Supplementary Information for "Behavioral Experiments in Email Filter Evasion"

Blind

Abstract

This Supporting Information document provides additional details alluded to in the main text. In particular, we present the English language test that was used to screen participants, the consent form, the text of all the "ideal" email templates used in the experiments (each is an actual spam or phishing email seen "in the wild"), additional material about the impact of randomization on performance, and additional details and material about the synthetic model.

1 User Interface Details

1.1 English Test

We screened participants to ensure English language proficiency by using the online English test http://www.easyenglish.com/index.asp. We took three questions from this test, shown in Figure 1. To pass, the subjects had to answer two of these correctly.

1.2 Consent Form

Prior to starting the experimental tasks, the subjects were asked to read a consent form (Figure 2) and indicate agreement to participate, as well as indicate that they were at least 18 years old, but clicking the "Agree" button.

2 "Ideal" Email Templates

Each task involved a random assignment of one of 10 "ideal" email templates which the subjects needed to subsequently modify (each of these was filtered by our classification-based filter). In this section we present the text of all of these email instances (all are actual spam/phishing emails). 4 of the 10 instances were spam emails, and the remaining 6 were phishing emails.

2.1 Instance 1 (Spam)

Save 80% discount on drugs . . . save 80% on every order !

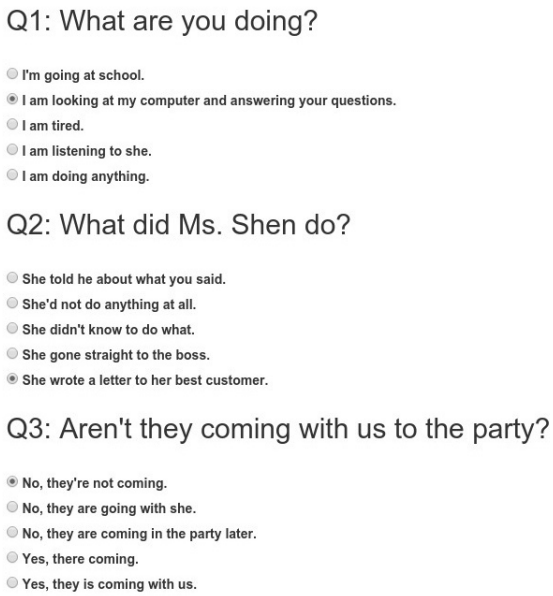


Figure 1: English language test questions.

We are the number one online retailler for dozens of medications. Our customers save 80 cents out of every dollar, every time, compared to the industry price. Yes, that is less than quarter - price

We have all the products that our customers have asked for, including new super-viagra soft-tabs that work in just 15 minutes ! This is the next-generation of sexual improvement wonder - drugs , far more effective than viagra - half a pill will last for 36 hours !

Get all the information on superviagra here : ***malicious url***

Our keys to keeping customers satisfied are:

Easy ordering online Save 80% on regular price

We have massive stocks of drugs for same day dispatch Fast delivery straight to your door with discrete packaging We are the biggest internet retailler with thousands of regular customers

No consultation fee

No intimate questions or examinations

No appointment

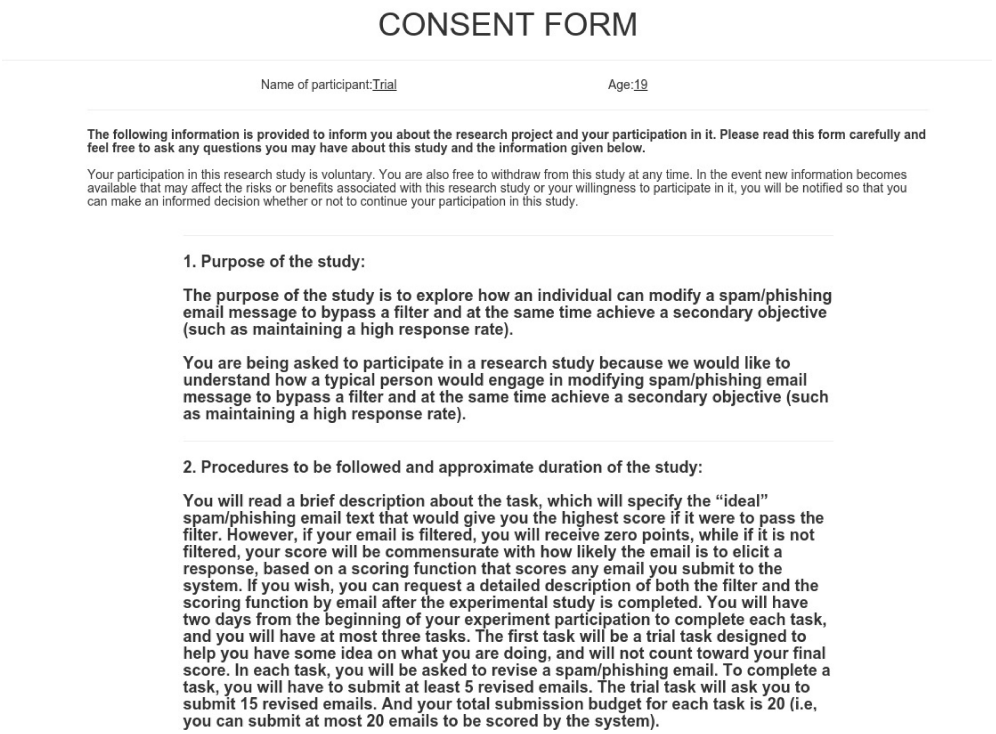
No prior prescription needed

private and confidential service

Please come by our shop, see for yourself the massive range of products that we have available. we do have the lowest price and huge stocks ready for same - day dispatch.

Two million customers can't be wrong !

See our full range at ***malicious url***



CONSENT FORM

Name of participant:Trial Age:19

The following information is provided to inform you about the research project and your participation in it. Please read this form carefully and feel free to ask any questions you may have about this study and the information given below.

Your participation in this research study is voluntary. You are also free to withdraw from this study at any time. In the event new information becomes available that may affect the risks or benefits associated with this research study or your willingness to participate in it, you will be notified so that you can make an informed decision whether or not to continue your participation in this study.

1. Purpose of the study:

The purpose of the study is to explore how an individual can modify a spam/phishing email message to bypass a filter and at the same time achieve a secondary objective (such as maintaining a high response rate).

You are being asked to participate in a research study because we would like to understand how a typical person would engage in modifying spam/phishing email message to bypass a filter and at the same time achieve a secondary objective (such as maintaining a high response rate).

2. Procedures to be followed and approximate duration of the study:

You will read a brief description about the task, which will specify the "ideal" spam/phishing email text that would give you the highest score if it were to pass the filter. However, if your email is filtered, you will receive zero points, while if it is not filtered, your score will be commensurate with how likely the email is to elicit a response, based on a scoring function that scores any email you submit to the system. If you wish, you can request a detailed description of both the filter and the scoring function by email after the experimental study is completed. You will have two days from the beginning of your experiment participation to complete each task, and you will have at most three tasks. The first task will be a trial task designed to help you have some idea on what you are doing, and will not count toward your final score. In each task, you will be asked to revise a spam/phishing email. To complete a task, you will have to submit at least 5 revised emails. The trial task will ask you to submit 15 revised emails. And your total submission budget for each task is 20 (i.e, you can submit at most 20 emails to be scored by the system).

Figure 2: The online consent form.

2.2 Instance 2 (Phishing Email)

Dear Sir / Madam:

Always looks forward for the high security of our clients. Some customers have been receiving phishing emails claiming to be from barclays and advising them to follow a link to what appear to be a barclays web site, where they are prompted to enter their personal online banking details. Barclays is in no way involved with this email and the web site does not belong to us.

For your security, we updated our new ssl servers. Barclays is proud to annouce about the new secure system, which will give our customers a better, fast and secure online banking service.

Due to the recent update of the servers, you are requested to update your account info at the following link.

J.S. Smith. Security Advisor Barclays Bank PLC. Please do not reply to this e - mail . mail sent to this address cannot be answered. For assistance , log in to your barclays online bank account and choose the " help " link on any page.

2.3 Instance 3 (Phishing Email)

Because someone has reported your actions, your account will be deactivated. Maybe you have written content that is abusive or uploaded a picture that can be insulting or harmful to other users. You must confirm your account to stop the warning on your account.

To stop the suspension of your account, please click the link below: ***malicious url***

Facebook Game Network Inc.
Phone: 650.543.4800 fax: 650.543.4801

2.4 Instance 4 (Spam)

Search Engine Position!

Be the very first listing in the top search engines immediately . Our company will now place any business with a qualified website permanently at the top of the major search engines guaranteed never to move: (E. google, yahoo!, msn, alta vista, etc.).

This promotion includes unlimited traffic and is not going to last long. if you are interested in being guaranteed first position in the top search engines at a promotional fee, please contact us promptly to find out if you qualify via email at searchl1@telefonica.net.pe. It's very important to include the url (s) if you are interested in promoting ! ! ! this is not pay per click. Examples will be provided.

P.S. This promotion is only valid in the usa and canada.

Sincerely ,
The Search Engine Placement Specialists.

If you wish to be removed from this list, please respond to the following email address and type the word "remove" in your subject line: search6@speedy.com.pe

2.5 Instance 5 (Phishing Email)

Greetings,

After reviewing your LinkedIn profile, our company would like to present you a part-time job offer as a finance officer in your region. This job does not require any previous experience. Here is a list of tasks that our employee should accomplish:
1. Receive payment from our customers into your bank account
2. Keep your commission fee of 10% from the payment amount
3. Send the rest of the payment to one of our payment receivers in Europe via Moneygram or Western Union.

For more details about this job offer, click ***here***

After enrolling to our part-time job you will be contacted by one of our human resource staff.

Thanks.
Karen Hoffman,
Human Resource Manager.

2.6 Instance 6 (Phishing Email)

Bank of America Online Banking

Message from Customer Service

To stop the suspension of your account, please click the link below: ***http://www.facebook.com/account-suspend/*** *A message from Customer Service is waiting in your Online Banking mailbox.*
If you havent already read it:
Sign in to Online Banking at ***malicious url***
Select Mail at the top of the page.

This alert relates to your Online Banking profile, rather than a particular account.

Want to confirm this email is from Bank of America? Sign in to Online Banking and select Alerts History to verify this alert.

Want to get more alerts? Sign in to your online banking account at Bank of America and within the Accounts Overview page select the "Alerts" tab.

Because email is not a secure form of communication, this email box is not equipped to handle replies.

If you have any questions about your account or need assistance, please call the phone number on your statement or go to Contact Us at ***www.bankofamerica.com****.*

2.7 Instance 7 (Spam)

Discover you made money while you were sleeping!

You must read this word for word ! Information that you may not receive again so please take it seriously ! ! Would you like to . . . . receive thousands in cash daily?

If yes, go here now !

People are making real fortunes, no hype - no false predictions ! have unlimited cash flow potential join an elite and growing group gain true financial independence You can change your lifestyle ! and we can prove it ! !

Our generating leveraging system has been proven 100% effective. Totally duplicable for anyone! The serious money is right here ! ! Do yourself a favor and take a close look at this, you' ll be thankful you did !

This is not sales, our private website will give you all the details. Go here to get them now !

If you received this by error or wish to be excused from our list, simply click here.

2.8 Instance 8 (Phishing Email)

Dear Customer,

You have received this email because we have strong reason to believe that your Amazon account had been recently compromised. In order to prevent any fraudulent activity from occurring we are required to open an investigation into this matter.

If your account is not confirmed, we reserve the right to terminate your Amazon subscription. If you received this notice and you are not an authorized Amazon account holder, please be aware that it is in violation of Amazon policy to present oneself as an Amazon user. Such action may also be in violation of local, national, and/or international law. Amazon is committed to assist law enforcement with any inquires

related to attempts to misappropriate personal information with the intent to commit fraud or theft. Information will be provided at the request of law enforcement agencies to ensure that perpetrators are prosecuted to the full extent of the law.

To confirm your identity with us click the link bellow: ***malicious url***

We apologize in advance for any inconvenience this may cause you and we would like to thank you for your cooperation as we review this matter.

2.9 Instance 9 (Phishing Email)

Dear Taxpayer,

I am sending this email to announce: After the last annual calculation of your fiscal activity, we have determined that you are eligible to receive a tax return of: $273.48

In order for us to return the excess payment, you need to create a e-Refund account after which the funds will be credited to your specified bank account.

Please click ***Get Started*** *to claim your refund:*

2.10 Instance 10 (Spam)

Ink prices got you down?

*Would you like to save up to 80 % on printer ,fax and copier supplies? On brands like epson, canon, hewlett, packard, lexmark and more ! 100 % quality satisfaction guarantee or your money back ! Free same day shipping on all us orders ! We'll beat any price on the internet - guaranteed ! * * Click here to order now ! or Call us toll - free at 1 - 800 - 758 - 8084 ! * Free shipping only on orders of $ 40 or more . * * We beat any online retailer's price by 5 % . * Call us with any other source advertising a lower price and once we verify the price, we will beat it by 5 %! (must be same manufacturer) You are receiving this special offer because you have provided permission to receive email communications regarding special online promotions or offers . If you feel you have received this message in error , or wish to be removed from our subscriber list , Click Here.*

Thank you and we apologize for any inconvenience.

3 Additional Material about the Impact of Randomization

While randomization has a significant effect on the ability of subjects to evade email as described in the main document, we found little evidence of impact on either score (of non-filtered submissions) or time taken to submit. The score for tasks including randomization and those which did not is shown as a function of submission sequence in Figure 3. The differences in the scores are not statistically significant, and there does not appear to be any systematic difference or trend with experience (unlike the ability to evade the filter, which demonstrates clear improvement over time).

Similarly, the differences in time spent on a submission are not statistically different between randomized and noise-free settings (Figure 4). Here, the initial few

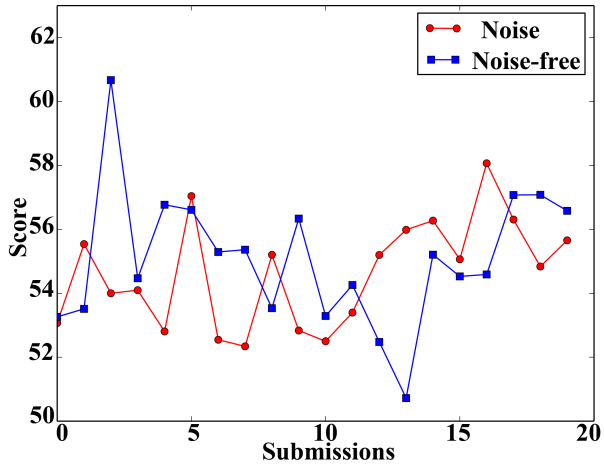


Figure 3: The scores of non-flipped submissions in the "noise" treatments compared to the scores in the "noise-free" treatment, as a function of the submission sequence. There is no clear difference between the two sets of treatments, or a clear trend.

submissions clearly took the longest, but thereafter the trend is quite weak (although submission time does appear to decrease slightly with experience).

4 Additional Details for the Synthetic Model of Human Evasion Behavior

As described in the main text, our synthetic model has two pieces: (1) the model to predict, for each feature, whether or not it will be changed in the next submission, and (2) if a feature is deleted, whether or not it is substituted for by another.

For model (1), a submission is modeled entirely as a feature vector x corresponding to the features in our classifier (filter). We treat each feature in the submission vector x as independent, and train n independent Support Vector Machine models with a radial basis function kernel [**?**], one for each submission feature (for features of the subsequent submission we are trying to predict). For each of these, the predicted variable is a binary indicator whether or not the corresponding feature is changed. Features of each of these n models (as opposed to the predicted outputs themselves) include: a) feature vectors corresponding to the two prior submissions ("ideal" email is used for this purpose for the first two submissions), b) gender of the participant, c) participant education level, d) age of the participant, e) English test score of the participant, and f) scores earned by the two prior submissions (the submission is indicated to be filtered by the classifier when the score is 0).

For model (2), we train a Support Vector Machine model for each deleted word to determine whether or not it is substituted (again, a binary classification problem). We use the same feature vector as for model (1), as well as two additional features, each

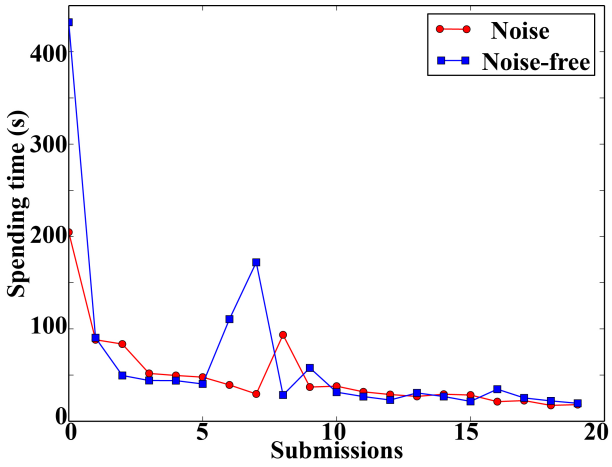


Figure 4: Time spent on submissions in the "noise" and "noise-free" treatments as a function of the submission sequence. There is little difference between the two treatments (the difference is not significant), and only a weak trend (submission time is decreasing with experience).

corresponding to a binary indicator whether or not a specific feature word was substituted in the two prior submissions. For the first two submissions, these two features represent whether or not the word is a "substituted word" based on the training data information. We define the Substituted Ratio of a word as $S_R = \frac{N_{sub}}{N_{del}}$, where N_{sub} and N_{del} correspond to the number of substitution and deletion times of the word within the training submissions, respectively. If $S_R > 0.5$, the feature word is considered as a "substituted word", and the corresponding feature value is 1; otherwise 0.

Figure 5 shows accuracy as a function of the submission sequence (the overall accuracy was 97%). Figures 6 and 7 show average score as a function of the submission sequence in the noise-free and noisy treatments, respectively, comparing the performance for the two scoring functions, S_1 and S_2. In all, the evidence clearly indicates that our synthetic model performs extremely well both in predicting individual behavior as well as in synthetically replicating experimental observables.

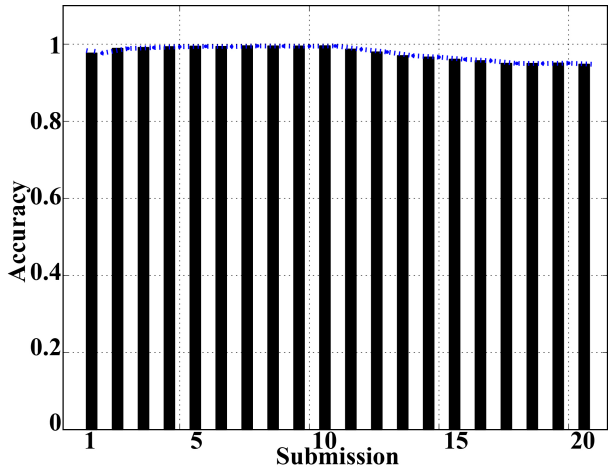


Figure 5: The average cross-validation accuracy shows that the synthetic model is able to predict the next submission vector based on the previous two submissions accurately.

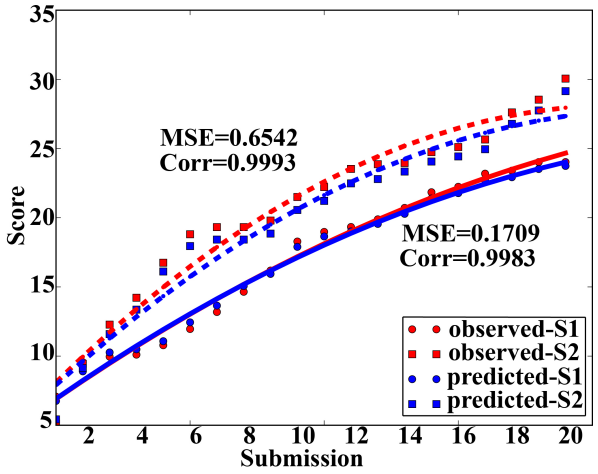


Figure 6: Comparison between experimentally observed scores and those based on the synthetic model of behavior as a function of the submission sequence for S_1 and S_2 scoring function treatments in the "noise free" setting.

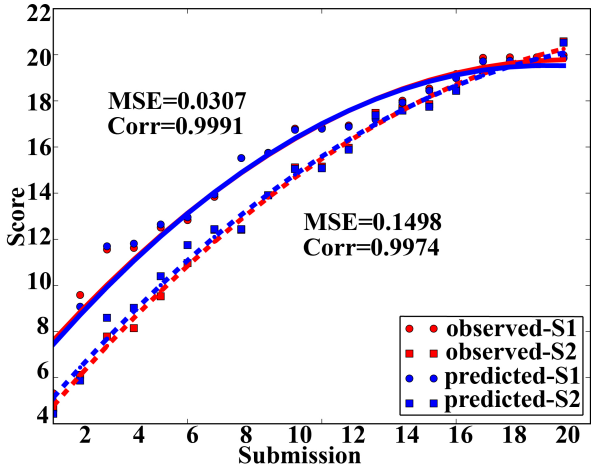


Figure 7: Comparison between experimentally observed scores and those based on the synthetic model of behavior as a function of the submission sequence for S_1 and S_2 scoring function treatments when filter randomization was used.