Innovations of Phishing Defense: The Mechanism, Measurement and Defense Strategies

Kutub Thakur[1], Juan Shan[2], and Al-Sakib Khan Pathan[3]

[1]Cyber Security, Department of Professional Security Studies, New Jersey City University, New Jersey, USA
[2]Department of Computer Science, Pace University, New York, USA
[3]Department of Computer Science and Engineering, Southeast University, Bangladesh
kthakur@njcu.edu, jshan@pace.edu, spathan@ieee.org

***Abstract*:** Now-a-days, social engineering is considered to be one of the most overwhelming threats in the field of cyber security. Social engineers, who deceive people by using their personal appeal through cunning communication, do not rely on finding the vulnerabilities to break into the cyberspace as traditional hackers. Instead, they make shifty communication with the victims that often enable them to gain confidential information like their credentials to compromise cyber security. Phishing attack has become one of the most commonly used social engineering methods in daily life. Since the attacker does not rely on technical vulnerabilities, social engineering, especially phishing attacks cannot be tackled using cyber security tools like firewalls, IDSs (Intrusion Detection Systems), etc. What is more, the increased popularity of the social media has further complicated the problem by availing abundance of information that can be used against the victims. The objective of this paper is to propose a new framework that characterizes the behavior of the phishing attack, and a comprehensive model for describing awareness, measurement and defense of phishing based attacks. To be specific, we propose a hybrid multi-layer model using Natural Language Processing (NLP) techniques for defending against phishing attacks. The model enables a new prospect in detection of a potential attacker trying to manipulate the victim for revealing confidential information.

***Keywords*:** Social Engineering, Information System, Network Security, Text Analysis, Phishing.

1. Introduction

For the past decades, cyber security remained a prime focus of every individual and business organization connected to the Internet. It may be in the form of defending the information of employees or the personal information of clients. Indeed, organizations are intensely targeted via the cyberspace and they are increasingly becoming more aware of protecting the sensitive information from unauthorized access over the Internet [1]. Nevertheless, the malicious attackers are constantly attempting to access the sensitive information for devastating the organizations and individuals for their personal gains. Traditionally, a computer hacker sitting in a dark room is treated as a major threat to cyber security. Nowadays, a cordial social engineer is a real threat to cyber security. With a simple psychological-based conversation and behavior manipulation, the social engineers are capable to breach the cyber security of any organization or individual. What is worse, the problem becomes more critical with the devolvement of social media and the availability of large amount of private information on the web.

According to the OWASP Top 10 Most Critical Web Application Security Risks 2017 [2], Broken Authentication takes the 2nd place. Broken authentication often occurs when application functions related to authentication and session management are implemented incorrectly which allows the attackers to compromise passwords, keys, or session tokens, or to exploit other implementation flaws to assume other users' identities temporarily or permanently. While the web application development flaw is the reason for that, it is also noted in this latest report that, "*Attackers can detect broken authentication using manual means, but are often attracted by password dumps, or after a social engineering attack such as phishing or similar.*" Hence, phishing and various forms of social engineering could be used to detect a broken authentication case and thus that could incur great harm on the victim. With the remarkable growth and prominence of social networking sites, the social engineering sites have also evolved rapidly. The web today has become a very convenient yet a very precarious environment to work in. Therefore, a trend of robust security policies combined with protective technologies is needed to provide organizations and individuals with a reliable cyber defense.

In the last decade, it was noticed that the organizations were increasingly becoming concerned about their cyber security due to a huge increase in security breaches reported by various organizations [3], [4], [5]. A report published in 2006 by DTI, UK [6] shows that 62% of UK companies had a security incident in 2005 though it was less than that of the previous years. However, the average cost of a UK company's worst security incident of that year was roughly £12,000 (up from £10,000 two years ago). Hence, even with relatively lesser number of security breaches in the subsequent years, as the time goes forward, the significance of the security breach incidents is becoming higher. The same trait in security breaches has been still continuing in spite of the development of new tools, techniques and strategies for defending against cyber attacks [7]. In fact, the financial loss of even a small number of incidents today would be greater than that of the past years due to the increased connectivity and heavy reliance on the Internet based communications for many trades, industries and organizations.

Given the reality today, organizations are willing to spend more resources in protecting their cyber resources and sensitive information. The irony is, in spite of spending the amount of around $77 billion for cyber security in 2015, the attackers were able to breach the security defenses regularly [2], [8]. In 2016, cybercrime cost the global economy over $450 billion, over 2 billion personal records were stolen and in the U.S. alone, over 100 million Americans had their medical records stolen [9], [33]. In fact, some experts opine that the next global financial crisis could be caused by a cyber attack [10]. The reality is, today even a naïve user with just a set of attack tools, coupled with an internet connection

can perform an attack against the intended target (victim). If any information system is compromised in this way, it would eventually increase fraudulent activities.

In the field of cyber security, people are often more focused on technical side for example, intrusion detection and prevention software and firewalls [11], [34]. However, as the social engineers deceive people, they use their personal appeal through tricky communication often without relying on finding the vulnerabilities to break into the cyberspace. They make such shifty communication with the victims that it often enables them to gain confidential information like their credentials to compromise cyber security. Since the attacker does not rely on technical vulnerabilities, these social engineering attacks cannot be tackled using traditional cyber security tools like firewalls, IDSs, etc. Hence, we look for new strategies and defensive methodologies to tackle social engineering attacks.

This paper focuses on an important element of the cyber security, which is the human element. This is an issue which was made notorious by Kevin Mitnick et al. in the book "The Art of Deception: Controlling the Human Element of Security" [12]. The deception techniques are used to gain someone's trust by lying to them and then abusing that trust for fun and profit. Hackers often use the euphemism "social engineering" and till today, many of those deceitful techniques are still used. For instance, a very sophisticated identity phishing campaign targeted Gmail users in the recent times, seeking to gain control of their entire email histories and spread itself to all of their contacts [13]. As the case was reported, "*The worm — which arrived in users' inboxes posing as an email from a trusted contact — asked users to check out an attached "Google Docs," or GDocs, file. Clicking on the link took them to a real Google security page, where users were asked to give permission for the fake app, posing as GDocs, to manage users' email account. To make matters worse, the worm also sent itself out to all of the affected users' contacts — Gmail or otherwise — reproducing itself hundreds of times any time a single user fell for it.*" In this particular case, a worm was used to launch further attacks but things started with a simple clicking on a link that came via an email. Smart people could also fall victim of such attacks. In another case for instance, the head of Austrian aerospace parts maker FACC was fired after the company was hit by a cyber fraud that cost it 42 million Euros ($47 million). [14] The hoax email asked an employee to transfer money to an account for a fake acquisition project. This is a kind of scam known as a "*fake president incident*".

As it can be understood, social engineering attack is quite different from the mainstream cyber attacks. The attack vectors are mostly computer programs, which are malicious malware [11], [12]. However, social engineers always have a clear purpose to acquire sensitive information. They can be either *white hackers* who are doing it just for their interest or curiosity, or black hackers, who actually are trying to steal information to cause harm [12].

With this introductory text, the rest of paper is organized as follows: Section 2 introduces the background and literature review on social engineering and phishing attacks. Section 3 presents our approach to defend against phishing attack. Experimental design and result analysis are presented in Section 4. Finally, Section 5 concludes the paper with our findings and future research directions.

2. Background and Related Work

Social engineering is a series of activities that manipulate people and mislead them to give in confidential information. Particularly, phishing, by sending fake emails that persuade people to reveal personal information (such as passwords) has become one of the most commonly used social engineering methods. Usually, social engineers use strategies to gain people's trust rather than hacking a program. Indeed, it is much easier to convince someone to reveal his or her password than cracking the password itself. Hence, this field basically associates sociology with cyber security. Social behavior of the human beings, which is often translated to online activities, influences the issue to a great extent.

Bjorck introduces a simple classification model for research in the field of information security [15]. The work proposes the key dimensions in the model for research as a level of abstraction ("theories and models", "empirical world") and domain (technical, formal, informal) of researches in the field. The study concludes that more emphasis should be exerted on the works on research issues in the information security education area so that people would learn and gain security awareness.

Indeed, there is a huge scope to study the art of deception in the cyberspace. In order to get a good insight in this area, information security researchers also need to gain knowledge of sociology to some extent. Nevertheless, one of the key issues of our work is to understand the important factors that could manipulate people in the cyberspace.

[16] presents an analysis of social engineering principles for effective phishing. This work's phishing is such an attack for which a company needs to worry about all emails that the employees receive while the attacker only needs to get a response from a key resource person. There may be many types of employees in a company and that makes it difficult to control their online behavior. Some, even after knowing may fall into the traps set by a cunning social engineer (i.e., attacker). One such case is reported in [17], which states that clever Gmail phishing Scam tricked even the technical users. Five principles of persuasion in social engineering as mentioned in [16] are: Authority (society generally trains people not to question authority); Social Proof (people tend to mimic what the majority of people do); Liking, Similarity, & Deception (people prefer to abide to whom (they think) they know or like), Commitment, Reciprocity, & Consistency (people feel more confident in their decision once they commit to a specific action and need to follow it until the end); and Distraction (people focus on one thing and ignore other things that may happen without noticing those). This paper also proposes some kind of manual strategy to sieve phishing from honest emails, based not on textual analysis, but more on semantics and goal processing.

[18] proposes a technique to identify phishing pages according to the visual similarity of webpage components, which are difficult to evade by attackers. The technique called Phishing-Alarm is based on CSS (Cascading Style Sheet) features of web pages. The authors propose techniques to identify effective CSS features, as well as algorithms to efficiently evaluate page similarity. For their evaluation of real-life phishing attacks, the authors use the Phishing-Alarm prototype as an extension with Google Chrome browser. Though the work is good and gives good insight into the issue, such kind of suspiciousness ratings of webpages based on the similarity of visual appearance

between the webpages may not be always effective. In fact, a human user may still continue browsing a website. Also, this work does not explicitly talk about email based phishing attacks that often reach one's inbox.

[19] presents an approach for detecting spam and phishing emails using SVM (Support Vector Machine) and Obfuscation URL Detection algorithm. Though, the authors claim some gains, there are various issues in the work that are not clear. The authors mention that SVM is supervised learning technique and provides binary classification and thus, training SVM is easy. However, it is a fact that SVM cannot deal with large number of input data. In practical case, the mechanism may not have the required efficiency. Moreover, the treatment of the subject seems to be a bit superficial in the work.

[20] is an interesting work where the authors propose a new approach for detecting phishing webpages in real-time as they are visited by a browser. The technique relies on modeling inherent phisher limitations stemming from the constraints they face while building a webpage. The authors note that the external hyperlinks and external content sources on a phish point to domains are typically outside the control of the phisher. Again, phishers may freely change most of the phishing page but the latter part of its domain name is constrained as it is limited to those domains that are generally controlled by phishers. Hence, by measuring differences in the composition and consistency of term usage in constrained/unconstrained and controlled/uncontrolled sources, they improve the effectiveness of phish detection. Though the work is good, this does not answer the behavioral issues of the users who would click the email and then get into the trap.

Studying various past works, in this work, our key focus is on the ethics and behavioral side of human being. We understand that just tools and techniques employed on a certain operating system or for some website or webpage would not be able to stop phishing but rather the main defenders must be the human beings who fall victims and then could let an entire company get compromised because of a personal action online. As far we have studied the area, there are very few works in this domain. One such work in [21] describes an approach to combating phishing by classifying phishing controls into relationships (human and technology (HT), human and organization (HO) and, organization and technology (OT).). They consider human, technology, and organization and note that the relationships need to be improved through educational strategies. The work talks about managing various factors that are crucial for minimizing security risks for the organization. The factors include recruitment of new staffs, job description, skilled staff, employment contract, orientation program, fair compensation, monitoring and evaluation, and termination of employment. While this work gives more insight into the issue, our position is that the company's policy is specific to a company and there are different ways to meet the demands of the employees. What we need to tackle phishing attack is an approach that combines the human behavior with some technical support [22], [23], [24]. Of course, company's recruitment, package, dealing with outgoing staffs, etc. would help improve the efficiency of the defense mechanism.

3. Our Proposed Mechanism

In this section, we propose an effective model for protecting sensitive information from social engineering attacks. We name our model, Security Training and Processing Evaluation (STPE). The model is basically a cycle with five stages as shown in Figure 1. The description of each stage is given in the subsequent subsections.

Figure 1. STPE Model Flow Chart.

A. Course Training

This stage is intended to make the users aware of social engineering. The training courses should provide enough real-life (that already occurred) as well as potential (that could happen) cases that deal with social engineering attacks. Courses should teach the mechanism and factors that influence people as well.

i. Course Topics - Many users are either careless or not aware of email links and attachment. In sociology terms, this is explained by our human nature, curiosity. However, it is not a good choice to click suspicious file coming through the cyberspace. An email scammer or hacker uses this method to send a large volume of phishing emails to people. If one of the users reads the email and is convinced by the content, many other users would be under attack or that would at least open the path for further manipulation and spamming. If one user is compromised by this, his or her friends in his email contact list would also be compromised by this attack. This is because people always trust their friends than random people they e-meet on the Internet. The course topics should cover these issues. The users should be given the idea that personal meeting on one-to-one basis (physically) and communication via cyberspace are not the same.

ii. Procedures to Defend against a Potential Social Engineering Attack - There are some easy-to-follow procedures to defend against a general phishing attack. These should be taught to the users. Some of the effective steps are:

1. Calm down – When dealing with emails, a user should always keep calm than jumping into actions. When email communication is done, there should not be rush to click a weblink or attachment that comes with an email even if the email indicates the matter to be urgent. This is because, in case of real urgency or high-level emergency, the sender of the message could use other communication methods like direct phone call or in-person meeting and talk. If it happens that a social engineer calls over phone and there is also an email with urgency that the user needs to check where there is a malicious weblink or attachment, the user could make sure that the caller is the genuine person who he/she knows. Sometimes, email

could be used for real urgency when the recipient is unreachable via phone or in-person talk. Even in this case, before clicking the attachment or weblink of an email, the user should calm down and assess the situation as the email has been used as a last resort in this case or there is only one way to reach him then. Still, some time could be taken for verification of the identity of the sender. Email is not expected to be replied or worked on within 2 (two) minutes for instance, though some professional people could do that. Hence, the recipient could make a phone call or personally meet the sender. The key issue here is taking some time to avoid immediate disclosure of critical information or damage caused by phishing.

2. Take a simple search/research approach on everything – In the cyber world, a user can search public information easily. Thus, if an email is from some agency or company that confuses the user, the user should first search the public information about this and check their email address, company name, location, and any other related information to confirm the identity. Often, a simple search with a portion of the email text reveals scam emails. There would be some forums and expert talks online that the user could read and accordingly act.

3. Say 'no' to request - This is very common that often phishing attackers ask for help from user. It may be quick financial help or bank account related information or some critical information like password or keyword. However, a professional company or industry people will not ask help without any referral. Again, a system administrator will never ask for the password of a user as with administrative access, all passwords are available to the administrator. Therefore, it is not hard to say 'no' to that kind of request and user should feel easy with this decision.

4. Install some anti-phishing software - Usually, most of anti-virus software provide the anti-phishing functions and it is a second alert to protect user from clicking or running unsigned program or malicious links. Hence, a user should always install some kind of anti-virus or anti-phishing software on the system. If Internet based tasks are done, using anti-virus software is a must.

B. Operation Monitoring

In our model, the operation monitoring is done by a 24/7 (running) server in a company that keeps monitoring all the network devices. The server side is solidly equipped and could store and organize all the log data automatically. The log is recorded on a daily basis and Redundant Array of Independent Disks (RAID) is used to keep data safe and for high speed. Today, many network and online monitoring tools are available, such as famous Wireshark, and others like GFI LanGuard, Capsa Free, Fiddler, Pandora FMS, Zenoss Core, Splunk, NetXMS, etc. The whole monitoring system would collect various types of log files and aggregate them with security criteria. The efficiency and availability of the monitoring system (Figure 2) is tested further in the experiment stage (described later).

C. Text Collecting

During the monitoring stage, all string types of data such as searching queries, external URL links, email links and contents are collected. These data can be used for both full-text searching and updating dictionary in the later stage. Full text collection is an important stage in our approach.

D. Model Analyzing/Analysis

Another well-equipped server collects data from the previous stage. The processing steps are depicted in Figure 3. In general, this stage has four phases: (i) Pre-preprocessing data, (ii) Building up action-resource pair, (iii) Keywords checking, and (iv) Attack detecting. In the pre-processing data phase, we collect data from monitoring server and then do several data cleaning and data format extraction so that it could be used in the next stage. In the meantime, we also aggregate those from the same IP address. In the Building up action-resource pair phase, we extract the tokenized words into pairs using NLP tools. In the third phase, whenever we find any malicious action word pair in our malicious dictionary, we assign a risk level on this particular source and then, in the phase four, we use the score to indicate whether a potential malicious attack has happened. For instance, the server raises an alert notification if the risk level score from one IP address exceeds a threshold level. The four phases are described in the subsequent paragraphs.

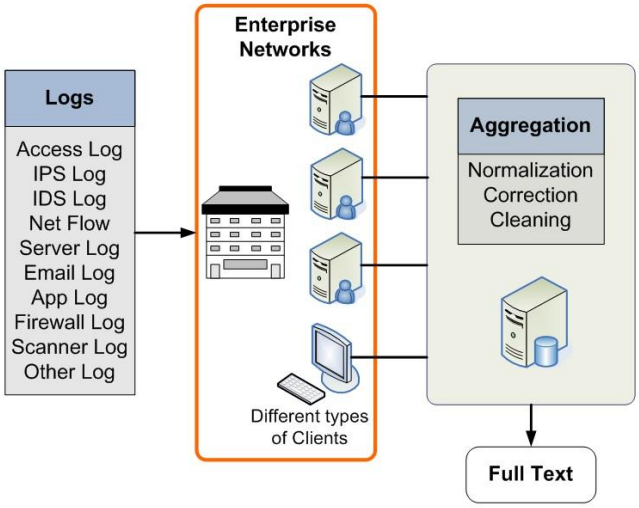

Figure 2. Log Aggregation.

i. Pre-processing Data - In the preprocessing phase, we collect all the text strings from text sources such as searching queries, email contents, SNS (Simple Notification Service) messages and text attachment. Then, we perform several common text analysis techniques, such as cleaning up stop words, TF-IDF (Term Frequency–Inverse Document Frequency). In particular, we first implement the noisy filtering approach. Our input is several samples of malicious social engineering text. There are several non-English texts and non-alphabet characters in the content of the text (such as special characters ~! @#$%^&* ()_+...). We exclude them and make that full paragraph into English-only sentences. Next, we convert the text to all lowercase letters and remove the punctuation from it. The objective of this step is to get prepared for the next stage.

ii. Building up Action-Resource Pair - Analysis of a sentence using Natural Language Processing (NLP) helps extract the verb (V) and object (Noun Phrase - NP) easily. Thus, it is more focused on the verb and the object of every single sentence and we do some analysis on them.

Topic Blacklist (TBL) model is introduced by Bhakta in 2015 [25]. The list covers topic of information security violations. Every single topic describes a series of actions,

either data or an operation. Particularly, one topic is consisted of two main attributes: *an action* and *a resource*. The resource stands for the information assets in a company and it should have policy and related permission restricted on it. However, the TBL uses a manually inputted restriction that is extracted from a company's policy document and related rules.

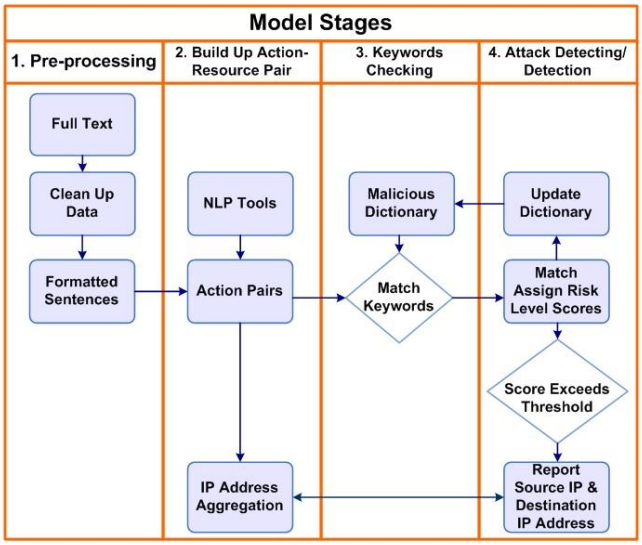

Figure 3. Model Flow Chart.

There are two key factors of this social engineering attack text sentence: the verbs and the related objects. The method in TBL model is to search general topic action pair, such as {"click", "link"} or {"provide", "account number"}. For this instance, the social engineering attacker is to mislead user's behavior and steal his or her confidential information. As noted above, TBL model uses a manually inputted dictionary based on company's policy rules and uses this dictionary to generate a topic blacklist. Let us consider a statement, "*Networking resources must not be manipulated*". When the action and resource pair is manually extracted, the action is here "*manipulate*" and the resource is "*networking resource*" (see Table 1). This kind of text based social engineering can be performed via not only email but also on some text-based platform such as twitter, social media or via simple text message.

Let us take another example. There is a dialog on a text-based social medium (an online chat incident). The attacker asked the user: "*Please provide your password.*" This is a very commonly used notification for the users who work with account login interfaces. This kind of message may come into the inbox in one's email account (perhaps, many have already received such text via email). We capture it in Table 1. The "*provide*" is a type of misleading behavior. Moreover, "*your password*" is a type of confidential information that should not be publicly saved. If the system finds the matching between sentence words and the *malicious dictionary* list, a notification would be prompted to alert the user that there could be a potential social engineering risk involved.

In the meantime, we also aggregate the IP addresses for those text/word sources. In this way, we could know which text comes from which source, especially which group of specific source IP and destination IP. It helps us identify the victim of social engineering attack and can be used in the attack detection stage (a sample is shown in Table 2).

iii. Keywords Checking - Figure 4 shows the pseudocode for an algorithm that checks whether the keywords in the detection table match with those in the malicious-dictionary database.

```
01    score = 0
02    for i in token.size:
03      if line[i] == dictionary.action:
04        for i+1 token.size:
05          if line[i+1] ==
dictionary.resource:
06              score = 1
07          elif:
08              score = 0.5
09  return score
```

Figure 4. Pseudocode for keyword matching.

The nested loop in the pseudocode is to check whether every word in a sentence is matched with the words in a *malicious-dictionary*. If a pair of the malicious words (resource and action) is found (i.e., both match), then the function would assign a risk score of +1. If none of the keywords is found, risk score would be kept the same (i.e., zero). Otherwise, the risk score is set to 0.5 as summarized in Table 3. Using this keyword match function, Table 4 lists several examples of score updating.

Table 1. Action-Resource Table.

Sr. No.	Action	Resource
1	Manipulate	network resource
2	Send	money
3	Send	data
4	call	number
5	provide	password
6	visit	web link

Table 2. Resource Table with IP Address.

IP_Source	IP Destination	Action	Resource
192.168.1.0	192.168.1.1	manipulate	network resource
192.168.1.0	192.168.1.3	send	money
192.168.1.4	192.168.1.2	send	data
192.168.1.4	192.168.1.1	call	number
192.168.1.4	192.168.1.1	provide	password
192.168.1.4	192.168.1.2	visit	web link

Table 3. Risk Score Table.

	Both Match	One Match	None Match
Changes of Risk Score	+1	+0.5	+0

iv. Attack Detecting/Detection - We build up a *malicious-dictionary* based on some social engineering emails and messages. After extracting a table of action-resource pairs from the full-text email, we search how many of the action-resource pairs matched in our malicious-dictionary.

Unlike the TBL model (as noted before, it uses a manually derived action-resource dictionary), to detect social engineering text, we propose a better mechanism to build up the action-resource pair table. The mechanism contains the following strategies:

Morphology: In linguistics, a verb phrase could have many forms. However, there is always a basic tense form which is called *Lemma*. In Table 5, for example, "*gave*", "*given*", and

"*giving*" are all different forms of the lemma "*give*". In order to learn from a dictionary so that a model could automatically use the dictionary, we find a verb's lemma form. This technique is often used in nature language processing by converting a text into its most fundamental meaning. For example, in Table 5, the first entry word is "*resetting*". Before analyzing the topic and action-resource pair, we replace "*resetting*" by "*reset*".

Table 4. Action-Resource Table Details.

IP_Source	IP Destination	Action	Resource	Risk Score
192.168.1.0	192.168.1.1	manipulate	network resource	+1
192.168.1.0	192.168.1.3	send	money	+1
192.168.1.4	192.168.1.2	tell	story	+0
192.168.1.4	192.168.1.1	say	hello	+0
192.168.1.4	192.168.1.1	thank	help	+0
192.168.1.4	192.168.1.2	give	feedback	+0.5

Table 5. Morphology Table.

Lemma	Morphology	Replace By
reset	resetting	reset
give	giving	give
send	sent	send
provide	provided	provide

Pluralization: In English language, there is a similar rule used for noun, which is called *pluralization*. Every noun has a singular form and a plural form. Even there is a scenario that both plural and singular are the same form but usually, there is a rule to follow (see Table 6). Thus, doing the same way, we replace plural words by their regular forms. For example, in Table 6, for the first entry, we replace "*accounts*" by its regular form "*account*".

Table 6. Plural Table.

Regular form	Plural Form	Replace
account	accounts	account
password	passwords	password
knife	knives	knife
child	children	child

Pronoun: A limitation of the TBL model [25] is that the pronoun is filtered out. For example, the sentence "*Go to BOA and send it to us*". This is not detected by TBL dictionary because of no matching with the word "*it*" in the sentence. This mapping issue could be addressed using *anaphora resolution* techniques in NLP [26]. Thus, we propose to provide converted pronoun before the text input. This method replaces the entire necessary pronoun with the related noun as depicted in Table 7.

Table 7. Pronoun Table.

No.	Action	Pronoun	Replaced
1	send	it	money
2	send	it	data
3	call	them	number
4	provide	them	password
5	visit	these	links
6	reset	it	password

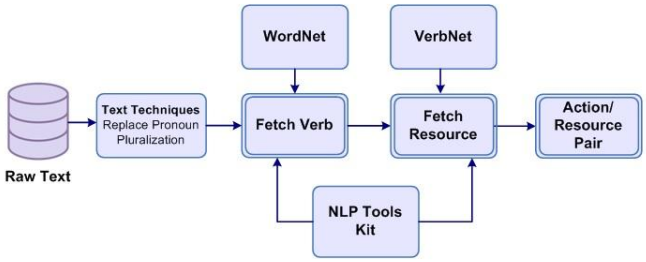

Figure 5. Flow Chart of Text Analysis.

Figure 5 summarizes these features in a flow chart of our model. After raw text processing, we need to run several text analysis preparations. Text analysis is all about deriving structured data from unstructured text. A well-understood process for text analytics includes the following steps:

1. Extracting raw text
2. Tokenizing the text—that is, breaking it down into words and phrases
3. Detecting sentence boundaries
4. Tagging parts of speech—words such as nouns and verbs
5. Parsing—for example, extracting facts and entities from the tagged text
6. Extracting knowledge to understand concepts such as a personal injury within an accident claim

With the help of NLP, we compare word phrases with WordNet and VerbNet databases, and find whether there is a match and create related action-resource pair for each IP package. In the next step, action-resource pair associated with its IP address would be aggregated.

Table 8. Hypothetical Accumulated Risk Score Summary.

IP Victim	Accumulated Risk Score
192.168.1.1	245
192.168.1.2	305
192.168.1.25	11
192.168.1.4	403
192.168.1.3	201
192.168.1.11	50

E. Score Updating

Since the manager of a team could get a risk score report periodically, the manager can decide what to do next. In fact, he/she could know when this social engineering attack happened and the accurate attack time of it. Accordingly, the manger could schedule a social engineering defense training course. For this, all the target machines with high-risk level scores could be notified and reminded by the IT (Information Technology) training team. More caution with well-prepared training course is always the most important weapon to defend against social engineering attacks. By employing the score updating mechanism described before (see Table 4), a table like Table 8 could be generated which shows the accumulated risk scores after a certain period of monitoring. The IT team could use this table to alert the victims or potential victims.

4. Experimental Settings, Results, and Analysis

We use Vmware [27] to deploy a client-server local network. We host two servers, one is for monitoring the traffic data, and the other one is acting as the attacking server that

forwards the malicious data to the clients as depicted in Figure 6. The clients are normal machines that might become social engineering victims. These machines are manually setup in a local network and assigned with specific IP addresses. For example, 192.168.1.0/24.

A. Experiment Tools

In the experiments, we use a couple of deployment tools, namely, NLTK and TextBlob as described below.

i. NLTK **-** Nature Language Toolkit (NLTK) is one of the most popular Python libraries in Nature Langue Processing field [28], [29]. NLTK provides user-friendly interfaces and includes more than 50 professional text corpora and related lexical resources.

ii. TextBlob **-** TextBlob [30] is another popular Python package in the text analysis area. It provides an API (Application Programming Interface) for diving into common NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation.

B. Databases

i. WordNet **-** For a database to analyze text, especially English, WordNet is one of the best tools. It has a database including English nouns and verb and related synonyms with specific category. Furthermore, WordNet is compatible with NLTK and all the tasks such as tokenizing and calculating *term frequency–inverse document frequency*.

ii. VerbNet **-** VerbNet is a database that classifies verbs according to their semantics and syntactic behavior.

iii. PhishMonger **-** Contains ~286,000 phishing websites collected between November 2015 and September 2017. This research is ongoing and more websites will be added to this portal as the researchers make them available to the public [31].

C. Experimental Design

The attacking server sends social engineering emails from PhishMonger database to five client machines in the same local network. This server runs an email service and sends malicious emails. The monitoring server runs an intrusion detection system (IDS) that keeps monitoring all those local network machines using a virtual router controller combined with a small port mirrored virtual switch. The server also has some commonly used log colleting tool installed, such as ***Nagios*** and it has email content's fetch environment setup so that it sends raw data to the analysis server. The analysis server aggregates those raw documents from monitoring server via analysis tools such as Nagios or Logstash as depicted in Figure 6. The analysis server computes text corpus experiments with Intel® Xeon® Processor E5-2687 v4 3GHz with 12 cores.

Corpus I contains a series of emails from actual phishing attacks. Corpus II is a set of normal emails that we use as a benchmark group. After we clean up the raw data, the Corpus I for demonstration contains a set of four different contents of phishing emails, which we refer to as Email #1, 2, 3 and 4. Each of the emails is the transcript of a phishing attack. The attacker's objective is to get confidential information. In order to gain user's (victim's) trust, the attacker made up some background stories and deceived the user to click a phishing link. In some previous works, the attacker often used many strategies to mislead user [12], [32].

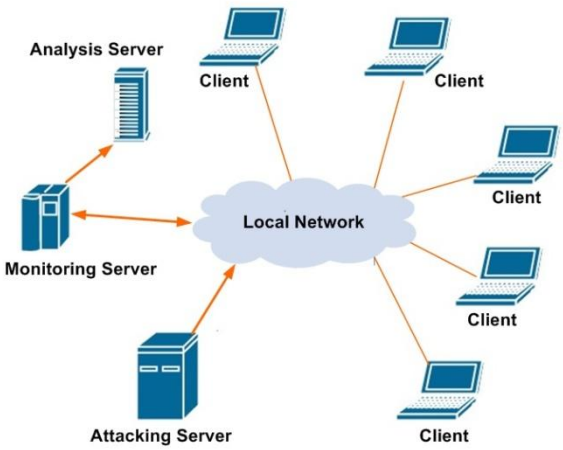

Figure 6. Experimental Network Setup.

We did the same for Corpus II which contains four different non-malicious emails, and we refer to them as Email #5, 6, 7, and 8. We also ran cleaning up and import operation into our model so that we can check false positives in the experimental results. It should be noted that we have received all necessary approvals to evaluate anonymized versions of the dialog transcripts. We removed all the confidential and personalized information in the example. Figure 7 shows a sample malicious email.

Hello.

Thank you for subscribing to this valuable information.
If this message arrived in error, please report it by clicking on the link below.
If, on the other hand, you would like to know why 98% of affiliates do not make money online, subscribe to the information.
HERE

Figure 7. Sample of a Phishing Email.

```
"subscribe", "information", "message"
"arrive", "report", "message", "click",
"link", "know", "why", "make", "money"
"subscribe", "information"
```

Figure 8. Tokenized Converted Words.

Figure 8 shows the result of our text processing using NLTK and WordNet. We converted the sample email into tokenized words and removed unnecessary words. Then, we transformed morphology into lemma form, substituted plural and replaced pronoun.

Table 9. Phishing Email Summary.

Corpus	Email #	Sentences	Converted Words	Identified Malicious #
I	1	6	14	2
I	2	37	20	4
I	3	27	19	3
I	4	19	15	3
II	5	8	14	1
II	6	9	8	0
II	7	7	5	0
II	8	6	5	0

Information about each email is shown in Table 9. The first column is the corpus group. Corpus I stands for malicious group, and Corpus II is the benchmark group with regular non-malicious emails. The second column shows the email

number. The third column is the number of sentences used by the attacker in the email. The fourth column is the number of tokenized converted phrases in each email. The fifth column summarizes the number of malicious action-resource pair that our model identified.

4.4 Analysis of the Results

The results of the experiments using our model on email text corpus are shown in Table 9. In summary, there are 8 emails (4 malicious and 4 non-malicious), 119 sentences and 100 converted phrases in our experiment. The threshold of malicious is whether the number of *Identified Malicious* is greater than zero. If *Identified Malicious* number is greater than zero, it indicates that the email is malicious. If the *Identified Malicious* number is equal to zero, our model identified the email as non-malicious, which is a normal email.

Table 10. Confusion Matrix.

		Actual class		
		Malicious	**Non-malicious**	**Total**
Identified	**Malicious**	4	1	5
Identified	**Non-malicious**	0	3	3
	Total	4	4	8

Table 10 displays the confusion matrix result of our model. The equations used for our calculations are:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$FPR = \frac{FP}{FP + TN} \quad (2)$$

$$TNR = \frac{TN}{FP + TN} \quad (3)$$

We assume that "*true*" means an email is actually malicious. TP is true positive and FP is false positive. FN is false negative and TN is true negative. Our True Positive Rate (TPR) is 100% (4 out of 4), which indicates that we have identified all the malicious phishing emails successfully. We misclassified one non-malicious email as malicious, so the False Positive Rate (FPR) is 25%. Thus, True Negative Rate (TNR) is 75%. In the next step, we analyze IP packet header and extract each IP address for text packet. Table 11 presents an example of risk score.

Table 11. Result of Single Client.

IP Victim	Accumulated Risk Score
192.168.1.2	13

Table 11 is based on the tracking Table 9. It presents a sample of action-resource case associated with IP address and risk score of our test result with one round of attack. We find that our model captures the sample attack in the client 192.168.1.2. The accumulated risk score is: (2+4+3+3+1)=13 for one round of attack. Thus, client (192.168.1.2) is identified as the victim in this case.

5. Conclusions and Future Research Directions

In this paper, we designed a framework that measures the behavior of the social engineers, and a comprehensive model for describing awareness, measurement and defense of social engineering based attacks. We proposed a hybrid multi-layered model using natural language processing techniques for defending the social engineering based attacks. The model enables the quick detection of a potential attacker trying to manipulate the victim for revealing confidential information.

In particular, we improved the Topic Blacklist (TBL) model with new features. Our model addresses the drawback of traditional TBL model dealing with pronoun words. We provided more features such as plural and morphology to have a more general approach to detect social engineering. Also, we designed a framework to analyze the language of real attacks, including a more visualized way to aggregate IP address of victim machine. The accumulated risk score gives a new monitoring perspective and offers a key component for building network-based detectors. Furthermore, emphasizing on the human element of social engineering, we suggest some pragmatic strategies to train the concerned people with the required information so that their online behaviors could act as a layer of defense against social engineering.

In future, we plan to work on a petabytes scale of database and observe the cases for a longer period with multiple clients. During the observation, we will compare accumulated risk scores and summarize efficiency of our model such as performance time. One issue that remains for our investigation is that we have a matching function which may not be efficient enough. Thus, we plan to look into this specific issue in a more detailed way and use machine-learning methods to compare whether commonly used classification algorithms would help increase the efficiency of malicious word matching.

References

[1] C. Snyder, "Handling Human Hacking: Creating a Comprehensive Defensive Strategy Against Modern Social Engineering," Thesis document, Liberty University, 2015.

[2] OWASP Top 10 2017: The Ten Most Critical Web Application Security Risks. available at: https://www.owasp.org/images/b/b0/OWASP_Top_10_2017_RC2_Final.pdf [last accessed: 24 October, 2017]

[3] G. Kumar, and K. Kumar, "Network security–an updated perspective," Systems Science & Control Engineering: An Open Access Journal, 2(1), pp. 325–334, 2014.

[4] K. Thakur, M. Qiu, K. Gai, and M.L. Ali, "An investigation on cyber security threats and security models," Proceedings of the 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud'15), pp. 307–311, 2015.

[5] K. Thakur, M.L. Ali, N. Jiang, and M. Qiu, "Impact of Cyber-Attacks on Critical Infrastructure," 2016 IEEE 2nd International Conference onBig Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE

International Conference on Intelligent Data and Security (IDS), pp. 183–186, 2016.
[6] Information Security Breaches Survey 2006. DTI, available at: http://webarchive.nationalarchives.gov.uk/+/http:/www.dti.gov.uk/files/file28343.pdf [last accessed: 25 October, 2017]
[7] C. Fung, and R. Boutaba, Intrusion Detection Networks: A Key to Collaborative Security. ISBN 9781138198890, Auerbach Publications, 2017.
[8] J.T.G. Kelsey, "Hacking into International Humanitarian Law: The Principles of Distinction and Neutrality in the Age of Cyber Warfare," Michigan Law Review, pp. 1427-1451, 2008.
[9] L. Graham, "Cybercrime costs the global economy $450 billion: CEO," Available at: https://www.cnbc.com/2017/02/07/cybercrime-costs-the-global-economy-450-billion-ceo.html (Available at: 28 October, 2017)
[10] "Runaway algorithms' and the cyber risks facing the global financial system," Australian Broadcasting Corporation, Published on 20 March, 2017, Available at: http://www.abc.net.au/news/2017-03-20/a-cyber-attack-could-cause-the-next-global-financial-crisis/8370860 [last accessed: 25 October, 2017]
[11] A.-S.K. Pathan, The State of the Art in Intrusion Prevention and Detection. (edited and contributed volume), ISBN 9781482203516, CRC Press, Taylor & Francis Group, USA, January 2014.
[12] K.D. Mitnick, W.L. Simon, and S. Wozniak, The Art of Deception: Controlling the Human Element of Security, ISBN-13: 978-0764542800, Wiley; 1st edition, October 17, 2003.
[13] A. Johnson, "Massive Phishing Attack Targets Gmail Users," NBC News, Published on May 4,2017, Available at: https://www.nbcnews.com/tech/security/massive-phishing-attack-targets-millions-gmail-users-n754501 [last accessed: 25 October, 2017]
[14] "Austria's FACC, hit by cyber fraud, fires CEO," May 25, 2016, Available at: https://www.reuters.com/article/us-facc-ceo/austrias-facc-hit-by-cyber-fraud-fires-ceo-idUSKCN0YG0ZF [last accessed: 27 October, 2017]
[15] F.J. Bjorck, "Discovering information security management," Ph.D. thesis, Stockholm University & Royal Institute of Technology, 2005.
[16] A. Ferreira, and G. Lenzini, "An Analysis of Social Engineering Principles in Effective Phishing," 2015 Workshop on Socio-Technical Aspects in Security and Trust (STAST), DOI: 10.1109/STAST.2015.10, July 2015.
[17] D. Bisson, "Clever Gmail Phishing Scam Tricked Even Technical Users," The State of Security: News. Trends. Insights. March 20, 2017, Available at: https://www.tripwire.com/state-of-security/latest-security-news/clever-gmail-phishing-scam-tricked-even-technical-users/ [last accessed: 25 October, 2017]
[18] J. Mao, W. Tian, P. Li, T., Wei, and Z. Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity," IEEE ACCESS, Volume 5, pp. 17020-17030, 2017.
[19] P. Patil, R. Rane, and M. Bhalekar, "Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm," 2017 International Conference on Inventive Systems and Control (ICISC), 19-20 Jan. 2017.
[20] S. Marchal, G. Armano, T. Gröndahl, K., Saari, N. Singh, and N. Asokan, "Off-the-Hook: An Efficient and Usable Client-Side Phishing Prevention Application," IEEE Transactions on Computers, Volume: 66, Issue: 10, Oct. 1, pp. 1717-1733, 2017.
[21] E.D. Frauenstein, and R. von Solms, "Combatting phishing: A holistic human approach," Information Security for South Africa (ISSA), DOI: 10.1109/ISSA.2014.6950508 2014.
[22] M.A. Sasse, S. Brostoff, and D. Weirich, "Transforming the 'Weakest Link' — a Human/Computer Interaction Approach to Usable and Effective Security," BT Technology Journal, Volume 19, Issue 3, pp 122–131, 2001.
[23] Thakur, K., "Analysis of denial of services (DoS) attacks and prevention techniques," International Journal of Engineering Research & Technology, 4 (7), 2015, pp. 171–176.
[24] K. Thakur, M.L. Ali, K. Gai, and M. Qiu, "Information security policy for e-commerce in Saudi Arabia," Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016 IEEE 2nd International Conference on, pp. 187-190, 2016.
[25] R. Bhakta, and I.G. Harris, "Semantic analysis of dialogs to detect social engineering attacks," 2015 IEEE International Conference onSemantic Computing (ICSC), DOI: 10.1109/ICOSC.2015.7050843, 2015.
[26] K. Dhole and H. Kohli, "Document categorization using semantic relatedness & Anaphora resolution: A discussion," 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), DOI: 10.1109/ICRCICN.2015.7434279, 20-22 Nov. 2015.
[27] Vmware. Available at: https://www.vmware.com/in.html (last accessed: 28 October, 2017)
[28] S. Bird, E., Klein, and E. Loper, "Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit," O'Reilly Media, 2009.
[29] Natural Language Toolkit, Available at: http://www.nltk.org/ (last accessed: 28 October, 2017)
[30] TextBlob: Simplified Text Processing, Available at: https://textblob.readthedocs.io/en/dev/ (last access: 28 October, 2017)
[31] Internet Phishing Websites, Available at: http://www.azsecure-data.org/phishing-websites.html (last accessed: 28 October, 2017)
[32] S. Grazioli, "Where Did They Go Wrong? An Analysis of the Failure of Knowledgeable Internet Consumers to Detect Deception Over the Internet," Group Decision and Negotiation, Volume 13, Issue 2, pp. 149-172, 2004.
[33] G. Aaron and R. Rasmussen, "Global Phishing Survey: Domain Name Use and Trends in 2016," APWG, Available at: https://apwg.org/resources/apwg-reports/domain-use-and-trends (last accessed: 28 October, 2017)
[34] A. Abduvaliyev, A.-S.K. Pathan, J. Zhou, R. Roman, and W.-C. Wong, "On the Vital Areas of Intrusion Detection Systems in Wireless Sensor Networks," IEEE Communications Surveys & Tutorials, Volume: 15, Issue: 3, pp. 1223-1237, Third Quarter 2013.