NISTIR 8194

Exploratory Lens Model of Decision-Making in a Potential Phishing Attack Scenario

Franklin P. Tamborello, II
Kristen K. Greene

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8194

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8194

Exploratory Lens Model of Decision-Making in a Potential Phishing Attack Scenario

Franklin P. Tamborello, II
Cogscent, LLC

Kristen K. Greene
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8194

October 2017

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Under Secretary of Commerce for Standards and Technology

Abstract

Phishing, the transmission of a message spoofing a legitimate sender about a legitimate subject with intent to perform malicious activity, causes a tremendous and rapidly-increasing amount of damage to information systems and users annually. This project implements an exploratory computational model of user decision making in a potential phishing attack scenario. The model demonstrates how contextual factors, such as message subject matter match to current work concerns, and personality factors, such as conscientiousness, contribute to users' decisions to comply with or ignore message requests.

Disclaimer

Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the products mentioned are necessarily the best available for the purpose.

The Rise of Information Systems and Information Theft

Beginning in the 1950s, information processing systems have held increasingly important economic roles at organizations. As that technology has improved, its economic importance has accelerated such that now information services themselves form a substantial proportion of the American economy (Gartner, 2013). As the amount of commerce we transact by computerized information system increases, the valuable information processed and contained within these systems has also become an increasingly tempting target for malicious actors intent on stealing information or performing other malicious acts (Anti-Phishing Working Group, 2016; Kaspersky, 2016; Phishlabs, 2016).

One way to gain unauthorized access to an information system is to attack the information systems' users, rather than attacking the information system itself. Attackers may attempt to lure users into a trap designed to steal authentication credentials such as user account names and passwords. "Phishing" is a set of malicious attack strategies designed around contacting users and persuading them to do something, much as "spam" is unsolicited advertising attempting to persuade users to click on unwanted ads. However, phishing tends to be a means to more sinister ends, such as to obtain information that may be itself valuable, such as credit card account information, or information that may lead to something else of value, such as information system account credentials.

Phishing attacks often take the form of messages directed to the user and transmitted through some computerized communication system that users use, such as email, Short Message Service (SMS), or social network services such as Facebook or Twitter. Like spam, these messages are financially-motivated, unsolicited attempts to persuade the user to take some action. But whereas spam typically is designed with nothing more malicious than advertising in mind, phishing is specifically designed to steal information from users and organizations.

Typically interventions designed to limit the impact of phishing attacks involve training individuals how to identify likely phishing messages and to delete such messages or refer them to their employer organization's information security team. However, some users are wary and savvy enough to avoid such attacks even without explicit training while others are more trusting of people and less aware of the dangers phishing can pose. These are personality traits that tend to be fairly stable independent of training (Parrish, Bailey, & Courtney, 2009; Vishwanath, Harrison, & Ng, 2016). Knowing an individual user's scores on a personality battery, training and technological interventions could adjust to optimize ratios of accepted and rejected messages, and on what bases.

We desired a computational model in which we could specify a variety of factors loading on constructs having to do with personality traits of the users as well as other factors of interest such as demographics, message properties, and the user's milieu that happens to be in place at the time the message is received. The Lens model (Brunswick, 1956) is a computational model of decision making which can be adapted for us in most decision-making contexts. The basic idea is that there are multiple pieces of information, "cues," upon which people can base their decisions. Cues and decisions probabilistically relate to each other, and we can measure their relationship statistically. Using multiple regression, we can construct a statistical model of the judgments users make. If we know what the judgments should have been then we can measure that relationship the same way, that is, we can construct a similar statistical model of the environment.

From these two regressions the model calculates quantities of the lens model, which serve as indicators of the judge's performance relative to the true state of the world and relative to the judge's own performance. "Knowledge" is the correlation of predicted judgment with predicted true state. "Achievement" is the correlation of judgment with true state. "Consistency" is correlation of predicted judgment with actual judgment. And finally, "predictability" is the correlation of predicted state with true state.

An Exploratory Lens Model of User Decision-Making in a Potential Phishing Attack Scenario

We identified from literature data factors of interest regarding user decision-making in a phishing cyber attack scenario. These factors of interest take the form of a statistical model of human decision making in a potential phishing attack scenario. We culled personality factors from the International Personality Item Pool (IPIP) which we assumed to be relevant to making judgments about the "phishiness" of messages: anxiety, self-consciousness, intellect, trust, cautiousness, and agreeableness. For each personality factor we generated values from a normally distributed random sampling process, for each half of cases so that the personality factor would correlate either positively or negatively with judgment and message state. For example, since we hypothesized anxiety to be negatively correlated with message judgment performance, its first half of cases averaged a score of -1, and its second half averaged 1. Since message judgment and true state were both first half "ham," (legitimate messages, as opposed to spam or phish) second half "phish," this made anxiety negatively correlated with judging "ham" when the message was actually ham.

The model uses a hypothetical dataset predicated on the assumption that stable personality traits contribute significantly to the variance in the response to potential phish messages. Parrish, Bailey, and Courtney (2009) proposed a framework using the Big-Five personality traits to explain why some people are more susceptible than others to phishing attacks. Other authors such as Vishwanath, Harrison, and Ng (2016) note that relatively suspicious individuals bias their message classification against compliance with the sorts of requests that phishing attack messages tend to include, such as to follow a link.

The hypothetical dataset implements a scenario in which personality traits such as anxiety and cautiousness do predict phishing susceptibility, but demographic traits such as gender and education do not. The data set encompasses distributions of scores taken from several scales of the International Personality Item Pool (2016): anxiety, self consciousness, intellect, trust, cautiousness, and agreeableness. The dataset also includes several demographic variables: gender, age, education, number of years at the organization, career path type, and also properties of the message.

We probabilistically generated a dataset of 1 000 hypothetical phishing judgments of 1 000 messages. We sampled one half of cases with replacement with a 90% chance of a ham judgment, then other half with replacement with a 90% chance of phish judgment. We performed the same sampling procedure for message true state, ham and phish, respectively, to simulate a scenario with significant knowledge and achievement.

Results of the Exploratory Lens Model

We generated synthetic data reflecting a scenario in which personality factors, but not demographic factors, predict phishing attack susceptibility. The R statistical programming language (https://www.r-project.org) contains several functions useful for generating distributions, such as sample and rnorm. We used those two functions to generate hypothetical judgments, message states, normalized personality scale scores, demographic data, and message property scores for Lens model ingestion. We manipulated probability distributions so that some factors would better predict judgments and states than other factors. Table 1 enumerates the lens model factors, their values, and their values' probabilities. We maintain source code for this project at https://github.com/tamborello/phishing-lens-model and https://github.com/usnistgov/PhishingLensModel. Tables 1 and 2 enumerate the model's independent factors and their weights for judgment and message type, respectively. Table 3 lists the four outputs of the lens model.

Table 1. Lens model independent factors and their weights for judgment outcomes, where judgment means classifying the message as phish or ham. Positive coefficients indicate ham judgment while negative coefficients indicate phish judgment. Asterisks indicate significance at the 0.05 level.

Factor	**Weight**
Context	0.019
Message Persuasiveness	0.002
Conscientiousness*	0.047
Anxiety	0.011
Self-Consciousness*	-0.003
Intellect*	-0.123
Trust*	0.151
Cautiousness*	-0.095
Agreeableness	0.017
Gender	-0.019
Age	0.000
Education	-0.005
Years at Organization	-0.012
Career Path	0.006

Table 2. Lens model independent factors and their weights for true states of the message types, where message type means ham or phish. Positive coefficients indicate the message is phish while negative coefficients indicate ham. Asterisks indicate significance at the 0.05 level.

Factor	**Weight**
Context	0.059
Message Persuasiveness	-0.001
Conscientiousness	0.008
Anxiety*	0.025
Self-Consciousness	0.015
Intellect*	-0.120
Trust*	0.141
Cautiousness*	-0.097
Agreeableness	0.009
Gender	-0.002
Age	-0.000
Education	-0.000
Years at Organization	-0.003
Career Path	-0.003

Table 3. Lens model measures and values for a hypothetical scenario in which personality factors, but not demographic factors, predict phishing attack susceptibility

Lens Measure	**Value**
Knowledge	0.997
Achievement	0.776
Consistency	0.890
Predictability	0.860

Discussion

The exploratory Lens model demonstrated a scenario in which some personality traits predict user response to phishing attacks. To the authors' knowledge this is the first computational model of its kind in this domain. Such personality profiling information could be used to target individual users for remedial training regimes and technological interventions. However, it could also put individuals at greater risk of phishing attack if phishing attackers can profile potential victims to selectively target those who may be more susceptible to that particular sort of persuasion.

Defensive personality profiling for phishing susceptibility could not only identify individuals at greater risk for succumbing to phishing attack, but also identify what kind of training to give those individuals. For instance, an individual scoring highly on the 'trust' trait would benefit from training emphasizing different skills and knowledge than individuals scoring highly on the 'conscientiousness' trait. Furthermore, such persons would likely benefit from more training tailored to those personality traits than would individuals scoring low on those traits.

Offensive personality profiling for phishing susceptibility may involve a tiered strategy of multiple phishing messages. First an attacker might send a seemingly innocuous request to respond to a questionnaire. The questionnaire's purpose would be to assess potential victims' susceptibility via their measured personality traits. Those users responding would by definition be more susceptible, but they might be susceptible for different reasons, e.g. because they are generally trusting of others or are conscientious about fulfilling perceived obligations. Phishers might then escalate their attack by spearphishing questionnaire respondents for more significant responses, but couched in terms tailored to victims' assessed personality type, e.g. "Send me your password, I promise to keep it safe," or "Send me your password so that I may update the profile that you must use to perform your duties."

As for technological interventions, these could be tailored to users' assessed personality as it pertains to likely phishing susceptibility. Most modern spam filters rely upon Bayesian classifiers to classify incoming messages as either ham or spam, phish of course being a subset of spam. However, signal detection theory teaches us that for all but the perfect classifiers, some signals will be misclassified and we can control the tradeoffs according to values of the four combinations of classification and true state. The point at which judgment is made according to relative weighting of outcome is called the bias. Users spam filters can have differing biases according to their assessed personality phishing susceptibility. For instance, trusting individuals might be equipped with a spam filter biased particularly against messages appealing to a sense of trust.

References

Anti-Phishing Working Group. (2016). Phishing attack trends reports. [Web page] Retrieved from http://www.antiphishing.org/resources/apwg-reports/

Brunswick, E. (1956). *Perception and the representative design of psychological experiments (2 ed.).* Berkeley, CA: University of California Press.

Forecast Alert: IT spending, worldwide, 4Q12 update. [Web page]. Retrieved from Forecast Alert: IT Spending, Worldwide, 4Q12 Update.

International Personality Item Pool. (2016). International personality item pool. [Web page] Retrieved from http://ipip.ori.org

Kaspersky Lab. (2016). Quarterly spam and phishing report. [Web page] Retrieved from https://securelist.com/all/?category=442

Parrish, J. L., Bailey, J. L., & Courtney, J. F. (2009). A personality based model for determining susceptibility to phishing attacks. In *Southwest Decision Sciences Institute Meeting*: Little rock: University of Arkansas. University of Arkansas. Retrieved from Google Scholar: http://www.swdsi.org/swdsi2009/

Phishlabs. (2016). Phishing trends & intelligence report. [Web page] Retrieved from https://info.phishlabs.com/pti-report-download

Vishwanath, A., Harrison, B., & Ng, Y. J. (2016). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research, 1-21*. doi:10.1177/0093650215627483

Appendix: Model Code

```
A# Author: Frank Tamborello
# contact: frank.tamborello@cogscent.com
# CC-BY-SA 2016 Cogscent, LLC
#
# This library is free software; you can redistribute it or
modify it under the terms of Creative Commons Attribute-
ShareAlike 4.0 (CC BY) International License:
http://creativecommons.org/licenses/by-sa/4.0/
#
# This library is distributed in the hope that it will be
useful, but WITHOUT ANY WARRANTY; without even the implied
warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR
PURPOSE.
#
# THE SOFTWARE IS PROVIDED 'AS IS' WITHOUT ANY WARRANTY OF
ANY KIND, EITHER EXPRESSED, IMPLIED, OR STATUTORY,
INCLUDING, BUT NOT LIMITED TO, ANY WARRANTY THAT THE
SOFTWARE WILL CONFORM TO SPECIFICATIONS, ANY IMPLIED
WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR
PURPOSE, AND FREEDOM FROM INFRINGEMENT, AND ANY WARRANTY
THAT THE DOCUMENTATION WILL CONFORM TO THE SOFTWARE, OR ANY
WARRANTY THAT THE SOFTWARE WILL BE ERROR FREE. IN NO EVENT
SHALL NIST BE LIABLE FOR ANY DAMAGES, INCLUDING, BUT NOT
LIMITED TO, DIRECT, INDIRECT, SPECIAL OR CONSEQUENTIAL
DAMAGES, ARISING OUT OF, RESULTING FROM, OR IN ANY WAY
CONNECTED WITH THIS SOFTWARE, WHETHER OR NOT BASED UPON
WARRANTY, CONTRACT, TORT, OR OTHERWISE, WHETHER OR NOT
INJURY WAS SUSTAINED BY PERSONS OR PROPERTY OR OTHERWISE,
AND WHETHER OR NOT LOSS WAS SUSTAINED FROM, OR AROSE OUT OF
THE RESULTS OF, OR USE OF, THE SOFTWARE OR SERVICES
PROVIDED HEREUNDER.
#
# Description: Hypothetical Brunswickian Lens model of
phishing attackee decision-making behavior highlighting
personality factors
#
# Acknowledgment: This research is sponsored by Measurement
Science and Engineering grant 70NANB15H252 from the
National Institute of Standards and Technology (NIST).
Special acknowledgements are due to Mary Theofanos and
Kristen K. Greene of NIST's Information Technology
Laboratory. The authors also thank anonymous reviewers for
their comments.
#
#
# Revision 5
#
# Revision History
# 2016.07.08 fpt 1
# Init
#
```

```
# 2016.07.09 fpt 2
# Added header
#
# 2016.11.21 fpt 3
# Deleted the random data scenario
#
# 2016.12.03 fpt 4
# Added factors for user work state context, such as the
accountant who receives a phish ostensibly regarding an
unpaid invoice, and the personality factor
conscienscıousness.
#
# 2016.12.07 fpt 5
# Manipulate factor weights



# Test Data, Hypothetical Scenario: Personality traits
predict phishing detection performance
# Execute this first, then linear models, then Lens model
measures, to see a hypothetical effect of personality
factors.
phishData <-
    data.frame(
        # 0: Reject phish message; 1: Respond to phish (ie
succumb to attack)
        "judgment"=c(sample(c(0,1), size=500, replace=T,
prob=c(.95, .05)),
                     sample(c(0, 1), size=500, replace=T,
prob=c(.05, .95))),
        # 0: Ham; 1: Phish
        "state"=c(sample(c(0,1), size=500, replace=T,
prob=c(.95, .05)),
                  sample(c(0, 1), size=500, replace=T,
prob=c(.05, .95))),
        "messageProperties"=c(rnorm(n=500, mean=0.5),
                              rnorm(n=500, mean=-0.5)),
        "context"=c(sample(c(0,0.5,1), size=450, replace=T,
prob=c(.9, .05, .05)),
                    sample(c(0,0.5,1), size=100, replace=T,
prob=c(.05, .9, .05)),
                    sample(c(0,0.5,1), size=450, replace=T,
prob=c(.05, .05, .9))),
        "conscientiousness"=c(rnorm(n=500, mean=-1,
sd=0.5),
                               rnorm(n=500, mean=1,
sd=0.5)),
        "anxiety"=c(rnorm(n=500, mean=-0.75, sd=0.75),
                    rnorm(n=500, mean=0.75, sd=0.75)),
        "selfConsciousness"=c(rnorm(n=500, mean=-0.75,
sd=0.75),
                              rnorm(n=500, mean=0.75,
sd=0.75)),
        "intellect"=c(rnorm(n=500, mean=1, sd=0.2),
                      rnorm(n=500, mean=-1, sd=0.2)),
        "trust"=c(rnorm(n=500, mean=-1, sd=0.2),
                  rnorm(n=500, mean=1, sd=0.2)),
        "cautiousness"=c(rnorm(n=500, mean=1, sd=0.25),
                         rnorm(n=500, mean=-1, sd=0.25)),
        "agreeableness"=c(rnorm(n=500, mean=-0.75,
sd=0.75),
                          rnorm(n=500, mean=0.75,
sd=0.75)),
        "gender"=sample(c(0, 1), size=1000, replace=T),
        "age"=rnorm(n=1000, mean=45, sd=5),
        "education"=rnorm(n=1000),
        "nYearsAtOrg"=rnorm(n=1000),
        "CareerPath"=rnorm(n=1000));



# Build the linear models
lm.judgment <- lm(judgment ~ context + conscientiousness
+anxiety + selfConsciousness + intellect + trust +
cautiousness + agreeableness + gender + age + education +
nYearsAtOrg + CareerPath + messageProperties,
data=phishData);
summary(lm.judgment);

lm.state <- lm(state ~ context + conscientiousness +
anxiety + selfConsciousness + intellect + trust +
cautiousness + agreeableness + gender + age + education +
nYearsAtOrg + CareerPath + messageProperties,
data=phishData);
summary(lm.state);

# Get the model coefficients
coef.judgment <- coefficients(lm.judgment); coef.judgment
coef.state <- coefficients(lm.state); coef.state


# Get the predicted judgment & state values
pred.judgment <- fitted(lm.judgment);
pred.state <- fitted(lm.state);


# Compute measures of the Lens model
knowledge <- cor(pred.judgment, pred.state) # correlation
between predicted judgment and state

achievement <- cor(phishData$judgment, phishData$state) #
correlation between actual judgement and state

consistency <- cor(pred.judgment, phishData$judgment) #
correlation between predicted and actual judgments
```

```
predictability <- cor(pred.state, phishData$state) #
corrlation between predicted and actual states

# See the values of the Lens model
knowledge; achievement; consistency; predictability



# plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(pred.judgment)
plot(pred.state)
plot(phishData$judgment)
plot(phishData$state)



# Reference
http://www.statmethods.net/stats/regression.html
```