2017

On Factors That Influence User Interactions with Social Media Spam: Empirical Exploration Based on a Survey and Experiment

Thomas J. Kyanko

On Factors That Influence User Interactions With Social Media Spam: Empirical Exploration Based On A Survey And Experiment

Thomas J. Kyanko

Thesis submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Computer Science

Katerina Goseva-Popstojanova, Ph.D., Chair
Saiph Savage, Ph.D.
YanFang Ye, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2017

Keywords: Empirical studies, Survey, Experiment

Abstract

On Factors That Influence User Interactions With Social Media Spam: Empirical Exploration Based On A Survey And Experiment

Thomas J. Kyanko

This thesis explores various factors that influence whether or not users of social media platforms will interact with spam. The research is based on using survey and experimental approaches. The survey looked at several spam related behaviors: ability to identify spam, tendency to interact with spam, and tendency to report spam. In total 256 responses were analyzed, which were collected by an online survey system. Results of the survey show that education about spam did not correlate with changes in behavior, even when users reported that the education had an effect on them. Several other factors, commonly thought to be related to interaction with spam, such as technical background, also showed little correlation to spam related behaviors. It was also found that users tend to have similar behaviors for social media spam and email spam. The experiment involved sending mock spam messages to Facebook, LinkedIn, and Twitter users. 1,200 messages per social media platform were sent. The factors studied were: social media platform, message content, matching the message content to the sending profile, and method of selecting message content. The experimental results showed that overall the highest interaction rate was on Twitter and the lowest was on Facebook. Matching the sender's profile to the content of the messages sent and matching the content of spam to recipient interests did not lead to higher interaction rates than randomly selected messages and sending profiles. Additionally, news related spam distributed on Twitter was most effective in tricking users.

Acknowledgments

Part of the work presented in this thesis was funded by Leidos, Morgantown, WV.

Special thanks to Carlos Toxtli for his work in implementing the experimental system, Dr. Saiph Savage for her work in the system design, Dr. YanFang Ye for her work on my advising committee, and Dr. Katerina Goseva-Popstojanova for advising me on research and writing throughout my degree.

This thesis is dedicated to my family and girlfriend for their support and encouragement throughout my work.

Contents

Abstract ii

Acknowledgments iii

List of Tables vi

List of Figures vii

1 Introduction 1

2 Related Work 4

3 Survey on Social Media and Email Spam 8
3.1 Introduction 8
3.2 Basic Survey Facts 9
3.3 Results 10
3.3.1 Respondent Profile 11
3.3.2 Analysis of Spam Encounters, Identification, and Reporting 17
Factors affecting ability to identify spam 18
Factors affecting spam interaction 20
Factors affecting reporting spam 21
3.4 Summary of the Results & Discussion 23
3.5 Threats to Validity 24
3.6 Conclusion 26

4 Social Media Spam Experiment **27**
4.1 Introduction 27
4.2 Experiment 28
4.2.1 Design of Experiment: Factors and Levels 28
4.2.2 Overview of System Design & Experimental System 29
4.2.3 Implementation & Running the Experiment 30
Message creation 30
Persona account creation 32
Selection of recipients 32
Sending messages 33
Collecting message interactions 34

Data preprocessing 35
4.3 Results 36
4.3.1 Empirical Results 37
4.3.2 Analysis Based on Logistic Regression 40
Overview of logistic regression 41
Results of logistic regression 42
4.3.3 Summary of the Results & Discussion 45
4.4 Threats to Validity 47
4.5 Conclusion 49

5 Conclusion 50

References 52

A Survey Questionnaire 56

B Explanation of Contingency Correlation Coefficient 63

C Debrief Page for the User Interaction with Social Media Spam 65

D Messages Sent as Spam via Social Media Platforms 67

List of Tables

3.1 Education major 12
3.2 Combinations of social media access methods 14
3.3 Combinations of email access methods 15
3.4 Education impact on respondents 16
3.5 Report Spam 21
3.6 Summary of the results 25

4.1 Results for each cell 38
4.2 Summary results by network 39
4.3 Summary results by persona 39
4.4 Summary results by message selection type 39
4.5 Summary results by message content type 39
4.6 Results for message content type by platform 40
4.7 Logistic regression results of main effects 43
4.8 Logistic regression results of main effects by platform 44
4.9 Summary of the main findings of the experimental study 45

List of Figures

3.1 Distribution of respondent ages 11
3.2 Social media platforms used by respondents 12
3.3 How frequently respondents use social media and email 13
3.4 Frequency of social media and email use by college major 13
3.5 Methods of accessing social media and email 14
3.6 Sources of spam education 15
3.7 Ratios of impact of spam education compared to sources of education 16
3.8 Frequencies of encountering spam 17
3.9 Ability to identify social media and email spam 18
3.10 Frequency of interacting with spam 18
3.11 Ratios of spam identification compared to access methods 19
3.12 Ratios of spam interaction compared to access methods 20
3.13 Numbers of respondents who report spam on social media, email, or both . . 21
3.14 Ratios of respondents who report spam on social media and email by access method 22

4.1 Conceptual overview of system design 30

Chapter 1

Introduction

As the Internet has advanced so have the methods to spread spam, and the purposes it is used for. This has become noticeable with the the creation and growth of social media platforms, where these platforms provide more opportunities for spreading spam. Spam is a significant security issue, as it can be used as part of processes for stealing login information [1], spreading malware [2], invading privacy [3], or other nefarious activities. Social media spam has also been used for "spamdexing", i.e. maliciously improving a web site's search engine rating by increasing the number of other sites that link to it, and spreading misinformation, often referred to as "fake news". Because of this, ensuring that users do not fall prey to malicious spam is of paramount importance. Due the several decades that email has existed, spam spread via email has been used for far longer than spam on social media sites. Because of this social media spam has been studied less than email spam. Additionally, due to the recent rise in social media platforms, users are likely to have less experience and knowledge related to the dangers on them. Therefore there is a need for research that will help to better understand spam and the behavior of computer users relating to it.

Empirical studies can be classified into three methodologies: surveys, observational studies, and experiments. Surveys question individuals about select topics and analyze the responses provided. Observational studies analyze data collected from a real-world source, e.g. case studies, though it can be difficult to draw conclusions as there are no researcher designed controls. Experiments study factors by designing the experimental system with controls such that only selected independent variables are changed to specific levels. This thesis is focused on empirical exploration of social media spam, using two methods: a survey and an experiment. The survey provides a comparison of user knowledge and behaviors between social media spam and email spam. Aspects compared include ability to identify spam, tendency to interact with spam, and tendency to report spam. The experiment provides a controlled method of collecting real world information on factors that affect user tendencies to interact with social media spam. Factors studied are the social media platform messages are sent on (e.g., Facebook, Twitter, and LinkedIn), the similarity between spam sending profiles and the content of the messages they send, the matching of spam message content to recipient interests, and the type of content included in spam messages.

This work has several contributions to the study of social media spam. First, using a survey approach, it provides a better understanding of the similarities and differences in user behavior related to social media and email spam. It also adds to the body of knowledge about how education affects or does not affect user behaviors. In [4], an online survey was used to study user knowledge of email phishing. That work focused on understanding users' past Internet experiences and was designed to understand how much knowledge email users have about computer security, while this thesis uses a survey approach to study user behaviors related to both social media spam and email spam, and factors that affect them. Surveys were also used in [5], which studied user behaviors related to several types of email scams. The study showed that the most successful scams are related to social influence, complying with authority figures, maintaining consistency, and are most effective on individuals who have low self-control.

Furthermore, in this thesis we use an experimental approach to explore and quantify the effect of several factors on the interactions of social media users with spam. To the best of our knowledge, this method is used for the first time to study social media spam. A similar experimental design and methods were used in [6], however, that work focused only on phishing using email and considered a different set of factors (e.g., social engineering influence techniques such as: liking, reciprocity, social proof, consistency, authority, and scarcity). Prior work done in [7] and [8] also studied the interaction rates with phishing emails through experiments, though the studies used different factors and the messages were sent only to university students, not to the general population via social media platforms as was done in this thesis. Other studies also used experimental methods, i.e. [9], though these were performed in a laboratory setting and therefore the individuals studied were less likely to act in realistic ways. Finally, the results of the survey and experiment presented in this thesis can be incorporated into education about social media spam with a goal to improve their ability to identify and avoid spam.

This thesis differs from prior work studying spam, in that it focuses on the human side of spam rather than building automated systems to classify spam, i.e. how social media users respond to spam. Those studies were observational in nature and studied attributes of both spam and spammers [10, 11, 12, 13]. Studying these attributes was done with the primary goal of building and testing spam classification algorithms. This work studies human behaviors related to spam and is therefore unique, though complementary, to such work.

The remainder of this thesis is organized as follows. Chapter 2 is a summary of related work. Chapter 3 the design and the results of the survey on social media spam. Chapter 4 the design and results of the experiment of social media spam. A summary of the results and main findings, as well as concluding remarks and potential areas of future work are given in Chapter 5.

Chapter 2

Related Work

Spam has been a problem for many years, adapting from postal mail to email and more recently to social media. During this time work has been done to analyze it, though due to the relatively recent advent of social media spam, as well as the rate at which social media changes, research has not kept pace. Spam is of concern from a cybersecurity perspective as it can be used as part of phishing attacks [1], to facilitate drive-by downloads [2], or as openings for scams [14].

Much of the research done thus far to study spam has been observational studies focused on spam attributes, both sent via email and social media, with the goal of designing and testing classification algorithms. Although this thesis is not focused on spam classification, such work is complementary. Research on email spam has been performed for some time, with work designing and testing machine learning algorithms ongoing for years [15] and new methods be tested today [16]. Due to the longevity of this line of research, there exist a variety of methods and algorithms, analyzing many email attributes, and implemented in a wide range of commercial products [17]. Prior work as studied attributes of social media spam and how spammers operate on social media platforms [10]. These attributes, such as number of messages posted per day and similarity of messages sent, were also found to be useful in automating the detection of spammer accounts. Theoretical work has also proposed methods of assisting email users in identifying likely phishing attacks through visual cues provided in the email client [11]. In recent years, the amount of automated bot traffic has increased, including bots used to spread spam [18]. Current research has also studied methods of performing automatic classification to build more accurate spam filters and detection of spam profiles on social media platforms [10]. Spam classification on social media platforms can involve studying text features, profile attributes, and network graphs for example [12]. Other detection methods include analyzing URL posting behavior [19, 20] and as well as how social media users interact with URLs [20]. [21] studied the behaviors of spammer accounts on Twitter, such as the ratio of followers to followings, message similarity, and API usage. The findings showed that several attributes are useful in automatically classifying spamming accounts. Work in [13] focused on spam language attributes, without focusing on spammer profiles, to build classifiers that are correctly classify over 90% of messages.

Additionally, research has been done to examine different types of social engineering attacks, such as in [22], to better understand how humans interact with spam. Part of this body of knowledge worked on relating psychological aspects of social engineering and how it pertains to spam and phishing, such as the common human weaknesses that are attacked to persuade users to interact with illicit messages [23]. These sorts of analysis have been part of a larger push in recent years to apply resources to empirical case studies and experiments [24], though true experiments are still far less common due to the ethical and resource requirements. Of particular interest to this research is how users perceive and understand spam and related issues, such as phishing and scams. Related work has shown that users have significant differences in their knowledge of cybersecurity topics. Research, such as a pair of surveys in [5], has also explored some of the most common reasons users fail to detect or mitigate various types of email scams. Those surveys showed the most successful attacks are ones related to social influence, complying with authority figures, maintaining consistency, and are most effective on individuals who have low self-control. For example, in a laboratory experiment performed in [25] studied user strategies for determining if emails are legitimate or spam. The results showed that users are able to manage risks they have a familiarity with, but are less able to do so for new attacks they are not aware of. As these types of attacks rely heavily on human reactions, education may reduce their effectiveness, as suggested by survey results detailed in [4]. In [26] it was shown that user actions are based on a combination of conceptual and practical knowledge of phishing attacks, and therefore education must focus on providing both to be effective. Research into the effectiveness of showing warnings to potentially harmful emails has found that, when shown immediately, warnings can reduce the chances a user will interact with emails, though this is not without its own set of challenges in designing an effective user interface in some cases [27]. Other work has looked into factors, such as gender and experience using computers, to relate them to user ability to distinguish between legitimate and illegitimate URLs by experimentally testing social media users [9].

Although not always used for spam, social media bots have a variety of uses and how human users interact with them has been a topic of research. Understanding this is important due to the ease of deploying large numbers of bots on social media platforms and the amount of interactions they can have, information they can gather, and potential malicious uses [28, 29]. Because of this there has been a wide range of methods developed to detect and classify social media accounts as bots [29]. An experiment, detailed in [30], studied how user attributes affect the likelihood of interacting with bots on social media. Some factors were found to be strong enough to be useful in predicting if a user will interact with a bot with some degree of certainty. In [31] it was shown that not only do users interact with bots, often mistaking them for humans, but they can also affect the development of the network and cause human users to emulate the behavior of bots to become more popular themselves. Other work has shown similar results, with human users often being unable to recognize social media bots and interacting with them [32]. Further work has shown that, not only will users interact with social media bots, they can be driven to collaborate on volunteer activities [33].

Work focused on social media spam has shown that due to the sharing focused nature of social media, a variety of new attack methods are used by criminals, including more advanced forms of spam that is not easily possible with email. Additionally, this type of spam can be used for various purposes, such as phishing and spreading malware [3]. Additionally, it has been demonstrated by crawling social media platforms that attackers are able to gather information that can be used to tailor spam messages to individuals, based largely on the information they share about themselves via social media [34].

Some current research has focused on understanding email spam and users understanding and reactions to it. Existing research performed by experiments in laboratory environments has shown that users prior knowledge of phishing and experience of having been phished were correlated with decreased interactions with spam and specifically phishing emails. However, general knowledge of computer security did not show the same effects [4]. Research has also looked into the attributes of emails that users use to make determinations between spam and non-spam messages [25]. Other studies have found that scams (one of the frequent uses of spam) are most effective when they focus on certain psychological aspects, such as using the appearance of authority, maintaining consistency, and against individuals with low self-control [5]. In a survey of user perceptions of the risk and severity of different cyber-crimes, online scams were ranked second in both categories [35].

In several cases researchers have devised experiments where spam or phishing messages were sent to users rather than testing users in a lab environment. One case dealing with students at the United States Military Academy, showed that 80% of students would follow a link in an email appearing to be sent by an officer [7]. Similar work was performed in [6], where different types of messages were sent by email to university students. The results showed that messages that appeared to be time-sensitive, sent from a likable sender, and did not reference prior events that did not actually occur, were the most likely to trick users into providing personal information. In [36] it was shown through a survey and experiment that differences in national culture can change what factors influence how computer users chose to interact or not with phishing emails. Additional experimental work involved sending spoofed information via email, such as [8], where messages were sent that masqueraded as real individuals and specifically selected the recipients of the spoofed messages, such that the recipients already knew the apparent message sender, thus increasing interaction rates, with up to 72% of users falling for the attack. Work explained in [37] showed by experiment that users are more likely to accept connection requests and interact with messages sent by accounts that appear to come from known individuals, despite the sending account being fake. In some cases, 50% of users will fall for these types of attacks. Prior work also found that even non-targeted connection requests are accepted by users. One experiment showed that 75,000 out of 250,000 users did so [38]. Similar work was done in [39], where fake phishing emails were sent to faculty members of a university. Results suggested that embedded education was effective in reducing the likelihood of users interacting with email phishing messages. Results also suggested that "pornographic scams" were the most likely users to fall for. An experiment that tested email recipients several times and explained in [40] showed that embedded education was not always effective after several months. Those results suggested that training may need to be refreshed more often than is desirable in a corporate setting.

Chapter 3

Survey on Social Media and Email Spam

3.1 Introduction

This chapter details the results of a survey on user behavior and knowledge related to email and social media spam, including the similarities and differences. Specifically explored are several factors, such as education, age, access methods, and others, and their correlations with spam identification ability, tendency to interact with spam, and tendency to report spam. Spam identification describes an individual's ability to recognize a message as spam, while interaction is the action of following a link or performing an action that is included in a spam message. Reporting spam is the action of using an email or social media platform's spam reporting feature. To assist with this, the following research questions were formulated:

RQ1 Do spam education, college education, age, or access methods affect an individual's ability to identify spam?

RQ2 Do spam education, college education, age, or access methods affect an individual's likelihood of interacting with spam?

RQ3 Do spam education, college education, age, or access methods affect an individual's likelihood of reporting spam?

In this study, spam is defined as "unsolicited messages or posts, that are delivered electronically". This was the definition provided to respondents when beginning the survey.

Email spam is spam sent by email, while social media spam is spam that is sent or posted on social media platforms. Due to the different natures of email and social media, spam can have different uses, such as how spam on social media can be used for “spamdexing”, i.e. maliciously improving a web site’s search engine rating by increasing the number of other sites that link to it.

The remainder of this chapter is organized as follows. Section 3.2 explains the design and implementation of the survey used to collect the data that are analyzed. The results are detailed in section 3.3. A summary of the results are given in section 3.4. The potential threats to validity are discussed in section 3.5. Concluding remarks are included in section 3.6. Finally, appendix A contains the survey questions and answers used.

3.2 Basic Survey Facts

The survey was conducted in the following manner: define research goals, create survey questions that can be used to collect information related to the goals, and analyze and describe the survey results. Research goals were based around areas that were not well explored in related research, either because of a lack of empirical results or because of lack of study in general. Survey questions and the survey design were based on specific areas that were likely have large effects on users knowledge and actions related to spam.

The final survey was 20 questions long, containing primarily multiple-choice questions. Some questions allowed for respondents to enter additional information if they chose to do so. A list of the questions and provided answers are given in Appendix A. The survey was broken into three sections:

1. Respondent attributes, such as age, educational background, and if they use social media sites.

2. Frequency of encountering social media spam, access methods, and other aspects of social media spam and using social media sites.

3. Frequency of encountering email spam, access methods, and other aspects of email spam and using email.

The survey did not have a specific target population, as it is intended to study a variety of users. One goal was that respondents would not be overwhelmingly college-aged. As the majority of advertisements for the survey were placed around West Virginia University, other methods of gaining respondents were needed. To reach the widest possible audience the following methods were used: snowballing (researchers notifying those they know and asking those persons to spread the word in turn), flyers posted on bulletin boards around campus, including different departments and colleges, and notifications in departmental email newsletters. Individuals employed outside the university were also notified about the survey and asked to spread information about it, which helped increase the number of responses from non-university individuals.

The survey included definitions of “spam” and “social media site”, to insure respondents would have a standard understanding of the questions and reduce variance caused by prior incorrect knowledge or assumptions. The definition of “spam” provided to respondents as part of the survey was “unsolicited messages or posts, that are delivered electronically”. “Social media site” was defined as “website where the primary content is created by users, and which is designed to facilitate communication between individuals or groups. Examples include Facebook, Twitter, Google+, and others”.

There were in total 281 responses to the survey. 256 respondents reported using social media, 25 reported only using email. It was assumed that all respondents would use email, so there was no question added to confirm email use. Due to the small number of users who did not use social media, only the 256 responses of those who used both social media and email are used for the analysis. The mean number of questions answered was 19 of the 20 total.

All responses were anonymous, and there was no option given for respondents to identify themselves. Questions related personally identifying information were limited to age and education major. This was an intentional choice made to increase the likelihood of respondents completing the survey and providing honest answers, as respondents tend to be more truthful when allowed to respond anonymously.

3.3 Results

This section explains the detailed results of the survey. First is an overview of the profile of the respondents, such as age and educational degrees. Following this are sections detailing the results related to social media and email spam, as they pertain to identifying, interacting with, and reporting both forms of spam. Note that some statistical test results given include the correlation coefficient C, and it's maximum and normalized versions, C_{max} and C^*. An explanation of the methods used to generate these are given in Appendix B.

3.3.1 Respondent Profile

In total, 256 respondents reported their age. The mean age was 33, and the median was 27, with a range of 18 to 74. The distribution of ages is shown in Figure 3.1. Note that 18 was the lowest age allowed to complete the survey, due to IRB requirements.

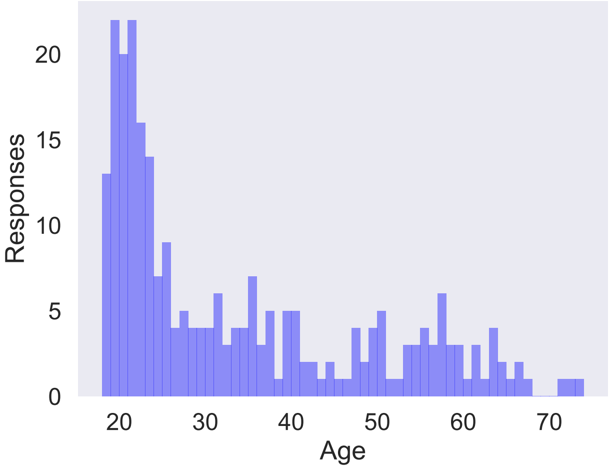


Figure 3.1: Distribution of respondent ages

In total, 253 respondents reported if they had received higher education and if so what their major was. Counts for each major are given in Table 3.1. Computer science/engineering was the most common major, with engineering being the second most common. These two majors account for 37% of responses. STEM (Science, Technology, Engineering, and Mathematics) majors account for 112, or 46%, of the responses. It is advantageous that 141, or 54%, of the responses are non-STEM majors, as it shows the survey contains responses from a wide variety of social media users, not only those who are likely to have received large amounts of training related to computer use. There was no "Other" or free-response selection to this question of the survey.

The most popular social media platform was Facebook, with 240 respondents reporting using it. Twitter and Instagram were the second and third most popular, with 126 and 123 responses, respectively. The results are shown in Figure 3.2. Note that multiple selections were allowed.

Frequencies of using social media and email were given by 253 respondents. Figure 3.3

Table 3.1: Education major

Education major	Count
Computer Science/Engineering	58
Engineering	37
Agriculture	29
Business	23
Education	17
Medicine	8
Biology	7
Communication	7
Arts	5
Journalism	5
Psychology	5
Statistics	5
Chemistry	4
Economics	4
Forestry	4
Public Administration	4
Social Work	3
Sociology	3
Earth Science	2
Geography	2
Literature	2
Political Science	2
Architecture	1
History	1
Philosophy	1
Physics	1
Mathematics	0
No Higher Education	13
Total	253

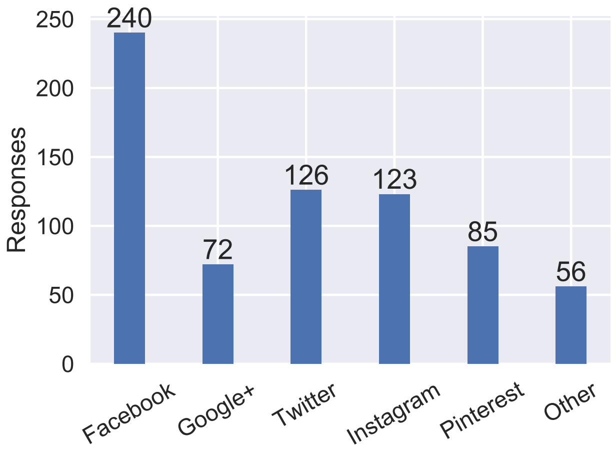


Figure 3.2: Social media platforms used by respondents

shows the frequencies that respondents use social media and email. Most respondents report using both several times per day, though there is a slightly more diverse range for social media. Figures 3.4a and 3.4b show respondent usage frequencies based on type of college major, showing similar usage frequencies. The frequencies were tested for correlation to the type of college major respondents had using a contingency table and Chi-squared test, though the results were not statistically significant.

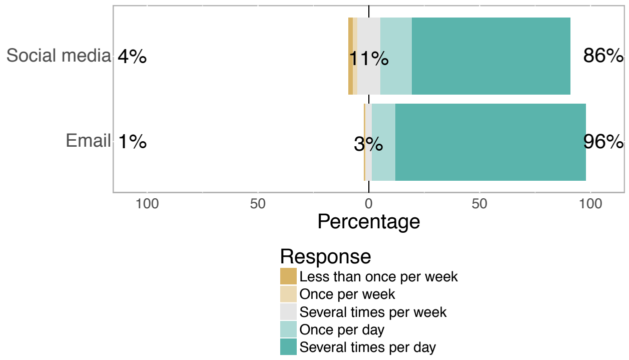
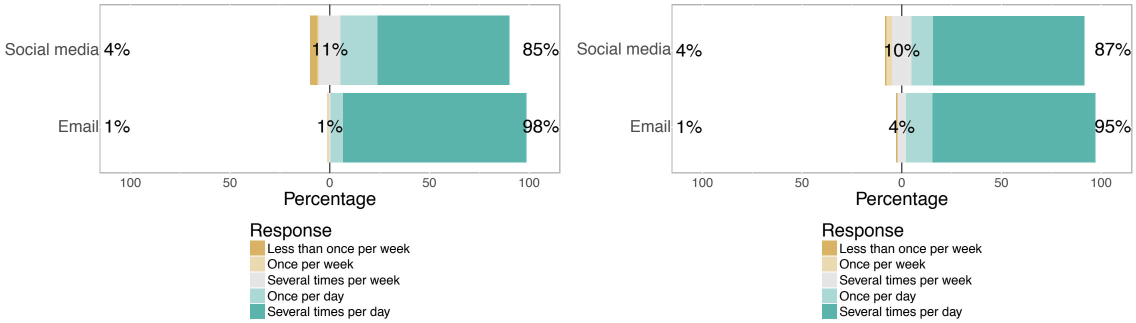


Figure 3.3: How frequently respondents use social media and email

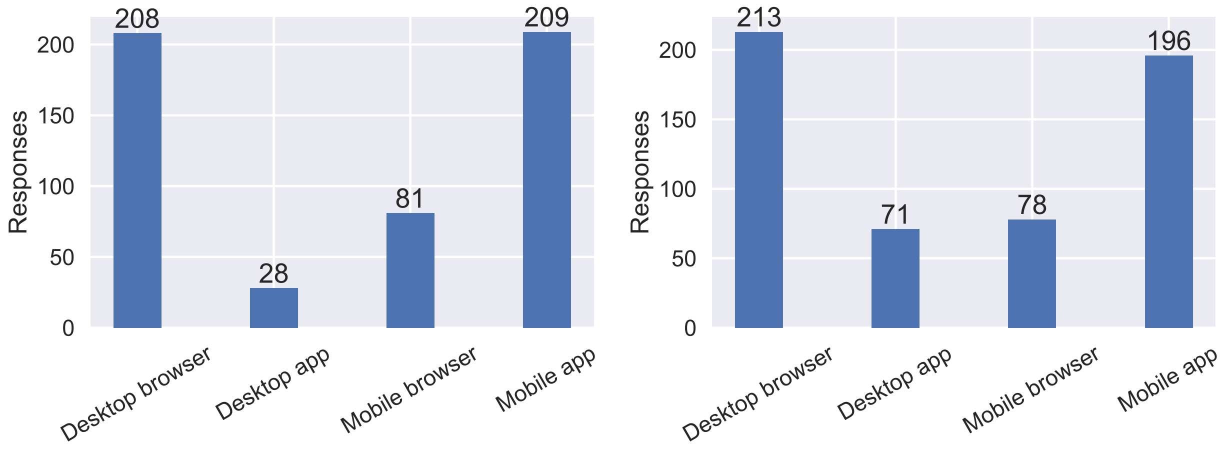


(a) Frequency of social media and email use by STEM majors

(b) Frequency of social media and email use by Non-STEM majors

Figure 3.4: Frequency of social media and email use by college major

Social media access methods were given by 255 respondents and email access methods by 249 respondents. Note that multiple choices were allowed. Figures 3.5a and 3.5b show the frequencies of access methods. Detailed lists of the combinations of access methods are given in Tables 3.2 and 3.3. Respondents had the option to give a non-listed method for social media, though only a single respondent did so (he or she reported using RSS feeds), so that response was excluded from further analysis. Most respondents reported using desktop browsers and mobile apps in both cases. However, the number of desktop app were used more than twice as often for email than social media.



(a) Methods of using social media

(b) Methods of using email

Figure 3.5: Methods of accessing social media and email

Table 3.2: Combinations of social media access methods

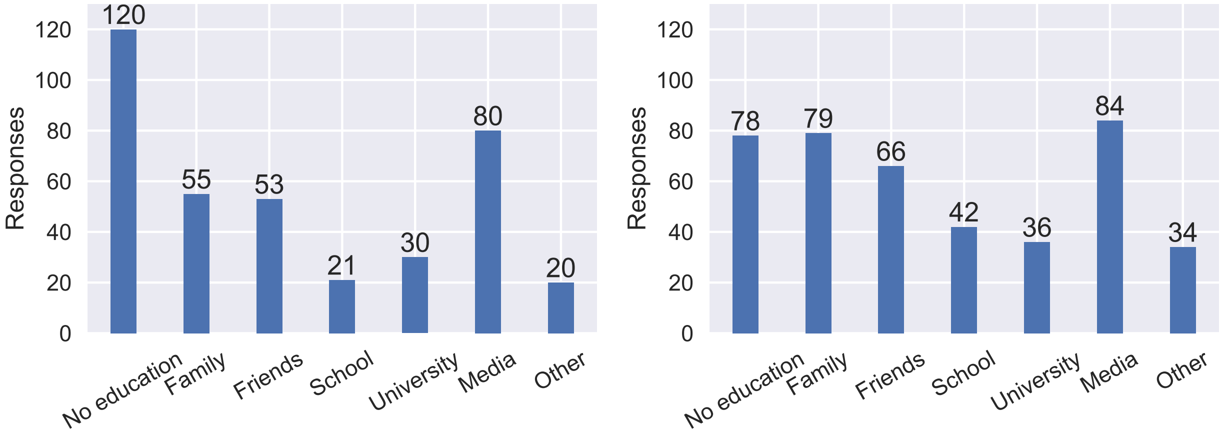
Access methods	Count
Desktop browser, Mobile app	110
Mobile app	37
Desktop browser, Mobile browser, Mobile app	31
Desktop browser, Mobile browser	24
Desktop browser, Desktop app, Mobile browser, Mobile app	20
Desktop browser	19
Desktop app, Mobile app	4
Desktop browser, Desktop app, Mobile app	4
Mobile browser	3
Mobile browser, Mobile app	3
None	1

Figures 3.6a and 3.6b present the sources of spam education reported. Note that some respondents selected more than one source. Interestingly, “no education” was the most common value for social media spam, while “media” was the most common value for email. This could be indicative of most spam education focusing on email spam, possibly as a results of email, and thus email spam and related education, existing for a longer period of time compared to social media spam. These results also show that many users of both email and social media are not receiving education about spam.

250 respondents reported if education about social media and email spam had an impact

Table 3.3: Combinations of email access methods

Access methods	Count
Desktop browser, Mobile app	97
Desktop browser, Mobile browser,	29
Desktop browser, Desktop app, Mobile app	26
Desktop browser, Desktop app, Mobile browser, Mobile app	23
Desktop browser, Mobile browser, Mobile app	18
Desktop browser	16
Desktop app, Mobile app	16
Mobile app	14
None	7
Desktop browser, Desktop app, Mobile browser	2
Desktop browser, Desktop app	2
Mobile browser	2
Desktop app, Mobile browser	2
Mobile browser, Mobile app	2



(a) Sources of social media spam education

(b) Sources of email spam education

Figure 3.6: Sources of spam education

on their behaviors. These results are given in Table 3.4. Most respondents reported that spam education had no impact on them, with this being more pronounced for social media. The correlation between the impact of education on email and social media spam was $\chi^2 = 79.93$, $\phi = 0.57$, $C = 0.49$, $C_{\max} = 0.71$, $C^* = 0.70$, $p = 3.9 \times 10^{-19}$, $n = 250$, based on a contingency table and Pearson's chi-squared test. This shows that respondents who reported spam education having an impact on them were moderately likely to perceive an impact for both types of spam. The sources of education and the ratios of respondents who stated that education had an effect on them compared to the total number of respondents who answered those questions is given in Figure 3.7. As shown in the figure, the distributions of respondents who reported an impact from spam education was similar for education about both social media and email spam.

Table 3.4: Education impact on respondents

Social media	Email education		
education	Impact	No impact	Total
Impact	61	11	72
No impact	40	138	178
Total	101	149	250

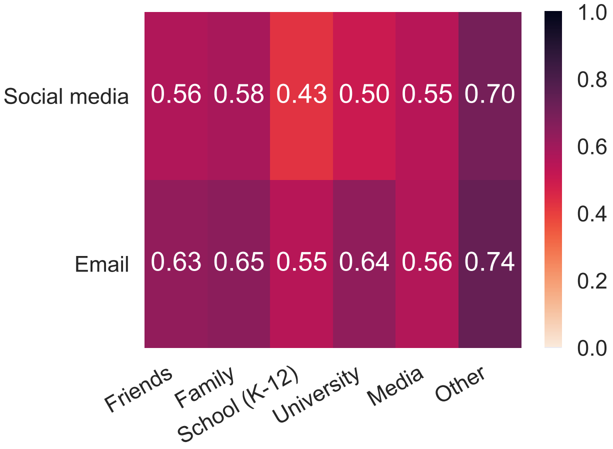


Figure 3.7: Ratios of impact of spam education compared to sources of education

Based on these results, it seems that the "Other" category was the most effective, though due to the small number of responses this is difficult to quantify. Interestingly, the most common "Other" responses referred to the respondents employers. Additionally, it should be noted that as respondents were not asked if education had an effect on their behavior for each source, only if their education in general had an impact. Therefore it is possible that some sources are more effective than others, yet were averaged with other, less effective, sources.

3.3.2 Analysis of Spam Encounters, Identification, and Reporting

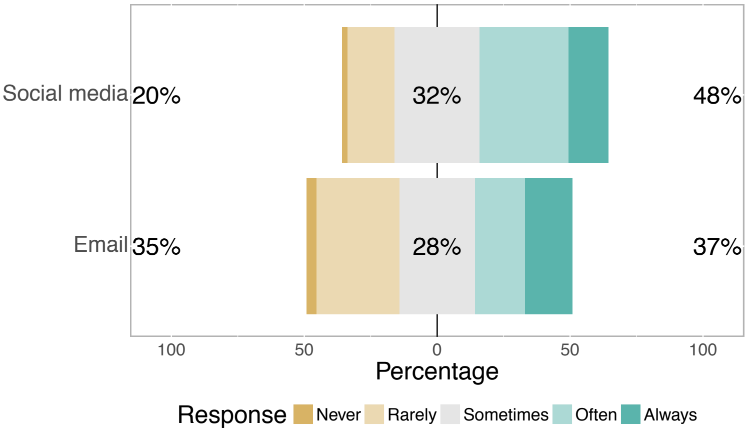


Figure 3.8: Frequencies of encountering spam

Respondents were asked to provide the frequency of how often they encounter spam on social media and email. The results are presented in Figure 3.8. In the figure, the percentages on the left and right sides are the percentage of responses that fall into the two values on either side, i.e. "Never" and "Rarely" or "Often" and "Sometimes". The percentage in the center is for the middle value, i.e. "Sometimes". The correlation between the frequencies of encountering spam on social media and email was $\chi^2 = 45.01$, $C = 0.39$, $C_{\max} = 0.89$, $C^* = 0.44$, $p = 1.4 \times 10^{-4}$, $n = 250$, based on a contingency table and Pearson's chi-squared test, and shows only a moderate correlation. In general respondents reported encountering spam on social media more often than in email. It is possible that this result is due to email spam filters blocking a higher percentage of email spam and respondents being instructed to only count spam that was not automatically sorted into a "spam folder". It is also possible that there exists a greater amount of spam on social media sites, due to additional types of social media spam, such as "spamdexing".

256 responses included their perceived ability to identify social media spam and email spam. Counts for responses are given in Figure 3.9. Most users reported they were able to identify spam, particularly email spam (as email spam has a higher ratio of "Strongly able" to "Somewhat able" than the social media results). Note that because surveys are based on respondent opinions, these results may be biased by respondents overestimating

their abilities.

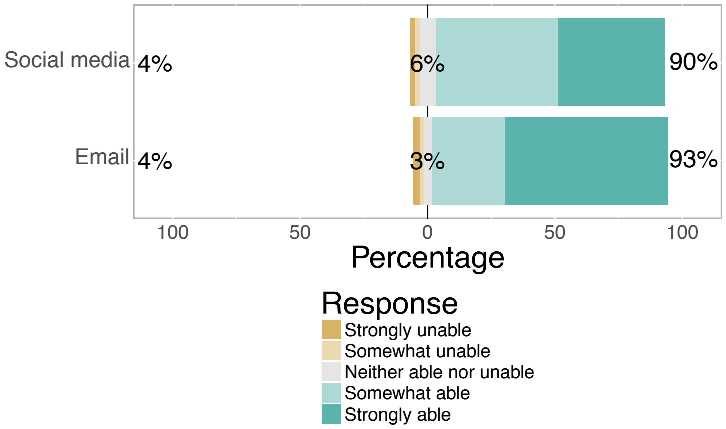


Figure 3.9: Ability to identify social media and email spam

The distribution of the responses related to how often users report interacting with spam messages is given in Figure 3.10. The majority of respondents reported never or rarely interacting with spam. Interestingly, slightly more users reported interacting with social media spam than email spam.

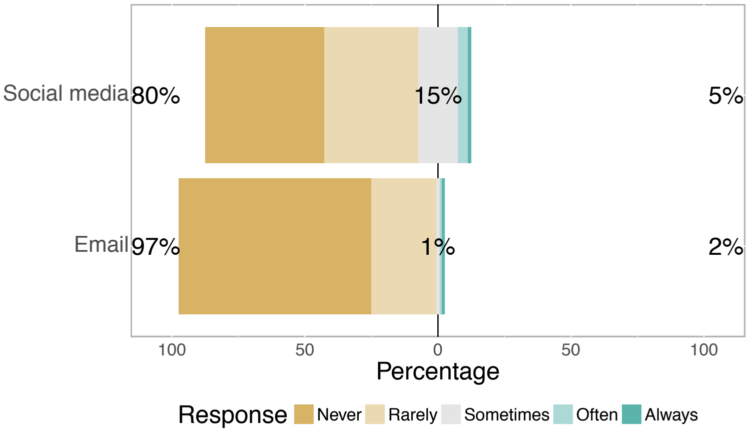


Figure 3.10: Frequency of interacting with spam

Factors affecting ability to identify spam

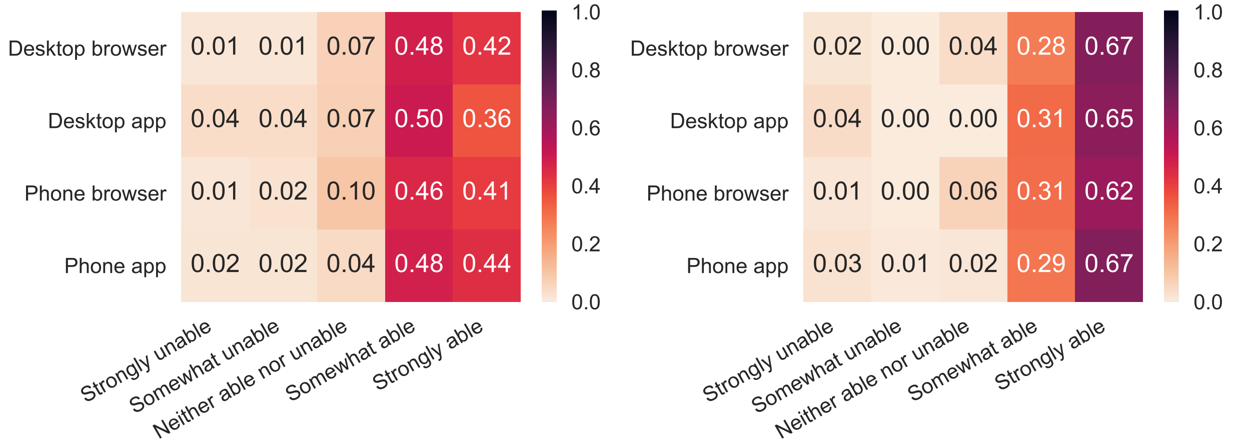
This section focuses on answering RQ1, and includes a detailed breakdown of the statistical tests used to do so and their results.

We compared respondents' ability to identify social media spam and ability to identify email spam. Using a contingency table and Chi-squared test provided results of $\chi^2 = 302.43$, $C = 0.74$, $C_{\max} = 0.89$, $C^* = 0.83$, $p = 8.0 \times 10^{-55}$, $n = 250$. This shows there is a strong correlation between the respondents' abilities to identify social media and email spam.

Statistical tests using contingency tables and chi-squared tests were performed between ability to identify social media and email spam and having received spam education, though the results were not significant at the $p = 0.05$ level. Tests also showed no significant correlation between education having a reported impact and ability to identify spam.

Next studied was the correlation between the ability to identify spam and the type of education a respondent had, STEM or Non-STEM, was also studied. It would be expected that those who studied STEM fields would report that they are more successful in identifying spam, as they likely have a stronger understanding of the potential threats related to spam. Results using contingency tables and chi-squared tests were $\chi^2 = 17.75$, $C = 0.26$, $C_{\max} = 0.80$, $C^* = 0.32$, $p = 1.4 \times 10^{-3}$, $n = 252$ for social media spam, and $\chi^2 = 5.56$, $C = 0.15$, $C_{\max} = 0.80$, $C^* = 0.19$, $p = 2.3 \times 10^{-1}$, $n = 248$ for email spam. These results show a weak correlation between educational major and ability to identify spam for social media, though no significant correlation for email.

The correlation between age and ability to identify spam was also studied, by applying the Spearman test, though the results were not statistically significant.



(a) Ratios of social media spam identification compared to access methods

(b) Ratios of email spam identification compared to access methods

Figure 3.11: Ratios of spam identification compared to access methods

A comparison between different access methods and the spam identification abilities reported for each method are given in Figures 3.11a and 3.11b. When analyzing the ratios of users who report different identification abilities to the number of respondents who use each access method, it can be seen that the access method is not related to identification ability. This could be because access methods do not influence how users identify spam, or that users with different levels of identification ability tend to be distributed the same way between access methods.

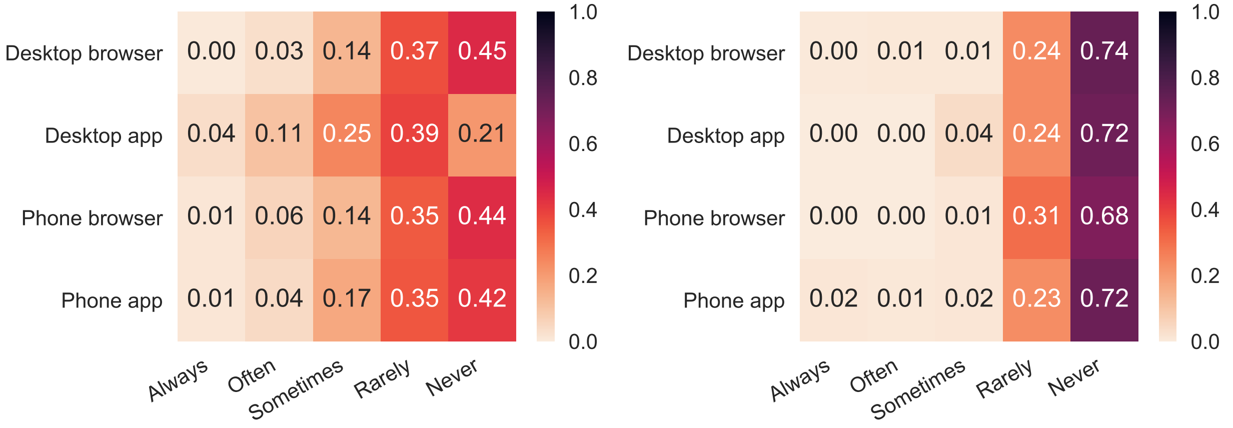
Factors affecting spam interaction

This section focuses on answering RQ2, and includes a detailed breakdown of the statistical tests used to do so and their results.

Using a contingency table and Chi-squared test, a moderate positive correlation was found between respondents' tendencies to interact with social media and email spam. The results were $\chi^2 = 89.97$, $C = 0.51$, $C_{\max} = 0.89$, $C^* = 0.58$, $p = 2.3 \times 10^{-12}$, $n = 250$.

To study the effects of spam education on the respondents' reported frequency of interacting with spam, we used contingency tables and chi-squared tests. Education on spam was found to be statistically insignificant in reducing social media and email spam interaction rates at the $p = 0.05$ level. This suggests that overall current education about spam is unlikely to change user behavior. Surprisingly, no correlation was found between respondents reporting spam education having an impact on them and their tendencies to interact with spam.

No statistically significant correlation was found between college major type and tendency to interact with spam when using contingency tables and chi-squared tests.



(a) Ratios of social media spam interaction compared to access methods

(b) Ratios of email spam interaction compared to access methods

Figure 3.12: Ratios of spam interaction compared to access methods

Also studied was the correlation between age and ability to identify spam, by applying

Spearman, though the results were not statistically significant.

Figures 3.12a and 3.12b show the ratios of different access methods and spam interaction rates. Note that these survey questions asked about spam interaction rates in general, not specifically for certain access methods. Social media rations are similar, except for desktop applications, where the rates increase somewhat. Email access methods have approximately the same ratios of interactions between method.

Factors affecting reporting spam

This section focuses on answering RQ3, and includes a detailed breakdown of the statistical tests used to do so and their results.

Table 3.5: Report Spam

Social media	Email report		
report	Yes	No	Total
Yes	78	24	102
No	101	47	148
Total	125	125	250

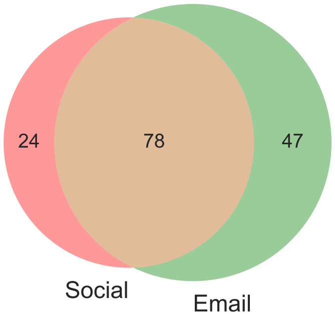


Figure 3.13: Numbers of respondents who report spam on social media, email, or both

In total, 250 respondents listed if they report spam or not. Based on the results in Table 3.5, it appears that an equal number of users report or do not report email spam. However, more users do not report social media spam than do. A graphical comparison between numbers of respondents who report social media and email spam is shown in Figure 3.13. The correlation between the reporting email and social media spam has $\chi^2 = 46.51$, $\phi = 0.43$, $C = 0.40$, $C_{\max} = 0.71$, $C^* = 0.56$, $p = 9.1 \times 10^{-12}$, $n = 250$, using a contingency table and Pearson's chi-squared test for correlation. This shows a moderate and statistically significant correlation between spam reporting behaviors for social media and email spam.

Spam education on its own does not seem to effect respondents' tendencies to report spam, and we find that there is unlikely to be a correlation. Contingency tables and Pearson's chi-squared test were used to compute the correlation between social media spam education and reporting social media spam, though no significant correlation was found at the $p = 0.05$ level. Additionally, there was no correlation found between respondents who report social media or email spam and stated that education had an impact on their behaviors. This seems to suggest that educating users about spam will not increase the likelihood they will report spam.

The type of major (STEM or non-STEM) that respondents reported was not found to be correlated to tendencies to report spam, using contingency tables and chi-squared tests, though no significant correlation was found for either social media or email.

Also studied was the correlation between age and tendency to report spam, by applying Spearman, though the results were not statistically significant.

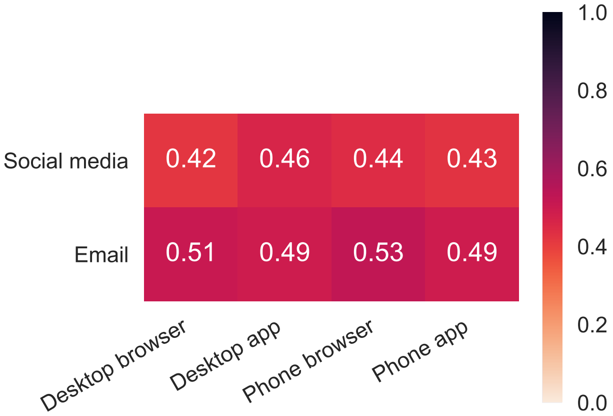


Figure 3.14: Ratios of respondents who report spam on social media and email by access method

Figure 3.14 shows the ratios of respondents who report spam compared to access methods they use, for both social media and email. As can be seen the ratios are similar across access methods, suggesting that there is no correlation between access methods and spam reporting.

3.4 Summary of the Results & Discussion

A summary of the results, organized by research question, and their pertinent sections are provided in Table 3.6.

An interesting result was the high number of respondents who reported never having received spam education. This is surprising given the frequency and threats posed by spam. As the majority of respondents reported having a university degree, this also suggests that universities are likely not educating students.

Surprisingly, many items that were expected to be correlated to spam identification ability, and tendencies to report or interact with spam, showed little to no correlation. For example, education about spam was not shown to have a correlation with reduced spam interactions. This was more surprising because as even respondents believing education had caused them to change their behaviors did not show significant differences in their behavior from those who did not. In addition, the type of college education that a respondent has is only weakly correlated with ability to identify both types of spam, and had no correlation with interaction or reporting rates. It is possible that individuals with a STEM education have been exposed to more education related to identifying spam. The lack of correlation for interaction or reporting rates could be caused either by respondents not internalizing their knowledge in a way that reduces their tendency to interact with or report spam, or over-confidence in their ability to not be harmed by spam.

Results did show that with regards to identifying, reporting, and interacting with spam, respondent actions are positively correlated between social media and email spam. This suggests that users have similar behaviors regarding spam on social media and email. However, the only other factors found to correlate with spam related behaviors were the type of college education a user had and, in the case of email spam, prior education about spam. In both of these cases the correlations were found to be weak. This suggests that the determining factors for respondent behaviors are not age or the method a respondent uses to access social media platforms or email.

That age was not shown to correlate with behaviors suggests that the belief in younger computer users being more knowledgable is a misnomer. Additionally, the methods respondents used to access social media or email did not correlate with the studied behaviors. This suggests that users will typically respond to spam the same way on both mobile and desktop experiences.

The correlations between behaviors across social media and email suggest that by influencing user behavior for either social media or email will have a positive influence on behavior on the other. However, as the correlations were not always strong, it is still important to focus on both types of spam when planing education. Additionally, respondents reported interacting with social media spam more often than email spam. This suggests that social media users may believe it to be less dangerous than email spam, and therefore work should be done to help users understand the true risks these behaviors pose.

3.5 Threats to Validity

Although the survey and analysis were designed to eliminate or mitigate threats to validity, there are some aspects of the study which could reduce the reliability or generalizability of the results. Those threats identified are included in this section. There are four types of threats, construct, internal, conclusion, and external, which are explained and enumerated here.

Construct threats are potential problems where what was measured is not what needed to be measured. In this thesis, the factors chosen to be studied may in fact not have the largest effects on respondent behaviors, while an unknown factor may be more important. For example, gender may be a factor in some behaviors and has been studied in related work.

Internal threats are unknown factors influencing results. As the questions rely on respondents self-reporting, they may have provided incorrect information themselves due to either misunderstanding, overconfidence, or other factors. To reduce the likelihood of this occurring, respondents were allowed and assured anonymity, which tends to lead to more honest responses in surveys.

Conclusion threats relate to if the conclusions have been correctly justified. To mitigate this type of threat, the data were provided as visualizations and appropriate statistical tests were used. In this thesis, factors were studied on their own and it is possible that the interactions between factors may have a greater effect on respondent behavior than the factors individually. However, the factors were chosen in part to minimize the likelihood that interactions would be significant.

External threats are those that affect the generalizability of the results. One example is that user actions may change as new websites, education, and other factors change. However, this is always a possible threat for any study of human behavior. More specifically to

Table 3.6: Summary of the results

Question number	Question statement	Result
RQ1	Do spam education, college education, age, or access methods affect an individual's ability to identify spam?	• Social media / email: Strong correlation between identification ability between social media and email spam. Results are $\chi^2 = 302.43$, $C = 0.74$, $C_{\max} = 0.89$, $C^* = 0.83$, $p = 8.0 \times 10^{-55}$, $n = 250$. • Spam education: No significant result. • College education: Significance for social media. Results were $\chi^2 = 17.75$, $C = 0.26$, $C^* = 0.32$, $p = 1.4 \times 10^{-3}$, $n = 252$ for social media spam, and $\chi^2 = 5.56$, $C = 0.15$, $C^* = 0.19$, $p = 2.3 \times 10^{-1}$, $n = 248$ for email spam. • Age: No statistically significant result. • Access methods: No difference between access methods. See Figures 3.11a and 3.11b.
RQ2	Do spam education, college education, age, or access methods affect an individual's likelihood of interacting with spam?	• Social media / email: Moderate correlation between interacting with social media and email spam. Results are $\chi^2 = 89.97$, $C = 0.51$, $C_{\max} = 0.89$, $C^* = 0.58$, $p = 2.5 \times 10^{-12}$, $n = 250$. • Spam education: No statistically significant result. • College education: No statistically significant result. • Age: No statistically significant result. • Access methods: No differences, except for slightly higher interaction rate for social media spam when using a desktop app. See Figures 3.12a and 3.12b.
RQ3	Do spam education, college education, age, or access methods affect an individual's likelihood of reporting spam?	• Social media / email: Moderate correlation between reporting social media and email spam. Results are $\chi^2 = 46.51$, $\phi = 0.43$, $C = 0.40$, $C^* = 0.56$, $p = 9.1 \times 10^{-12}$, $n = 250$. • Spam education: No statistically significant result. • College education: No statistically significant result. • Age: No statistically significant result. • Access methods: No difference between access methods. See Figure 3.14.

this study, a large number of responses were from young college students and almost all respondents reported having a higher education degree, which could effect the results.

3.6 Conclusion

In conclusion, provided in this chapter is a detailed analysis of user behaviors related to social media and email spam. Additionally, comparisons and correlations have been made between these behaviors and various related factors. Furthermore, cases where factors that did not show statistical correlations have also been provided, which allows for sorting between factors that warrant further study from those that are unlikely to require it. It is hoped that these findings prove useful as both a basis for further research, as well as useful in helping to better understand the user behavior related to spam online.

Chapter 4

Social Media Spam Experiment

4.1 Introduction

This chapter presents the design of experiment approach we used to study factors that affect if social media platform users interact with spam. Both details of how the system was created, the experiment executed, and an analysis of the results are included. To assist with this explanation, the following research questions are used:

RQ1 Do spam interaction rates differ by social media platform?

RQ2 Does matching the sending profile to the type of spam it sends affect interaction rate?

RQ3 Does tailoring spam to recipient interests affect interaction rate?

RQ4 Does the content of spam messages affect interaction rate?

The remainder of this thesis is organized in the following manner. Section 4.2 explains the design of the experiment and the creation and design of the system used to administer it. The results of the experiment and a discussion of the results are presented in Section 4.3. Threats to the validity are listed in Section 4.4 and concluding remarks and potential areas of future work are presented in Section 4.5.

4.2 Experiment

4.2.1 Design of Experiment: Factors and Levels

To improve the utility of the experiment results, we chose to use a multi-factor design. This design allows using statistical tests to quantify the effects of each factor. Four factors were studied: the social media platform, S, the interests of the profile (referred to as the "persona" in this thesis) the message was sent from, P, the method of selecting the message to send, referred to as the message selection type, M, and the content type of the message, m. These factors were chosen due the expectation that they are key determinators in users' choices whether or not to interact with social media spam. A list of the factors and their levels is as follows:

$$\begin{aligned}
S &= \{\text{``Facebook''}, \text{``LinkedIn''}, \text{``Twitter''}\} \\
P &= \{\text{``Generic''}, \text{``Adaptive''}\} \\
M &= \{\text{``Generic''}, \text{``Adaptive''}\} \\
m &= \{\text{``Generic''}, \text{``Ad''}, \text{``News''}\}
\end{aligned}$$

With this design of experiment, the final system contains 36 levels (the total number of combinations of each factor level, i.e. $3 \times 2 \times 2 \times 3$).

In this study, for factor P, "adaptive" refers to when the system uses a persona that had interests related to the content of the messages sent from it, and "generic" for when the persona had non-specific interests listed. For factor M, "adaptive" refers to when the content of the message was based on the interests of the receiving social media user, and "generic" for when the message content was not matched to the recipient interests (i.e. the message was chosen at random). For factor m, "generic" refers to a selection of message contents that are not related to either advertisements or news.

These factors and levels were chosen to explore aspects that have little to no prior study. The social media platforms were chosen based on their popularity and the fact that they have distinct uses and goals. The choice of whether or not to match personas with their messages was based on a theory that doing so would make the spamming appear more like a human than an automated system. Doing so would also make the sender seem more relatable to the receiver, similar to the "likability" aspect covered in [6]. Tailoring the persona interests and message selection are also unique abilities of social media spam compared to email spam, as the use of user profile data listed on social media sites is much more comprehensive and easy for spammers to obtain, compared to information relating to an email account user. The choices of message content were based on an analysis of real social media spam and the typical content of those messages.

4.2.2 Overview of System Design & Experimental System

We first designed and constructed the system used. A graphical overview of the system is shown in Figure 4.1. In the figure solid lines represent information flow over a social media platform, and the dashed line represents information flow outside a platform. The system functions can be broken down into the following steps:

1. Creating a selection of messages based on types of real spam.
2. Creating mock user accounts on three popular social media platforms.
3. Selecting message recipients.
4. Sending mock spam messages to the selected recipients.
5. Collecting social media users' responses
6. Performing an analysis on results.

A key aspect in designing the experiment was that it handle its tasks in an ethical manner. As this study involved interactions with individuals, Institutional Review Board (IRB) approval was required before running the experiment. The study obtained a waver for obtaining informed consent, which made it possible to send the messages without notifying the recipients before the messages were sent. Because the waver was required, the study was classified as expedited by the IRB, but was not except from exempt from review. As it was considered to involve minimal to no risk to individuals it did not require full board approval. The waiver of informed consent was required to prevent individuals from changing their behavior to be more careful, which could occur if they knew some spam they received was part of the experiment. An additional requirement was that participant personally identifiable information not be collected and stored, which was accomplished by not asking for information and using the encoded URL for each message of a factor level combination,

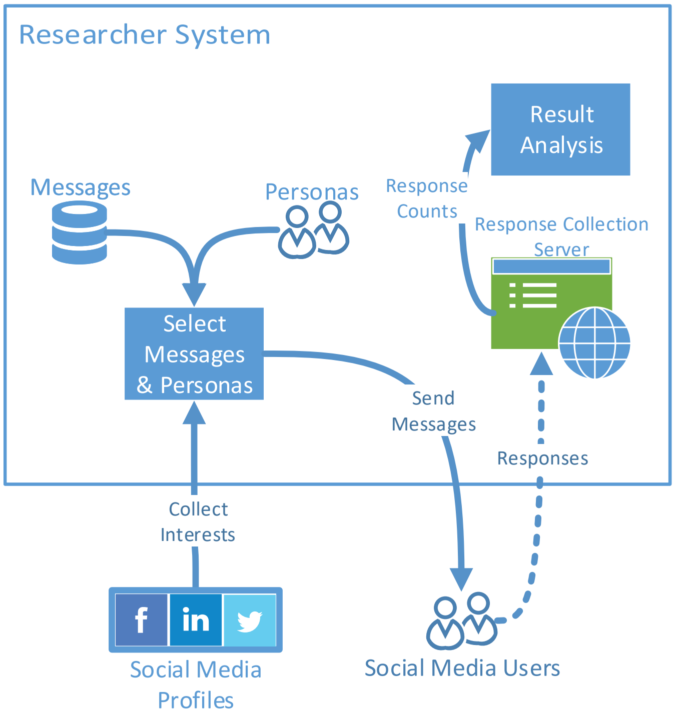


Figure 4.1: Conceptual overview of system design

so counts could not be traced back to individual users. This is described in more detail in Section 4.2.3.

A detailed explanation of the system is given in the remainder of this section. This includes details on the creation of messages, system architecture, and technical details of the implementation. The messages were sent over a period of approximately one and a half months. Interactions with messages were counted for a further two weeks, as it was believed that if users had not interacted with a message by that time they would be unlikely to do so in the future.

4.2.3 Implementation & Running the Experiment

Message creation

The design of the messages was based on an analysis of Twitter spam from a previously available dataset [41]. The analysis was primarily focused on finding common types of spam messages, focused primarily on their content. The techniques used were similar to previous research, that had been applied to web spam [42]. This choice was made so the current research is a further development of prior work, and because generalities were applicable to the types of spam seen. No datasets were available for Facebook or LinkedIn, so message design was generalized from the Twitter data. This also allowed for fair comparisons of the results across platforms.

To create the classes, 1,000 tweets were randomly selected from the original dataset and manually classified them into categories. Counts were made of the number of tweets in each category, with the most common categories being the used for the experiment. Modifications were made to the final classes to simplify the experiment design, such as combining related classes (e.g. "product advertisement", "business advertisement", "service advertisement", and several others were combined into a single "advertisement" class). The messages were then sorted into these new, combined classes. Generalizations were made about these combined categories, based on commonalities between messages in each class. The classes of messages used were:

- Advertisement: Messages related to advertisements for goods, such as golf equipment, music, et. al. E.g. "BEST golf gear here [URL]" or "BEST FREE Music Streaming [URL]".
- News: Headlines related to current politics, sports, or economic news. Note that all headlines were fictitious and designed to appear interesting to users. E.g "Drastic changes in global futures markets [URL]" or "2017 Grammy information leaks [URL]".
- Generic: Generic phrases, strings of random words, or strings of random characters. E.g "Good luck today! [URL]" or "Village did removed enjoyed [URL]". This class served as the control.

A list of all messages used for this experiment are given in Appendix D. Messages of the "news" class were updated to relate to events recent at the time the messages were sent, as that was considered necessary for their effectiveness. All messages created were in English, with the exception of the messages containing only random letters.

To avoid possible detection by spam filters, the exact messages included in the original Twitter dataset were not used for sending to participants. Instead, we manually created new messages based on commonalities in each class. Some of these messages did not contain a URL, but were miscellaneous quotes. These messages were only sent to collector accounts created to test and monitor the system. The no-URL messages were used as both a method of preventing the personas from being detected as bots as well as a method of insuring that the personas were sending messages. This is further explained in section 4.2.3.

Persona account creation

The implementation of the message sending accounts (the personas) and specific methods of selection selecting recipients and massages to be sent were developed by another team member. Overviews of these aspects of the system are given here for completeness.

Persona accounts were created for each social media platform. Three personas per class of message content were created for each platform. Personas had information added to their profile information, such as profile pictures and descriptions, so as appear more human-like. The information added to the persona profiles was same across platforms. The same set of interests, such as brands or promotional groups, news groups, et. al., were used for each type of persona and across platforms. These interests were matched to the message content class that the personas sent. Personas also liked or followed ten or more popular accounts, though were not set to follow the other personas. On LinkedIn, before sending a message, each persona would follow the message recipient as that is a requirement to send a private message on that platform. On Facebook and Twitter this was not a requirement, and so therefore not done.

Selection of recipients

The experiment was designed to mimic the sending of real spam messages, to elicit realistic responses that individuals would have from actual spam. To do this, user accounts were selected to receive the mock spam messages. Accounts were chosen at random from those that had desired interests (for adaptively selected messages) or interests unrelated to the message contents being sent (for generically selected messages). Additionally, only accounts that were set to use English as their language were selected, as all messages were written in English. Selection was intended to have a high degree of certainty that the accounts were used by human users and not automated bots themselves.

For Twitter, the selections were made by using Twitter's Streaming API, as it provides a random sample of current user activity, and therefore the accounts are known to be active. As the messages to be sent were all in English, the API call was set to only be streamed tweets with English as the language code. Selection by interests for Twitter accounts was handled by using the Streaming API's keyword filter feature, and providing a list of keywords related to the message classes. Verification of the accounts being non-bots was done using BotOrNot (now named Botometer) [43], which used a collection of heuristics to determine if an account is likely to be used by a human or bot.

Another project member performed the design of the recipient selection methods for Facebook and LinkedIn, though the selection of recipients for those platforms was still randomly performed. On Facebook, the selection recipient interests was based on membership in a set of fan pages for different news organizations, promotional pages, or when specific interests were not needed, generic groups such as social clubs or food related groups. LinkedIn used a similar selection method, using LinkedIn groups in place of fan pages. To select LinkedIn users for cases where their interests were not needed, they were randomly selected from users with common stop-words in their titles (e.g., "at", "for", "in", et. al.). In both cases only English language groups were used, which ensured that the recipients would be able to understand the sent messages. Selection of English language users is extremely important on LinkedIn, as approximately 70% of its users live outside the United States [44]. Although on Twitter, the selection was based on recent activity (as that what the Streaming API provides), activeness was not checked on the other platforms. Additionally, as there were not services equivalent to BotOrNot for Facebook and LinkedIn, there was no way to ensure that selected recipients were human controlled. However, these risks are considered to be mitigated due to the nature of those platforms. On Facebook by the selections being based on group membership and group administrators attempt to remove bots from the groups. On LinkedIn there is typically a low percentage of bots on the platform.

Sending messages

Sending messages was handled by the persona profiles we created to match the interests referenced by the messages the accounts sent. The exception was that no profile only sent "No-URL" messages, as those messages were not sent to recipient accounts, but instead used sent only to researcher controlled accounts. The profiles were controlled by software running on experimental setup computers.

Several methods were used to make the persona accounts appear more like human controlled accounts. This was done both to prevent the social media platforms from disabling access to the persona accounts and to make the persona accounts appear like other humans to the message recipients. Each profile was accessed via a unique IP address, to ensure the accounts appeared as normal user accounts to the social media platforms. Also, each persona sent messages at random intervals of time and varied the content sent to avoid sending many repeated messages at once. Further, personas were set to limit the number of messages sent in short periods of time. As a monitoring system, the profiles sent messages to researcher controlled accounts at regular intervals, to verify that the sending profiles were successfully sending messages. These accounts were only used as a monitoring system and did not send messages. Note that all messages were sent as private messages, not via any platform's public messaging method. This was done to maintain the privacy of any users who received a message.

The messages were selected for sending based on the profile of the persona sending the message, P, the interests of the receiving profile, M, and the content class of the message, m. Personas were designed to be either adaptive, which had interests related to the content of messages they sent, or generic, which sent messages that were not related to the persona interests. Messages were sent either adaptively, with the content class of message matched to the receiver's interests, such as "news" class messages sent to profiles that mentioned a news network, or generically, without checking the receiver's interests. Messages were sent such that an equal number each combination of levels were sent. Combined with the three social media sites, S, this gave 36 different combinations, with each combination being having messages sent 100 times, for a total of 3,600 messages. This number was selected to be large enough to perform statistical tests, while remaining low enough to be able to send the messages in a manageable amount of time.

Only a single message was sent to any given recipient. This was to insure that recipients could be counted for at most a single message class. Although it is theoretically possible that an individual recipient may have been sent messages on multiple social media platforms, this is considered to be unlikely due to the number of messages sent relative to the total number of users of the social media platforms.

Collecting message interactions

Because the study goal was to learn the social media platform users interactions with spam, each message contained a URL to a response collection server. If a user followed the link, it would present the user with a webpage explaining that the messages was part of a study and an internal counter on the server hosting it would increment. This method of counting user interactions was chosen for several reasons. First, it does not require injecting tracking code into the messages that runs when a user views the message. This was important both to simplify the technical implementation and to maintain the ethical considerations of the experiment and to adhere to IRB approval requirements. Due to requirements of the system design, personal information could not be stored, so to count the interactions each message contained a URL that encoded the factor levels. URLs were encoded in the form `http://domain.name/abcd` where `a` is the platform used to send the message, `b` is the type of the sending profile, `c` is the message selection type, and `d` is the message content type. For example, a message sent on Twitter, generic sending profile, adaptively selected message, with news content, would be encoded as `http://domain.name/tgan`. The server maintained counts of how many times participants viewed each URL, and thus how many responses were received for each of the 36 combinations of factor levels.

One of the IRB requirements was that individuals had to be informed of the study and given an opportunity to opt-out. This was accomplished by showing a debrief statement that explained the study to anyone who followed a URL included in the message. The text of the debrief page is included in Appendix C. This page also contained a button allowing individuals to opt-out. If an individual opted-out an opt-out counter was incremented, and these were subtracted from the total count for each message class and are not included in the analysis.

Data preprocessing

Any public web page receives traffic from automated systems, such as search engine crawlers indexing the page and other non-malicious systems, and from malicious systems such as scanners searching for common vulnerabilities. As the system for counting the number of times a URL was accessed, such automated systems accessing the website would be counted. These extraneous accesses must to be removed to have accurate counts of how many times humans interacted with the mock spam. The methods used to filter out automated accesses are explained in this section.

While testing the functionality of the system, accesses to the response collection server were monitored, to study what types of automated access attempts would be made to it. Both legitimate and malicious automated accesses were found. Legitimate accesses were social media platforms following links that were sent via their services, likely as part of malware detection systems, e.g. scanning sent links to provide some assurance that their users are not spreading malware. Legitimate web crawlers also accessed the collection server, as part of their purpose of analyzing websites for inclusion in search engine results. In addition, methods were required to detect accesses from social media bot accounts that may have been mistakenly sent messages, as bot accounts are allowed by the Twitter API and terms of service and also exist on Facebook and LinkedIn. To avoid mistakingly counting these accesses as humans during the experiment, blacklists user-agent strings were implemented. A "robots.txt" file was used on the server for cases where a crawler was not included in a blacklist, as ethical crawlers are expected to respect it.

Additional measures needed to be taken to account for malicious accesses, as they vary more widely in their intended purposes and sometimes attempt to obfuscate their actions. For these cases a preselected blacklist of known malicious IP addresses and user-agent strings were automatically blocked and not counted. This blacklist also included user-agent strings that were not necessarily malicious programs, but were not standard web browsers, e.g. tools such as the program "curl". Attempts to access URLs that were not part of the experiment were also not counted, and the source IP addresses or user-agent strings were marked to not be counted in the event they accessed a used URL. User-agents or IPs that accessed the response collection server a significant number of times at the same URL were also discarded. Attempts to access the website multiple times in a short period of time were also discarded, and those IP addresses and user-agents were blocked. A random selection of responses were also analyzed to validate the quality of the final data.

4.3 Results

This section explains the results of the experiment and statistical methods used. First, the result values are explained for each factor level, followed by an empirical analysis using logistic regression. Finally, a discussion of the results is explained and a summary of the key findings is presented.

4.3.1 Empirical Results

In total 3,600 messages were sent, 1,200 per network. Each of the 36 combinations of factor levels had 100 messages sent. Only responses that were not opted-out from are used for the analysis. Note that very few respondents opted-out, with no more than 5 for any given combination of factor levels and most combinations had no optouts, so this is unlikely to significantly alter the results. In the following tables, "Total messages" refers to the number of messages used for the analysis. Table 4.1 shows the number of interactions for each combination of factor levels. However, in this form it is difficult to make conclusions so the remainder of this section provides more easily analyzed forms of the results.

An overview of counts by platform are given in Table 4.2. Of the total counts by platform, Twitter had the highest interaction rate at 20.9% and LinkedIn had the second at 14.3%. Facebook at 4.3% was the lowest with less than one-quarter of the rate of Twitter and approximately one-third the rate of LinkedIn. The much higher interaction percentage for LinkedIn compared to Facebook was unexpected. When comparing the interaction rates between platforms, it is important to note the differences between them. One potential reason for this result may be LinkedIn's more professional nature, which causes users to place greater trust in what they view on LinkedIn, despite messages coming from unknown sources. Meanwhile, Twitter is commonly associated with openness and often used to disseminate information to a wide audience, which can partly explain the high interaction rates for messages sent on that platform. Additionally, users of each platform likely have different expectations regarding the number of bot accounts they may interact with. For example, an estimated 15% of Twitter accounts are bots [45], while this is true of only 3% of accounts on Facebook [46]. Because of these differences, message recipients may be more comfortable interacting with messages sent from accounts they suspect are not humans on Twitter than they are on the other platforms. Response rates may also be partly due to age differences of users on each platform. 65% of Facebook users are age 35 or older [47] and almost 80% of LinkedIn users are over 30 [48], while Twitter is most popular with users aged 18 to 29 [47].

Results by persona type are shown in Table 4.3 and results by message selection type are shown in Table 4.4. Comparing the persona types, both the adaptive and generic personas had similar interaction rates, with generic personas having slightly higher rates than adaptive ones. Similar results were found for message types, where again generic messages had slightly higher interaction rates than their adaptive counterparts. Interestingly, tailoring

Table 4.1: Results for each cell

Network	Persona	Message selection	Message content	Interactions
Facebook	Generic	Generic	Generic	8
			Ad	1
			News	2
		Adaptive	Generic	7
			Ad	3
			News	6
	Adaptive	Generic	Generic	2
			Ad	1
			News	4
		Adaptive	Generic	13
			Ad	1
			News	3
LinkedIn	Generic	Generic	Generic	36
			Ad	5
			News	4
		Adaptive	Generic	18
			Ad	10
			News	12
	Adaptive	Generic	Generic	10
			Ad	12
			News	18
		Adaptive	Generic	26
			Ad	15
			News	3
Twitter	Generic	Generic	Generic	31
			Ad	0
			News	53
		Adaptive	Generic	8
			Ad	4
			News	49
	Adaptive	Generic	Generic	11
			Ad	2
			News	39
		Adaptive	Generic	9
			Ad	1
			News	43

either the sending profile or the selection of the message was not correlated with an increase in interactions, and in fact showed a slight decrease. In both cases the interaction rate decreased, although only slightly. This would seem to indicate that social media users do not closely examine profiles that send them messages, even when they are not known individuals. Additionally, it appears that users interacted with messages that were not related to their personal interests. It is possible, however, that if messages had been tailored specifically for individuals those messages may have generated more interactions.

Table 4.2: Summary results by network

Platform	Interactions	Total messages	Interaction %
Facebook	55	1196	4.3%
LinkedIn	189	1180	14.3%
Twitter	251	1197	20.9%

Table 4.3: Summary results by persona

Persona	Interactions	Total messages	Interaction %
Generic	271	1784	14.4%
Adaptive	224	1789	11.9%

Table 4.4: Summary results by message selection type

Message selection	Interactions	Total messages	Interaction %
Generic	252	1785	13.4%
Adaptive	243	1788	12.9%

Table 4.5: Summary results by message content type

Message content	Interactions	Total messages	Interaction %
Generic	196	1183	15.1%
Ad	58	1195	4.6%
News	241	1195	19.8%

Results by message content type are shown in Table 4.5. The “news” messages had the highest interaction rate at 19.8%, while “ad” messages had the lowest of only 4.6%. Generic messages were close to news messages with a 15.1% rate. The low results for “ad” messages may be due to long running education about the spam as an advertising medium used for fake or illegal products. In contrast, education may not focus on the potential dangers of other types of spam. However, the relatively high interaction rate for “generic” messages is surprising, given that many of the messages are nonsensical phrases.

Results for message content types broken down by social media platform are given in Table 4.6. When interaction rates are broken down by message content classes for each platform, some noticeable differences can be seen. On Twitter the “news” messages had the highest rates, with almost one-half of messages sent receiving an interaction. While on Facebook and LinkedIn, “generic” messages had the highest rates. One explanation for these results may be the differing reasons that individuals use each platform. Twitter has grown in popularity as a means of receiving news stories, while this may have been a less common use case of the other platforms at the time of the study.

Table 4.6: Results for message content type by platform

Network	Message content	Interactions	Total messages	Interaction %
Facebook	Generic	30	397	7.6%
	Ad	6	400	1.5%
	News	15	399	3.8%
LinkedIn	Generic	90	386	23.3%
	Ad	42	397	10.5%
	News	37	397	9.3%
Twitter	Generic	59	400	14.8%
	Ad	7	398	1.8%
	News	184	399	46.1%

4.3.2 Analysis Based on Logistic Regression

Due to the factorial design of experiments used, as explained in section 4.2.1, it is possible to analyze the effects of each individual factor, as well their interactions. Most commonly this is done using ANOVA F statistics [49]. However, as the response variable is binomial, either no interaction or an interaction of the social media user with the spam message, the ANOVA F statistic cannot be used. Instead we used a logistic regression, which allows us to explore and quantify the influences of the factors on the response variable.

Overview of logistic regression

A brief overview of the logistic regression method is provided in the following.

The general form of the logistic function is given as:

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{4.1}$$

where z represents the dependent variable, X_i represents the value of a given independent variable, α is the intercept, and β_i represent unknown parameters that determine the effect each independent variable has on the dependent. To obtain the logistic model, the function must be changed into the form:

$$\begin{gathered}\mathrm{P}(\mathrm{R} = 1 | X_1, X_2, ..., X_k) = \\ \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}\end{gathered} \tag{4.2}$$

where $\mathrm{P}(\mathrm{R} = 1 | X_1, X_2, ..., X_k)$ represents the probability that a response variable $\mathrm{R} = 1$, given specific values for each X_i [50].

To simplify the description of the following formulas, the value $\mathrm{P}(\mathrm{R} = 1 | X_1, X_2, ..., X_k)$ for some particular set of values for each X_i will be denoted $\mathrm{P}(\mathbf{X})$. Because the relationship between the values X_i and the resulting probability are non-linear, a transformation must be performed. The most common transformation is the logit, or log-odds. The logit $P(\mathbf{X})$ is given by the formula:

$$\text{logit } \mathrm{P}(\mathbf{X}) = \ln\left(\frac{\mathrm{P}(\mathbf{X})}{1 - \mathrm{P}(\mathbf{X})}\right) \tag{4.3}$$

From this the values of each β can be computed, by holding the values of each X constant except for a particular X_i, then computing $\beta_i = \text{logit } \mathrm{P}_{i1}(\mathbf{X}) - \text{logit } \mathrm{P}_{i0}(\mathbf{X})$.

To compare the relative effects of each factor X_i, the odds ratio is used. This is expressed as:

$$\begin{aligned}\mathrm{OR}_{\mathbf{X}_1,\mathbf{X}_0} =& \frac{\text{odds for } \mathbf{X}_1}{\text{odds for } \mathbf{X}_0} \\ =& \frac{e^{(\alpha + \sum \beta_i X_{1i})}}{e^{(\alpha + \sum \beta_i X_{0i})}} \\ =& e^{\sum \beta_i (X_{1i} - X_{0i})}\end{aligned} \tag{4.4}$$

Because all independent variables in this analysis are categorical, the odds ratio calculation can be simplified to $\mathrm{OR} = e^{\beta_i}$ [50]. The odds ratio describes the relative effects of different values of a variable X_i. For example, an odds ratio of 0.5 would mean that when $X_i = 1$, then $\mathrm{P}(\mathbf{X})$ has one-half the likelihood compared to when $X_i = 0$. If the odds ratio is greater than 1 the likelihood increases, e.g. an odds ratio of 3.0 specifies an three-fold increase in likelihood. An odds ratio of 1 specifies no difference in likelihoods. An odds ratio less than 1 signifies a reduction in likelihood [51].

In this analysis the independent variables are categorical with either 2 or 3 possible values. This requires the use of a coding scheme with n-1 regressors for each categorical variable. Therefore, for this analysis the logistic model for the main effects is given as:

$$\begin{split} z = \alpha + \beta_S X_{S=\text{LinkedIn}} + \beta_S X_{S=\text{Twitter}} + \\ + \beta_P X_P + \beta_M X_M + \beta_m X_{m=\text{Ad}} + \beta_m X_{m=\text{News}} \end{split} \tag{4.5}$$

The X values for when S = Facebook, P = Generic, M = Generic, and m = Generic are used as the reference values for their variables, and thus are not shown independently in the equation. $X_{S=\text{LinkedIn}} = 1$ for cases where LinkedIn was the network the message was sent on, and 0 otherwise. The same method is used for $X_{S=\text{Twitter}}, X_{m=\text{Ad}}$, and $X_{m=\text{News}}$. X_P and X_M are equal to 1 when the persona or message type are “Adaptive”, respectively. The logistic model therefore is:

$$\begin{split} \mathrm{P}(\mathrm{R} = 1 | X_{S=\text{LinkedIn}}, X_{S=\text{Twitter}}, X_P, X_M, \\ X_{m=\text{Ad}}, X_{m=\text{News}}) = \\ \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \end{split} \tag{4.6}$$

Where $i = \{S = \text{LinkedIn}, S = \text{Twitter}, P, M, m = \text{Ad}, m = \text{News}\}$. The odds ratio for each variable is therefore the ratio of the odds for a particular value compared to the reference value.

Results of logistic regression

The results of the regression using only the main effects is given in Table 4.7. Based on these results Twitter had the highest overall interaction rate, as users were 6.3 times more likely to respond than users on Facebook. Users on LinkedIn were 3.9 times more likely to follow links than Facebook. In both cases p-values are extremely low and thus statistically significant, suggesting that these results are highly unlikely to be due to random chance. Adaptive personas were found to have only 0.78 times the likelihood of users following links compared to generic personas, with a 98% likelihood these results are not caused by chance.

Changes in message selection method were shown not to be statistically significant, with $p = 0.65$. Interestingly, the "Ad" messages were the least likely of the three classes to solicit following links, at only 0.26 times the likelihood of generic messages and a p-value of almost 0. This is interesting as generic messages ranged from short phrases to random strings of words or characters. "News" messages had the highest interaction likelihood, at 1.4 times that of generic messages and a 99.8% chance the results are not by chance.

Table 4.7: Logistic regression results of main effects

Regression Coefficient	Estimated Regression Coefficient	Estimated Odds Ratio	p-value
α	-2.84524	0.05812041	$< 2e-16$
$\beta_{S=\text{LinkedIn}}$	1.35864	3.89087880	4.6e-16
$\beta_{S=\text{Twitter}}$	1.83674	6.27601531	$< 2e-16$
β_P	-0.23969	0.78687004	0.02103
β_M	-0.04639	0.95467337	0.65435
$\beta_{m=\text{Ad}}$	-1.34615	0.26023922	$< 2e-16$
$\beta_{m=\text{News}}$	0.34285	1.40895415	0.00225

A list of values from performing a linear regression on the values for each platform is given in Table 4.8. On Facebook, the "adaptive" message selection method has close to twice the likelihood of eliciting a user to follow a link than "generic" messages, though the results for the other networks are not shown to be statistically significant. Messages with "ad" content were also shown have less than 0.2 times the likelihood and "news" less than 0.5 of being interacted with compared to "generic" message contents. Results for LinkedIn show that "news" and "ad" messages were less likely to be interacted with than "generic" messages. On Twitter, "adaptive" personas are shown to have little more than half the likelihood compared to "generic" personas. Additionally, it can be seen that "news" messages on Twitter have over 5 times the likelihood of users following links compared to "generic" messages, while "ad" have only 0.1 times the chance. The high interaction likelihood for "new" messages is also only seen on Twitter, but not the other two platforms

Table 4.8: Logistic regression results of main effects by platform

Regression Coefficient	Estimated Regression Coefficient	Estimated Odds Ratio	p-value
Facebook			
α	-2.8135	0.05999719	$<$ 2e-16
β_P	-0.1141	0.89218567	0.6937
β_M	0.6436	1.90326669	0.0326
$\beta_{m=\text{Ad}}$	-1.6869	0.18509234	0.0002
$\beta_{m=\text{News}}$	-0.7435	0.47544463	0.0223
LinkedIn			
α	-1.16666	0.3114042	5.96e-12
β_P	-0.03346	0.9670890	0.843
β_M	-0.01414	0.9859629	0.933
$\beta_{m=\text{Ad}}$	-0.94401	0.3890662	3.23e-06
$\beta_{m=\text{News}}$	-1.08480	0.3379706	2.54e-07
Twitter			
α	-1.3827	0.2508936	1.36e-15
β_P	-0.5125	0.5990076	0.0016
β_M	-0.2917	0.7470236	0.0712
$\beta_{m=\text{Ad}}$	-2.2774	0.1025500	2.21e-08
$\beta_{m=\text{News}}$	1.6254	5.0806500	$<$ 2e-16

4.3.3 Summary of the Results & Discussion

A summary of the results and their relationships to previous research are provided in Table 4.9.

Table 4.9: Summary of the main findings of the experimental study

Question number	Question statement	Result	Relevant tables
RQ1	Do spam interaction rates differ by social media platform? Levels: $S = \{$“Facebook”, “LinkedIn”, “Twitter”$\}$	• Yes, spam interaction rates differ by platform, with Twitter being the highest and Facebook the lowest.	4.2, 4.7
RQ2	Does matching the sending profile to the type of spam it sends affect interaction rate? Levels: $P = \{$“Generic”, , “Adaptive”$\}$	• On Facebook and LinkedIn, using an adaptive sending profile does not have a statistically significant effect. • On Twitter, using an adaptive sending profile, surprisingly, reduces the likelihood of a user interacting with spam.	4.3, 4.7, 4.8
RQ3	Does tailoring spam to recipient interests affect interaction rate? Levels: $M = \{$“Generic”, “Adaptive”$\}$	• On Facebook tailoring the spam to recipient interests may slightly reduce the likelihood of user interaction. • On LinkedIn and Twitter tailoring the spam to the recipient does not have a statistically significant effect.	4.4, 4.7, 4.8
RQ4	Does the content of spam messages affect interaction rate? Levels: $m = \{$“Generic”, “Ad”, “News”$\}$	• Yes, interaction rates differ by message content, though the rates are dependent on the platforms the messages are sent on. • Sending “news” messages on Twitter leads to the highest interaction rate of any class. • “Ad” messages are in general the least likely to generate interactions.	4.5, 4.6, 4.7, 4.8

The different interaction rates between social media platforms are interesting and, in the cases of Facebook and LinkedIn, are unexpected. Explanations for this result may be that because Facebook is designed for a network of users who know each other, users are more suspicious of messages from unknown individuals. Additionally, Facebook stores messages from non-friended users in a separate message inbox and does not provide a notification. Therefor some recipients may not have known of the message sent to them at all. If this is the case, it shows that a design choice employed by Facebook is effective in reducing the number of users falling prey to spam, even if reducing the effectiveness of spam was not a design goal. Because LinkedIn is designed for business communication, including recruiting, users may be accepting of messages from unknown accounts. Additionally, there may be a feeling that messages on LinkedIn are more likely to be legitimate due to the business nature of the site compared to the more social aspect of Facebook. Further, the average age of users being older on these platforms compared to Twitter could have an effect that reduces their interaction rates.

The high interaction rate on Twitter was driving mostly by the “news” class of messages. A possible reason for this is that users are more likely to expect news to be shared via Twitter, compared to the other platforms. Additionally, the extremely open nature of Twitter, where almost all profiles are publicly viewable, may actually make users less likely to be cautious of messages they receive from unknown accounts. It is also possible that the high number of bots in use on Twitter causes users to be more accepting of messages sent from accounts they suspect of being bots.

The “ad” class had the lowest overall response rate, with the exception of on LinkedIn where it was comparable to “news”. The most likely reason for this that users have been educated before that unsolicited ads are a common form of spam. Ads overall are also the type of message that seems “to good to be true”. “News” had the highest interaction rate, largely driven from Twitter. This may be due to the mistaken belief that there is little to no harm possible from following a link to a news story. Interestingly, the “news” class was only the highest interaction class on Twitter, which suggests that users are likely expecting to be sent and to follow linked to unsolicited news stories. “Generic” messages had a moderate response rate overall, though had the highest response rates on Facebook and LinkedIn. This could be due to the nice, generic messages seeming harmless. Users could also feel a social pressure to interact with messages on those platforms. Another factor may be some level of curiosity about the messages comprised of random words or characters. Although a common aspect of spam education is to look for poor grammar or spelling, users may interpret that only is a factor where messages do have some semblance of a meaning rather than true randomness. There is also the possibility that the results are a combination of both effects.

The adaptiveness of both sending personas and message content did not seem to be useful in increasing interaction rates. This may be due to the methods used in the study, which is possible as the methods were relatively basic compared to the amount of detail that targeted advertising users. Another possible reason is that when reading spam, users do not tend to examine the source of the messages and the content of the messages sent was generic enough that tailoring the recipient was not needed. This contrasts somewhat with the results in [6], which showed appearing more "likable" in a phishing message increased the likelihood of a recipient interacting with a message.

Overall, the relatively high interaction rates should perhaps have been expected, given the high rates that users will interact with bots [31, 32]. Analyzing the results of our study provide several possible methods for improving education that will reduce the tendencies for users to interact with spam. First, education must focus on both email and social media spam, as the types of messages sent on each can differ. Education must also include information on the potential dangers of social media spam. An interpretation of the low interaction rate of "ad" messages can be that years of education on spam being used for advertisements has led to users transferring their knowledge about email spam to social media spam. Education should also teach users that even platforms focused on relatively closed networks, where users tend to interact with other users they personally know, such as Facebook or LinkedIn, are in fact still open enough that spammers can still reach them and to not be completely trusting on those platforms. The high interaction rate of "news" messages, suggests that education should include information about this type of spam. However, teaching how to identify this type of spam is likely to be difficult, as the spam can use real headlines and therefore have few details to use to identify it.

4.4 Threats to Validity

Although efforts were made throughout the study to eliminate or mitigate threats to validity, there are some aspects of the study which could reduce the reliability or generalizability of the results. A selection of threats we identified are included in this section. There are four types of threats, construct, internal, conclusion, and external, which are enumerated here.

Construct threats are potential problems where what was measured is not what was needed to be measured. Examples in this study are that the sent messages may be different than actual spam in subtle ways that changed the experimental outcomes, even though the messages were based on real spam. The headlines created for the news class of spam may be more time sensitive than expected, which given the time required to create the messages and have them approved by the IRB could make them less likely to be interacted with. Additionally, some responses counted could be caused by automated systems that were not filtered out of the results. For example, some messages may have been mistakenly sent to bot accounts that then interacted with the URLs. This is important to note when comparing results between networks, as only Twitter allowed and provided an API for implementing bots at the time the experiment was performed and therefore likely had a higher percentage of bot accounts. To mitigate this, accounts being sent messages were selected based on having a high probability they were human controlled (e.g. by using BotOrNot). Also, the response collection server employed several methods to avoid counting non-human accesses and therefore the results are considered valid. Finally, although user profiling was used in the selection of some messages, the methods used may not be precise enough to affect the interaction rates of the adaptively selected messages.

Internal threats are unknown factors influencing results. Some participants could have viewed a URL multiple times, which would appear as multiple participants viewing it and artificially inflate some counts. However, this is likely a rare occurrence and therefore would not affect the results. Some participants may have used third-party programs to control their accounts, which may not show the sent messages, yet still appear as active accounts.

Conclusion threats relate to if the conclusions have been correctly justified. To mitigate threats of this type, a large sample was collected and appropriate statistical methods were performed.

External threats are those that affect the generalizability of the results. Because the results of the experiment are so heavily affected by user actions, changes to user behavior can change the generalizability. Tendencies of users may change in the future as new websites, better user education, and other factors change. Tendencies may also be affected by differences between social media platforms that were not studied as part of this experiment, or by changes in user interfaces made by the current sites.

4.5 Conclusion

In conclusion, this chapter presented the results of an experimental study on factors influencing social media users' interaction rates with spam, based on a survey of 256 respondents. This included descriptions of the design of the experimental system used to perform the experiment. Additionally, the results included both the interaction rates and an analysis of the factors, including possible causes for results. It is hoped that the findings are useful in providing a better understanding of user behaviors that can be used to build actionable methods of helping users and social media platforms maintain security online.

Chapter 5

Conclusion

A major goal of research is to discover knowledge that can be applied to help solve practical problems. Based on the results of this thesis, several such possibilities are given here.

First, current education to users about spam may not be sufficient and effective in changing their behaviors. It appears that long term education has been effective (due to low response rates of advertisement spam), but more recent types of spam (twitter news spam) is still not focused on enough. This suggests that spam education must be updated to reflect new developments, such as the increase in news spam on social media that was less common with email. This also suggests that education must update quickly enough for information to be disseminated and applied by social media users. A possible addition that can help reduce the need to constantly update education for specific threats, is to design education to focus on how the various threats of spam function at a conceptual level. This would allow users to apply their knowledge to types of spam they may encounter before they have been educated about it. In addition, a large number of survey respondents reported having no prior education about spam, which suggests that many social media and email users have a severe lack of understanding of threats posed to them.

The high interaction rates for some types of spam suggest that given enough time an individual will be tricked into interacting with a spam message, in spite of any education they may have. To help mitigate this, organizations should implement security plans that reduce the amount of access individuals have, i.e. only having access to information that is needed to fulfill specific roles. Organizations should also work to ensure that individuals who have fallen for a spam message are able to quickly notify security departments who can mitigate potential damage.

The low interaction rates for Facebook may have been partly caused by the messages being sent to an separate “non-friends” inbox rather than the standard “friends” inbox. Assuming this was a cause for the results, it suggest that social media platforms can design user interfaces in a way that makes spam more obvious and less enticing to users.

In summary, this thesis describes a survey of user behaviors and an experiment testing user behaviors related to social media spam. Results of the survey show that user behavior sometimes runs contrary to what is expected, such as the lack of correlation between spam education and improved spam behaviors. Additionally, it was found that user spam behaviors tend to be similar across social media and email. Results of the experiment show that some factors of social media spam can increase or decrease the likelihood that users will interact with it. For example, spam sent via Twitter, particularly with news related information, has significantly higher interaction rates compared to spam sent via Facebook or LinkedIn.

As with all research, there are limits to what can be accomplished in a single project. Several potential areas for future work to expand upon this thesis are identified here. First is the use of larger scale studies, both in terms of the number of respondents for surveys and for experiments with messages sent to larger amount of social media users. Surveys can be designed to focus on both additional factors, but also to focus more heavily on specific factors and include questions designed as validation of each other. Additional work can also focus on specific populations, and potentially use information from social media profiles to help identify users who match specific criteria, such as select age ranges, jobs, or geographic areas, or to send messages at specific times or days. Other factors can also be studied as well. Studies can also work with social media platforms, both improve the running of experiments but also to work together to help educate users and improve user interfaces with a focus on security. Finally, longitudinal experiments could also be performed, wherein message recipients are sent messages over time to study the effects of different message types on the same individuals or given education and tested to see how effective the education is.

References

[1] C. Hassold, "The Mobile Phishing Threat You'll See Very Soon : URL Padding," 2017.

[2] Google, "Stay safe online: Malware," 2011.

[3] H. Gao, J. Hu, T. Huang, J. Wang, and Y. Chen, "Security issues in online social networks," *IEEE Internet Computing*, vol. 15, no. 4, pp. 56–63, 2011.

[4] J. S. Downs, M. B. Holbrook, and L. F. Cranor, "Behavioral response to phishing risk," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit on - eCrime '07*, pp. 37–44, ACM, 2007.

[5] D. Modic and S. Lea, "Scam Compliance and the Psychology of Persuasion," *Journal of Applied Social Psychology*, vol. 304, no. January 2013, pp. 1–34, 2013.

[6] R. T. Wright, M. L. Jensen, J. B. Thatcher, M. Dinger, and K. Marett, "Influence Techniques in Phishing Attacks : An Examination of Vulnerability and Resistance," *Information systems research*, vol. 25, no. 2, pp. 385–400, 2014.

[7] B. A. J. Ferguson, "Fostering E-Mail Security Awareness :," *EDUCASE Quarterly 1.*, no. 1, pp. 54–57, 2005.

[8] B. 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.

[9] C. Iuga, J. R. C. Nurse, and A. Erola, "Baiting the hook: factors impacting susceptibility to phishing attacks," *Human-centric Computing and Information Sciences*, vol. 6, no. 1, p. 8, 2016.

[10] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *ACSAC '10 Proceedings of the 26th Annual Computer Security Applications Conference*, pp. 1–9, ACM, 2010.

[11] A. Lötter and L. Futcher, "A Framework to Assist Email Users in the Identification of Phishing Attacks," in *Proceedings of the Eighth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2014)*, pp. 42–52, 2014.

[12] X. Jin, "A Data Mining-based Spam Detection System for Social Media Networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 12, 2011.

[13] J. Martinez-Romo and L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language," *Expert Systems with Applications*, vol. 40, no. 8, pp. 2992–3000, 2013.

[14] BullGuard, "Socializing with malware on Facebook and Twitter...."

[15] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos, "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 24–28, 2000.

[16] J. J. Sheu, K. T. Chu, N. F. Li, and C. C. Lee, "An efficient incremental learning mechanism for tracking concept drift in spam filtering," *PLoS ONE*, vol. 12, no. 2, pp. 1–17, 2017.

[17] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63–92, 2008.

[18] I. Zeifman, "Bot Traffic Report 2016."

[19] E. Tan, L. Guo, S. Chen, X. Zhang, and Y. Zhao, "UNIK: unsupervised social network spam detection," in *Proceedings of the 22nd ACM international Conference on information & knowledge management - CIKM '13*, pp. 479–488, 2013.

[20] C. Cao and J. Caverlee, "Detecting Spam URLs in Social Media via Behavioral Analysis," in *37th European Conference on IR Research*, pp. 703–714, 2015.

[21] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013.

[22] K. Ivaturi and L. Janczewski, "A taxonomy for social engineering attacks," in *International Conference on Information Resources Management (CONF-IRM)*, 2011.

[23] N. L. Muscanell, R. E. Guadagno, and S. Murphy, "Weapons of influence misused: A social influence analysis of why people fall prey to internet scams," *Social and Personality Psychology Compass*, vol. 8, no. 7, pp. 388–396, 2014.

[24] R. E. Crossler, A. C. Johnston, P. B. Lowry, Q. Hu, M. Warkentin, and R. Baskerville, "Future directions for behavioral information security research," *Computers and Security*, vol. 32, pp. 90–101, 2013.

[25] J. S. Downs, M. B. Holbrook, and L. F. Cranor, "Decision strategies and susceptibility to phishing," *SOUPS '06: Proceedings of the second symposium on Usable privacy and security*, vol. 15213, pp. 79–90, 2006.

[26] N. A. G. Arachchilage and S. Love, "Security awareness of computer users: A phishing threat avoidance perspective," *Computers in Human Behavior*, vol. 38, pp. 304–312, 2014.

[27] N. Stembert, A. Padmos, M. S. Bargh, S. Choenni, and F. Jansen, "A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence," in *European Intelligence and Security Informatics Conference*, pp. 113–120, IEEE, 2015.

[28] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in *Proceedings of the 27th Annual Computer Security Applications Conference*, pp. 93–102, 2011.

[29] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[30] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Predicting susceptibility to social bots on Twitter," *Proceedings of the 2013 IEEE 14th International Conference on Information Reuse and Integration, IEEE IRI 2013*, pp. 6–13, 2013.

[31] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, "People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks," in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pp. 10–17, 2014.

[32] C. A. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso, "Reverse Engineering Socialbot Infiltration Strategies in Twitter," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 25–32, 2015.

[33] S. Savage, A. Monroy-hernandez, and T. Hollerer, "Botivist : Calling Volunteers to Action using Online Bots," *19th ACM conference on Computer-Supported Cooperative Work and Social Computing*, 2016.

[34] G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders, "Social networks and context-aware spam," in *CSCW '08 Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp. 403–412, ACM, 2008.

[35] S. Yu, "Fear of cyber crime among college students in the United States: An exploratory study," *International Journal of Cyber Criminology*, vol. 8, no. 1, pp. 36–46, 2014.

[36] W. R. Flores, H. Holm, M. Nohlberg, and M. Ekstedt, "Investigating personal determinants of phishing and the effect of national culture," *Information & Computer Security*, vol. 23, no. 2, pp. 178–199, 2015.

[37] L. Bilge, T. Strufe, D. Balzarotti, E. Kirda, and S. Antipolis, "All Your Contacts Are Belong to Us : Automated Identity Theft Attacks on Social Networks," *WWW 2009*, pp. 551–560, 2009.

[38] K. Jump, "A new kind of fame," jul 2008.

[39] K. Jansson and R. Von Solms, "Phishing for phishing awareness," *Behaviour and Information Technology*, vol. 32, no. 6, pp. 584–593, 2013.

[40] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson, "Going spear phishing: Exploring embedded training and awareness," *IEEE Security and Privacy*, vol. 12, no. 1, pp. 28–38, 2014.

[41] K. Lee, B. D. Eoff, and J. Caverlee, “Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter,” in *International AAAI Conference on Web and Social Media*, Citeseer, 2011.

[42] J. Hernandez and K. Goseva-Popstojanova, “Empirical Analysis of Spammers Activity on Web 2.0 Applications,” *Submitted for publication.*

[43] Center for Complex Networks and Systems Research and Indiana University Network Science Institute, “BotOrNot.”

[44] Leverage, “Social Media Comparison Infographic,” 2017.

[45] A. Bessi and E. Ferrara, “Social bots distort the 2016 U.S. Presidential election online discussion,” *First Monday*, vol. 21, no. 11, 2016.

[46] A. Heath, “Facebook quietly updated two key numbers about its user base,” 2017.

[47] Meyer Foundation, “Social Media Platform Comparison,” 2017.

[48] D. Fontein, “The Top Social Media Sites That Matter to Marketers,” 2016.

[49] D. C. Montgomery, *Design and analysis of experiments.* John Wiley & Sons, 2017.

[50] D. G. Kleinbum and M. Klein, *Logistic Regression.* Springer Science + Buisness Media, LLC, third ed., 2010.

[51] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Regression Models.* Irwin, third ed., 1996.

[52] N. Blaikie, *Analyzing Quantitative Data.* SAGE Publications, 2003.

Appendix A

Survey Questionnaire

Note that as part of the survey design, if a respondent selected "No" to question #3, the next question presented would be #14.

1. Please select the closest choice for the most recent higher education major you have completed or are enrolled in.

    - No Higher Education
    - Agriculture
    - Anthropology
    - Archeology
    - Architecture
    - Arts
    - Astronomy
    - Biology
    - Business
    - Chemistry
    - Communication
    - Computer Science/Engineering
    - Cultural Studies
    - Earth Science

- Economics
- Education
- Engineering
- Exercise Physiology
- Forestry
- Gender Studies
- Geography
- History
- Journalism
- Law
- Linguistics
- Literature
- Mathematics
- Medicine
- Military Science
- Philosophy
- Physics
- Political Science
- Psychology
- Public Administration
- Religion
- Social Work
- Sociology
- Statistics
- Transportation

2. What is your age?

- FREE RESPONSE

3. Do you use social media sites?
   - Yes
   - No
4. What social media sites do you currently use? Please check all that apply.
   - Facebook
   - Google+
   - Twitter
   - Flickr
   - Instagram
   - Pinterest
   - Tumblr
   - Other (please specify)
5. On average, how often do check email? Please list the total across all accounts.
   - Less than once per week
   - Once per week
   - Several times per week
   - Once per day
   - Several times per day
6. On average, how often do you use social media sites? Please list the total across all accounts.
   - Less than once per week
   - Once per week
   - Several times per week
   - Once per day

- Several times per day

7. What methods do you use to interact with social media sites? Please check all that apply.

    - Desktop / Laptop (web browser)
    - Desktop / Laptop (dedicated app)
    - Smartphone / Tablet (web browser)
    - Smartphone / Tablet (dedicated app)
    - Other (please specify)

8. On average, how often do you encounter spam on social media sites? Please note that each spam post, status, upload, etc., counts as an occurrence.

    - Never
    - Rarely
    - Sometimes
    - Often
    - Always

9. How often do you interact with spam on social media sites? Examples of interaction include clicking a link, watching a video, or some other action. However, simply reading spam is NOT considered interacting.

    - Never
    - Rarely
    - Sometimes
    - Often
    - Always

10. Have you ever received education and/or information about social media spam, and if so from where? Please select all that apply.

    - Never

- Family members
- Friends
- Teachers (before university)
- Instructors (during university)
- Media (TV, magazines, online, etc.)
- Other (please specify)

11. Has the education and/or information you received about social media spam had an impact on your behaviors related to social media spam?
    - No
    - Yes (please explain)
12. Do you believe that you are able to identify social media spam?
    - Strongly able
    - Somewhat able
    - Neither able nor unable
    - Somewhat unable
    - Strongly unable
13. If you encounter social media spam, do you report it?
    - Yes
    - No (please explain why not)
14. What methods do you use to interact with email? Please check all that apply.
    - Desktop / Laptop (web browser)
    - Desktop / Laptop (dedicated app)
    - Smartphone / Tablet (web browser)
    - Smartphone / Tablet (dedicated app)
    - Other (please specify)

15. On average, how often do you encounter email spam in your inbox? Please note that each spam email counts as an occurrence. Also note that spam automatically sent to a spam folder does NOT count.

    - Never
    - Rarely
    - Sometimes
    - Often
    - Always

16. How often do you interact with email spam? Examples of interaction include clicking a link, watching a video, or some other action. However, simply reading spam is NOT considered interacting.

    - Never
    - Rarely
    - Sometimes
    - Often
    - Always

17. Have you ever received education and/or information about email spam, and if so from where? Please select all that apply.

    - Never
    - Family members
    - Friends
    - Teachers (before university)
    - Instructors (during university)
    - Media (TV, magazines, online, etc.)
    - Other (please specify)

18. Has the education and/or information you received about email spam had an impact on your behaviors related to email spam?

- No
- Yes (please explain how)

19. Do you believe that you are able to identify email spam?

    - Strongly able
    - Somewhat able
    - Neither able nor unable
    - Somewhat unable
    - Strongly unable

20. If you encounter email spam, do you report it?

    - Yes
    - No (please explain why not)

Appendix B

Explanation of Contingency Correlation Coefficient

After performing a Chi-squared test on a contingency table, it is necessary to find the correlation between the variables studied. One method to do so is to compute the contingency coefficient, C. A useful property of C is that it is not reliant on the order that category rows and columns are arranged in the table. The contingency coefficient can be expressed as:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \tag{B.1}$$

where χ^2 is the value of the Chi-squared statistic and N is the total number of observations included in the contingency table. Because C is only reliant on the value of χ^2, then the value of C will significant if the null hypothesis is rejected. A value of 0 signifies no correlation, while larger values signify progressively higher degrees of correlation.

One issue with using C is that the maximum possible value is determined by the size of the table is is calculated from, rather than 1 as is common with other correlation measures. The maximum value, $C_{\max}$, is defined as:

$$C_{\max} = \sqrt[4]{\frac{m-1}{m} \times \frac{n-1}{n}} \tag{B.2}$$

where m is the number of rows and n the number of columns in the table. Therefore the range of possible values of C is between 0 and $C_{\max}$. Because of the different ranges of possible values, different values of C cannot be compared directly between tables. To account for this, values of C must be normalized to between 0 and 1 [52]. Here this is referred to as the normalized correlation coefficient, C^*, and is expressed as:

$$C^* = \frac{C}{C_{\max}} \tag{B.3}$$

Appendix C

Debrief Page for the User Interaction with Social Media Spam

The following text appeared as the debrief page that was viewed by recipients who followed a link in a message:

You have been directed to this site as part of an ongoing study in cybersecurity in social media and spam. The message containing the link to this site is a mock spam message sent as part of this study. The purpose of the study is to explore factors that influence interactions with spam on social media. Results of this study will be used for master's theses and published research by researchers at the Lane Department of Computer Science and Electrical Engineering at West Virginia University. This study has been approved by the Internal Review Board (IRB) at West Virginia University's Office of Research Integrity and Compliance.

All interactions, including this one, are anonymous and no personal information is recorded. The only information recorded as part of this study is that this page has been accessed, though there is no record of who may have viewed it. Participation in this study is completely voluntary, and will not affect class standing, grades, or job status in any way.

If you wish to opt-out of having taken part in this study, please select so below.

[Opt-out Button]

Some tips for avoiding real spam messages include:

1. Take note of poor spelling and grammar, as these are common for spam messages

2. Beware of links or attachments in messages you were not expecting or from people you

do not know

3. If you receive an unusual message that appears to be from someone you know, contact them through another method to verify the information

If you have questions about this study, please contact the researchers:
[Researcher contact information]

Appendix D

Messages Sent as Spam via Social Media Platforms

1. Advertisements

   (a) All networks

   - BEST golf gear here [URL]
   - Fly fishing gear [URL]
   - GREATEST book series of all time: [URL]
   - TOP TABLET REVIEWS -¿ [URL]
   - The BEST MUSIC STORE!!! Just OPENED [URL]
   - FREE game have to play! [URL]
   - BEST FREE Music Streaming [URL]
   - Funniest celeb fail videos [URL]
   - FREE WEB HOSTING!! [URL]
   - FREE Hotel Rooms - multiple Cities - countries - beaches [URL]
   - Watch these videos earn money at home! [URL]
   - John Byrance's secret tips for making money [URL]
   - Work AT home $25 hour [URL]
   - The latest news on making money at home: [URL]
   - Get Stocks -¿ MAKE MONEY [URL]

   (b) Twitter only

- Join the BEST MUSIC site ever! [URL]
- Join our group get followers! [URL]
- I got 103 followers in 24 HOURS from this group! :) [URL]
- Get the most followers!!! Our Service Better than the rest -¿ [URL]
- Meet all the new PEOPLE! around YOU[URL]

(c) Facebook only

- Join the best music streaming service for free! [URL]
- Join our group and double the number of likes you get! [URL]
- I got 500 likes in 24 hours from this group! [URL]
- Get the most likes!!! Our service is better than the rest: [URL]
- Meet the best people around you! [URL]

(d) LinkedIn only

- Join the best music streaming service for free! [URL]
- Looking for more profile reviews? Create a free account, and at least 20 employers will view your profile per day! [URL]
- Join [URL], 500 profile views in 24 hours!
- Maximize profile views! Our service is the best in class! [URL]
- Contact the best people in your industry: [URL]

2. News

- Brexit fallout: Greek banks on verge of default [URL]
- News: Markets tremble as Pound falls 12% [URL]
- Drastic changes in global futures markets [URL]
- Is this the end of the Eurozone? [URL]
- You'll never guess who is on track for a Grammy [URL]
- 2017 Grammy information leaks [URL]
- 2016's Hottest Movies! [URL]
- Michael Phelps breaks 45 year record!!! [URL]
- New hurdles for Rio as games approach [URL]

- Latest medal predictions for 2016 Olympics [URL]
- New info leaked from Clinton email scandal [URL]
- You won't believe what Trump said now! [URL]
- Latest poll results for major swing states [URL]
- Just announced: France to hold EU referendum [URL]

3. Generic

- Thanks for the ride last night! [URL]
- Great job! [URL]
- This might cheer up! [URL]
- Hope today goes well! [URL]
- How is your day going? [URL]
- Good luck today! [URL]
- Don't be sad it's over, be happy it happened [URL]
- Big join soda love [URL]
- Run Mississippi large goal today down [URL]
- Jsdhafjieo [URL]
- Faiooiaqfirhhao [URL]
- Village did removed enjoyed [URL]
- Advantages prosperous remarkably my [URL]
- my mile sold four. Need miss all four [URL]
- so reasonably be if [URL]
- sir curiosity discovery extremity [URL]
- followed learning prepared [URL]
- yet forfeited prevailed [URL]
- Belonging sir curiosit[URL]
- Expenses own moderate day [URL]

4. No-URL

   - “Don’t cry because it’s over, smile because it happened.” – Dr. Seuss
   - “Be yourself; everyone else is already taken.” – Oscar Wilde
   - Loving the weather today
   - “Fun is one of the most important and underrated ingredients in any successful venture.” – Richard Branson
   - Every strike brings me closer to the next home run. – Babe Ruth
   - The most common way people give up their power is by thinking they don’t have any. – Alice Walker
   - Either you run the day, or the day runs you. – Jim Rohn
   - The best revenge is massive success. – Frank Sinatra
   - Life shrinks or expands in proportion to one’s courage. – Anais Nin
   - An unexamined life is not worth living. – Socrates