Users Really Do Answer Telephone Scams

Huahong Tu, *University of Maryland;* **Adam Doupé,** *Arizona State University;* **Ziming Zhao,** *Rochester Institute of Technology;* **Gail-Joon Ahn,** *Arizona State University and Samsung Research*

https://www.usenix.org/conference/usenixsecurity19/presentation/tu

This paper is included in the Proceedings of the 28th USENIX Security Symposium.

August 14–16, 2019 • Santa Clara, CA, USA

978-1-939133-06-9

Open access to the Proceedings of the 28th USENIX Security Symposium is sponsored by USENIX.

Users Really Do Answer Telephone Scams

Huahong Tu[1], Adam Doupé[2], Ziming Zhao[3], and Gail-Joon Ahn[2,4]

[1]*University of Maryland, hh2@umd.edu*
[2]*Arizona State University, {doupe, gahn}@asu.edu*
[3]*Rochester Institute of Technology, zxzics@rit.edu*
[4]*Samsung Research*

Abstract

As telephone scams become increasingly prevalent, it is crucial to understand what causes recipients to fall victim to these scams. Armed with this knowledge, effective countermeasures can be developed to challenge the key foundations of successful telephone phishing attacks.

In this paper, we present the methodology, design, execution, results, and evaluation of an ethical telephone phishing scam. The study performed 10 telephone phishing experiments on 3,000 university participants without prior awareness over the course of a workweek. Overall, we were able to identify at least one key factor—spoofed Caller ID—that had a significant effect in tricking the victims into revealing their Social Security number.

1 Introduction

The rise of telephone spam, scams, fraud, phishing, or vishing, is a significant and growing problem. According to FTC reports for 2018, phone impersonation scams have increased significantly in the recent years. The national Do-Not-Call Registry received more than 5.78 million unwanted call complaints [1], with fraud and imposer scam in the top spots and more than 69% of all reported frauds were attempted over the phone [2].

With the growing dissatisfaction of telephone scams, however, little research has been done to study *why* people fall for telephone scams and how to combat the problem. In this paper, inspired by the work of Tischer et al. [3] on USB drives, we present the results of an empirical telephone phishing study, designed to systematically measure different attributes in relation to the success rate of telephone scams. Although the current understanding of telephone scams might be accepted as conventional wisdom, no prior work has specifically validated such claims with a systematic study. From this study, we hope to dispel some myths about what is "scammy" and what is not. With the understanding of the key attributes that make a scam convincing, the research community can focus on developing prevention methods to challenge the fundamentals of telephone phishing attacks. The key takeaway from this study is that caller ID spoofing is an incredibly effective feature in telephone scams, and, therefore, authenticated caller ID [4,5] is likely to be an important countermeasure.

The main contributions of this paper are the following:

- We describe a systematic approach to test the significance of various telephone phishing scam attributes and conduct an empirical study.
- We present our evaluation of the phishing study and provide our recommendations for combating the telephone phishing problem.

2 Background

With the emergence of distribution technology, decreasing economic cost, high reachability, and automation, the telephone has become an attractive medium for disseminating unsolicited information. As with any form of spam, there are three key ingredients: the recipient list, the content, and the distribution channel [6]. Telephone scams rely on distributing *deceitful* voice content, whereas telephone spam or telemarketing primarily distributes marketing and advertising content. In telephone scams, fraud, phishing, or vishing, the goal of the voice content is to trick the human victim into performing harmful actions for the benefit of the attacker (while other types of fraud are possible on telephone networks [7–9]).

Compared to other forms of phishing, such as email and website phishing [10–15], telephone phishing differs by having the potential to make the scam more convincing by falsifying both visual and auditory perceptions to induce the victims into falling for the scam. Visually, the scam can be made more convincing by altering the caller ID, such as by spoofing the caller ID, manipulating the area code (e.g., in "neighbor spoofing" attacks [16]), and impersonating a familiar contact name. Once the recipient has answered the call, the attacker then switches to using deceitful voice content to exploit the human recipient [17,18]. Within the voice content, an attacker can spoof or duplicate the speech from a known organization

or a familiar personal contact. To provide a motivation for the recipient to divulge confidential or personal information, the scammer can present a demanding scenario that forces the victim to divulge sensitive information.

By looking at telephone phishing from a perspective that can be characterized by the visual and voice attributes which it embodies, a systematic approach can be used to study and understand why some scams work better than others. Understanding why telephone phishing works can help us design solutions that challenge the core foundations of telephone scams.

3 Study Design

The goal of the study is to design a systematic approach that can reveal the effective factors in telephone scams by conducting our own telephone phishing scam. Our approach to designing the study is to first identify the attributes that could lead to an effective telephone phishing scam. After that, we design a set of experiments and procedures that allow comparison of different variations of an attribute. Each experiment followed a standardized procedure that was conducted on each group simultaneously (all calls were distributed in a randomized order throughout the experiment). Finally, we provide a discussion on what could be learned from the analysis and provide our recommended solutions for combating the telephone phishing problem. The study was conducted with significant ethical consideration and with IRB approval (see Section 3.5 for an in-depth discussion of ethics).

3.1 Attributes

To identify the telephone scam attributes, we gathered and reviewed more than 150 existing real-world telephone scam samples from various Internet sources, including the FTC website, IRS website, news websites, YouTube, SoundCloud, user comments, and industry surveys. While reviewing the scams, we identified the following attributes used in telephone scams:

Area Code: In North America, the area code is the first three digits on the caller ID. The area code specifies the geographic location associated with the caller's phone number, e.g., 202 is associated with Washington, DC. In addition, a toll-free phone number is also identified by the three-digit prefix similar to a geographic area code, e.g., 800, 888, 877, etc. According to reports of real-world IRS impersonation scams [19, 20], many scammers appeared to have either spoofed or obtained a 202 area code or toll-free area code on their caller IDs to make it appear as if the IRS is calling. To test the hypothesis that the area code could effect telephone phishing success, in our experiments we varied the caller ID area code between: 202 (Washington, DC), 800 (Toll-free), and 480 (local area code of the university location).

Caller Name: Today, most telephone terminals have the capability of associating a name with a telephone number. With a stored contact, an incoming call from the stored contact would show the name associated with the caller ID. To perform a spear phishing attack [21, 22], a malicious caller could spoof the caller ID of a known stored contact. A known stored contact can be identified for an organization by studying the publicly available phone numbers or for an individual by manually analyzing social network information. For legal, ethical, and IRB approval reasons, we did not actually spoof a known caller name. Instead, we asked our telephone service department to temporarily create a new contact in the university's internal phone directory and associated a legitimate sounding name with the telephone number. We used that telephone number in our scam experiments to produce a similar effect to caller name spoofing.

Voice Production: According to reports of real-world telephone scams, some used a robotic (synthesized) voice, while others used a pre-recorded human voice [20, 23]. To test the effect of synthesized voices vs. human voices, we recreated known scams using a text-to-speech synthesizer to generate a speech similar to the real-world scams. To mimic the human voice version of the scams, we recorded human voices speaking the exact same announcement message.

Gender: From listening to recordings of actual telephone scams, some used a male voice, and some used a female voice. To test if the vocal gender of the voice could have an effect on the telephone scam, we varied the voice gender between male and female in the text-to-speech synthesizer.

Accent: From the reports of telephone scams, some spoke with an Indian accent, and some others spoke with an American accent. It seems possible that recipients would be more wary of scams that speak in a foreign accent, and would be less suspicious of scams that speak in an American accent. To test if this could have an effect on the telephone scam, we varied the recorded voice accent between Indian and American in our experiments.

Entity: From gathering real-world telephone scams, two types of scams stood out in terms of the number of reports: IRS impersonation scams [24] and HR impersonation scams [25]. In these scams, the scammer claimed to be from the IRS or the company's HR department. While the IRS scams can affect any taxpayer in the US, the HR scams are usually targeted toward people in a specific company. Intuitively, it seems that a more targeted attack would have more success. Thus, we varied the impersonated entity of our scams between the IRS and ASU's HR department[1]. To simulate the real-world HR scams as closely as possible, we initially wanted to impersonate our university's HR department. However, our HR department had strong objections about using their name to conduct the scam experiments. As a compromise, our experiments claimed to be from a fake

[1] ASU is the university acronym for Arizona State University

but legitimate-sounding HR-like department called the "W-2 Administration"[2].

Scenario: Real-world telephone scams create various scenarios to motivate their victims to fall for the scam, such as tax lawsuits, payroll issues, or credit card verification. The type of motivation are generally either *fear-based* or *reward-based*. In our study, we crafted a fear-based and a reward-based scenario related to each entity. These scenarios were inspired by real-world IRS scams and HR scams. To test each type of scenario, our message announcements varied between Tax Lawsuit (IRS fear-based), Unclaimed Tax Return (IRS reward-based), Payroll Withheld (HR fear-based), and Bonus Issued (HR reward-based).

3.2 Experiments

To test these attributes, we designed the experiments such that variations of each attribute can be compared under similar environmental conditions. When performing experiments under the same environmental conditions, one of the design issues is to decide whether to counterbalance the environmental conditions such that all variations of background attributes are tested. This would theoretically avoid possible interference due to a specific set of background conditions.

However, performing a counterbalanced measures design does not come without costs. Counterbalancing the conditions is performed by splitting the experiments into groups of every possible order of attribute conditions. Given the large number of attributes that we have identified, and of each attribute with 2–4 variations that we have identified, would require us to create 384 separate groups of experiments. This is unfeasible for an empirical study with real-world time and resource constraints.

As a solution to this problem, instead of experimenting with a large number of background conditions, we compare variations of each attribute under a specific set of background conditions that seem to be the most popular in the real world. We decided on a standard background condition: a phishing scam with area code 202, with no caller name, speaking in a synthesized, male voice, in an American accent, impersonating the IRS, motivating the recipient with a tax lawsuit. The set of 10 experiments and the variations of each attribute are listed in Table 1.

3.3 Population

To comply with legal requirements [26], our own ethical considerations, and our IRB (Section 3.5), we conducted experiments on our university's internal population (rather than the general population). This population was unaware of our study (and we discuss the ethical implications of this deceptive non-consent study in Section 3.5). The population of the study were work telephone numbers that are associated with university staff and faculty. We decided on a population of 3,000 recipients (300 per experiment) for the study. To compile the list of telephone numbers, we wrote a custom tool to download the university's internal phone directory. For a real-world scammer, our university's phone directory is also publicly available for crawling.

To minimize selection bias, the telephone numbers were randomly chosen from the university telephone directory, and then the chosen contacts were randomly put into one of the 10 experiment groups. The sample selection procedure was as follows: (1) Compile the list of work telephone numbers associated with university staff and faculty, (2) remove telephone numbers of people already aware of the study, and (3) randomly assign 300 numbers to each of the 10 experiments.

3.4 Procedure

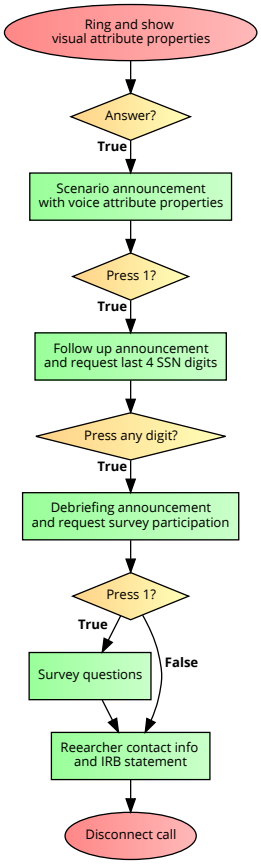


Figure 1: Procedure of each experiment.

Several considerations went into the design of the procedure. First, we need to ensure that the procedure is standardized across all experiments, such that the results are directly comparable to each other. Second, we need to ensure that the process minimizes false positives and false negatives, otherwise, the study results could be unreliable. Finally, the

[2]The W-2 is the income tax form currently used in the United States, so this name has associations with payroll and taxes.

No.	Caller ID	Area Code Location	Caller Name	Voice Production	Gender	Accent	Entity	Scenario
E1	202-869-XXX5	Washington, DC	N/A	Synthesizer	Male	American	IRS	Tax Lawsuit
E2	800-614-XXX9	Toll-free	N/A	Synthesizer	Male	American	IRS	Tax Lawsuit
E3	480-939-XXX6	University Location	N/A	Synthesizer	Male	American	IRS	Tax Lawsuit
E4	202-869-XXX0	Washington, DC	N/A	Synthesizer	Female	American	IRS	Tax Lawsuit
E5	202-869-XXX2	Washington, DC	N/A	Synthesizer	Male	American	IRS	Unclaimed Tax Return
E6	202-849-XXX7	Washington, DC	N/A	Human	Male	American	IRS	Tax Lawsuit
E7	202-869-XXX4	Washington, DC	N/A	Human	Male	Indian	IRS	Tax Lawsuit
E8	480-462-XXX3	University Location	N/A	Synthesizer	Male	American	ASU	Payroll Withheld
E9	480-462-XXX5	University Location	W-2 Administration	Synthesizer	Male	American	ASU	Payroll Withheld
E10	480-462-XXX7	University Location	N/A	Synthesizer	Male	American	ASU	Bonus Issued

Table 1: Table of all experiments and their attributes.

procedure also must be carried out ethically and minimize potential harm to the participants.

To ensure that the procedure is standardized, we used an autodialer to automate the process of sending out the telephone calls and collecting the recipients' responses.

Every experiment followed a standard procedure that is summarized in Figure 1. The procedure has several steps that require inputs from the recipient. The purpose of this action is to reduce the likelihood of recipients making random input actions without hearing the announcement. The action also helps to filter out answers from answering machines. Note that a recipient could break off from the procedure at any point by simply disconnecting the phone, hence not every recipient follows the procedure until the end.

The procedure first begins with a ring on the recipient's work phone (the recipient does not expect the call). When the phone is ringing, the incoming call screen shows the caller ID and, in experiment E9, the caller name. An example of the incoming call screen is shown in Figure 2a. In all of our experiments, the caller ID showed up as 91XXXXXXXXXX, where XXXXXXXXXX is the caller ID used in the respective experiment. Our university's work phone adds a 91 prefix to every incoming phone call from an external source as all of the calls were distributed from an external telephone service provider, similar to what a real-world scammer would do.

For Experiment 9, the incoming call screen also shows a caller name as shown in Figure 2b. This experiment was designed to simulate a scammer spoofing a known caller name. For legal and ethical reasons, we did not actually spoof a phone number. Instead, we asked our telephone service department to temporarily create a new contact in the university's internal phone directory and associated a legitimate sounding HR department name "W-2 Administration" with the telephone number. In a normal external call, there is no caller ID displayed, however, IT was able to help us create the caller ID shown in Figure 2b. While a scammer would not be able to create a new name, they can spoof the caller ID of a known caller with a targeted spearphishing scam.

If the call is answered, it starts by playing a prerecorded scenario announcement message (which is different for each scenario). The prerecorded scenario announcement message incorporates the voice attribute properties of each particular experiment. We crafted the four different announcement messages to mimic what a real-world scammer would say by using words and sentences from our collected scam samples.

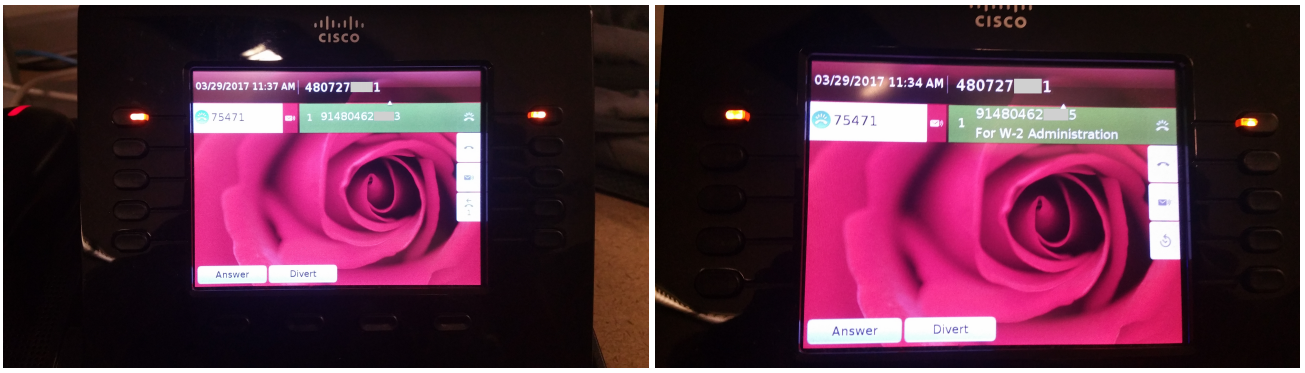
In the *Tax Lawsuit* scenario, we claimed to be the IRS and presented a scenario where the recipient had to act because of a tax lawsuit. The transcript of the announcement message is in Appendix A.1. In the *Unclaimed Tax Return* scenario, we claimed to be the IRS and presented a scenario where the recipient had to act because of an unclaimed tax return. The transcript of the announcement message is in Appendix A.2. In the *Payroll Withheld* scenario, we claimed to be ASU "HR" department and presented a scenario where the recipient had to act because pay would be withheld. Our university has a publicly available payroll calendar on the HR department's website[3], hence a real-world scammer could also use this information to craft an announcement message based on the payroll information. The transcript of the announcement message is in Appendix A.3. In the *Bonus Issued* scenario, we claimed to be ASU "HR" department and presented a scenario where the recipient had to act because a performance bonus was issued. The transcript of the announcement message is in Appendix A.4.

Every scenario announcement message requests the recipient to enter `1` to continue to the next step for a follow-up message (same for every participant). After pressing `1`, the follow-up message asks the recipient to enter the last four digits of their Social Security number and mimics the process of connecting the phone call to a live agent. The transcript of the follow-up announcement message is in Appendix B.

In the real world, the last four digits of the Social Security number can be used to perpetrate financial and identity fraud [27]. Other parts of the Social Security number can also be inferred from the recipient's phone number [28]. To minimize potential risk to the recipient (with cooperation and consultation with our IRB), *we did not record which digits were pressed*, we instead recorded only if any digit was pressed.

This then led to a debriefing announcement and a request to participate in our phone survey. The transcript of the debriefing message is in Appendix C. To emphasize the fact that whatever they listened to was not a real scam, the debriefing announcement and survey questions were recorded with the researcher's real voice. The post-debriefing survey

[3] https://cfo.asu.edu/payroll-calendars



(a) All experiments except experiment E9

(b) Experiment E9 with caller name displayed

Figure 2: Incoming call screen of different experiments.

consisted of two questions: (1) a survey question that asked whether the recipient was convinced by the scam (transcript in Appendix D.1) and, depending on how they responded, (2) asked what factor convinced them of the scam (Appendix D.2) or convinced them not to believe the scam (Appendix D.3). We recorded the participant's voice recording for the second question. After the second survey question, the autodialer system plays an ending message stating the researcher's contact information (transcript in Appendix E).

In summary, during each step of the procedure, the autodialer was configured to collect the following inputs from the recipient: Continued, Entered SSN, Convinced, Unconvinced, and Recording.

3.5 Ethics

These experiments were a deceptive study on involuntary participants, and therefore we deeply considered the ethical issues. To address the ethical issues inherent in our experiments, we carefully designed the experiments and worked with our university's IRB, to not simply obtain approval but to conduct the study minimizing harm. This is important because, to have scientifically valid results, we could not obtain informed consent (this would bias the results of the study) and we must deceive the participants (they would need to believe that the call was an actual scam call). To protect our participants, we implemented several safeguards in the experimental design.

The nature of this experiment, studying telephone phishing attacks, involves deception as well as involuntary participation. Both aspects are critical to receiving scientifically valid results—informing the participants of the study would significantly bias the results. However, the use of deception can harm the recipients, by wasting their time, confusing them, or leading them to believe they fell victim to a scam. Therefore, our debriefing served to not only inform the participants of the study, but to also educate recipients about the dangers of telephone scams. In addition, we only called each participant once throughout the entire study duration (to minimize the disruption).

Before proceeding with the study, we also worked with our university's IT security group to provide them with information that would help to alleviate the concerns of our participants. This IT security group at ASU is responsible for the security of all aspects of the university. We shared with the security group the experiment contact list, the experimental design, and the incoming phone numbers (that we used to send the calls) so that the help desk personnel could be prepared to handle any requests and reports. In this way, our participants who reported the scam calls to IT would be assured that it was part of a study.

In recording the results, we also strove to do so ethically and in accordance with established IRB protocols. One of the major safeguards is that we did not record the Social Security number. While a spammer would typically want the Social Security number, all that we record is the fact that they pressed any digit. In fact, we did not even ask for the full Social Security number, and we performed no analysis to see if they provided nonsensical last four Social Security numbers. This has the drawback of decreasing the validity of our data—participants may have felt safe to input only the last four of their Social Security number (when they would not input the full number) or they input fake last four digits of their Social Security number. Although these measures may diminish the strength of our data, we believe ethics is a more important aspect of designing a telephone phishing study.

3.6 Dissemination

We ran the previously described procedure using the 10 described experiments during a workweek in the late March of 2017, during core working hours of 10:00am–5:00pm each day. We used an Internet-hosted autodialer[4] to automate the process of sending out the telephone calls to the 3,000 recipients. Each experiment's calls were simultaneously distributed

[4] https://www.callfire.com/

during the experiment period at a rate of 1–3 live calls per experiment.

We associated each experiment with a unique caller ID. In all experiments, the vast majority of the outbound calls did not reach a live recipient and were answered by a voicemail answering machine. If a recipient could not answer the phone, the recipient could use the caller ID in their call history to call us back. As each experiment had a unique caller ID, the return call would be directed to that particular experiment's procedure. When a recipient called back, the same procedure was administered where a prerecorded scenario announcement message is first played.

While disseminating the phone calls, several unexpected events impacted our study.

The ASU school of journalism and mass communication identified the scam call incidents only 2 hours and 45 minutes from the launch of the experiments on the first day. Instead of reporting it to the university help desk (who were prepared and aware of our study), the school sent out a mass email warning all journalism staff and faculty at 4 hours 28 minutes from launch. However, we did not notice a significant dip in the number of recipients that continued with our scam calls as the portion of work phones at the journalism department represents less than 2% of our sample population.

At 4 hours and 22 minutes from the launch of the experiments, our university's telephone service office also started blocking some of our phone calls as they were receiving system alerts of too many incoming phone calls exhausting the telephone trunk routes. We worked with the telephone service office to get our calls unblocked within the next 4 hours as we decreased the simultaneous call rate of our phone calls to one per experiment.

The IRB office also received some complaints (we were not told exactly how many) regarding the scam call experiments, which resulted in our experiments being paused for roughly 12 hours (start to finish) starting on day 2, as we waited for the IRB committee to review the complaints. The IRB examined our procedures and decided that, as our study was originally designed, the beneficence outweighed the harm (as evidenced by the complaints) and allowed the study to proceed.

A summary showing how these events affected our calls is shown in Figure 3. In the end, despite the unexpected events, we finished sending out the telephone calls to the 3,000 recipients as planned before the end of the workweek.

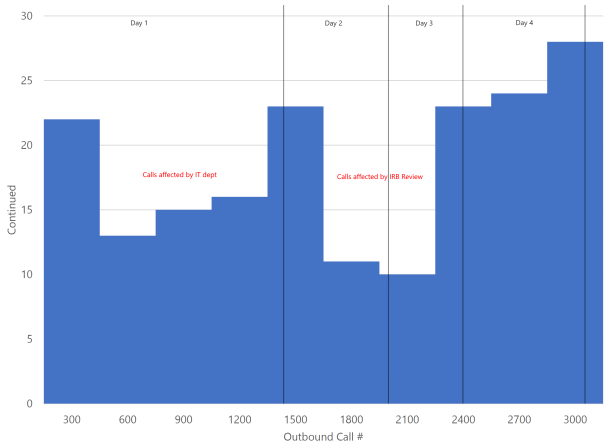


Figure 3: No. of recipients pressed 1 to continue the received calls over the experiment time.

4 Results and Analysis

The input data collected from the 3,000 recipients are presented in Table 2. Across all 10 experiments of 3,000 total recipients, 8.53% (256/3000) of all recipients continued after listening to the scam scenario announcement, 3.73% (112/3000) of all recipients called back after receiving the initial call from us, 4.93% (148/3000) of all recipients entered at least a digit when requested to enter the last four digits of their Social Security number, 1.17% (35/3000) of all recipients explicitly stated that they were convinced by the scam, and 1.23% (27/3000) of all recipients explicitly stated that they were not convinced by the scam.

Before presenting our analysis of the experiments, we first discuss our methodology to systematically analyze their relative effectiveness. The first step of performing the analysis is to decide on metric(s) that will be used as the standard of measurement. To chose an ideal metric, we believe a good metric should not only be quantifiable but also be a proxy for what ultimately matters. From the telephone scammers' perspective, the ultimate goal is to collect as many Social Security numbers as possible for the purpose of conducting identity fraud.

We could use the metric of *Entered SSN*, which is the number of participants that entered any value for their Social Security number (SSN). However, as discussed in Section 3.5, we did not collect the SSNs input by the user. Although this seems to be an ideal metric to estimate the number of SSNs collected, there is still the possibility that the recipient may have tried to enter a fake Social Security number. In fact, in some of the recordings, a few recipients stated that they did not enter their real Social Security number information.

Therefore, we need to derive a metric that could provide a reasonable estimate of the actual number of real SSNs given to us in each experiment. *Convinced* is the metric of the number of recipients that explicitly stated that they were convinced by the scam after the first survey question. This metric is the most conservative for estimating attack success. However, with the low number of responses, participants rarely made it to that step. Using this metric would exclude a large number of recipients that fell for the scam but declined to participate in the phone survey after the debriefing announcement.

Because we cannot assume that all SSNs entered were real, to reduce these types of false positives, we could create a new metric and remove the participants that entered their SSNs and then subsequently stated that they were unconvinced by

No.	Continued		Callbacks		Entered SSN		Convinced		Recordings		Unconvinced		Recordings	
E1	12	4.00%	7	2.33%	6	2.00%	0	0.00%	0	0.00%	4	1.33%	2	0.67%
E2	19	6.33%	7	2.33%	15	5.00%	3	1.00%	0	0.00%	3	1.00%	3	1.00%
E3	13	4.33%	6	2.00%	8	2.67%	1	0.33%	1	0.33%	2	0.67%	1	0.33%
E4	23	7.67%	14	4.67%	13	4.33%	2	0.67%	0	0.00%	3	1.00%	2	0.67%
E5	9	3.00%	3	1.00%	2	0.67%	1	0.33%	0	0.00%	1	0.33%	1	0.33%
E6	9	3.00%	7	2.33%	8	2.67%	2	0.67%	2	0.67%	2	0.67%	1	0.33%
E7	13	4.33%	8	2.67%	9	3.00%	3	1.00%	1	0.33%	5	1.67%	4	1.33%
E8	53	17.67%	22	7.33%	30	10.00%	8	2.67%	3	1.00%	9	3.00%	8	2.67%
E9	60	20.00%	15	5.00%	35	11.67%	7	2.33%	3	1.00%	4	1.33%	3	1.00%
E10	45	15.00%	25	8.33%	22	7.33%	8	2.67%	7	2.33%	4	1.33%	2	0.67%
Total	256	8.53%	112	3.73%	148	4.93%	35	1.17%	17	0.57%	37	1.23%	27	0.90%

Table 2: Summary of recipient inputs from all experiments.

the scam during the survey process. This metric, which we call *Possibly Tricked*, provides a reasonable estimate of the actual number of recipients that fell for the scam by entering the last four digits of their Social Security number. Compared to the previous metrics, this metric provides a good balance of conservativeness and sample size, and, therefore, we use this metric for our analysis.

No.	Entered SSN	Unconvinced	Possibly Tricked	
E9	35	4	31	10.33%
E8	30	9	21	7.00%
E10	22	4	18	6.00%
E2	15	3	12	4.00%
E4	13	3	10	3.33%
E3	8	2	6	2.00%
E6	8	2	6	2.00%
E7	9	6	3	1.00%
E1	6	4	2	0.67%
E5	2	1	1	0.33%
Total	148	37	111	3.70%

Table 3: Estimating the number of recipients possibly tricked into entering their real SSN information

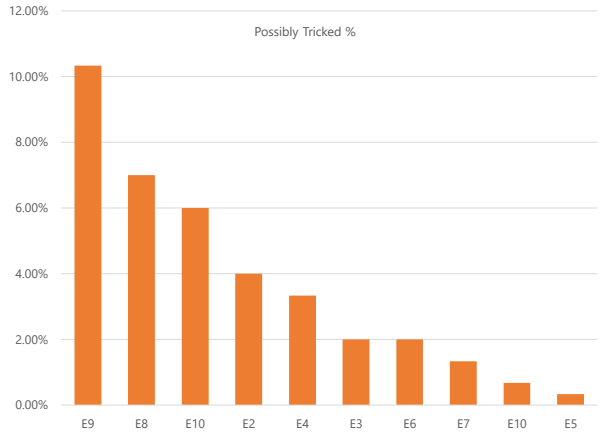


Figure 4: Recipients possibly tricked into entering their real SSN information.

Figure 4 presents a view of the number of *possibly tricked* recipients for each experiment, ranked from most successful to least successful. The tabulated data is in Table 3. Comparing the possibly tricked result between experiments, experiment E9 (spoofed caller ID) had the highest possibly tricked rate among all experiments, with an estimate of 10.33% (31/300) of recipients possibly tricked into entering the last four digits of their Social Security number. Experiment 5 (202 area code, unclaimed tax return) had the lowest success rate among all experiments, with an estimate of only 0.33% (1/300) of recipients possibly tricked into entering the last four digits of their Social Security number.

Attribute	Property	Linear Regression Coefficient
Area Code	Washington, DC	-2.22
	Toll Free	7.78
	Local	1.78
Caller Name	Unknown	-1.32
	Known	8.68
Voice Production	Synthetic	1.68
	Human	5.68
Gender	Male	-0.32
	Female	7.68
Accent	American	5.18
	Indian	2.18
Entity	IRS	-0.99
	ASU	8.34
Scenario	Tax Lawsuit	0.00
	Unclaimed Tax Return	-1.00
	Payroll Withheld	5.67
	Bonus Issued	2.67

Table 4: Linear regression coefficients of all attribute properties overfitted on possibly tricked.

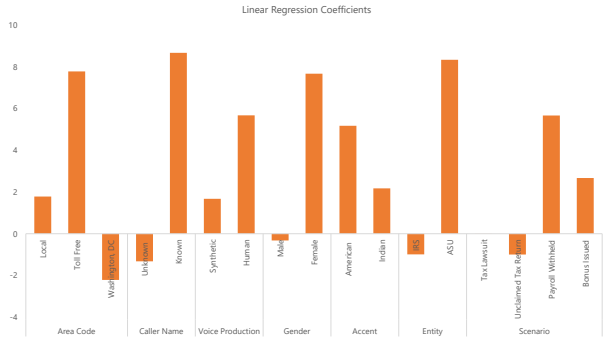


Figure 5: Linear regression coefficients of all attribute properties overfitted on possibly tricked.

The next step is to decide on an appropriate method of data analysis on the chosen metric. With a myriad of possible data analysis methods, we decided to use both linear regression and statistical hypothesis testing analysis. Linear regression is a model-based analysis can produce a model that can fit an optimal mapping of attribute properties to the results (i.e. possibly tricked). However, such method tend to overfit the spurious correlations that occur in training data since it is a

Small Data problem [29]. Furthermore, the attribute properties used in our experiments are also not conditionally independent. Nonetheless, the results of linear regression analysis are shown in Table 4 and Figure 5.

Alternatively, we used a statistical hypothesis testing approach for analysis. Before doing statistical hypothesis testing, we asked, "what are the hypothesis questions that our data can provide an answer for?" We will provide a discussion on the hypothesis questions we decided to ask and how we applied a data analysis method to provide a contextual answer to the hypothesis questions. Because we are testing several hypotheses, we perform the Holm-Bonferroni step-down correction [30] on the significance tests. The results are shown in Table 5 sorted by the individual p-value.

Can manipulating the area code have a significant effect on the attack success of a telephone scam?

In the real world, we observed that telephone scammers used area code manipulation in many instances (in particular in Neighbor Spoofing scams [16]). To provide an answer to this question, we can compare the number of possibly tricked between similar experiments that used different area codes, i.e., E1, E2, and E3. We see that E1 had 0.67% possibly tricked, E2 had 4% possibly tricked, and E3 had 2% possibly tricked.

In our question concerning the significance of area code, since E1 and E2 have the greatest difference in the number of possibly tricked recipients, we test if using a toll-free area code is significantly more effective than Washington, DC area code in the context of the IRS scam example. So we perform a right-tailed p-value hypothesis testing approach on the chosen experiment groups (E1 vs. E2) using the adjusted p-value corrected with Holm-Bonferroni's step-down method [30].

The use of right-tailed p-value statistical hypothesis testing approach is a method to answer if it is "likely" or "unlikely" to observe the improved alternative hypothesis (i.e. E2 possibly tricked) – assuming that the null hypothesis is true (i.e. probability distribution of E1 possibly tricked).

With regards to the choice of using Bayesian vs. Frequentist methods, since we are aware of no similar prior experiments, we can only use Frequentist methods to calculate the statistical significance on the underlying truths using only data from the current experiment.

In addition, not only do we want to know if the improvement to attack success is significant, it is also important to know the magnitude of improvement. To avoid making statements such as "E2 is 5 times more effective than E1", instead of measuring the relative difference, we calculated Cohen's d to measure the effect size for comparison between the two groups.

Using the right-tailed p-value approach, we have a χ^2 statistic of 7.314 and an adjusted p-value of 0.00684. Using an arbitrary confidence level of 95%, it is very likely that using a toll-free area code can result in a more successful attack than using a Washington, DC area code in the context of the IRS scam example. The two groups also have a Cohen's d of 0.222, which suggests it has a small effect according to Cohen [31] and has a somewhat educationally significant effect according to Wolf [32]. Therefore, we could say that the area code can have a statistically significant yet somewhat minor effect on the attack success of telephone phishing scam.

Can manipulating the type of voice production have a significant effect on the attack success of a telephone scam?

To provide an answer to this question, we can compare the number of possibly tricked between similar experiments that used different types of voice production, i.e., E1 and E6. In our question concerning the significance of voice production, we test if using a recorded human voice is significantly more effective than using synthesized voice in the context of the IRS scam example.

Using the same right-tailed p-value approach, we have a χ^2 statistic of 2.027 and an adjusted p-value of 0.155. Using an arbitrary confidence level of 95%, we cannot conclude that using a recorded human voice can result in a more successful attack than using synthesized voice in the context of the IRS scam example. The two groups have a Cohen's d of 0.117, which also suggests the effect size is very small and not educationally significant. Therefore, we are not able to conclude at this time if the type of voice production has a significant effect on the attack success of a telephone phishing scam.

Can manipulating the voice gender have a significant effect on the attack success of a telephone scam?

For the telephone scammer, the voice gender of the voice synthesizer can be easily changed with a simple option click in the autodialer. To provide an answer to this question, we compare the number of possibly tricked between similar experiments that used different voice genders, i.e., E1 and E4. In our question concerning the significance of voice gender, we test if using a female synthesized voice is significantly more effective than using male synthesized voice in the context of the IRS scam example.

Using the same right-tailed p-value approach, we have a χ^2 statistic of 5.442 and an adjusted p-value of 0.0197. Using an arbitrary confidence level of 95%, it is unlikely that using a female synthesized voice can result in a more successful attack than using a male synthesized voice in the context of the IRS scam example. The two groups have a Cohen's d of 0.192, which suggests the effect size is small and not educationally significant. Therefore, it is difficult for us to conclude at this time if the voice gender has a significant effect on the attack success of a telephone phishing scam.

Can manipulating the voice accent have a significant effect on the attack success of a telephone scam?

To provide an answer to this question, we compare the number of possibly tricked between similar experiments that used different accents, i.e., E6 and E7. In our question concerning the significance of voice accent, we test if speaking with an American accent is significantly more effective than speaking with an Indian accent in the context of the IRS scam example.

Hypothesis	Group A	Possibly Tricked	Group B	Possibly Tricked	p-value	Adjusted p-value[1]	Significant[1]	Cohen's d	Effect Size[2]	Conclusive
Entity Scenario (IRS vs. HR)	E1 + E5	3/600	E8 + E9	39/600	1.56E-8	1.09E-7	Yes	0.331	Small & educationally significant	Yes
Area Code (202 vs. 800)	E1	2/300	E2	12/300	0.00684	0.0410	Yes	0.222	Small & somewhat educationally significant	Somewhat
Voice Gender (Male vs. Female)	E1	2/300	E4	10/300	0.0197	0.0985	No	0.192	Small & not educationally significant	No
Caller Name (Unknown vs. Known)	E8	21/300	E9	31/300	0.147	0.588	No	0.119	Very small & not educationally significant	No
Voice Production (Synthetic vs. Human)	E1	2/300	E6	6/300	0.155	0.465	No	0.117	Very small & not educationally significant	No
Voice Accent (Indian vs. American)	E7	3/300	E6	6/300	0.314	0.628	No	0.082	Very small & not significant	No
Motivation (Reward vs. Fear)	E5 + E10	19/600	E1 + E8	23/600	0.530	0.530	No	0.036	Very small & not significant	No

Table 5: Summary of statistical hypothesis testing results ordered individual p-value.
[1]Using p-values corrected with Holm-Bonferroni's step-down method [30].
[2]Using effect size descriptors by Cohen [31] & Wolf [32]

Using the same right-tailed p-value approach, we have a χ^2 statistic of 1.015 and an adjusted p-value of 0.314. Using an arbitrary confidence level of 95%, we cannot conclude that speaking with an American accent can result in a more successful attack than speaking with an Indian accent in the context of the IRS scam example. The two groups also have a Cohen's d of 0.082, which suggests the effect size is very small and not educationally significant. Therefore, we are not able to conclude at this time if the voice accent has a significant effect on the attack success of a telephone phishing scam.

Can spoofing a known caller name have a significant effect on the attack success of a telephone scam?

To provide an answer to this question, we compare the number of possibly tricked between similar experiments that show a difference in the display of a caller name, i.e., E8 and E9. In our question concerning the significance of spoofing caller name, we test if displaying a HR-department caller name "W-2 Administration" is more effective than not displaying a caller name in the context of the HR scam example.

Using the same right-tailed p-value approach, we have a χ^2 statistic of 2.106 and an adjusted p-value of 0.147. Using an arbitrary confidence level of 95%, we cannot conclude that displaying a HR-department caller name can result in a more successful attack than displaying a caller name in the context of the HR scam example. The two groups also have a Cohen's d of 0.119, which suggests the effect size is very small and not educationally significant. Therefore, we are not able to conclude at this time if spoofing a known caller name has a significant effect on the attack success of a telephone phishing scam.

Can manipulating the entity scenario have a significant effect on the attack success of a telephone scam?

Any form of spear phishing involves impersonating an internal entity that the recipient is familiar with. The scammer has to create a spoofed caller ID and devise a scenario that is tailored to the entity, as the "Entity" cannot be set independently from "Scenario". To provide an answer to the hypothesis question, we compare the number of possibly tricked between similar experiments that used different entity-scenarios, i.e. comparing E1 and E5 with E8 and E10. In our question concerning the significance of impersonating an internal entity, we test if impersonating an internal entity is more effective than impersonating the IRS with the context of the scenarios tested.

Using the same right-tailed p-value approach, we have a χ^2 statistic of 31.976 and an adjusted p-value of 1.56E-8. Using an arbitrary confidence level of 95%, it is likely that impersonating an internal entity can result in a more successful attack than impersonating the IRS with the context of the scenarios tested. The two groups also have a Cohen's d of 0.331, which suggests the effect size is small and educationally significant. Therefore, we could say that impersonating an internal entity had a significant effect on the attack success of a telephone phishing scam.

Can manipulating the type of motivation have a significant effect on the attack success of a telephone scam?

To motivate the recipient into taking some harmful action, the scammer could either use fear or reward. To provide an answer to the hypothesis question, we compare the number of possibly tricked between similar experiments that used different types of motivation, i.e., comparing E1 and E8 with E5 and E10. In our question concerning the significance of the type of motivation, we test if fear-based scenarios are more effective than reward-based scenarios the context of the entities tested.

Using the same right-tailed p-value approach, we have a χ^2 statistic of 0.395 and an adjusted p-value of 0.530. Using an arbitrary confidence level of 95%, we cannot conclude that fear-based scenarios can result in a more successful attack than reward-based scenarios with the context of the entities tested. The two groups also have a Cohen's d of 0.036, which suggests the effect size is very small and not educationally significant. Therefore, we are not able to conclude at this time if manipulating the type of motivation has a significant effect on the attack success of a telephone phishing scam.

Summary

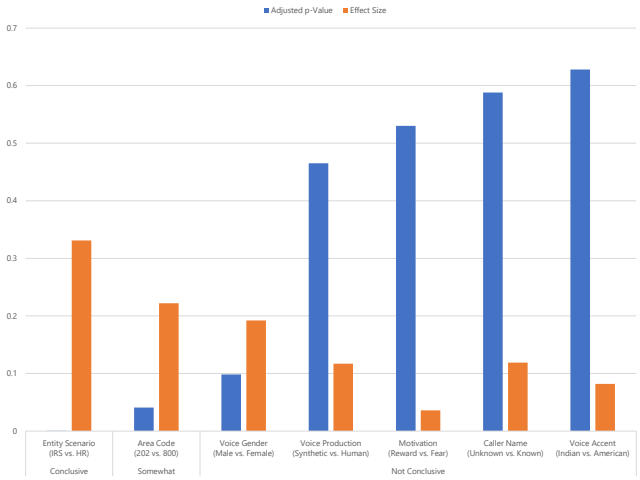


Figure 6: Summary of statistical hypothesis testing results.

The summary of our statistical hypothesis testing results is shown in Figure 6. Based on the statistical hypothesis results, we found that impersonating an internal entity had the most significant effect on the attack success of a telephone phishing scam. We also found that manipulating the area code (using a toll-free vs. a 202 area code) can have a somewhat significant effect.

On the contrary, manipulating the type of motivation, voice production, voice accent, and caller name, individually had an insignificant effect on the attack success. It is also difficult for us to conclude whether manipulating the voice gender has a significant effect even though the result was statistically significant.

5 Survey Responses

In this section, we highlight the recorded survey responses that asked the participants for the reasons they were convinced or unconvinced to enter the last four digits of their Social Security number. We listened to all 44 recorded voice responses and tabulated their responses in Table 6.

Based on the voice responses from the survey respondents, no one provided an explicit voice response on why they were convinced by the IRS scams. The four recordings we received were either silent or contained no useful information. We believe that participants were less willing to report the reasons why they were convinced by the scam after they were explicitly told that they had fallen victim to an attack.

On why the IRS scams were unconvincing, most of the survey respondents stated that they already knew that the IRS would not make a call like this or that they were already vigilant about IRS scam calls. This is understandable because there are numerous media reports about the IRS scams, and the IRS posted many public warnings not to trust these types of scams. This further supports the hypothesis that the impersonated identity and the corresponding scenario was the most significant factor. In experiment E7, two respondents also mentioned that the Indian accent was one of the reasons they were unconvinced.

On why the ASU imposer scams convinced them, most of the survey respondents described something related to the scam scenario, which means that the impersonated entity and the scenario were the key factors. Three respondents also believe that the caller ID was from ASU and stated caller ID was one of the reasons they believed in the scam, even though none of the caller IDs were actually from ASU.

On why the ASU scams did not convince them, most of the survey respondents stated that they were quite certain that ASU would not make a call like this or they were already vigilant about giving their SSN information over an incoming call. Two respondents in experiment E9 mentioned that the scenario only asked for the last four digits of their SSN, and should have asked for their complete SSN if it was really payroll related, which quite possibly meant that those two might have given out their complete SSNs if the phishing scam had asked for it. The external caller ID and synthetic voice were also mentioned as factors that made the survey respondents suspicious.

6 Limitations

The experiments were conducted in a university setting where the recipients are university staff and faculty. The demographics of the recipients in our study are not representative of the general population of telephone users in the US.

The experiments only sent out calls to a specific brand of work phones. The type of phone in our study is not representative of the entire population of telephones in the US. The vast majority of telephones in the US are mobile phones [33], and it is possible that these have different actual tricked rates than work phones. In addition, the participants had to be in their office when receiving the phone call (or to return our call if they listened to the voicemail), which is a different usage behavior compared to mobile phones.

The experiments requested only partial SSN information without storing it. The experiments had several safeguards, and the process was carefully designed and tightly regulated to ensure risks and harm to the human research subjects were minimized. This prevented us from collecting any actual Social Security numbers from the recipients. Collecting actual Social Security numbers might have changed the results of the study: more people might be willing to give out their full Social Security numbers, or more people could be skeptical of providing their full Social Security number.

As we did not collect the Social Security numbers directly, we derived a metric called “possibly tricked.” While the goal is to provide an estimate of the number of Social Security numbers that a real scammer would collect, this metric may

No.	Reasons Convinced	Reasons Unconvinced
E1		Would never enter SSN on incoming call; No name mentioned for the charge
E2		IRS won't make a call like this (x2); Already aware of scams like this
E3		IRS won't make a call like this
E4		IRS won't make a call like this; Didn't sound legitimate
E5		IRS won't make a call like this
E6		IRS won't make a call like this; Already aware of scams like this
E7		IRS won't make a call like this (x4); Indian accent (x2)
E8	To get paid (x2); Sounded legitimate; Trusted work phone; Only asked for last 4 SSN; Caller ID showed local ASU number	ASU won't make a call like this (x5); Not from ASU number (x2); Synthetic voice;
E9	Sounded legitimate; Only asked for last 4 digits of SSN; Caller ID showed ASU W-2	Should have asked for complete SSN (x2); Would never enter SSN on incoming call
E10	To get bonus (x2); Trusted work phone; From ASU number; Asked to do so	ASU won't make a call like this; Not ASU number

Table 6: Summary of recorded survey responses.

be under or overestimating the number of real collected Social Security numbers. With the data presented in this paper (Table 2), others can choose to use different metrics to calculate significance. These new metrics and hypotheses should be corrected to prevent *p*-hacking.

7 Discussion

Our results show that automated telephone phishing attacks can be effective. One experiment, E9, which simulated a targeted phishing attack with caller name spoofing, achieved a 10.33% possibly tricked rate, where recipients possibly divulged the last four digits of their Social Security numbers.

We have also validated some potential key attributes that can have a significant effect on the scam effectiveness: impersonating an internal entity and announcing a relevant scenario. Manipulating the caller ID to a toll-free area code may also somewhat improve the scam effectiveness for certain scams. Other attribute properties such as human voice, female voice, American accent, caller name spoofing, and fear-based scenario also improved the scam effectiveness in our empirical study, however, at this time we are not able to conclusively demonstrate that they have a significant effect. Nonetheless, given how easy it is for a scammer to manipulate all these attributes, a scammer would seek to incorporate all attribute properties that made an improvement to the attack success, i.e. a phishing scam with toll-free area code, spoofing known a caller name, speaking in a recorded human, female voice, in American accent, impersonating an internal entity, motivating the recipient with a relevant fear-based scenario.

To prevent falling victim to these types of phishing scams, we believe that the key is to target and prevent *impersonation*. Our statistical results have shown that impersonating an internal entity had a significant effect on the scam. To address the impersonation issue, feedback from our survey participants suggests that vigilance was an important reason for not falling for a scam. Many surveyed subjects expressed distrust towards our scam calls when they were already vigilant about the scam scenario. Therefore, we recommend education and awareness of telephone phishing as a countermeasure.

On technical solutions, we recommend a similar approach to help the subjects stay vigilant against phishing calls. There are solutions that can provide early warnings against impersonated calls, such as, caller ID authentication [4, 34, 35], which has strong safeguards against caller ID impersonation and could help to warn the users against malicious calls with a reputation system.

8 Related Work

To our knowledge, there have been no prior empirical user studies on telephone phishing. The most similar work we found was by Aburrous et al. [36], who performed a phone phishing experiment on a group of 50 employees contacted by female colleagues assigned to lure them into giving away their personal e-banking usernames and passwords. They were able to deceive 32% of the employees to give out their e-banking credentials. In the experiment, the 50 employees already knew the female colleagues that contacted them, which suggests an insider attack rather than an impersonation attack.

Other related work studied phishing using different channels. Dhamija et al. performed a website phishing study on 22 university participants and their best phishing site was able to fool more than 90% of participants [37]. Egelman et al. performed an email and website phishing experiment on 60 in-person participants to test the effectiveness of various web browser phishing warnings at that time, and it was found that 79% of Internet Explorer 7.0 participants heeded the active phishing warnings and only 13% of them obeyed the passive warnings [38]. Jagatic et al. performed a social media spear phishing study on 481 targeted Indiana University student emails obtained by crawling social network websites and it had a 72% success rate of recipients authenticating themselves on a redirected website [13]. Vidas et al. performed a QR code phishing study where the experiment distributed 139 posters containing QR codes at various locations at Carnegie Mellon University and the city of Pittsburgh, the experiment was able to trick 225 individuals to visit the associated website in four weeks [39].

9 Conclusion

This paper presented the methodology, design, execution, results, analysis, and evaluation of a study exploring why tele-

phone phishing works and how to defend against it. The study was executed using 10 experiments simulating telephone phishing attacks, administered to 3,000 work phones of university staff and faculty over the course of a workweek. The results were collected from the inputs and survey responses of the phone recipients. We analyzed the results by performing linear regression and statistical hypothesis testing methods on a chosen metric derived from the inputs, and we were able to identify at least one attribute that had a significant effect. We provided a discussion on how to effectively prevent such types of telephone phishing scams, and we believe that the best countermeasures should target impersonation and instill vigilance.

References

[1] Federal Trade Commission, "National Do Not Call Registry Data Book for Fiscal Year 2018," 2018.

[2] Federal Trade Commission, "Consumer Sentinel Network Data Book for January - December 2018," 2019.

[3] M. Tischer, Z. Durumeric, S. Foster, S. Duan, A. Mori, E. Bursztein, and M. Bailey, "Users Really Do Plug in USB Drives They Find," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2016.

[4] H. Tu, A. Doupé, Z. Zhao, and G.-J. Ahn, "Toward Authenticated Caller ID Transmission: The Need for a Standardized Authentication Scheme in Q.731.3 Calling Line Identification Presentation," in *ITU Kaleidoscope 2016 - ICTs for a Sustainable World*, ITU Telecommunication Standardization Sector (ITU-T), Institute of Electrical and Electronics Engineers (IEEE), Nov. 2016.

[5] I. S. working group, "Secure telephone identity revisited (stir)." https://datatracker.ietf.org/wg/stir/about/. (Accessed on 04/30/2019).

[6] H. Tu, A. Doupé, Z. Zhao, and G.-J. Ahn, "SoK: Everyone Hates Robocalls: A Survey of Techniques against Telephony Spam," in *Proceedings of the 37th IEEE Symposium on Security and Privacy*, IEEE, 2016.

[7] M. Sahin, A. Francillon, P. Gupta, and M. Ahamad, "SOK: Fraud in telephony networks," in *Proceedings of the 2nd IEEE European Symposium on Security and Privacy (EuroS&P'17), EuroS&P*, vol. 17, 2017.

[8] C. Peeters, H. Abdullah, N. Scaife, J. Bowers, P. Traynor, B. Reaves, and K. Butler, "Sonar: Detecting SS7 Redirection Attacks Via Call Audio-Based Distance Bounding," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2018.

[9] B. Reaves, L. Blue, D. Tian, P. Traynor, and K. R. Butler, "Detecting SMS Spam in the Age of Legitimate Bulk Messaging," in *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pp. 165–170, ACM, 2016.

[10] A. Oest, Y. Safaei, A. Doupé, G.-J. Ahn, B. Wardman, and G. Warner, "Inside a Phisher's Mind: Understanding the Anti-phishing Ecosystem Through Phishing Kit Analysis," in *Proceedings of the Symposium on Electronic Crime Research (eCrime)*, May 2018.

[11] R. C. Dodge Jr, C. Carver, and A. J. Ferguson, "Phishing for user security awareness," *Computers & Security*, vol. 26, no. 1, pp. 73–80, 2007.

[12] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge, "Protecting people from phishing: the design and evaluation of an embedded training email system," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 905–914, ACM, 2007.

[13] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Communications of the ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[14] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching johnny not to fall for phish," *ACM Transactions on Internet Technology (TOIT)*, vol. 10, no. 2, p. 7, 2010.

[15] N. Miramirkhani, O. Starov, and N. Nikiforakis, "Dial one for scam: A large-scale analysis of technical support scams," in *Proceedings of the Symposium on Network and Distributed System Security (NDSS)*, 2017.

[16] E. Fletcher, "That's not your neighbor calling." https://www.consumer.ftc.gov/blog/2018/01/thats-not-your-neighbor-calling, Jan. 2018.

[17] V. B. Payas Gupta, Bharat Srinivasan and M. Ahamad, "Phoneypot: Data-driven Understanding of Telephony Threats," in *Proceedings of the Symposium on Network and Distributed System Security (NDSS)*, 2015.

[18] A. Marzuoli, H. A. Kingravi, D. Dewey, A. Dallas, T. Calhoun, T. Nelms, and R. Pienta, "Call me: Gathering threat intelligence on telephony scams to detect fraud," *Black Hat*, 2016.

[19] I. R. Service, "Irs repeats warning about phone scams." https://www.irs.gov/uac/newsroom/irs-repeats-warning-about-phone-scams, 8 2014. (Accessed on 04/20/2017).

[20] Andrew Johnson, Division of Consumer and Business Education, FTC, "Voicemail from an irs imposter? | consumer information." https://www.consumer.ftc.gov/blog/voicemail-irs-imposter, 9 2016. (Accessed on 04/20/2017).

[21] G. Stringhini and O. Thonnard, "That ain't you: Blocking spearphishing through behavioral modelling," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pp. 78–97, Springer, 2015.

[22] J. Hong, "The state of phishing attacks," *Communications of the ACM*, vol. 55, no. 1, pp. 74–81, 2012.

[23] J. Pavia, "Sadly, irs phone scams are very successful 'businesses'." http://www.cnbc.com/2016/10/18/sadly-irs-phone-scams-are-very-successful-businesses.html, 10 2016. (Accessed on 04/20/2017).

[24] I. R. Service, "Irs warns of pervasive telephone scam." https://www.irs.gov/uac/newsroom/irs-warns-of-pervasive-telephone-scam, 10 2013. (Accessed on 04/17/2017).

[25] I. R. Service, "IRS Alerts Payroll and HR Professionals to Phishing Scheme Involving W-2s." https://www.irs.gov/uac/newsroom/irs-alerts-payroll-and-hr-professionals-to-phishing-scheme-involving-w2s, 3 2016. (Accessed on 04/17/2017).

[26] Federal Communications Commission, "Telephone Consumer Protection Act 47 U.S.C. § 227," 1991.

[27] K. Queen, "Guard the last 4 digits of your social security number: they're all id thieves need." http://blogs.creditcards.com/2015/11/social-security-last-4-digits.php, 19 2015. (Accessed on 08/30/2017).

[28] A. Acquisti and R. Gross, "Predicting social security numbers from public data," *Proceedings of the National academy of sciences*, vol. 106, no. 27, pp. 10975–10980, 2009.

[29] A. E. Deeb, "What to do with "small" data? rants on machine learning." https://medium.com/rants-on-machine-learning/what-to-do-with-small-data-d253254d1a89, 5 2015. (Accessed on 09/25/2017).

[30] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[31] J. Cohen, "Statistical power analysis for the behavioral sciences (revised ed.)," 1977.

[32] F. M. Wolf, *Meta-analysis: Quantitative methods for research synthesis*, vol. 59. Sage, 1986.

[33] "CTIA - The State of Wireless 2018." https://www.ctia.org/news/the-state-of-wireless-2018. (Accessed on 05/02/2019).

[34] C. Jennings, J. Peterson, and E. Rescorla, "Authenticated Identity Management in the Session Initiation Protocol (SIP)," *IETF*, 2016.

[35] B. Reaves, L. Blue, and P. Traynor, "AuthLoop: End-to-End Cryptographic Authentication for Telephony over Voice Channels," in *Proceedings of the USENIX Security Symposium (USENIX)*, 2016.

[36] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Experimental case studies for investigating e-banking phishing techniques and attack strategies," *Cognitive Computation*, vol. 2, no. 3, pp. 242–253, 2010.

[37] R. Dhamija, J. D. Tygar, and M. Hearst, "Why Phishing Works," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 581–590, ACM, 2006.

[38] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: an empirical study of the effectiveness of web browser phishing warnings," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1065–1074, ACM, 2008.

[39] T. Vidas, E. Owusu, S. Wang, C. Zeng, L. F. Cranor, and N. Christin, "Qrishing: The susceptibility of smartphone users to qr code phishing attacks," in *International Conference on Financial Cryptography and Data Security*, pp. 52–69, Springer, 2013.

Appendix

A Announcement Messages

A.1 Tax Lawsuit scenario

This is an urgent notice! This is a final notice from the IRS. The reason for this call is to inform you that the IRS is filing a lawsuit against you. Your action is required immediately, or a penalty will be assessed. To speak to an IRS agent and get more information about this case, please press 1 on your phone now.

A.2 Unclaimed Tax Return scenario

This is an urgent notice! This is a final notice from the IRS. The reason for this call is to inform you that the IRS has an unclaimed tax return for you that is due to expire within three days. Your action is required immediately. To speak to an IRS agent and get more information about claiming your tax refund, please press 1 on your phone now.

A.3 Payroll Withheld scenario

Dear ASU employee. This is an urgent notice! This is a final notice from the ASU W-2 administration office. The reason for this call is to inform you that to process your next Friday payroll, you are required to update your 2017 tax information immediately. To speak to a staff agent and get more information, please press 1 on your phone now.

A.4 Bonus Issued scenario

Dear ASU employee. This is an urgent notice! This is a final notice from the ASU W-2 administration office. The reason for this call is to inform you a performance bonus has been issued to your account. Your action is required immediately. To speak to a staff agent and get more information, please press 1 on your phone now.

B Follow-up Message

Please wait for the next available agent. Thank you for holding. Your call will be connected shortly. Please enter the last four digits of your Social Security number on your phone now.

C Debriefing Message

Hi, I am [redacted for anonymity] in the Department of Computer Science at Arizona State University. I am conducting a research study to measure the effectiveness of telephone phishing. The reason you are receiving this message is because I would like to inform you that what you just did could potentially lead you becoming exploited in a real telephone scam. However, I would like to assure you that this is not an actual scam, none of your social security information was actually collected.

We would like to invite you to participate in our phone survey, to help us better understand your thoughts about the scam. You will be able to listen to the survey questions right after this message. Your participation in this survey is voluntary. There are no foreseeable risks for your participation. If you choose not to participate or to withdraw from the survey at any time, there will be no penalty. Your responses will be anonymous. The results of this study may be used in reports, presentations, or publications but your identity will not be used. Please press 1 to listen to the survey questions or participate in the phone survey.

D Survey Questions

D.1 Did the scam convince you

Thank you. Could you please help us understand if the scam was able to convince you to enter your Social Security number? Please use the number on your keypad to answer this question. If "yes", please press 1. If "no" please press 0.

D.2 What factor made the scam convincing

Thank you. Could you please help us understand what was the most important factor that made the scam convincing? We will record your voice response for this question. At the tone, please state briefly what you thought was the most important factor. When you are finished, please press the pound key to end recording.

D.3 What reason made the scam unconvincing

Thank you. Could you please help us understand what was the most important reason you did not believe in the scam? We will record your voice response for this question. At the tone, please state briefly what you thought was the most important reason. When you are finished, please press the pound key to end recording.

E Ending Message

Thank you. This is the end of the research experiment. If you have any questions concerning the research study, please contact the research team at [redacted for anonymity]. If you have any questions about your rights as a participant in this research, or if you feel you have been placed at risk, you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU [redacted for anonymity], at [redacted for anonymity]. Thank you for your participation. Goodbye.