Scaling and Effectiveness of Email Masquerade Attacks: Exploiting Natural Language Generation

Shahryar Baki
University of Houston
shahryar@cs.uh.edu

Rakesh Verma
University of Houston
rmverma@cs.uh.edu

Arjun Mukherjee
University of Houston
arjun@cs.uh.edu

Omprakash Gnawali
University of Houston
gnawali@cs.uh.edu

ABSTRACT

We focus on email-based attacks, a rich field with well-publicized consequences. We show how current Natural Language Generation (NLG) technology allows an attacker to generate masquerade attacks on scale, and study their effectiveness with a within-subjects study. We also gather insights on what parts of an email do users focus on and how users identify attacks in this realm, by planting signals and also by asking them for their reasoning. We find that: (i) 17% of participants could not identify any of the signals that were inserted in emails, and (ii) Participants were unable to perform better than random guessing on these attacks. The insights gathered and the tools and techniques employed could help defenders in: (i) implementing new, customized anti-phishing solutions for Internet users including training next-generation email filters that go beyond vanilla spam filters and capable of addressing masquerade, (ii) more effectively training and upgrading the skills of email users, and (iii) understanding the dynamics of this novel attack and its ability of tricking humans.

Keywords

Phishing Email; Social Engineering; Experimental Study; Within-Subject Study; Hillary Clinton Emails; Sarah Palin Emails

1. INTRODUCTION

Security, so far, has been largely a reactive field wherein attackers expose new vulnerabilities, which are then patched by defenders. Another problem has been that the solutions have been to a large extent one-size-fits-all. For example, in the case of spam, phishing and malware-containing emails, organizations have installed email filters, which are typically based on machine learning techniques. The problem with machine learning techniques is well-known. They work well when the instance in question is similar to the historical data on which they have been trained [39]. Knowing this, the attackers constantly change the attack, so that the attacks escape the email filters. These new attacks, when they reach the inboxes of unsuspecting users, cause havoc, which periodically makes it into the news headlines, but most often is kept under wraps by companies worried about tarnishing their reputations. We need to change the playing field. Everyone understands this, but the crucial question is how.

We believe that a multi-pronged approach is needed. First, we need to give defenders insights and tools that will make them proactive rather than reactive so that they can improve their email filters to anticipate the next generation of attacks. Second, we need to equip Internet users and employees with defense-in-depth. After the generic email filter, which will invariably fail at some point *even if it is much better than before*, should come the second level of defense: a customized solution that: (i) takes into account the context and past behavior of the user to warn and defend the user from the attacks that escape the cookie-cutter filter, and (ii) constantly trains and upgrades the skills of the user. In this paper, we show how attackers could generate masquerade attacks on scale and gather insights and tools to: help defenders improve their machine learning filters, generate customized solutions, and train/upgrade user skills.

We focus on the realm of email-based attacks, which could be phishing/spear-phishing, malware. As others have observed, e.g., see [7], "email-based attacks are probably one of the most effective in today's hacker bag of tricks." Our key findings and contributions are as follows:

1. We show how to use current Natural Language Generation (NLG) to generate several variants of masquerade attacks with a modest amount of manual effort (5-6 hours for someone with a basic knowledge of NLG techniques). We then study their effectiveness with a within-subjects study. It is well known that there is a division of labor in the phishing ecosystem [19], so although not every phisher will spend this kind of effort or have the needed knowledge, but a subgroup could spend that effort so that more could reap the rewards.
2. By planting signals in the generated emails and explicitly asking for their reasoning, we gather insights on the parts of the email that users focus on and how users identify attacks in this realm.
3. Real world experiments with a varied participant population indicate that the proposed masquerade email attack is non-trivial: for humans, detection rates only

ASIA CCS '17, April 02-06, 2017, Abu Dhabi, United Arab Emirates

DOI: http://dx.doi.org/10.1145/3052973.3053037

slightly better than random (50%), and also for state-of-the-art automatic approaches. We manipulate some variables in the emails (signals, real/fake and reading level) and study how they affect the detection rate. We also investigate whether unmanipulated variables such as, demographics, email knowledge/experience, time spent, personality, and strategy used, make a difference or not. We find that none of these make any difference, except for personality traits of extraversion and conscientiousness. (Section 4.4.1).

4. The proposed masquerade attack can deceive 74% of participants when there is no embedded signal (Section 4). Masquerade attack on Hillary Clinton's emails fooled 34% of the people and on Sarah Palin's fooled 71% of the people (Section 4.2).

Our work suggests that current NLG techniques are already effective even for: experienced users of emails, and users with different ages and gender (Section 4). Therefore, they can be used by defenders to improve their filters by subjecting them to freshly synthesized attacks. Similarly, they can also be used to train Internet users/employees by sending users newly generated attacks. Our work also yields insights into reasoning strategies used by users to identify email attacks, which can be used to design better training tools and materials (Section 4.4.3). Regardless of how sophisticated were the strategies used, our method achieves similar success in deceiving users, with performance ranging from 46% to 57% (Section 4.4.3) Our work could spur further research in NLG techniques, which will further help defenders and Internet users/employees. Because we focus more on how users identify deceptiveness/impersonation and not on characteristics of specific attacks, our work is more generally applicable to any attack involving an email, be it a company representation fraud, phishing, malware, spear-phishing, or a totally new attack involving an email.

2. MASQUERADE ATTACK

Email spam [8] and phishing [26] have been studied extensively in the literature. However, the problem of masquerade in emails has not received much attention. To our knowledge, the closest works on masquerade are those in the context of a user impersonating another (e.g., by a compromised account) and the detection mechanism tries to find anomalies in the character streams of commands issued during the masquerader's session [31, 30, 36]. In [41] the problem was addressed using recursive data mining and author identification methods. Other flavors of the problem in computer security are those appearing in [35] that study the problem in the context of search behaviors and [47] that employ statistical learning and one-class classification.

Email masquerade refers to the situation where an adversary after gaining access to the email of a (potentially prominent) person, scans previous emails and learns about the writing style of the compromised account. The masquerader also learns about the contacts of the compromised account and context in which previous conversations took place. The masquerader then uses these to his advantage to send out new emails to the various contacts of the compromised account. The emails sent by the masquerader have fake content (yet they simulate the same style and context of the compromised account). The attack could be used for miscommunication, disharmony, discord, phishing, malware downloads via clickjacking/link spam or simply email abuse and fooling people. It can also be exploited with modest human effort to create misinformation and political campaigns, given that so many emails of politicians are now available on the web. The problem is important and has recently been covered in the news [14], particularly, in the banking industry where the problem is referred to as "executive impersonation." Here, we consider a specific version of masquerade attack in which the impersonator copies the style and grammar of the person with whom the victim had communicated before.

2.1 Data

To understand the dynamics of the masquerade problem and its detection hardness, we need actual samples of masquerade. However, the inherent nature of the problem prohibits obtaining large-scale samples of real-world cases of masquerade emails. One main reason for this is that one can get real-world cases of email masquerade only by either (1) confession of the hacker or (2) an affirmation of the compromised account holder that he/she was a victim of email masquerade, as the receivers of the masqueraded emails will never know for sure. Clearly the above factors may not be possible to obtain thereby prohibiting us to explore large-scale real-world cases of masquerade emails. The problem gets harder if we are trying to understand masquerade on eminent personalities due to censorship. Nonetheless, we devise a novel scheme to simulate the masquerade attack. We generated deceptive masquerade emails using a Natural Language Generation tool, specifically the Dada Engine [5].

We decided to explore the attack for two eminent personalities: Hilary Clinton (HC) and Sarah Palin (SP). Their emails were obtained from the archives released in [42, 48]. After parsing the content, style and structure of the original emails, we created a grammar that simulates emails as if being written by the original sender (HC/SP).

Natural Language Generators and the Dada Engine. One of the aims of computational linguistics and natural language generation is to facilitate the use of computers by allowing the machine and their users to communicate using natural language. Usually, a generator works by having a large dataset of knowledge to pull from that is then manipulated by programmable grammatical rules in order to generate readable text. This process is usually broken down into two steps: text planning, which is concerned with deciding the content of the text, and realization, which is concerned with the lexigraphy, and syntactic organization of the text. Some of the variations of these approaches appear in [15]. For our purposes, we used the Dada engine, which has been successfully used to construct the academic papers on postmodernism [6].

The Dada engine is a natural language generator tool that is based on the principle of recursive transition networks, or recursive grammars. A recursive transition network (RTN) can be thought of as a schematic diagram of a grammar, which shows the various pathways that different yields of the grammar can take. For example, in the construction of a sentence, one may choose to follow the RTN has been shown in Figure 1.

If one follows the RTN in Figure 1 from the start to the end, one passes through states (boxes) representing the various elements which make up the sentence in sequence: first a preposition, then an optional adjective, then a noun and

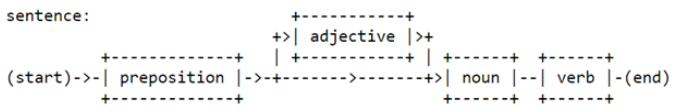

Figure 1: Example of Recursive Transition Network

then a verb. After the preposition, the network branches into two paths, one of which leads to the adjective and the other which bypasses it and goes straight on to the noun.

2.2 Generating Masquerade Emails with Dada

The goal is to tune the grammar of Dada that respects key stylistic elements of the original (compromised account) author yet induce deception via content to simulate masquerade behavior. Further, we mandated that the generated email should be similar to real emails in all aspects: time at which the email was sent, the way that the actual author uses punctuation, sophistication level of grammar rules that he/she uses, etc. This process entailed manually combing (2-3 hours of effort) through a dataset of true emails in order to construct a structure for the generated text. This included creating an email header that contained the names of the recipients of the email. These names were stored in variables and used throughout the script in order to maintain consistency. Other constraints were added in order to ensure the generated emails matched in terms of style. This included creating methods that allowed for occasional misspelling and abbreviations, if the actual author has misspellings or uses abbreviations in his/her writing. Dada follows the programmed grammatical rules to select the appropriate subject and organization to create coherent sentences.

For writing the grammar (approximately three hours of effort), we went over real emails and tried to find key features in their writing style. Misspelling, abbreviation, prepositional phrases, conjunctions, correct/incorrect usage of punctuations are some of these features. Then we generated all fields of an email, subject, to, date/time, and body of the email. There are some semantic constraints in generating all these fields automatically. Data/time should be crafted carefully in order to have a correct day of the week. For example, we should not generate something like "Sun Nov 1 01:10:03 2012", since November 1 is Thursday. The most important constraint is the relevance of the body of the email to the subject line. Solving date/time issue was done by generating a list of valid date and time and using that list in the grammar. For the subject-body relevance, we made our grammar hierarchical by grouping the rules. Each group has its own topic and each group has its own rules for generating the subject line. We illustrate this process with an example.

First, here are sample sentences that discuss a meeting:

- "Could you pls bring a copy of the strategy memo with you."
- "Where is the meeting?"
- "Has the meeting been canceled?"

And the following sentences are talking about a report:

- "It is markedly better."
- "But as I'm reading thru it, I notice missing words"

Now, if the selected topic is "meeting", here is the procedure for generating an email: The nonterminal *Request* will search through different text literals such as "Is the" and "When will". The nonterminal *Subject* will be replaced by the text literals such as "memo" or "meeting." The nonterminal *Transition* contains transitional phrases that can be used such as "be at" or "be after" and the nonterminal *Time* will return a time. The grammatical rules of our script construct sentences in this way to generate a full text. Full text can be seen below in generated emails. For Hilary Clinton:

> **From**: H [HDR22@clintonemail.com]
> **To**: Wilhelm Hamburger
> **Cc**: Paul Buxton, Stephen T. Drucker
> **Sent**: Thu Nov 1 01:10:03 2012
> **Subject**: Strategy memo
> Pls type this out in BIG print for me. Can you please contact Paul Buxton to see what we have to do to get them on board. Thx.

In this text, the use of the name variable is apparent as the name in the CC section is mimicked in the text body. This ensures a convincing deception. For generating names, we used a list of first and last names of different countries and chose from them randomly.

Although we used NLG to generate the masquerade samples, it should be noted that it is ground truth (we know those emails are fake as they were machine generated and not from the original author). One can argue it lacks the mindset of an actual masquerader which is true and a factor very difficult if not impossible to simulate. We believe that it nevertheless provides us a decent opportunity to explore this novel attack and provides us sufficient ground truth masquerade samples for an eminent email author that are otherwise extremely difficult to procure. As we will see in the result section, even expert human detection performance was close to random in detecting true vs. masquerade email showing that the data generated is reasonable.

We intentionally added some fake signals into output of our system to see: which parts of an email people focus on and how sophisticated a system needs to be to deceive people. Particularly, we considered these errors: Fake name, repeated sentence, and incoherent flow of idea. "Captain America" and "Humpty Dumpty" are the fake names that we used. *Incoherent flow of idea* is the case that the email talks about two different topics. Here, we show a sample question with *incoherent flow of idea*.

> **From**: Person2
> **To**: Wilhelm Tournier
> **Date**: Thursday, August 02, 2007 01:46 PM
> **Subject**: Meeting
> Thomas coverage of the ethics issue made Martin look like she was on the high rd by basically claiming I was inaccesible and we just don't understand the process. Can you get back to them and let her know if someone is available to go the event?

We also see in this script the misspelling of the word "inaccessible" which is characteristic of Palin's emailing habits. We will later see the impact of each of the above three fake signals towards the effectiveness of attack in the results section. It will also show what features of an email do humans focus on when they are engaged in identifying deception.

It is worth noting that in some emails, we as a sender ask

for some information from the receiver, which is common in spam and phishing emails, but we do not have any email that asks the receiver to download an attachment. The reason is the dataset that we used for the NLG. In Hillary Clinton and Sarah Palin's emails, there is no email that asks for downloading an attachment. So if we add it manually as a fake signal to the generated emails, it is a big deviation from the real emails (we do not have any similar real email to show the participants at the beginning).

3. EXPERIMENT SETUP

To evaluate effectiveness of our proposed attacks and investigate the reasoning/strategies that people use in their decision-making process, we perform an experimental study. We use some real emails and some fake ones (created by the attack introduced in Section 2 to generate the questions. Figure 2 shows the process flow diagram of the experiment. First, all participants did a personality test before coming to the lab. This was done to reduce the time spent in the lab and the potential for fatigue. We used the Big Five Personality test for measuring personality traits of participants. The Big Five Personality Traits, is a well known and widely accepted model for measuring and describing different aspects of human personality and psyche [9]. This model describes human personality from five broad dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. We check later whether there is any correlation between personality traits and performance of participants.

In the lab, we started by asking participants some basic questions about their email experience: approximate number of emails received each day, years of email use, and the spam filter used if any. Next, we ask them about education and computer background. Once the participants answer these questions, the main part of the experiment starts (gray area in the flowchart). We divided the questions into two parts: Hillary Clinton's emails and Sarah Palin's emails in this order. In each part, before the questions, we gave six real emails of the actual author, then we show the eight different questions and for each we asked them to decide if it is fake or real. The participants were given sufficient time to become familiar with the styles of the emails sent by the senders to emulate a real world scenario in which the victims are deceived by not only the sender name and email address but also the styles with which they are familiar. We ask each participant "Do you think this email is Real or Fake?" exact question from our survey, and after that we ask about the reasoning, "If you think this email is fake, please list ALL the reasons that made you think the email is fraudulent. Otherwise, justify why you think this email is real."

In addition to the questions, we also asked participants to indicate their confidence level, "How confident are you about your answer?" to know if they got lucky. Range of confidence is from 1 (least confident) to 5 (extremely confident).

During the experiment, a PhD student was assigned to keep track of the responses (in real time) given by the participants. We have two reasons for this real time checking. First, to make sure all the reasoning and responses are clear and unambiguous. Second, to avoid human mistakes. We had two cases in which, during the checking of responses, the student found an inconsistency between the participant's response and his/her reasoning, e.g. one of the participants chose "fake" as a response but in the reasoning part we got

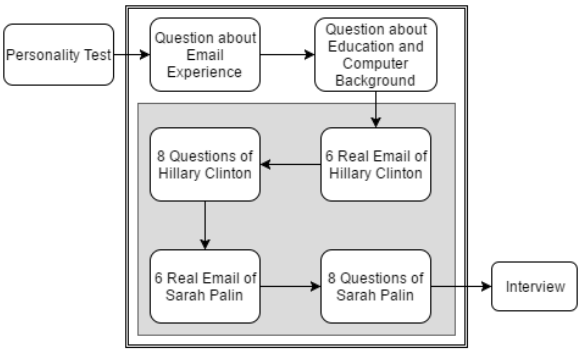

Figure 2: Flowchart of entire phishing experiment

"his/her language is similar to Person 1 real emails." The student noted all these issues and asked the participant(s) about them after the experiment.

At the end, we interviewed the participants to evaluate their knowledge and experience about email. "What are the different parts of the email" and "How much do they rely on sender's email address in order to decide if an email is fake or not" are some examples of these questions. We also asked them to show the full header of an email in a client that they use.

We have 16 questions in the study for which participants need to write down their complete reasoning besides choosing their answers: fake/real. Such an experiment may make the participants bored or tired, and they could start hurrying through the last few questions. The time that they spent on each question is a good indicator to see if they are really got bored/fatigued or not. We divided the 16 questions into four parts: first, second, third and fourth quarter and then calculated the average spent time in each part. We found that there is no significant difference between these groups (p-value=0.549) and the averages (stddevs) are 128.15 (78.28), 115.47 (71.94), 126.93 (55.84) and 108.09 (57.31) seconds respectively. We can say that participants gave the same amount of attention since they spent similar amounts of time on first group as the last group of questions.

While some might argue, such as the work of [29] albeit conducted in a financial setting, that the lack of any risk in a lab setting may make the subjects behave very differently from a real-world scenario, others have argued for more laboratory experiments not less [13].

3.1 Datasets

Building block of the NLG is a dataset of actual texts which is needed to extract the grammar rules. As mentioned in section 2.1, for the masquerade attack we use Hillary Clinton and Sarah Palin as the persons whose identity is being masqueraded, and we use Dada engine in order to generate fake emails that are similar to their actual emails. To remove any effect from familiarity with Hillary Clinton or Sarah Palin, we removed names of the actual authors (HC and SP) and used Person1 and Person2 instead.

The Flesch-Kincaid readability test was used in order to measure the readability level of Hillary Clinton and Sarah Palin. This test has two parts, Flesch Reading Ease, and the Flesch-Kincaid Grade Level. In the first part, the higher the measure means the easier it is to read and understand.

The Range of this score is usually between 0 to 100. Flesch-Kincaid Grade Level shows how many years of education are needed to understand a text. For example, if the grade level a text is 10, it means 10 years of U.S. schooling is needed.

We randomly chose 50 emails (25 for each author) and calculated the two aforementioned scores for each of them. Flesch Reading Ease for Hillary Clinton is 79.78 and for Sarah Palin 66.81, and Flesch-Kincaid Grade Level is 3.78 and 7.35 respectively. These results show that Sarah Palin's emails are harder to read than Hilary Clinton's emails. We study how this affects performance of participants in detecting real and fake emails of SP and HC.

3.2 Participants

Before running the experiment we requested IRB approval. After receiving the approval, we conducted a small pilot study with three participants before running the actual survey. During the pilot study we found and fixed a few problems in the experiment design. Below is the list of major improvements that we made:

- **Fake Name**: We found that the fake names we had used were hard to detect. We had combined first name and last name from different languages (e.g. French and German) and included more than one middle name. We decided to change these into more obvious fake names, e.g., *Captain America* and *Humpty Dumpty*.
- **Proper Real Emails**: At the beginning of each part (HC/SP), we give six sample emails of each actual author to the participants, and there were four real emails in the questions. We had some signals (e.g., abbreviations) in the real emails of HC/SP in the questions part that did not exist in the six sample real emails, so participants marked them as fake because of these signals. We checked these signals and chose six sample real emails that corresponded more closely to the real emails in the questions.

After the pilot study, a recruitment email was sent to all the students at the college of Natural Sciences and Mathematics, which includes six departments (and over 20 majors): Biology & Biochemistry, Chemistry, Computer Science, Earth & Atmospheric Science, Mathematics, and Physics. To further diversify the participant pool, we also recruited staff so we have some majors from other colleges also. We had 34 participants, of which 15 were female (44%) and 19 male (56%). The majors of our participants are: computer science (33%), Biology (26%), Chemistry (12%), finance (9%) and others (20%). From the academic degree aspect, four of them are Ph.D. student, two are Masters student, 26 Undergrad student and two of them have High School Diploma.

For their spam filter, 30 participants (88%) use the Gmail spam filter. Kaspersky, McAfee, Yahoo, and Web Of Trust chrome extension is each used by one participant (3% each). Below is some other information about characteristic of the participants.

- Age ranges from 18 to 64 years (Mean = 25, stddev = 10.7, var = 114.61).
- Number of emails received daily ranges from three to 100 (Mean = 19.66, stddev = 19.74, var = 389.70).
- Years of email usage range from four to 20 (Mean = 10.55, stddev = 3.76, var = 14.17).
- Social network usage per week (number of times checked) ranges from zero to 1200 (Mean = 71.81, stddev = 222.96, var = 49712.66)

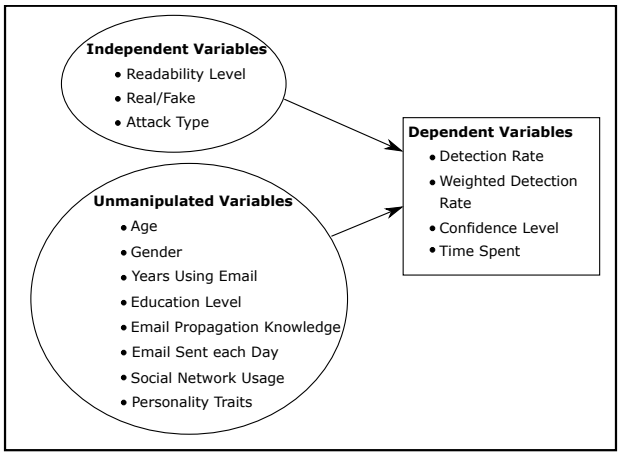

Figure 3: Relationship Between Variables

3.3 Variables

Several variables exist in the study. Figure 3 shows the variables and their relationships. Independent variables are those that we manipulated on different questions to see how they affect performance. Dependent variables are the observed variables and the goal of this study is to find how they depend on the other variables. Most of the dependent variables are related to the performance of the participants.

Detection rate (DR) is proportion of emails that are detected correctly, real detected as real and fake detected as fake. *Confidence level* is the degree of confidence that participants have in their answers. *Weighted detection rate* is the combination of detection rate and confidence level. The reason for having weighted detection rate (WDR) in addition to detection rate is that DR only considers the final output of participants for each question, either zero or one. We need a way to combine the final output with participants' confidence level. WDR does this by multiplying confidence level with +1/-1 for correct/wrong answers. *Time spent* is the time that participants spent for answering each question.

Table 1 explains the independent variables. Hillary Clinton and Sarah Palin are used as the authors with two different readability levels. The questions part includes eight questions for each of them, four generated (fake) and four real emails. Three of those four fake emails have embedded signals of falseness and one is without any embedded signal.

Unmanipulated variables are variables over which we have no control. We check if there is any relationship between these variables and dependent variables. By randomly choosing the participants from the population we reduce the effect of these variables on our study. Table 2 shows unmanipulated variables and their description (Age and gender are removed from the table). For *Email Propagation Knowledge*, we defined four levels of knowledge: (i) those who do not know anything about it, (ii) those who think there is a server that receives from the sender and sends it to the receiver, (iii) those who think there is more than one server in the middle (sender mail server and receiver mail server), and (iv) those who have fairly comprehensive knowledge.

Table 1: Independent Variables and Description

Independent Variable	Description
Readability Level	Readability level of the actual author. Flesch-Kincaid test used for calculating readability
Real and Fake	Actual emails and masquerade attack emails
Attack Type	Different fake signals added to the emails: fake name, repeated sentence, incoherent flow of idea

Table 2: Unmanipulated Variables and Description

Unmanipulated Variable	Description
Years Using Email	Number of years using email
Education Level	Participants seeking degree
Email Propagation Knowledge	Level of knowledge about how emails are sent through the Internet
Email Sent Each Day	Number of emails sent each day
Social Network Usage	How many times per week they use any social network
Personality Traits	Five personality traits defined by Big Five Personality Test

4. RESULTS

In this section, we present the effectiveness of the attack and whether any variables had any statistically significant effect on the detection rate or WDR.

First, we study the effect of independent variables on dependent ones, then we check the effect of unmanipulated variables on dependent variables. Table 3 shows detailed description of all questions in the masquerade attack with performance of participants and their average confidence level for each question. Participants' detection rate ranges from 5/16 (31.25%) to 12/16 (75%) (mean=0.529, stddev=0.099, var=0.0098). Their performance on correctly identifying real emails ranges from 1/8 (12.5%) to 6/8 (75%) (mean = 0.452, stddev=0.125, var=0.016) and in correctly identifying fake emails range from 2/8 (25%) to 8/8 (100%) (mean=0.607, stddev=0.160, var=0.026). Average confidence level of participants in masquerade attack ranges from 2.81 to 5 (mean=3.61, stddev=0.454, var=0.206).

Automatic Detection Technique. We also implemented a state-of-the-art one-class authorship verification (AV) technique [27] and deception detection (DD) [1]. We used the emails entire datase of Hillary Clinton and Sarah Palin and 100 NLG generated emails for each of them as training set, except the 16 real emails that are used in our survey, and tested the real and fake emails with these technique. The results are in the last two columns of the Table 3, 62.5% accuracy for AV and 75% accuracy for DD. For SP's emails, the AV technique considered all emails as real (50% accuracy), and DD technique considered only two fake email as real (75% accuracy). For HC's emails, both AV and DD misclassified two fake emails as real (75% accuracy). So, even with training on the entire dataset (except the test emails) these technique could not identify fake emails correctly.

Table 3: 16 questions, their features and percentage of right and wrong answers (average confidence level). n- question number. R/F- the correct answer. G- real email. FN- Fake name, IFI- Incoherent, RS- Repeated Sentence and NE- No Error. AV and DD are the outputs of authorship verification

	n	R/F	Feature	Right	Wrong	AV	DD
Hillary Clinton	1	R	G	88.23 (4.1)	11.76 (3.5)	R	R
	2	F	FN	82.35 (3.25)	17.64 (3.6)	R	R
	3	R	G	73.52 (3.96)	26.47 (2.88)	R	R
	4	R	G	67.64 (3.6)	32.35 (3.09)	R	R
	5	F	IFI	61.76 (3.9)	38.23 (3.61)	F	R
	6	R	G	67.64 (4.08)	32.35 (3.27)	R	R
	7	F	RS	76.47 (3.46)	23.52 (3.62)	R	F
	8	F	NE	55.88 (3.73)	44.11 (3.73)	F	F
Sarah Palin	9	R	G	5.8 (5)	94.11 (4.34)	R	R
	10	F	RS	47.05 (3.06)	52.94 (3.94)	R	F
	11	R	G	26.47 (3.33)	73.52 (3.44)	R	R
	12	R	G	20.58 (3.28)	79.41 (4.03)	R	R
	13	F	IFI	67.64 (3.08)	32.35 (3.45)	R	R
	14	F	FN	50 (3)	50 (3.52)	R	R
	15	F	NE	44.11 (2.53)	55.88 (3.42)	R	F
	16	R	G	11.76 (3.75)	88.23 (3.83)	R	R

Table 4: Detection rate (DR), average confidence level (ACL) and weighted detection rate (WDR) of participants for all 16 questions

	DR	ACL	WDR
Overall	52.94%	3.54	0.134
High Conf.	46.52%	4.34	-0.091

Table 4 provides overall performance of participants on masquerade attack. Weighted detection rate, detection rate, and average confidence level are shown in the first row. Detection rate of responses that have confidence level at least four (High Conf.) have shown in the second row of the table.

Detection rate is about 50% and WDR is close to 0. This means that the participants are not able to reliably distinguish between fake and real emails. We should keep in the mind that over all 8 fake questions that we have, six questions have obvious embedded signals of falseness (fake name, repeated sentence, and incoherent flow of idea), two questions per signal. So if we exclude those fake emails that have embedded signals, we might have an even more effective attack although it will increase attack generation time a little. To answer this question, we need the performance of participants on each type of fake email separately to see how much complexity is needed for an effective attack. In the rest this section, we analyze the effect of unmanipulated and independent variables on our dependent variables.

We have three different independent variables: real and fake email, readability level and fake signals. We want to know how these variables can affect performance of participants, in general how they affect the dependent variables.

4.1 Real and Fake

Half of the questions in the experiment are real and the other half are generated by masquerade attack (fakes, for short). Comparing these two groups of questions gives us an insight about the detailed performance of the participants.

Table 5: Detection rate (DR), average confidence level (ACL) and weighted detection rate (WDR) of participants for 8 real and 8 fake questions

Questions	Metric	Overall	High Conf.
Real	DR	45.22%	43.53%
	ACL	3.79	4.39
	WDR	-0.301	-0.383
Fake	DR	60.66%	50.36%
	ACL	3.41	4.3
	WDR	0.56	0.409

Table 5 shows the performance of participants on real and fake emails separately. The results illustrate that participants do better in detecting fake emails than real emails. This could have happened because of an in-lab experiment effect, viz., a slight bias towards marking emails fake. We perform a significance test to see whether the difference in performance is by chance or not. To answer this question we test these two groups to find a significant difference between various variables (performance, confidence, spent time):

- Comparing mean DR of participants on real and fake emails (p=6.979e-05, df=62.523, t=-4.2611)
- Comparing mean WDR of participants on real and fake emails (p=0.002 , df=62.157, t=-3.1272)
- Comparing mean confidence of participants on real and fake emails (p=0.005, df=62.26, t=-2.8977)
- Comparing average spent time of each participants on real and fake emails (p=0.3378, df=65.197, t=-0.965)

P-value for first three tests are less than 0.05. Since we are running the t-test on each response variable more than one time (three times in this case), this increases probability of type 1 error (multiple comparison problem). To avoid this issue, we utilize Benjamini-Hochberg [18] procedure to find significant p-values. First three tests are still significant after applying the Benjamini-Hochberg procedure. So, there is significant difference between performance and confidence of participants on real and fake emails. Fake emails are easier to detect with higher confidence. This shows that the difference we saw in the previous section is significant and not by chance. This suggests that participants have a slight bias towards choosing fake, so our attack could have higher success rates in a non-lab setting.

4.2 Hillary Clinton and Sarah Palin

Hillary Clinton and Sarah Palin's writings are different from each other from the readability aspect. In section 3.1, we showed that Sarah Palin's emails are harder to read. We now study how this difference can affect performance of participants in detection of real and fake emails. For example, if we know that participants are more easily deceived when the attacker mimics the email of a writer whose emails are simpler, then it would be better to design attack vectors for easy-to-read authors.

Table 6 shows performance of participants on masquerade attack for Hillary Clinton and Sarah Palin separately (both real and fake). Detection rates of those responses with confidence level at least four (High Conf.) have been shown in the fourth column of the table. Comparing performance of participants on detecting Sarah Palin and Hillary Clinton's emails yields an interesting result. There is a huge

Table 6: Detection rate (DR), average confidence level (ACL) and weighted detection rate of participants for Hillary Clinton's and Sarah Palin's questions

Questions	Metric	Overall	High Conf.
HC	DR	71.69%	74.35%
	ACL	3.66	4.04
	WDR	1.72	2.12
SP	DR	34.19%	23.41%
	ACL	3.56	4.13
	WDR	-1.45	-2.56

gap between participants' detection rate of Hillary Clinton (71.69%) and Sarah Palin (34.19%). So our results suggest that masquerade attacks at higher Flesh-Kincaid level may be harder to detect.

We now use t-test to study whether the difference between the performance of participant on each part is significant or not. Below is the result of significance test on all dependent variables (performance, confidence, spent time):

- Comparing mean DR of participants on HC and SP emails (p=6.1e-16, df=65.868, t=10.638)
- Comparing mean WDR of participants on HC and SP emails (p<2.2e-16 , df=65.959, t=11.452)
- Comparing mean confidence of participants on HC and SP emails (p=0.4242, df=60.75, t=0.223)
- Comparing average time spent by participants on HC and SP emails (p=0.8238, df=60.75, t=0.223)

Based on the significance test results and after applying Benjaminin-Hochberg, we found a significant difference between performance (DR and WDR) of participants on detecting Hillary Clinton and Sarah Palin's emails. This supports what we mentioned earlier about the difference between HC and SP's detection rate (34.19% for SP and 71.69% for HC).

4.3 Embedded Fake Signals

There are three different embedded signals in our attack emails. Table 7 presents the performance of participants on each type of signal for falseness separately. As expected, detection rate of participants on the questions with no signal is much lower than the detection rate on the other three types, 26.4% compared to at least 61%. This means that our attack successfully deceived 74% of participants. Among all 34 participants, six participants (17%) could not detect any of the signals. As shown above, by using a higher Flesch-Kincaid level we can have a higher performance on deceiving people, but even with lower Flesch-Kincaid level and simpler grammar (the one with incoherent flow of idea) we can have a good performance (about 40% success rate in deception).

We should note that among those participants who detected emails with embedded signals as fake, some of them chose *fake* because of the reasons other than actual embedded signals, Table 7 is just based on their answer choice (real or fake) not their reasoning.

Since our independent variable (attack type) takes four different values, we use the ANOVA test instead of the t-test. ANOVA reveals that average detection rate of each groups (p-value = 1.19e-06) differed significantly as a function of fake signals (F(4, 136) = 11.31, p-value = 1.19e-06).

Table 7: Detection rate for each type of attack

	Attack Type	Detection Rate
1	Fake Name	66.17%
2	Incoherent	64.70%
3	Repeated Sentence	61.76%
4	No Error	26.47%

Table 8: T-test for comparing the performance of participants based on each trait

Trait (median)	DR	WDR	Time	ACL
Extraversion (27)	0.208	0.145	0.19	0.398
Agreeableness (33)	0.762	0.922	0.081	0.726
Conscien. (32)	0.672	0.666	0.436	**0.007**
Neuroticism (23)	0.922	0.922	0.365	0.316
Openness (35)	0.847	0.961	0.831	0.992

A Tukey post-hoc test reveals that questions with "No Error" are significantly different from other questions and there is no significant difference between other questions. So, all three signals of attack are similar to each other from difficulty point of view, and questions that do not have any attack signals are harder to detect.

4.4 Unmanipulated Variables

Now we show whether or not there is any relation between unmanipulated variables and dependent variables. We have different kinds of unmanipulated variables: some of them can be grouped together, e.g., age and sex are both part of participant demographics. Based on these similarities, we categorize the variables and show the correlation for each category separately. We have these four categories:

- *Personality*: Five different traits defined by Big Five Personality Test
- *Demographics*: Age, gender
- *Strategies*: Strategies that each participants uses to make decision about emails
- *Knowledge and Background*: Participants' background knowledge on computer/email, and education level (includes Emails sent each day, Email propagation knowledge, Social network usage, and Education level).

4.4.1 Personality Traits

Personality traits reveal internal aspects of each person's mind. In this study, we check whether these traits have any effect on the detection rate of participants. Since the performance of participants and the personality trait scores are not continuous variables, it is not a good idea to use the pearson correlation test on each trait and the performance value directly. For each trait, suppose M is the median. We divide the participants into two groups based on the value of M; those who are greater than M and those who are less than or equal to M. Then we apply t-test to compare their performance between each group. Personality scores of our participants approximately range from lowest to highest score for all the traits. This means that our sample of participants is not skewed in any direction from the personality perspective. More details about the personality scores are in the Appendix.

Table 8 presents p-values of t-test on different groups of participants. There is a significant difference between confidence level of participants as a function of conscientiousness. Those who have lower conscientiousness have higher confidence in their answers.

Table 9: Strategy and number of participants (N) that used it

Strategy	N	**Strategy**	N
Style	34	Mechanics	31
Grammar, Capitalization, Punctuation	30	Incoherent	18
Fake Name	17	Topic	17
Repeated Sentences	14	Asking for action/info	14
Time format	12	Different time	9
Subject	6	Similar/Different recipients	4
Over-thinkers	4	Spelling Issue	3
Sender does not give information	3		

4.4.2 Demography

Now we show the effect of demographic features of participants on their responses. For this we check the significance of difference on mean values of dependent variables as a function of age and sex. For *age*, we divided the participants into two groups: older than 21 years old (median) and younger than 21. Based on t-test results there is no statistically significant difference in participants' average performance with different age and sex. P-values are in the Appendix.

4.4.3 Reasoning Analysis

In each question of the survey, we asked participants to write down their entire reasoning for choosing fake or real. Strategies that participants used to distinguish attacks from real emails are another important aspect of this study. We can use these strategies to understand how we can improve the attack, or from the defender/training point of view what are the things that email users do not pay attention to and need help with.

We categorized participants' strategies into 14 groups (each group is in bold font). Below is the list of signals with a brief description (our embedded signals are excluded).

- **Style**: The way of writing, tone of talking, and level of formality, e.g., referring by first name or last name
- **Mechanics**: Length of sentences and Email, using abbreviation, choice of words
- **Grammar**, **capitalization**, and **punctuation**
- **Topic**: Topic of email is not related to the real emails
- **Asking for action/info**: Does the email ask for doing an action? or give any information?
- **Time Format**: The time format that original author's used in real sample emails. 24-hour or 12-hour or using date and time together
- **Different Time**: The email sending time is unusual compared to the time in real sample emails
- **Subject**: The subject contains *Reply* or *Forward* in it
- **Similar/Different recipients**: Recipient never exist in real emails or it exists
- **Over-thinkers**: Participants who argued that spelling issue and repeated sentence show that the email is real, since if it was from an attacker it would not have such

Table 10: Average detection rate and WDR of participants on each cluster. N is the number of participants

Cluster	N	% Avg DR	Avg WDR	ACL
0	12	55.20	0.3385	3.77
1	6	45.83	-0.4270	3.53
2	16	53.90	0.1914	3.52

mistakes

- **Spelling Issue**: Misspelled words in the Email
- **No Info From Sender**: Sender starts without giving any information about the previous discussion

Table 9 shows how many participants used each type of signal. All participants consider style. Mechanics, grammar, capitalization, and punctuation issues are next most used strategies. These strategies along with their importance for users' decision-making process can be used to improve effectiveness of the our attack or to train people about phishing emails.

Studying the effect of strategies that participants used on their performance is difficult. To achieve this goal we need to find a way to group participants based on their strategies or their decision-making process (similarity of strategies that they used). Once we group them together, we can compare performance of users in each group.

We use different methods for grouping participants together. In the first method we used the 14 extracted strategies as boolean features for participants and put participants into different groups based on their feature vector. Since we were not sure how many clusters we need, to group the participants, hierarchical agglomerative clustering was used to map each feature vector to a cluster (group). The cosine similarity is used as a metric for defining the distance between the feature vectors. The output of hierarchical clustering is a tree whose leaves are the individual observations. By cutting the tree at every height we can get different clustering output, some of them may have an individual observation as a cluster depending on the cutting height. We try different cutting points giving different numbers of clusters and keep only those clusterings that do not have any cluster of size one for further analysis.

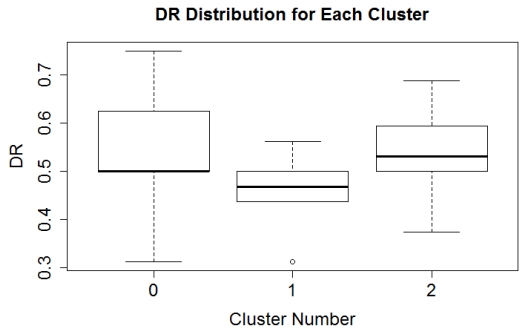

Figure 4: Clusterwise participants' DR distribution (WDR in the appendix)

We use one of the clustering outputs that has three classes *as an example here*. The results for other cutting points are similar to this one. Table 10 shows average performance of participants in each group. In Figure 4, the distribution of performance for each cluster is depicted. Cluster zero does perform slightly better than two other clusters. We check later if the difference is statistically significant or not.

Table 11: Average performance of participants and their confidence level on each sophistication level. N is the number of participants

Sophistication Level	N	Avg DR	Avg WDR	ACL
4	2	46.87	-0.3437	3.46
5	9	55.55	0.3611	3.54
6	8	49.21	-0.1562	3.76
7	8	54.68	0.2734	3.53
8	4	57.81	0.4687	3.56
9	3	47.91	-0.2708	3.81

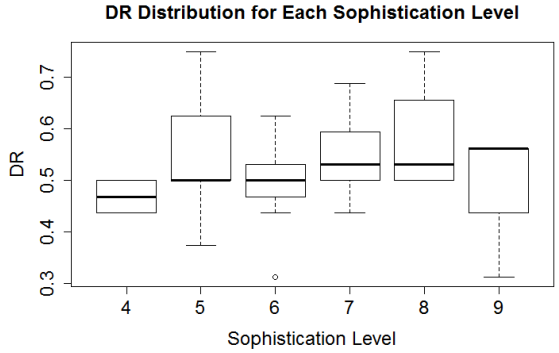

Figure 5: Participants' DR distribution (WDR in the appendix) on different sophistication level

Second method of grouping uses *sophistication level* to separate groups from each other. *Sophistication level* is defined as the total number of strategies that a participant used (among those 14 groups). Table 11 shows average performance of participants based on their sophistication level. The difference in the performance of each group is quite small. In Figure 5, the distribution of performance for each sophistication level is depicted. Among these levels, sophistication level eight is the best performing category.

The last method for categorizing participants is by considering our built-in signals and grouping them based on the number of embedded signals they caught. Six participants did not catch any of the signals, 11 caught one, 13 caught two, and four participants caught all of them. The reason that these numbers are different from what we had in table 5 is that here we consider a signal "has been caught" if the corresponding email is detected as fake (user may consider it as fake because of other reasons).

Table 12 shows average performance of participants in each group. In Figure 6, distribution of performance for each group is shown. Among these groups, those participants who caught one of our signals seem to perform as well as those who caught two or more, and all participants who caught at least one signal seem to perform better than those who did not detect any.

We apply ANOVA test on these three aforementioned grouping methods to see whether or not the performance

Table 12: Average performance and confidence level of participants on each group (based on our signals only). N is the number of participants

Caught signals	N	% Avg DR	Avg WDR	ACL
0	6	46.87	-0.364	3.67
1	11	55.11	0.284	3.64
2	13	53.36	0.177	3.63
3	4	54.68	0.328	3.35

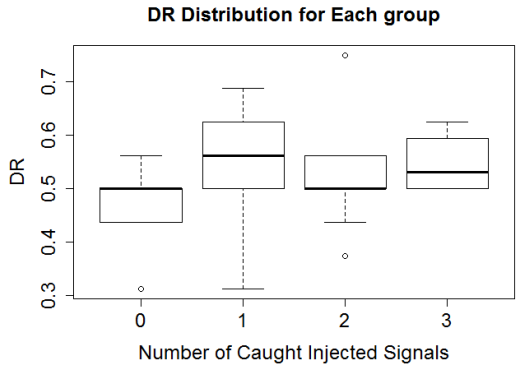

Figure 6: Participants' DR distribution (WDR in the appendix)on different groups based on injected signals

of each group differ significantly. All the p-values are bigger than 0.05 so we can infer from our results that it does not seem to matter which strategies were used by the participants, they could not change their performance significantly in detecting fake emails. However, we cannot generalize this conclusion since there are lots of other features that we did not consider in our study: IQ, native English speaking, etc. Detailed tables of ANOVA tests are in the Appendix.

4.4.4 *Knowledge and Background*

Users' background knowledge and experience in working with email and other applications like social networks may have an effect on their performance. Hence we test the relation between these variables and performance indicators. Using correlation test between performance indicators and these variables (email propagation knowledge, email sent each day, social network usage, education level, and years using email) does not show any significant correlation between them. More details of the correlation test are in the Appendix.

4.5 Summary of Findings

We analyzed the effect of different variables (independent and unmanipulated) on performance of participants to learn which variables can improve the effectiveness of the attack. We also studied different strategies and signals that people use to detect attack emails. We found that:

- Participants perform close to random in detecting fake and real emails, 75% in best case and 52% on average.
- Performance in detection of real emails is worse than fake emails. This could be because participants were a little bit more willing to choose fake than real.
- Masquerade attack on emails with higher Flesh-Kincaid level (harder to read) are harder to detect.
- Among the four types of attacks: *fake name*, *incoherent flow*, *repeated sentence* and *no error*, first three are at the same level of difficulty, but *no error* is significantly harder than all the others.
- There is significant correlation between conscientiousness of participants and their confidence level in answering the questions.
- Strategy analysis showed that there is no difference between performance of different group of people with different detection strategies.
- Their knowledge and background on using email did not have any significant effect on their performance.

We also did more detailed analysis on real/fake and HC/SP emails. For lack of space, we put the details in the appendix, but list important results below:

- Participants with lower extraversion than the median perform better.
- Younger participants perform better in detecting real emails.

Limitation of the Study. The way that we chose our participants may cause a concern that the participants are not extremely familiar with Hillary Clinton and Sarah Palin writing style, and a trained person might behave differently than an untrained person. First, it is not easy to find a person familiar with their writing style (like their secretary). Second, in our attack model, the goal is to have a mass phishing/deception attack. In this case, the victims are not necessarily well trained on the sender's writing style. They might have received a few emails from sender (Clinton or Palin in our case).

5. RELATED WORK

Phishing email detection is one of the areas in computer security that Natural Language Processing is applicable. Semantic features [43], syntax feature [33], and contextual features [44] previously have been used in this area. But, nobody utilized NLP techniques in the other direction, i.e., creating attack.

There are several user studies *comparing* phishing/spoofed and genuine websites, e.g., [2, 3, 10, 32, 25, 45], a few works on *comparing* phishing and legitimate emails [34, 24] (although there are several studies that examine various aspects of interaction with phishing emails alone, e.g., [11, 12], which are well summarized in [25]), and quite a few on the effectiveness of training programs, e.g., [7, 28, 38]. The work on training programs is not directly within the scope of this research. Work is scant on emails containing malware, but effectiveness of malware warnings was studied in [32]. Most studies have been conducted in laboratory settings, notable exceptions to this are [7, 25, 23, 46]. Masquerade attacks on written documents have been studied previously [1], but *in the email context* have not been studied previously to our knowledge. Cues in the context of phishing and spear-phishing were studied previously in [46, 40]. The authors in [46] also studied cognitive load indirectly through a survey question. Cognitive load in a lab setting was examined in [32]. There are quite a few studies on factors affecting the strength of phishing attacks. Social trust in the form of friend request was exploited in an actual attack scenario by [4]. A demographic analysis of phishing

susceptibility and effectiveness of interventions was studied in [37]. Because we focus more on how users identify deceptiveness/impersonation and not on characteristics of specific attacks, our work is more generally applicable to any attack involving an email, be it an IRS Scam, a company representation fraud, phishing, malware or spear-phishing.

Phishing versus Legitimate Emails. In [34], 50 actual emails on a variety of topics, 25 legitimate and 25 phishing, were analyzed by five "experts" for the presence or absence of 11 types of cues. T-tests were conducted to determine which cues could distinguish phishing emails from legitimate ones. Next, data collected in an earlier role-playing laboratory experiment was analyzed to determine which cues were used by the 59 student participants. Since the emails were actual examples taken from various sources, researchers could not control the number of cues in each email and since the objective of the earlier experiment was not to elicit cues, students probably indicated a subset of the cues in each email. Another group or researchers ran a similar experiment with a higher number of participants (179) and showed that people are not good at finding the cues for illegitimate emails [24].

Spear-phishing. In [46], researchers sent out a single actual spear phishing email containing a mix of "visceral" cues (e.g. urgency, other emotional appeals) and visual phishing cues (spelling, grammar, sender address) to business/communication majors and then surveyed them. The survey had a 5-point Likert scale response variable (Not At All Likely to Respond to Very Likely to Respond), and other questions to determine attention to indicators and cognitive load. Analysis of 267 data points, showed that attention to visceral indicators increased the likelihood of response and attention to phishing indicators decreased it. Since they used an actual email, they could not control and vary the presence of cues and their measurements were all indirect.

Researchers in [40] used user behavioral model for distinguishing a real email and an email which is written by someone pretending to be the actual author. This model includes frequent interactions with certain people, sending emails at specific hours of the day, and using certain greetings and modal words in their emails. This model is not effective for our attack model since we consider the attacker has access to the victims emails and create the masquerade emails based on the actual emails.

Security researchers have found context-aware phishing to be more effective at making users fall for phishing attacks [20, 21], however, with some caveats [22]. The new attacks we study in this paper explore previously unexplored angles in context-aware attacks. Impersonation of a familiar person goes beyond mining social network data for name linkages or including public activity/social circle to increase email credibility. Since previous studies indicate that participants pay attention to minute details such as spelling and grammar, impersonating styles of writing is likely to further increase the authenticity of the phishing emails.

Role of personality in Phishing. This is a recent area of study. Neuroticism was correlated with response to a phishing email involving a prize [16]. Generalized communicative suspicion (a derived metric of personality) was correlated with efficacy in phishing email detection [17].

6. CONCLUSIONS

We introduced a scalable masquerade attack using Natural Language Generation. We studied its effectiveness in deceiving people using Hillary Clinton and Sarah Palin's emails. We also analyzed the participants' reasoning among other variables. Our study showed that the attack is successful and detection rate is only slightly better than random guessing. Our attack fooled 74% of participants using a relatively more complex grammar. State-of-the-art author verification tools can be used to verify the style of the senders. On our dataset, the best automatic approach was able to catch 50% of the fake emails suggesting the in-effectiveness of the approach to the attack we present in this paper.

7. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive feedback on this work. We also thank Josh Davis for help with the grammar, Mahsa Shafaei for data cleaning and pilot, Nour Smaoui for data conversion and Niloofar Samghabadi for the pre-pilot. This research was partially supported by NSF grants CNS 1319212, DUE 1241772, DGE 1433817 and CNS 1527364.

8. REFERENCES

[1] S. Afroz, M. Brennan, and R. Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *IEEE S&P*, pages 461–475, 2012.

[2] H. Almuhimedi, A. P. Felt, R. W. Reeder, and S. Consolvo. Your reputation precedes you: History, reputation, and the chrome malware warning. In *Proc. SOUPS 2014*, pages 113–128, 2014.

[3] M. Alsharnouby, F. Alaca, and S. Chiasson. Why phishing still works: user strategies for combating phishing attacks. *Int'l Journal of Human-Comp Stud.*, 82:69–82, 2015.

[4] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proc. 18th WWW*, pages 551–560, 2009.

[5] A. C. Bulhak. The dada engine. *Available at dev.null.org/dadaengine/*, 1996.

[6] A. C. Bulhak. On the simulation of postmodernism and mental debility using recursive transition networks. *Monash Univ. Tech. Report*, 1996.

[7] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson. Going spear phishing: Exploring embedded training and awareness. *IEEE Security Privacy*, 12(1):28–38, Jan 2014.

[8] P.-A. Chirita, J. Diederich, and W. Nejdl. Mailrank: using ranking for spam detection. In *Proc. of 14th ACM CIKM*, pages 373–380. ACM, 2005.

[9] L. A. Clark. Assessment and diagnosis of personality disorder: Perennial issues and an emerging reconceptualization. *Annu. Rev. Psychol.*, pages 227–257, 2007.

[10] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *Proceedings of the CHI*, pages 581–590. ACM, 2006.

[11] J. S. Downs, M. B. Holbrook, and L. F. Cranor. Decision strategies and susceptibility to phishing. In *Proc. of 2nd SOUPS*, pages 79–90, 2006.

[12] J. S. Downs, M. B. Holbrook, and L. F. Cranor. Behavioral response to phishing risk. In *Proc. 2nd Annual eCrime Researchers Summit*, pages 37–44, 2007.

[13] A. Falk and J. J. Heckman. Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952):535–538, 2009.

[14] C. Fraud. Business email masquerading: How hackers are fooling employees to commit fraud. http://www.bankinfosecurity.com/webinars/business-email-compromise-achilles-heel-w-664, 2015.

[15] D. Galanis, G. Karakatsiotis, G. Lampouras, and I. Androutsopoulos. An open-source natural language generator for owl ontologies and its use in protégé and second life. In *Proc. 12th EACL: Demonstrations Session*, pages 17–20, 2009.

[16] T. Halevi, J. Lewis, and N. Memon. A pilot study of cyber security and privacy related behavior and personality traits. In *22nd WWW*, 2013.

[17] B. Harrison, E. Svetieva, and A. Vishwanath. Individual processing of phishing emails: How attention and elaboration protect against phishing. *Online Information Review*, 40(2):265–281, 2016.

[18] Y. Hochberg and Y. Benjamini. More powerful procedures for multiple significance testing. *Statistics in medicine*, 9(7):811–818, 1990.

[19] J. Hong. The state of phishing attacks. *Communications of the ACM*, 55(1):74–81, 2012.

[20] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, 2007.

[21] M. Jakobsson. Modeling and preventing phishing attacks. In *Financial Cryptography*. IFCA, Springer Verlag, Feb. 2005.

[22] M. Jakobsson and J. Ratkiewicz. Designing ethical phishing experiments: a study of (ROT13) rOnl query features. In *WWW '06*, pages 513–522. ACM, 2006.

[23] M. Jakobsson, A. Tsow, A. Shah, E. Blevis, and Y.-K. Lim. What instills trust? a qualitative study of phishing. In *FC'07/USEC'07*, pages 356–361, Berlin, Heidelberg, 2007. Springer-Verlag.

[24] A. Karakasiliotis, S. Furnell, and M. Papadaki. Assessing end-user awareness of social engineering and phishing. In *Australian Information Warfare and security conference*, 2006.

[25] T. Kelley and B. I. Bertenthal. Real-world decision making: Logging into secure vs. insecure websites. In *USEC*, 2016.

[26] M. Khonji, Y. Iraqi, and A. Jones. Phishing detection: a literature survey. *IEEE Communications Surveys & Tutorials*, 15(4):2091–2121, 2013.

[27] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *Proc. 21th ICML*, page 62. ACM, 2004.

[28] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. Blair, and T. Pham. School of phish: a real-world evaluation of anti-phishing training. In *Proc. 5th SOUPS*, pages 1–12. ACM, 2009.

[29] S. D. Levitt and J. A. List. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2):153–174, 2007.

[30] R. A. Maxion. Masquerade detection using enriched command lines. In *DSN*, volume 3, pages 5–14, 2003.

[31] R. A. Maxion and T. N. Townsend. Masquerade detection using truncated command lines. In *Proc. DSN*, pages 219–228. IEEE, 2002.

[32] A. Neupane, M. L. Rahman, N. Saxena, and L. Hirshfield. A multi-modal neuro-physiological study of phishing detection and malware warnings. In *Proc. of the 22nd ACM SIGSAC*, pages 479–491, 2015.

[33] G. Park and J. M. Taylor. Using syntactic features for phishing detection. *arXiv preprint*, 2015.

[34] K. Parsons, M. Butavicius, M. Pattinson, D. Calic, A. Mccormac, and C. Jerram. Do users focus on the correct cues to differentiate between phishing and genuine emails? *arXiv preprint*, 2016.

[35] M. B. Salem and S. J. Stolfo. Modeling user search behavior for masquerade detection. In *RAID*, pages 181–200. Springer, 2011.

[36] M. Schonlau, W. DuMouchel, W.-H. Ju, A. F. Karr, M. Theus, and Y. Vardi. Computer intrusion: Detecting masquerades. *Statistical science*, 2001.

[37] S. Sheng, M. B. Holbrook, P. Kumaraguru, L. F. Cranor, and J. S. Downs. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proc. 28th CHI*, pages 373–382, 2010.

[38] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. Cranor, J. Hong, and E. Nunge. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proc. 3rd SOUPS*, pages 88–99. ACM, 2007.

[39] R. Sommer and V. Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *31st IEEE S&P*, pages 305–316, 2010.

[40] G. Stringhini and O. Thonnard. That ain't you: Blocking spearphishing through behavioral modelling. In *DIMVA*, pages 78–97. Springer, 2015.

[41] B. K. Szymanski and Y. Zhang. Recursive data mining for masquerade detection and author identification. In *Proc. IAW 2004*, pages 424–431. IEEE, 2004.

[42] T. N. Y. Times. Sarah Palin emails: The Alaska archive. http://documents.latimes.com/sarah-palin-emails/, 2011.

[43] R. Verma and N. Hossain. Semantic feature selection for text with application to phishing email detection. In *International Conference on Information Security and Cryptology*, pages 455–468. Springer, 2013.

[44] R. Verma, N. Shashidhar, and N. Hossain. Detecting phishing emails the natural language way. In *European Symposium on Research in Computer Security*, pages 824–841. Springer, 2012.

[45] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Y. Zhao. Social turing tests: Crowdsourcing sybil detection. *arXiv preprint*, 2012.

[46] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao. Phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE Trans. Prof. Communication*, 2012.

[47] K. Wang and S. J. Stolfo. One-class training for masquerade detection. In *DMSEC*, pages 10–19, 2003.

[48] Wikileaks. Hillary Clinton Email Archive. https://wikileaks.org/clinton-emails/, 2016.

APPENDIX

A. BIG FIVE PERSONALITY

We created the personality test using Google Form and asked the participants to do it before coming to the lab for the actual experiment. Table 13 presents the range of possible values and range of the participants' scores in each trait. It shows that our participants' scores approximately covers from lowest to highest score for all the traits. That means that our samples from the population are not skewed in one or more directions from the personality point of view.

Table 13: Range of possible(poss.) scores in each trait of Big Five Personality test and also for our participants (partic.)

Traits	Minimum	Maximum	Median
	Poss./Partic.	Poss./Partic.	Partic.
Extraversion	8/14	40/39	27.5
Agreeableness	9/14	45/43	33
Conscien.	9/19	45/44	32
Neuroticism	8/11	40/38	23
Openness	10/21	50/46	35

B. MASQUERADE ATTACK: DETAILED BREAKDOWN

In Section 4 we separately studied the effect of independent and unmanipulated variables on the dependent variables, but we did not combine them together. Do people with more knowledge about email perform better on detecting fake emails? Do younger people perform worse on detecting emails with higher readability level? These are the questions that remain unanswered in the previous section. Here, we study the effect of unmanipulated variables by taking into account the independent variables.

B.1 Personality Traits

Table 14 provides the p-values of t-test between performance of participants on real/fake and HC/SP emails as a function of the five personality traits. Same as previous section for each personality trait, we divide participants into two groups based on the median of each trait. The only significant difference (Benjamini-Hochberg procedure applied) in this table is for extraversion and detection of fake emails. Participants with lower extraversion than the median perform better in detecting attacks than those with higher extraversion.

B.2 Knowledge and Experience

The next analysis is about the relation between the knowledge and experience of participants and their performance for real/fake and HC/SP separately.

Table 15 presents the results of applying the correlation tests. As shown, all the p-values are larger that 0.05, so there is no significant difference between performance based on the knowledge and experience of our participants.

B.3 Demography

Table 16 presents p-values of t-test between performance of participants on real/fake and HC/SP emails as a function of gender and age. Same as before, for the *age*, we divide the participant into two groups: less than 21 years old and older than 21 years old.

There is a significant difference in performance of young and old participants (DR) in detecting real emails. Younger participants have better performance in detecting real emails. Although the p-value for WDR on real emails is less than 0.05 but it is not found to be significant when we use the Benjamini-Hochberg procedure.

C. SIGNIFICANCE TESTS

C.1 Demography

Here, we show the result of significance test between age and gender as unmanipulated variables and dependent variables in our study. For the *age*, we divide the participants into two groups, those who are older than 21 (the median) and those who are younger than 21. Table 17 shows the results for significance test of dependent variables for the two demographic features.

C.2 Reasoning Analysis

In Section 4.4.3 we defined three different ways of grouping participants together based on their strategies in detecting fake emails. Figures 7, 8 and 9 show the distribution of performance for different grouping methods.

Figure 7: Clusterwise participants' WDR distribution

Here we compare the performance of users in each group to see whether the strategies can affect the performance or not. Table 18 shows p-value of ANOVA tests on different grouping method introduced in Section 4.4.3.

All the p-values are bigger than 0.05 so there is no significant difference between the groups in each grouping methods. This means using different strategies do not affect detection performance.

C.3 Knowledge and Experience

Table 19 shows the results of correlation test between dependent variables and variables related to knowledge and background on all 16 questions. We use Pearson correlation test for all variables except those that are nominal. Spearman correlation has been used for nominal variables. Here we have three p-values less than 0.05 but they are not significant after applying Benjamini-Hochberg.

Table 14: T-test for comparing the participant based on each trait on real/fake and HC/SP emails

Trait (median)	DR				WDR			
	Real	Fake	HC	SP	Real	Fake	HC	SP
Extraversion (27)	0.07	**0.001**	0.882	0.113	0.123	**0.001**	0.669	0.122
Agreeableness (33)	0.504	0.868	0.406	0.236	0.492	0.666	0.37	0.325
Conscientiousness (32)	0.806	0.492	0.37	0.787	0.855	0.515	0.319	0.689
Neuroticism (23)	0.209	0.264	0.945	0.947	0.158	0.218	0.933	0.961
Openness (35)	0.391	0.641	0.358	0.272	0.338	0.475	0.469	0.456

Table 15: P-value of correlation test between unmanipulated variables and dependent variables for real/fake and HC/SP separately. *Spearman correlation instead of Pearson

Variable	DR				WDR			
	Real	Fake	HC	SP	Real	Fake	HC	SP
Email Propagation Knowledge*	0.533	0.771	0.579	0.828	0.557	0.533	0.882	0.78
Email Sent Each Day	0.659	0.841	0.273	0.379	0.65	0.829	0.235	0.373
Social Network Usage	0.209	0.16	0.402	0.058	0.175	0.158	0.195	0.143
Education Level *	0.506	0.221	0.445	0.746	0.967	0.144	0.107	0.9198
Years Using Email	0.221	0.273	0.912	0.798	0.347	0.289	0.704	0.462

Table 16: P-value (t) of significance test between demographic features and dependent variables separately on real/fake and HC/SP

	Questions	Gender	Age
DR	Real	0.24	**0.004 (-3.038)**
	Fake	0.288	0.746
	HC	0.067	0.11
	SP	0.578	0.566
WDR	Real	0.154	0.034
	Fake	0.412	0.631
	HC	0.054	0.137
	SP	0.81	0.949

Table 17: P-value of significance test between demographic features and dependent variables

Demography	DR	WDR	Time Spent	ACL
Gender	0.09	0.122	0.685	0.785
Age	0.123	0.318	0.166	0.125

Table 18: P (F) value of ANOVA tests on different grouping methods

Method	DR	WDR	ACL
Clustering	0.157 (1.963)	0.136 (2.129)	0.337 (1.126)
Sophistication	0.532 (0.841)	0.521 (0.859)	0.858 (0.38)
Our signals	0.439 (0.929)	0.387 (1.045)	0.718 (0.452)

Table 19: P-value of correlation test between unmanipulated variables and dependent variables for all 16 questions. *Spearman correlation instead of Pearson

variable	DR	WDR	Time	ACL
Email Propagation Knowledge*	0.807	0.806	0.344	0.334
Email Sent Each Day	0.904	0.847	0.812	0.409
Social Network Usage	0.045	0.04	0.874	0.741
Education Level*	0.662	0.242	0.92	0.66
Years Using Email	0.912	0.788	0.32	0.037

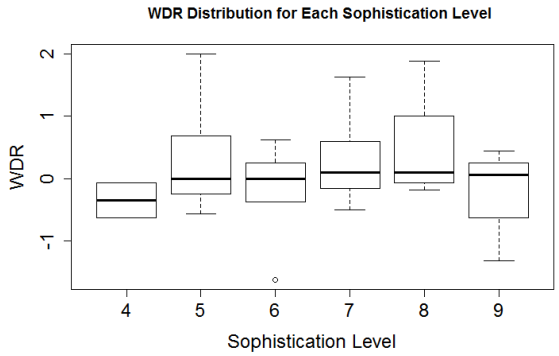

Figure 8: Participants' WDR distribution on different sophistication level

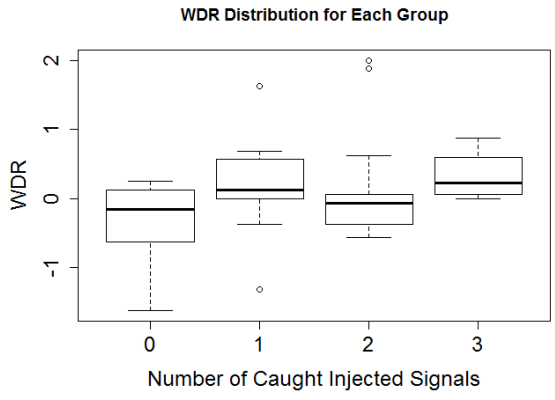

Figure 9: Participants' WDR distribution on different group based on injected signals