There were 517 people who take the survey in total, after data cleaning, the number decreased to 497. Our conclusion is based on data analysis and verification of hypothesis.

The environmental factors include age, sex, education level, yearly income, whether work in IT related area, the time spent on computer at work, the time spent on computer at home, the time spent on smartphone. To conclude whether these environmental factors groups have different performance towards how likely one scenario can be a social engineering attack, we combine basic descriptive analysis and ANOVA which is often used for determining the difference between groups. In addition, we also propose the hypothesises based on individual's past experience and their performance on social engineering attacks. Since the ANOVA is used for groups more than two, t-test is used for determining whether there are variance and mean difference between two groups. PCA and UMAP are machine learning algorithms that visualize the high-dimensional data, and judge the difference between groups to support the former conclusion.

This study is based on social cognitive theory, which states that individuals' behaviors are based on their environment and past experience other than motivation and emotion. According to [1] there are three main factors that can affect human behaviors, including cognitive factors, environmental factors, and behavioral factors. The cognitive/personal factors include knowledge, expectations, and attitudes. The environmental factors include social norms, access in community, and influence on others. The behavioral factors include skills, practice, and self-efficiency.Since past experience can has significant consequence on individuals' behavior, we also collected the data from their past social engineering attack related experience. These questions include: a) Have you ever had a virus/ransomware on their machine. b) Have any of your accounts ever been hacked or compromised? c) Have you ever willingly shared credentials (username/password/answers to security questions) for any password protected account? d) Have you ever been a victim of identity theft? e) Have you ever lost money as a result of any online or phone scam? f) Were you taught any computer science, cybersecurity or related topics from the ages of 5-18?

The results from descriptive statistics suggest that different age groups have obvious different mean values towards phishing, vishing, and ad fraud. At the same time, with the increasing of age, the performance of the ability of recognizing social engineering attacks is getting better. Esposito states that older participants tend to perform more safely towards online security. However the results of [64] shows the opposite, they argued that age cannot affect the treatments' performance. This maybe because this research more focus on the result of education, during which age cannot affect people's learning ability. In addition, 80 percent of people who take our questionnaire are under 45, which could be a factor that why we have different conclusion.

[35] shows that males tend to have better awareness of online safety. But in our study, the male group and the female group have no obvious mean difference in identifying SEA, and even women have better performance on some questions. In the ANOVA results, we can only accept the hypothesis that the two groups have no obvious mean difference and variance difference. At the same time, the results of PCA and UMAP also indicate that the performance of the two groups are similar.

[35] not only pointed out that gender affects people's awareness of cyber security, IT-related majors tend to have better performance in cyber security. In the meantime, it is also indicated that students at higher study level performs better than the students at lower study level. In our study, the participants who work or study in IT related area are better at recognizing a social engineering attack. The ANOVA results also supported this finding by suggesting that there is significant different between two groups. This conclusion is also supported by UMAP result. However, [38] stated that sometimes education can reduce the chance of an individual of recognizing social engineering attacks. Our study shows that it is true participants with higher education level does not perform better. According to ANOVA results, there is no significant difference between each group. Same as the UMAP and PCA results. In addition, yearly income level can also be highly related to the education level,

which is the reason why we also include the income level as an environmental factor. If we exclude groups with less than 10 people, we find that the groups with the least and the most salaries have relatively better ability to recognize cybercrime. In other words, it is not that people with higher income will perform better. The ANOVA results and PCA results also suggest the same thing. This finding is also consistent with the education level.

The ANOVA results from this research indicate that the time spent on computer at home and the time spent on smartphone can affect the SEA performance. Although through basic descriptive statistics, we cannot conclude that people who spend more time perform better or worse, that is to say, there is no clear linear relationship. However, according to UMAP, we can also find that there are obvious differences in dispersal areas between these groups. We can infer that these two factors can affect people's response to SEA. In the meanwhile, the time spent on computer at work cannot affect participants' performance. This finding also correspond to the [39] which illustrated that people at home have different performance from work.

[42] concluded that people who have been victims of cybercrime before have lower awareness. In the 4.1.4 section, we focus on whether an individual's past experience can influence his/her response towards recognizing social engineering attacks. According to t-test results, participants' past experience against cyber crime makes them perform differently when it comes to phishing, vishing, and ad fraud.Since there is research proved that by educating participants the online security awareness can be trained[37][40], we also applied t-test to analyze between the people who has been taught CS related topic and the people who has not. The result shows that two groups have different performance towards phishing and vishing.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

In this research, we proved that age, whether work in IT related area, the time spent on computer at home, time spent on computer at work, and time spent on smartphone are the factors that can affect participants' performance of recognizing social engineering attacks. Also, an individual's past experience can also affect his/her ability of recognizing SEA, including whether an individual ever shared credentials, ever been a victim of identity theft, ever lost money of scam, and ever been taught any computer science related topics. Because the questions that measure participants' performance are divided into three parts: phishing, vishing, and ad fraud, the influence of past experience also measured in three aspects. For example, people had shared credentials for password protected account or not only has significant differences on vishing and ad fraud. Among all the questions that related to past experience, only "whether an individual ever had a ransomware on their machine" and "whether an individual even lost money of any online or phone scam"has influence on all three (phishing, vishing, ad fraud). In the section of text analysis, we proposed a framework that can visualize the questions about a possible social engineering attack scenario, which was helpful for relate the keywords to how likely this question is SEA.

6.2 Future work

Some of environmental factors only have influence on partial SEA, more questions can be designed in the future to validate the results. The size of some groups is very small, more participants are needed in the future. For example, we can increase the number of people over 45. Most of the participants are Americans, and the geographical scope of participants can be expanded in the future. The reason we added text analysis is to try to find out why there are some scenarios that make participants more vigilant. By visualizing keywords and relationships, it is found that people are more guarded against giving information than receiving information. At the same time, we also found that even if the participant is required to provide sensitive information, a large number of people do not think it is a fraud. At the same time, we found that on average, everyone thinks that the problem of SEA is most likely to ask for personal information rather than bank information, which is obviously different from what we expected. After visual analysis, we speculate that it may be because winning lottery is involved in this scenario. So in the future, if the reason of choice can be further analyzed, for example, add a mechanism that allows participants to click on keywords, it will be of great help to this research.

REFERENCES

[1] F. Mouton, L. Leenen, and H. S. Venter, "Social engineering attack examples, templates and scenarios," *Computers & Security*, vol. 59, pp. 186–209, Jun. 1, 2016, ISSN: 0167-4048. DOI: 10.1016/j.cose.2016.03.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404816300268.

[2] F. Mouton, M. M. Malan, L. Leenen, and H. S. Venter, "Social engineering attack framework," in *2014 Information Security for South Africa*, ISSN: 2330-9881, Aug. 2014, pp. 1–9. DOI: 10.1109/ISSA.2014.6950510.

[3] F. Mouton, L. Leenen, M. M. Malan, and H. S. Venter, "Towards an ontological model defining the social engineering domain," in *ICT and Society*, K. Kimppa, D. Whitehouse, T. Kuusela, and J. Phahlamohlaka, Eds., ser. IFIP Advances in Information and Communication Technology, Berlin, Heidelberg: Springer, 2014, pp. 266–279, ISBN: 978-3-662-44208-1. DOI: 10.1007/978-3-662-44208-1_22.

[4] V. Thomas, "Chapter 7 - social engineering," in *Building an Information Security Awareness Program*, B. Gardner and V. Thomas, Eds., Boston: Syngress, Jan. 1, 2014, pp. 45–63, ISBN: 978-0-12-419967-5. DOI: 10.1016/B978-0-12-419967-5.00007-7. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124199675000077.

[5] S. M. Albladi and G. R. S. Weir, "User characteristics that influence judgment of social engineering attacks in social networks," *Human-centric Computing and Information Sciences*, vol. 8, no. 1, p. 5, Feb. 28, 2018, ISSN: 2192-1962. DOI: 10.1186/s13673-018-0128-7. [Online]. Available: https://doi.org/10.1186/s13673-018-0128-7.

[6] A. Algarni, Y. Xu, and T. Chan, "An empirical study on the susceptibility to social engineering in social networking sites: The case of facebook," *European Journal of Information Systems*, vol. 26, no. 6, pp. 661–687, Nov. 2017, ISSN: 0960-085X, 1476-9344. DOI: 10.1057/s41303-017-0057-y. [Online]. Available: https://www.tandfonline.com/doi/full/10.1057/s41303-017-0057-y.

[7] C. Iuga, J. R. C. Nurse, and A. Erola, "Baiting the hook: Factors impacting susceptibility to phishing attacks," *Human-centric Computing and Information Sciences*, vol. 6, no. 1, p. 8, Jun. 10, 2016, ISSN: 2192-1962. DOI: 10.1186/s13673-016-0065-2. [Online]. Available: https://doi.org/10.1186/s13673-016-0065-2.

[8] G. Saridakis, V. Benson, J.-N. Ezingeard, and H. Tennakoon, "Individual information security, user behaviour and cyber victimisation: An empirical study of social networking users," *Technological Forecasting and Social Change*, vol. 102, pp. 320–330, Jan. 1, 2016, ISSN: 0040-1625. DOI: 10.1016/j.techfore.2015.08.012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0040162515002590.

[9] R. T. WRIGHT and K. MARETT, "The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived," *Journal of Management Information Systems*, vol. 27, no. 1, pp. 273–303, 2010, Publisher: M.E. Sharpe, Inc., ISSN: 0742-1222. [Online]. Available: http://www.jstor.org/stable/25699619.

[10] Ashish Thapar. (2016). SOCIAL ENGINEERING an attack vector most intricate to tackle, [Online]. Available: https://www.coursehero.com/file/27129635/Social-Engineering-AThaparpdf/.

[11] C. Hadnagy, *Social Engineering*, 1st ed. John Wiley & Sons, Ltd, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119433729. DOI: 10.1002/9781119433729. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/9781119433729.

[12] (2020). Vishing and cybercriminals during COVID-19, [Online]. Available: https://www.securitymagazine.com/articles/92277-vishing-and-cybercriminals-during-covid-19?v=preview.

[13] (). Spam | definition of spam by merriam-webster, [Online]. Available: https://www.merriam-webster.com/dictionary/spam.

[14] F. Breda, H. Barbosa, and T. Morais, "SOCIAL ENGINEERING AND CYBER SECURITY," Mar. 1, 2017, pp. 4204–4211. DOI: 10.21125/inted.2017.1008.

[15] Raj Samani. (Apr. 13, 2015). Hacking the human operating system, McAfee Support Community. Section: Security Awareness Documents, [Online]. Available: https://community.mcafee.com/t5/Security-Awareness-Documents/Hacking-the-Human-Operating-System-Raj-Samani/ta-p/550808.

[16] (Dec. 31, 2020). 10 cyber security trends you can't ignore in 2021, PurpleSec, [Online]. Available: https://purplesec.us/cyber-security-trends-2021/.

[17] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," *Future Internet*, vol. 11, no. 4, p. 89, Apr. 2019, Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/fi11040089. [Online]. Available: https://www.mdpi.com/1999-5903/11/4/89.

[18] (Nov. 8, 2020). 2019 cyber security statistics trends & data, PurpleSec, [Online]. Available: https://purplesec.us/resources/cyber-security-statistics/.

[19] AranaM. (Oct. 24, 2017). How much does a cyberattack cost companies? | open data security, ODS - Cybersecurity, [Online]. Available: https://opendatasecurity.co.uk/how-much-does-a-cyberattack-cost-companies/.

[20] A. Berg, "Cracking a social engineer," *Computers & Security*, vol. 8, no. 14, p. 700, 1995, ISSN: 0167-4048. [Online]. Available: https://www.infona.pl//resource/bwmeta1.element.elsevier-80e104e1-fd04-39e8-b957-5b80efca285d.

[21] S. Granger, "Social engineering fundamentals, part i: Hacker tactics," p. 17, Dec. 18, 2001.

[22] D. Airehrour, N. Nair, and S. Madanian, "Social engineering attacks and countermeasures in the new zealand banking system: Advancing a user-reflective mitigation model," *Information*, vol. 9, p. 110, May 3, 2018. DOI: 10.3390/info9050110.

[23] N. J. Evans, "Information technology social engineering: An academic definition and study of social engineering - analyzing the human firewall," Pages: 2806871, Doctor of Philosophy, Iowa State University, Digital Repository, Ames, 2009. DOI: 10.31274/etd-180810-436. [Online]. Available: https://lib.dr.iastate.edu/etd/10709/.

[24] O. Media. (). 1. dive in a quick dip: Breaking the surface - head first java, 2nd edition, [Online]. Available: https://learning.oreilly.com/library/view/head-first-java/0596009208/ch01.html.

[25] O. Media. (). Social engineering - the art of deception: Controlling the human element of security, [Online]. Available: https://learning.oreilly.com/library/view/the-art-of/9780764542800/pr02.html.

[26] C. E. Morgan, "Social engineering and influence," ISBN: 9781687906342, M.S. Utica College, United States – New York, 2019, 39 pp. [Online]. Available: http://search.proquest.com/docview/2314064851/abstract/52C5685D7C294B5FPQ/1.

[27] A. Koyun and E. A. Janabi, "Social engineering attacks," vol. 4, no. 6, p. 6, 2017.

[28] E. Yeboah-Boateng and P. M. Amanor, "Phishing, SMiShing & vishing: An assessment of threats against mobile devices," *undefined*, 2014. [Online]. Available: /paper/Phishing%2C-SMiShing-%26-Vishing%3A-An-Assessment-of-Yeboah-Boateng-Amanor/7a271a3ff90b2a19d6b4f4ecc800e0aebdcda063.

[29] (). Wiesen test of mechanical aptitude (WTMA) - criteria corporation, [Online]. Available: https://www.criteriacorp.com/solution/wtma.php.

[30] K. Doria, "Identifying why social engineering continues to be successful and how the social engineering risk can be reduced," ISBN: 9781687985354, M.S. Utica College, United States – New York, 2019, 50 pp. [Online]. Available: http://search.proquest.com/docview/2318150051/abstract/A26025E3CB5A4590PQ/1.

[31] Cyberoam, "INTERNET THREATS TREND REPORT," Apr. 2014.

[32] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Computers & Security*, vol. 68, pp. 160–196, Jul. 1, 2017, ISSN: 0167-4048. DOI: 10.1016/j.cose.2017.04.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404817300810.

[33] A. Bandura, "Self-efficacy: Toward a unifying theory of behavioral change," p. 25,

[34] (). Social engineering tools, Security Through Education, [Online]. Available: https://www.social-engineer.org/framework/se-tools/.

[35] A. A. Aldossary and A. M. Zeki, "Web user' knowledge and their behavior towards security threats and vulnerabilities," in *2015 4th International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, Dec. 2015, pp. 256–260. DOI: 10.1109/ACSAT.2015.51.

[36] D. Jacobson, J. Rursch, and J. Idziorek, "Workshop: Teaching computer security literacy to the masses: A practical approach," in *2012 Frontiers in Education Conference Proceedings*, ISSN: 2377-634X, Oct. 2012, pp. 1–2. DOI: 10.1109/FIE.2012.6462423.

[37] (Apr. 13, 2011). Research: Education can't protect against social engineering attacks, Security Intelligence. Section: Advanced Threats, [Online]. Available: https://securityintelligence
com/rsa-and-epsilon-research-shows-education-cant-protect-against-new-social-engineering-attacks/.

[38] S. Hazari, W. Hargrave, and B. Clenney, "An empirical investigation of factors influencing information security behavior," *Journal of Information Privacy and Security*, vol. 4, no. 4, pp. 3–20, Oct. 1, 2008, Publisher: Routledge _eprint: https://doi.org/10.1080/2333696X.20
ISSN: 1553-6548. DOI: 10.1080/2333696X.2008.10855849. [Online]. Available: https://doi.org/10.1080/2333696X.2008.10855849.

[39] G. Kiss and A. Szasz, "Analysing of the information security awareness of the economic information technology students," in *2016 IEEE 17th International Symposium on Computational Intelligence and Informatics (CINTI)*, ISSN: 2471-9269, Nov. 2016, pp. 000 213–000 218. DOI: 10.1109/CINTI.2016.7846406.

[40] S. Talib, N. Clarke, and S. Furnell, "An analysis of information security awareness within home and work environments," presented at the ARES 2010 - 5th International Conference on Availability, Reliability, and Security, Mar. 18, 2010, pp. 196–203. DOI: 10.1109/ARES.2010.27.

[41] G. Öğütçü, Ö. M. Testik, and O. Chouseinoglou, "Analysis of personal information security behavior and awareness," *Computers & Security*, vol. 56, pp. 83–93, Feb. 1, 2016, ISSN: 0167-4048. DOI: 10.1016/j.cose.2015.10.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404815001406.

[42] Lance James, *Banking on phishing*. Syngress, Jan. 1, 2006, Pages: 1-35, ISBN: 978-1-59749-030-6. DOI: 10.1016/B978-159749030-6/50006-4. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9781597490306500064.

[43] U. L. John and L. Norman, *Master the Mechanical Aptitude and spatial relations*, 6th edition. Thomas Learning Inc., ISBN: 13: 978-0-7689-1699-7.

[44] G. Ho, A. Sharma, M. Javed, V. Paxson, and D. Wagner, "Detecting credential spearphishing in enterprise settings," presented at the 26th {USENIX} Security Symposium ({USENIX} Security 17), 2017, pp. 469–485, ISBN: 978-1-931971-40-9. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/ho.

[45] H. Wilcox and M. Bhattacharya, "A framework to mitigate social engineering through social media within the enterprise," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, ISSN: 2158-2297, Jun. 2016, pp. 1039–1044. DOI: 10.1109/ICIEA.2016.7603735.

[46] Bruce Schneier. (2000). Crypto-gram: October 15, 2000 - schneier on security, [Online]. Available: https://www.schneier.com/crypto-gram/archives/2000/1015.html.

[47] S. Purkait, "Phishing counter measures and their effectiveness - literature review," *Information Management & Computer Security*, vol. 20, no. 5, pp. 382–420, 2012, Num Pages: 382-420 Place: Bradford, United Kingdom Publisher: Emerald Group Publishing Limited, ISSN: 09685227. DOI: http://dx.doi.org/10.1108/09685221211286548. [Online]. Available: https://search.proquest.com/docview/1191823882/abstract/346D6618E05F47F8PQ/1.

[48] J. Rastenis, S. Ramanauskaitė, J. Janulevičius, A. Čenys, A. Slotkienė, and K. Pakrijauskas, "E-mail-based phishing attack taxonomy," *Applied Sciences*, vol. 10, no. 7, p. 2363, Mar. 30, 2020, ISSN: 2076-3417. DOI: 10.3390/app10072363. [Online]. Available: https://www.mdpi.com/2076-3417/10/7/2363.

[49] S. Hofbauer, K. Beckers, and G. Quirchmayr, "Defense methods against VoIP and video hacking attacks in enterprise networks," Nov. 23, 2015.

[50] K. F. Kok. (Apr. 16, 2020). Truecaller insights 2020 u.s. spam & scam report, Truecaller Blog, [Online]. Available: https://truecaller.blog/2020/04/16/truecaller-insights-2020-us-spam-scam-report/.

[51] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Computers & Security*, vol. 68, pp. 160–196, Jul. 1, 2017, ISSN: 0167-4048. DOI: 10.1016/j.cose.2017.04.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404817300810.

[52] S. E. Griffin and C. C. Rackley, “Vishing,” in *Proceedings of the 5th annual conference on Information security curriculum development*, ser. InfoSecCD ’08, New York, NY, USA: Association for Computing Machinery, Sep. 26, 2008, pp. 33–35, ISBN: 978-1-60558-333-4. DOI: 10.1145/1456625.1456635. [Online]. Available: http://doi.org/10.1145/1456625.1456635.

[53] K. S. Jones, M. E. Armstrong, M. K. Tornblad, and N. A. Siami, “How social engineers use persuasion principles during vishing attacks,” *Information & Computer Security*, vol. ahead-of-print, ahead-of-print Jan. 1, 2020, ISSN: 2056-4961. DOI: 10.1108/ICS-07-2020-0113. [Online]. Available: https://doi.org/10.1108/ICS-07-2020-0113.

[54] S. A. Alrwais, A. Gerber, C. W. Dunn, O. Spatscheck, M. Gupta, and E. Osterweil, “Dissecting ghost clicks: Ad fraud via misdirected human clicks,” in *Proceedings of the 28th Annual Computer Security Applications Conference on - ACSAC ’12*, Orlando, Florida: ACM Press, 2012, p. 21, ISBN: 978-1-4503-1312-4. DOI: 10.1145/2420950.2420954. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2420950.2420954.

[55] (Oct. 22, 2015). Report: For every $3 spent on digital ads, fraud takes $1, [Online]. Available: https://adage.com/article/digital/ad-fraud-eating-digital-advertising-revenue/301017.

[56] J. Crussell, R. Stevens, and H. Chen, “MAdFraud: Investigating ad fraud in android applications,” in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, Bretton Woods New Hampshire USA: ACM, Jun. 2, 2014, pp. 123–134, ISBN: 978-1-4503-2793-0. DOI: 10.1145/2594368.2594391. [Online]. Available: https://dl.acm.org/doi/10.1145/2594368.2594391.

[57] H. Haddadi, “Fighting online click-fraud using bluff ads,” *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 2, pp. 21–25, Apr. 9, 2010, ISSN: 0146-4833. DOI: 10.1145/1764873.1764877. [Online]. Available: https://dl.acm.org/doi/10.1145/1764873.1764877.

[58] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, “TIARA: Interactive, topic-based visual text summarization and analysis,” *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 2, pp. 1–28, Feb. 2012, ISSN: 2157-6904, 2157-6912. DOI: 10.1145/2089094.2089101. [Online]. Available: https://dl.acm.org/doi/10.1145/2089094.2089101.

[59] S. Coates-Stephens, “The analysis and acquisition of proper names for robust text understanding,” doctoral, City University London, Oct. 1992. [Online]. Available: https://openaccess.city.ac.uk/id/eprint/8015/.

[60] W. J. Black, F. Rinaldi, and D. Mowatt, “FACILE: DESCRIPTION OF THE NE SYSTEM USED FOR MUC-7,” p. 10, 2014.

[61] M. Marrero, S. Sánchez-Cuadrado, J. M. Lara, and G. Andreadakis, “Evaluation of named entity extraction systems,” p. 12, Oct. 11, 2008.

[62] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, “Analysis of named entity recognition and linking for tweets,” *Information Processing & Management*, vol. 51, no. 2, pp. 32–49, Mar. 1, 2015, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2014.10.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457314001034.

[63] K. Yeager. (). LibGuides: SPSS tutorials: Independent samples t test, [Online]. Available: https://libguides.library.kent.edu/SPSS/IndependentTTest.

[64] R. van Bavel, N. Rodríguez-Priego, J. Vila, and P. Briggs, “Using protection motivation theory in the design of nudges to improve online security behavior,” *International Journal of Human-Computer Studies*, vol. 123, pp. 29–39, Mar. 1, 2019, ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2018.11.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581918306475.

A. APPENDIX A

RESEARCH PARTICIPANT CONSENT FORM

Environmental Factors Influencing Phishing Attacks Department of Computer and Information Technology Purdue University

What is the purpose of this study? This study is being conducted to create a new measure assessing privacy needs in individuals.

What will I do if I choose to be in this study? You will be asked to read a number of statements related to privacy concerns. You will be asked rate each of the statements on how well you reflect to it: low versus high privacy needs.

How long will I be in the study? This survey should take less than 20 minutes.

What are the possible risks or discomforts?

Minimal: The risks are not greater than those ordinarily encountered in daily life.

Are there any potential benefits?

There are no direct benefits to you for participating, However, you may have the opportunity to gain insight into the nature of designing, running, and interpreting research on emotional responses to individuals in situations of need.

Will I receive payment or other incentive?

For participating in this study, you will receive $.70 credited towards your mTurk account.

Are there costs to me for participation?

There are no extra costs to you for participating in this study.

Will information about me and my participation be kept confidential? The project's research records may be reviewed by departments at Purdue University responsible for regulatory and research oversight.

What are my rights if I take part in this study?

Your participation in this study is voluntary. You may choose not to participate or, if you agree to participate, you can withdraw your participation at any time without penalty or loss of benefits to which you are otherwise entitled.

Who can I contact if I have questions about the study?

If you have questions, comments or concerns about this research project, you can talk to one of the researchers. Please contact Ida Ngambeki 496-6839. If you have questions about your rights while taking part in the study or have concerns about the treatment of research participants, please call the Human Research

Protection Program at (765) 494-5942, email (irb@purdue.edu)or write to: Human Research Protection Program - Purdue University Ernest C. Young Hall, Room 1032, 155 S. Grant St., West Lafayette, IN 47907-2114

Documentation of Informed Consent

I have had the opportunity to read this consent form and have the research study explained. I have had the opportunity to ask questions about the research study, and my questions have been answered. I am prepared to participate in the research study described above.

Click “I agree” indicates that you have read and understand the information provided above, that you willingly agree to participate, and that you are aware that you may withdraw your consent at any time and discontinue your participation without penalty. If you choose not to participate, simply close your web browser.

○ I agree

○ I do not agree

Demographics

How old are you?

○ 18-25

○ 26-35

○ 36-45

○ 46-55

○ 55+

What is your sex?

○ Male

○ Female

○ Oth

e
r

○ Choose not to disclose

What the highest level of education that you have completed?

○ High School
Graduate ○
GED or equivalent
○ Some College

○ Associate's degree: occupational, technical or vocational program ○
Associate's degree: academic program
○ Bachelor's degree

○ Master's Degree
Completion ○ Choose not
to disclose

What is your yearly income?

○ Less than
$15,000 ○
$15,000 -
$24,999 ○
$25,000 -
$34,999 ○
$35,000 -
$49,999 ○
$50,000 -
$74,999 ○
$75,000 -
$99,999 ○
$100,000 -
$149,999 ○
$150,000 -
$199,999 ○
$200,000+
○ Choose not to disclose

Are you currently a student or working in an information technology or computing related profession or field?

○
○ Yes
No

On an average day, how long do you spend on a computer at work or school?

○ less than one hour

○ 1 - 3 hours

○ 3 - 5 hours

○ more than 5 hours

○ I never use a computer at work/school

On an average day, how long do you spend on a computer at home?

○ less than one hour

○ 1 - 3 hours

○ 3 - 5 hours

○ more than 5 hours

○ I never use a computer at home

On an average day, how long do you spend using a smartphone/tablet?

○ less than one hour

○ 1 - 3 hours

○ 3 - 5 hours

○ more than 5 hours

○ I never use a smartphone/tablet

Knowledge/Validation Question

What is a reasonable definition of phishing?

○ An electronic communication pretending to be from a reputable source intended to make individuals reveal personal information or provide unsafe access to their computer

○ The action of hacking into telecommunications systems to obtain free calls

○ A fraudulent phone call or voicemail purporting to be from a reputable company in order to obtain personal information

○ An electronic communication from a company

○ I do not know

For this question please select windows

○ Mac OS X

○ Windows

○ Ubuntu

What is a reasonable definition of ransomware?

○ A program called ransom that runs in the background of your computer

○ Malicious software that threatens to publish the victim's data or perpetually block access to it unless a ransom is paid

○ A site or program that one can use to purchase bitcoin ○ I do not know

What is a reasonable definition of vishing?

○ A fraudulent phone call or voicemail pretending to be from a reputable company in order to obtain personal information

○ The action of hacking into telecommunications systems to obtain free calls

○ A fraudulent text sent to lure personal information such as passwords through forms such as e-mail

○ I do not know

What is a reasonable definition of ad fraud?

○ A phone call or electronic communication offering unexpected goods or services

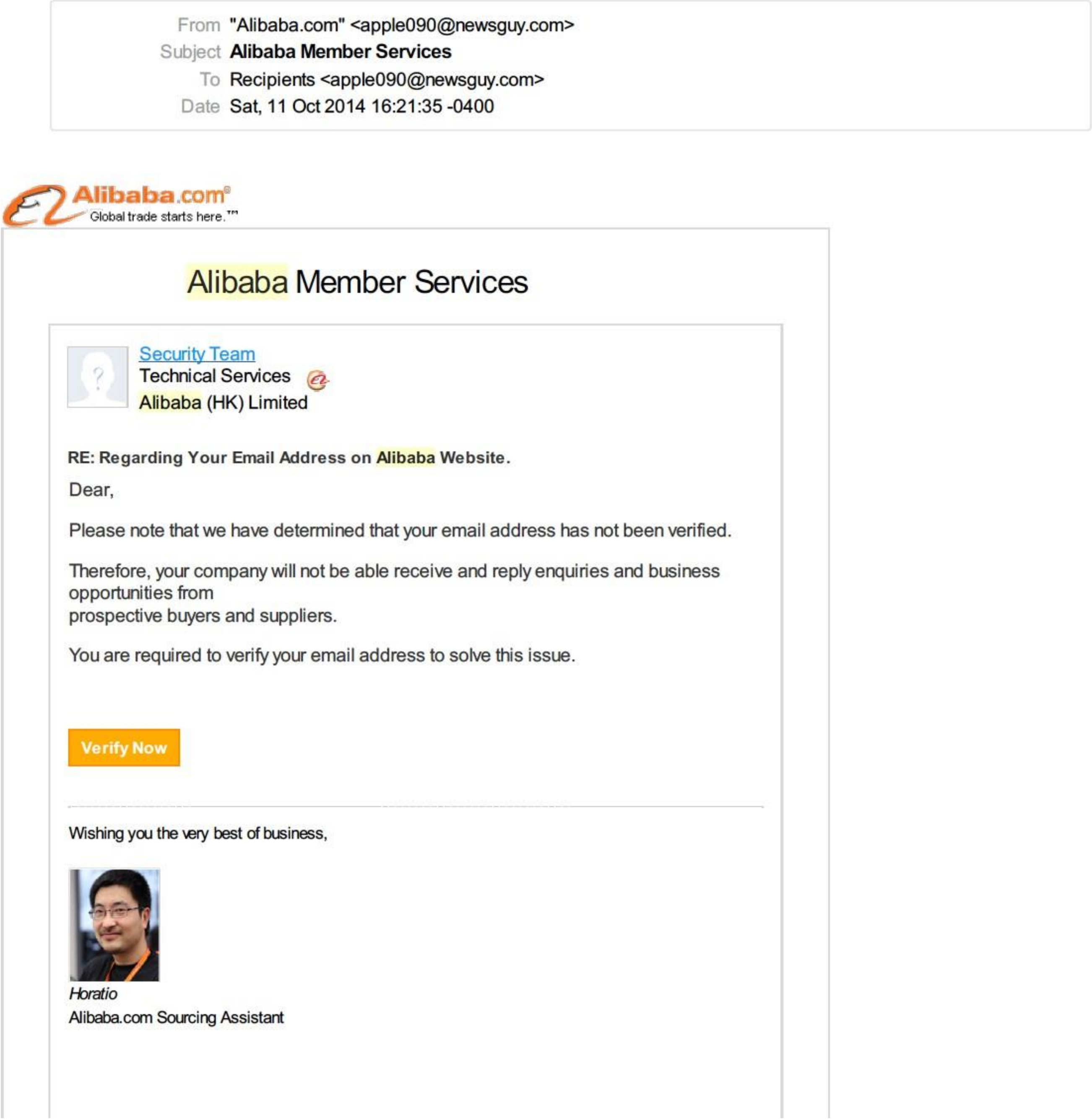
○ A phone call or electronic communication offering goods or services with the intent of getting you to respond in order to obtain personal information or compromise your computer

○ An electronic communication offering access to software

○ I do not know

Recognition Phishing

For the emails below please indicate how likely it is that each one is a phishing email. Phishing is a fraudulent email pretending to be from a reputable source intended to make individuals reveal personal information or provide unsafe access to their computer



From "Alibaba.com" <apple090@newsguy.com>
Subject **Alibaba Member Services**
To Recipients <apple090@newsguy.com>
Date Sat, 11 Oct 2014 16:21:35 -0400

Alibaba Member Services

Security Team
Technical Services
Alibaba (HK) Limited

RE: Regarding Your Email Address on Alibaba Website.

Dear,

Please note that we have determined that your email address has not been verified.

Therefore, your company will not be able receive and reply enquiries and business opportunities from
prospective buyers and suppliers.

You are required to verify your email address to solve this issue.

Wishing you the very best of business,

Horatio
Alibaba.com Sourcing Assistant

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○ ○	○	○	

Subject: Robinhood Password Has Been Reset

From: notifications@robinhood.com
To: [redacted]@yahoo.com
Date: Saturday, February 17, 2018, 2:47:43 PM EST

Robinhood Password Has Been Reset

Hi [redacted]

You recently changed the password to your Robinhood account [redacted]

If you did not request this change, please contact us immediately at support@robinhood.com. Please note we will never ask for your password over email or phone.

Questions? Check out our Help Center.

Sincerely,
The Robinhood Team
robinhood.com

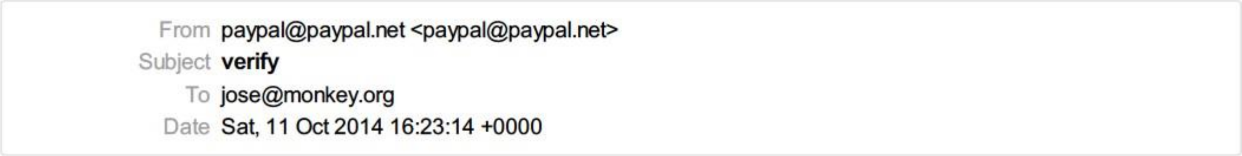
Free trading refers to $0 commissions for Robinhood Financial self-directed individual cash or margin brokerage accounts that trade U.S. listed securities via mobile devices. Relevant SEC & FINRA fees may apply. Please see our Commission and Fee Schedule.

Robinhood is currently registered in the following jurisdiction(s). This is not an offer, solicitation of an offer, or advice to buy or sell securities, or open a brokerage account in any jurisdiction where Robinhood is not registered.

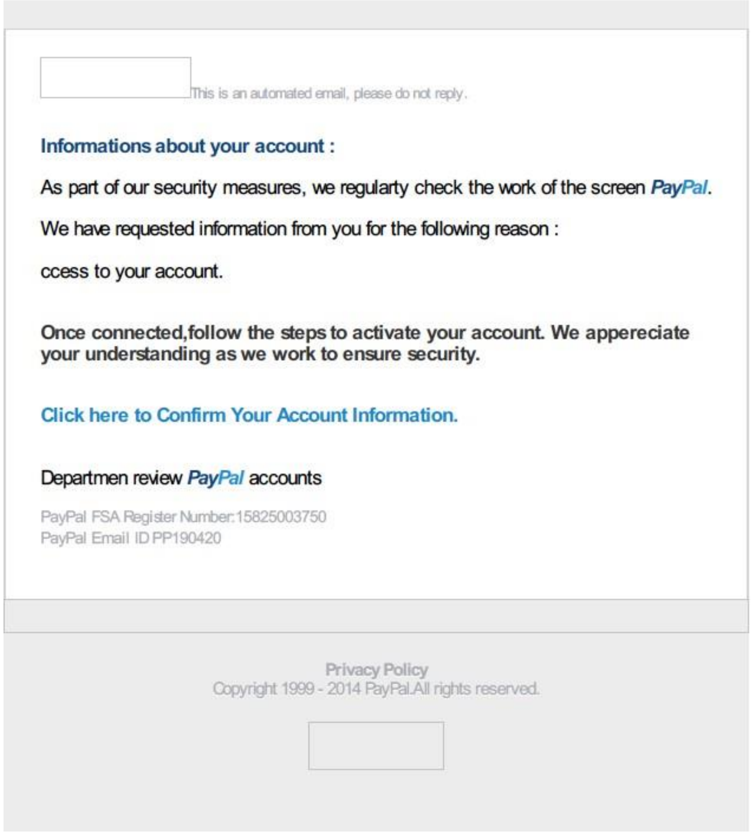
All investments involve risk and the past performance of a security, or financial product does not guarantee future results or returns. Keep in mind that while diversification may help spread risk it does not assure a profit, or protect against loss, in a down market. There is always the potential of losing money when you invest in securities, or other financial products. Investors should consider their

Extremely unlikely Somewhat unlikely Neither likely nor Somewhat likely
Extremely likely

○	○	unlikely ○	○	○



From paypal@paypal.net <paypal@paypal.net>
Subject **verify**
To jose@monkey.org
Date Sat, 11 Oct 2014 16:23:14 +0000



This is an automated email, please do not reply.

Informations about your account :

As part of our security measures, we regularty check the work of the screen *PayPal*.

We have requested information from you for the following reason :

ccess to your account.

Once connected,follow the steps to activate your account. We appreciate your understanding as we work to ensure security.

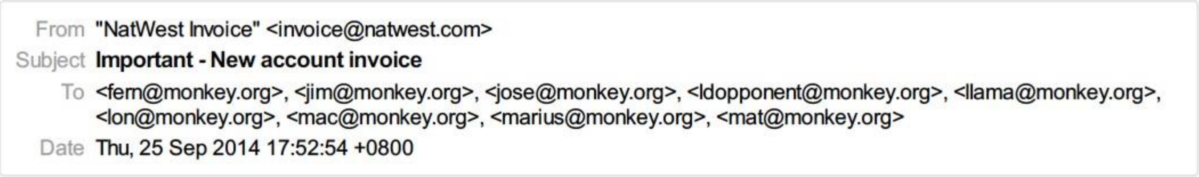
Click here to Confirm Your Account Information.

Departmen review *PayPal* accounts

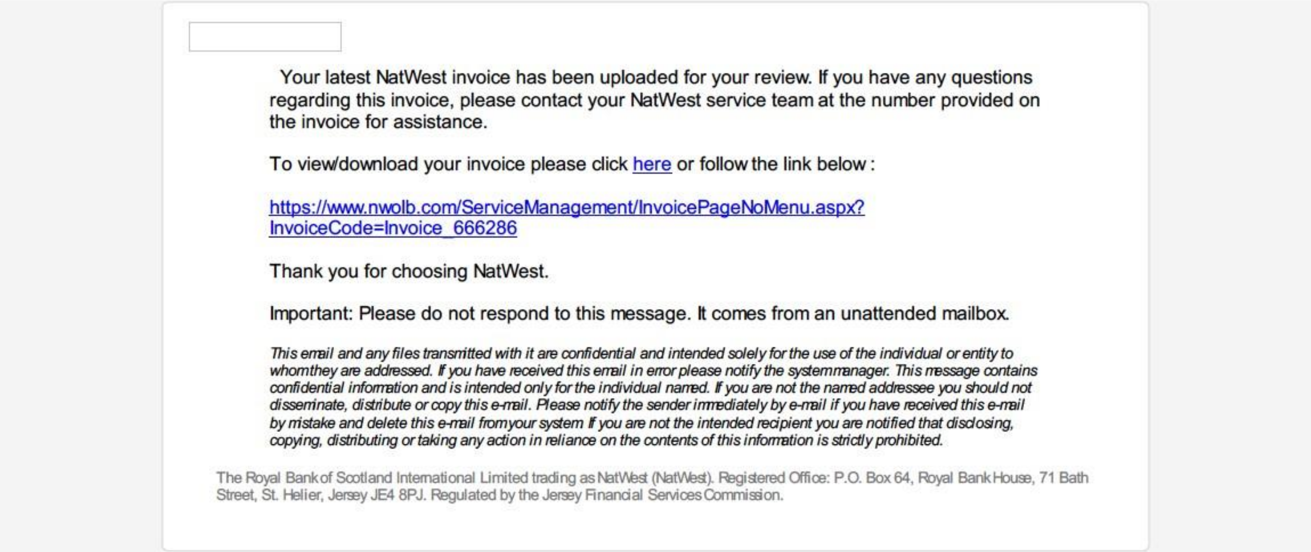
PayPal FSA Register Number:15825003750
PayPal Email ID PP190420

Privacy Policy
Copyright 1999 - 2014 PayPal.All rights reserved.

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○ ○	○	○	



From "NatWest Invoice" <invoice@natwest.com>
Subject **Important - New account invoice**
To <fern@monkey.org>, <jim@monkey.org>, <jose@monkey.org>, <ldopponent@monkey.org>, <llama@monkey.org>, <lon@monkey.org>, <mac@monkey.org>, <marius@monkey.org>, <mat@monkey.org>
Date Thu, 25 Sep 2014 17:52:54 +0800



Your latest NatWest invoice has been uploaded for your review. If you have any questions regarding this invoice, please contact your NatWest service team at the number provided on the invoice for assistance.

To view/download your invoice please click here or follow the link below :

https://www.nwolb.com/ServiceManagement/InvoicePageNoMenu.aspx?InvoiceCode=Invoice_666286

Thank you for choosing NatWest.

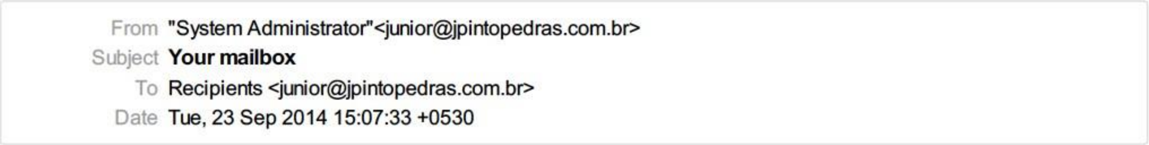
Important: Please do not respond to this message. It comes from an unattended mailbox.

This email and any files transmitted with it are confidential and intended solely for the use of the individual or entity to whomthey are addressed. If you have received this email in error please notify the systemmanager. This message contains confidential information and is intended only for the individual named. If you are not the named addressee you should not disseminate, distribute or copy this e-mail. Please notify the sender immediately by e-mail if you have received this e-mail by mistake and delete this e-mail fromyour system If you are not the intended recipient you are notified that disclosing, copying, distributing or taking any action in reliance on the contents of this information is strictly prohibited.

The Royal Bank of Scotland International Limited trading as NatWest (NatWest). Registered Office: P.O. Box 64, Royal Bank House, 71 Bath Street, St. Helier, Jersey JE4 8PJ. Regulated by the Jersey Financial Services Commission.

Extremely likely likely | Somewhat | Neither likely nor unlikely | Somewhat unlikely | Extremely unlikely

○ ○ ○ ○ ○



From "System Administrator"<junior@jpintopedras.com.br>
Subject **Your mailbox**
To Recipients <junior@jpintopedras.com.br>
Date Tue, 23 Sep 2014 15:07:33 +0530

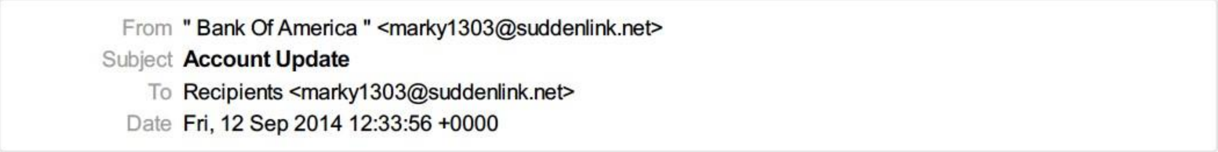
Your mailbox has exceeded the storage limit which set by your administrator,you may not be able to send or receive new mail until you re-validate your mailbox. To re-validate your mailbox please send the following details below:

Name:
Username:
Password:
Retype Password:
Email Address:
Phone Number:

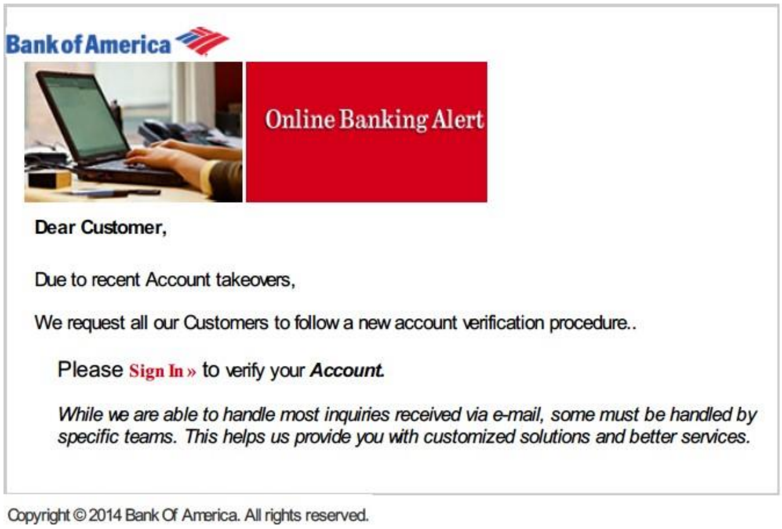
If you fail to re-validate your mailbox, your mailbox will be De-activated!!!

Thanks
System Administrator

○ ○ ○ ○ ○



From " Bank Of America " <marky1303@suddenlink.net>
Subject **Account Update**
To Recipients <marky1303@suddenlink.net>
Date Fri, 12 Sep 2014 12:33:56 +0000



Dear Customer,

Due to recent Account takeovers,

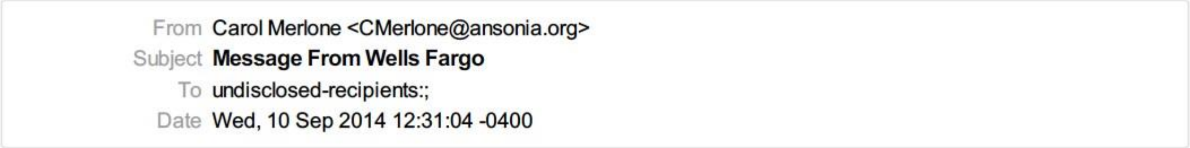
We request all our Customers to follow a new account verification procedure..

Please Sign In » to verify your ***Account.***

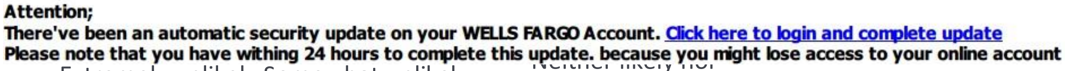
While we are able to handle most inquiries received via e-mail, some must be handled by specific teams. This helps us provide you with customized solutions and better services.

Copyright © 2014 Bank Of America. All rights reserved.

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○	○	○	○



From Carol Merlone <CMerlone@ansonia.org>
Subject **Message From Wells Fargo**
To undisclosed-recipients::
Date Wed, 10 Sep 2014 12:31:04 -0400

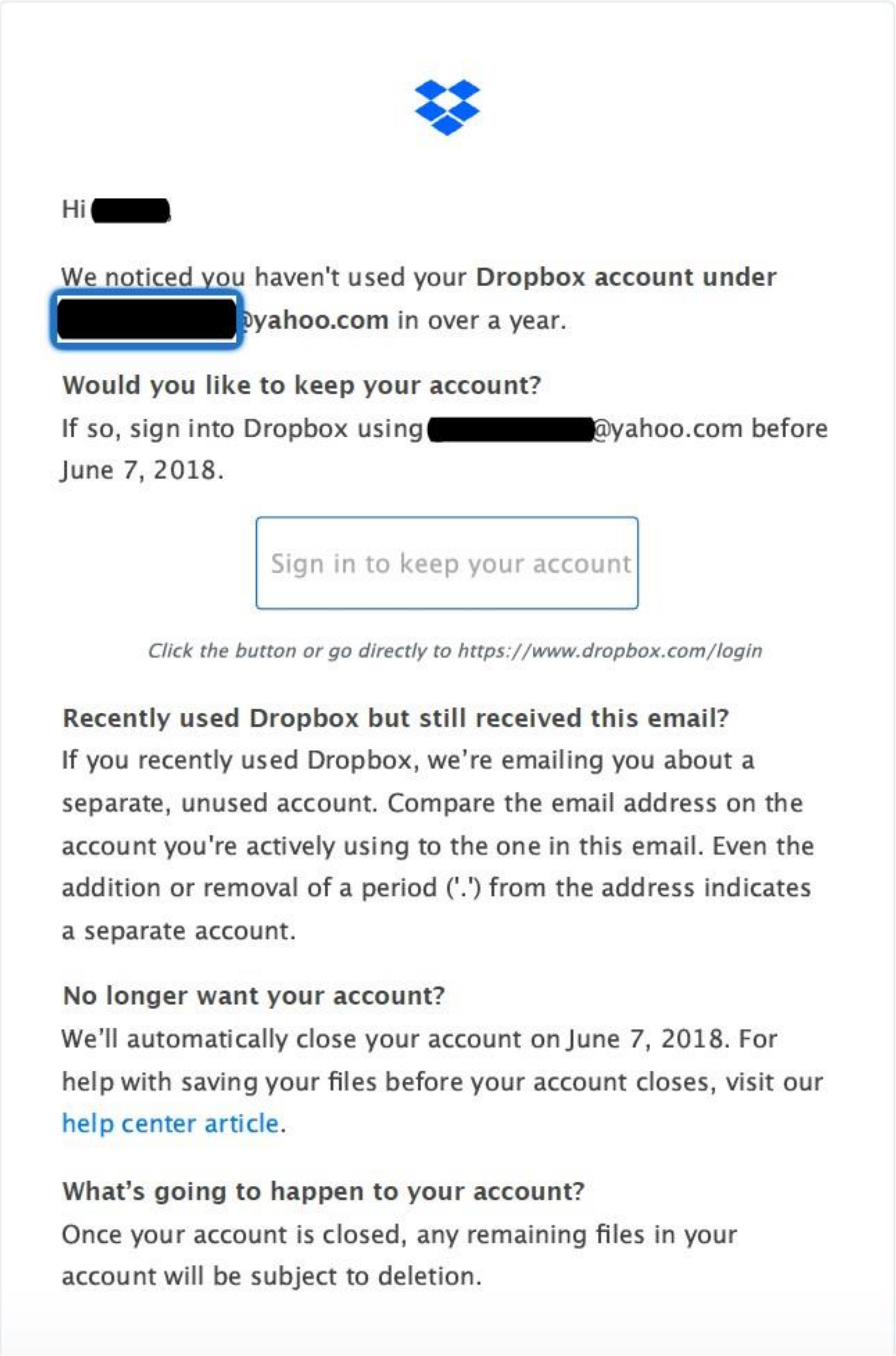


Attention;
There've been an automatic security update on your WELLS FARGO Account. **Click here to login and complete update**
Please note that you have withing 24 hours to complete this update. because you might lose access to your online account

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○ ○	○	○	

Subject: Hi [redacted] your Dropbox account is closing in 90 days

From: no-reply@dropboxmail.com
To: [redacted]@yahoo.com
Date: Thursday, March 8, 2018, 11:17:19 PM EST



Hi [redacted],

We noticed you haven't used your **Dropbox account under [redacted]@yahoo.com** in over a year.

Would you like to keep your account?
If so, sign into Dropbox using [redacted]@yahoo.com before June 7, 2018.

Sign in to keep your account

Click the button or go directly to https://www.dropbox.com/login

Recently used Dropbox but still received this email?
If you recently used Dropbox, we're emailing you about a separate, unused account. Compare the email address on the account you're actively using to the one in this email. Even the addition or removal of a period ('.') from the address indicates a separate account.

No longer want your account?
We'll automatically close your account on June 7, 2018. For help with saving your files before your account closes, visit our help center article.

What's going to happen to your account?
Once your account is closed, any remaining files in your account will be subject to deletion.

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely

○ ○ ○ ○ ○

Recognition Vishing

For the scenarios below please indicate how likely it is to be a vishing scenario. Vishing is a fraudulent phone call or voicemail pretending to be from a reputable company in order to obtain personal information

Yesterday a student from University of Arizona got a phone call from the company that she was interviewing with. The caller left a voicemail saying, "Hi Samantha, My name is Laura Smith, I am calling from PurpleBox regarding the internship you applied and interviewed for on February 9th at 2:00 P.M. I was just calling to confirm that this number 675- 823-2343 is the best way to contact you regarding your offer letter and 54 Flamingo Ct Denver CO, is the correct address where we can send the offer letter to. Please let me know if this information is correct."

How likely is this to be a vishing scenario?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○	○	○	○

Yesterday a student from University of Arizona got a phone call from the company that she had interviewed with several weeks earlier. She had already received an offer letter from the company in the mail. The caller left a voicemail saying, "Hi Samantha, My name is Laura Smith, I am calling from PurpleBox regarding the internship you applied and interviewed for. I was just calling to ask for your address and social security number so we can send you an offer letter. Please call me back with the information."

How likely is this to be a vishing scenario?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○	○	○	○

A student received a call from someone claiming to be a faculty member at the university in which he had just enrolled. The caller offered to hire him as a student worker. His job would be buying supplies for a startup. The caller asked for his address, bank routing number to send a check to cover the cost of supplies and his first salary payment.

How likely is this to be a vishing call?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely

○ ○ ○ ○ ○

The HR manager’s phone is ringing. She answers and the caller claims to be from the IT department and claims there is a possible data breach that might have made some of her contact details accessible. He asks the HR manager to give him her username and passwords to access her company account, so he can check if everything is fine.

How likely is this to be a vishing scenario?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○	○	○	○

Emily receives a call on her phone. the caller claims she has won a free vacation to the Bahamas in a random lottery. The caller asks for the name, address, and license number of herself and her travel companion. The caller claims to need this information so the tickets match to Emily's and her travel companion's IDs.

How likely is this to be a vishing scenario?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○	○	○	○

Mark receives a call from someone claiming to work for his bank. The caller says that they have observed suspicious behavior on Mark's account and lists a number of purchases totaling over $2000 made at online sex shops. He claims that unless Mark proves his identity he will be responsible for the charges. The caller asks him to provide his account number, date of birth and social security number in order to verify his identity and remove the charges from his account.

How likely is this to be a vishing call?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○	○	○	○

Julie receives a phone call from her bank. The caller claims that her credit card was used at a location where several breaches have occurred. As a result the bank will be sending her a new credit card. They ask her to verify the last four digits on her credit card.

How likely is this be a vishing call?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely

○ ○ ○ ○ ○

In June of last year, Emma Watson, a British businesswoman who was setting up a children's nursery, got a phone call from her bank's fraud team. They told her that they had stopped some unusual transactions on her account, but because her account had been compromised, she had to transfer her money into another account that they had set up in her name. The callers knew her name, account numbers, and social security number.

How likely is this to be a vishing scenario?

Neither likely nor unlikely Somewhat likely Extremely likely

Extremely unlikely Somewhat unlikely

○ ○ ○ ○ ○

John receives a call from someone claiming to be a student at his alma mater. The caller claims to be raising money for the school and requests a donation. The caller asks for John's credit card number, expiry date, and security code.

How likely is this to be a vishing call?

Neither likely nor unlikely Somewhat likely Extremely likely

Extremely unlikely Somewhat unlikely

○ ○ ○ ○ ○

Recognition Ad Fraud

For the scenarios below please indicate whether it is an ad fraud scenario or not.

You are streaming a movie on a third party website. An ad pops up for "Free Headphones No Purchase Necessary" from a website that appears with Apple's logo on it.

How likely is this to be ad fraud?

Extremely unlikely Somewhat unlikely Neither likely nor unlikely Somewhat likely Extremely likely

○ ○ ○ Somewhat likely Extremely likely

○ ○ ○ ○

You visit Abercrombie and Fitch's website to buy your favorite pair of jeans. The next day you log into Facebook.com in order to update your status. You notice on the panel bar there is a advertisement link for Abercrombie and Fitch's new line of jeans. You get excited and click the link.

How likely is this to be ad fraud?

Extremely unlikely Somewhat unlikely Neither likely nor unlikely

○ ○ ○

You are shopping on the website of an online electronics retailer. An ad pops up saying because you searched for certain items you have qualified for a free giveaway of a new Amazon Echo dot. The ad also says that you are receiving this offer as part of a test drive for the product and all you need to do to qualify is answer a short survey.

How likely is this to be ad fraud?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○ ○	○	○	

You see a flyer posted for a concert for a popular band. In the corner is a QR code (a square bar code that you can scan). The flyer says if you scan the QR code, you will get a discount on tickets. The code takes you to a website that has links to pictures and videos of the band and asks you to click on them to get the discount.

How likely is this to be ad fraud?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○ ○	○	○	

You are trying to buy something from a boutique online retailer. The website provides a link to PayPal which is a common online payment site.

How likely is this to be ad fraud?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
○	○	○	○	○

Exposure

Have you received a phishing email (a fraudulent email intended to obtain sensitive information or compromise your security) in the past week, month and year?

□ Yes, in the past week □ Yes, in the past month

□ Yes, in the past year

□ No, I have not received a phishing email

□ I do not know if I have received a phishing email

How many phishing emails have you received in the past month?

○ 1-5

○ 6-10

○ 11-20

○ More than 20

Have you fallen victim phishing email in the past week, month and year?

□ Yes, in the past week □ Yes, in the past month □ Yes, in the past year

□ No, I have not fallen victim a phishing email

□ I do not know if I have fallen victim to a phishing email

How many phishing emails have you ever fallen victim to?

○ 1

○ 2-3

○ 4-5

○ More than 6

Have you seen an ad that was probably ad fraud (advertisements offering goods or services with the intent of getting you to respond in order to obtain personal information or compromise your computer) in the past week, month, year?

□ Yes, in the past week □ Yes, in the past

month □ Yes, in
the past year

□ No, I have not seen an ad that was probably ad fraud

□ I do not know if I have seen an ad that was probably ad fraud

Have you been a victim of ad fraud in the past week, month and year?

□ Yes, in the
past week □
Yes, in the past
month □ Yes, in
the past year

□ No, I have not been a victim of ad fraud

□ I do not know if I have been a victim of ad fraud

How many times have you ever been a victim of ad fraud?

○ 1

○ 2-3

○ 4-5

○ More than 6

Have you received a vishing (a fraudulent phone call or voicemail pretending to be from a reputable company in order to obtain personal information) call in the past week, month and year?

□ Yes, in the
past week □
Yes, in the past
month □ Yes, in
the past year

□ No, I have not received a vishing call

□ I do not know if I have received a vishing call

How many vishing calls have you received in the past month?

○ 1-5

○ 6-10

○ 10-20

○ More than 20

Have you ever had a virus on your machine?

○ Yes

○ No

How many times have you had a virus on your computer?

○ 1

○ 2-3

○ 4-5

○ More than 6

Have you ever had ransom ware (malicious program that locks your computer until a fee is paid) on your machine?

○ Yes

○ No

How many times have you had ransomware on your computer?

○ 1

○ 2-3

○ 4-5

○ More than 6

Have any of your accounts ever been hacked or compromised?

○ Yes

○ No

What kind of account was hacked or compromised?

□ Social Media

□ Financial

□ Personal Email

□ School or Work

□ Health

□ Video streaming

□ Other

Have you ever willingly shared credentials (username/password/answers to security questions) for any password protected account?

○ Yes

○ No

What kind of account have you willingly shared credentials for?

□ Social Media

□ Financial

□ Personal Email

□ School or Work

□ Health

□ Video streaming

□ Other

Has any of your private data, photos, or video ever been obtained and/or shared without your knowledge and permission?

○

○ Yes

No

Have you ever been a victim of identity theft?

○ Yes

○ No

Have you ever lost money as a result of any online or phone scam?

○ Yes

○ No

How much money have you lost as a result of a phone or online scam?

○ less than $100

○ $100 - $500

○ $501 - $1000

○ $1001 - $5000

○ More than $5000

Environmental Factors

Where did you live (for the majority of the time) between the ages of 5-18?

○ Africa

○

Europe
○ Asia
○ North America
○ South America
○ Australia (or Oceania)
○ Antarctica

Which Country?

Which Country?

Which Country?

Which Country?

Which Country?

Which Country?

Which State?

Were you taught any computer science, cybersecurity or related topics from the ages of 5-18?

○ Yes

○ No

How much time did you spend learning about computer science, cybersecurity or related topics from the ages of 5-18?

○ less than a week ○ about a month

○ an entire term/semester ○ an entire year

○ multiple years

Were you taught about online safety in school from the ages of 5-18?

○ Yes

○ No

Were you taught about online safety at home from the ages of 5-18?

○ Yes

○ No

Were you in a computer science, cybersecurity or related club from the ages of 5-18?

○ Yes

○ No

Did your parents work in a computer science, cybersecurity or related field?

○ Yes

○ No

Did you have a computer/laptop/tablet/smartphone at home from the ages of 5-18?

○ Yes

○ No

How much average time did you spend per day on your home computer/laptop/tablet/smartphone from the ages of 5-11?

○ less than an hour

○ 1-2 hours

○ 3-5 hours

○ more than 5 hours

How much average time did you spend per day on your home computer/laptop/tablet/smartphone from the ages of 12-18?

○ less than an hour

○ 1-2 hours

○ 3-5 hours

○ more than 5 hours

Were there a lot of technology industries around the area where you spent the most time during the ages of 5-18?

○ Yes

○ No

Did you ever watch programs on TV that were computing or technology related?

○ Yes

○ No

Was there public cybersecurity or online safety education offered in the area where you spent the most time during the ages of 5-18, for example seminars at the public library?

○ Yes

○ No

Were there any billboards present in the area where you spent the most time during the ages of 5-18 that promoted online safety?

○ Yes

○ No

Have you ever received any training or information about how to stay safe online and protect your information within the last two years?