# Examining the Demand for Spam: Who Clicks?

**Elissa M. Redmiles**
University of Maryland
eredmiles@cs.umd.edu

**Neha Chachra**
Facebook Inc.
nchachra@fb.com

**Brian Waismeyer**
Facebook Inc.
briwais@fb.com

## ABSTRACT

Despite significant advances in automated spam detection, some spam content manages to evade detection and engage users. While the spam supply chain is well understood through previous research, there is little understanding of spam consumers. We focus on the demand side of the spam equation examining what drives users to click on spam via a large-scale analysis of de-identified, aggregated Facebook log data (n=600,000). We find (1) that the volume of spam and clicking norms in a users' network are significantly related to individual consumption behavior; (2) that more active users are less likely to click, suggesting that experience and internet skill (weakly correlated with activity level) may create more savvy consumers; and (3) we confirm previous findings about the gender effect in spam consumption, but find this effect largely corresponds to spam topics. Our findings provide practical insights to reduce demand for spam content, thereby affecting spam profitability.

### ACM Classification Keywords

H.1.2 User/Machine Systems: Human factors

### Author Keywords

spam, security behavior, social media, Facebook

## INTRODUCTION & BACKGROUND

Falling for spam can embarrass people, lead them to accidentally pollute online spaces by spreading spammy content, cause them to purchase low-quality or non-existent goods, and can threaten their digital security - forcing them to cope with the consequences of accidentally installing malware or giving up their account credentials. Automated spam detection efforts are increasingly robust, helping to preserve people's security and prevent them from seeing spam in the first place. However, a small amount of spam still elusively slips through even the best-designed detection mechanisms. To reduce the incidence of spam even further and ultimately keep people safe, we must disincentivize spammers. Toward this goal, prior work has examined the spam supply chain: investigating what spammers sell, their profit models, and their distribution channels in order to design interventions – such as proactively identifying and blocking spam from reaching targets – to disrupt their value chain [35, 28, 9]. We examine the other side of the coin: the *consumers* of spam. By intervening with people who consume (i.e. click on) spam and helping them avoid clicking, we can reduce the profitability of spammers and keep people safer.

Past research has explored the use of games [50, 31], comics [43, 51], and other training materials [7, 2, 32] that teach people how to recognize and avoid spam. While these approaches have been shown to be effective in small-scale evaluations, it is impractical to continuously surface education materials to all people, given people's limited compliance budget for security [5]. To determine how best to target spam education and explore new interventions for spam consumers, we must first understand how features of the spam viewer and the spam itself influence consumption.

To this end, prior work has focused on understanding the consumers of email spam [49, 26, 57, 13, 40]. While email spam is an important first step to examining this problem, social media spam is also of increasing concern. Indeed, prior work has found that Twitter spam has a click through rate 10x higher than email spam [20]. Social media spam has a number of unique components [20, 33] both in terms of topics (e.g., some Twitter spam was found to be aimed at helping people gain additional followers, a motivation not seen in email) and mechanism of distribution (e.g., spammers utilizing trending hashtags on Twitter). As such, it is important to understand consumption of social media spam, specifically, and not merely rely on prior findings regarding email spam consumption.

In this work we expand on these prior investigations by observing social media, rather than email, spam consumption behavior. Prior social media spam research has focused on characterization and classification of spam content and spammers [33, 20, 60, 39], compromised account detection [25, 55, 52], and the relationship between spam creation and compromised accounts [55, 54]. Our work, instead, focuses on what features of the user and of the spam itself affect consumption; not on the spammers.

Further, much of the prior work on email spam has been conducted with laboratory or survey data [49, 26, 57, 13], which may have limited generalizability. Our work examines spam consumption behavior in the wild. We use de-identified, aggregated log data (n=600,000) of spam and non-spam consumption patterns on Facebook to develop a model of consumption behavior and lay the groundwork for developing large-scale, targeted interventions for reducing spam click-through rates.

Examples of content that Facebook considers spam include attempts to illicit illegitimate financial gain, e.g., by gathering account credentials (phishing); distributing malware or attempting to gain control of a person's Facebook account; or failing to deliver on a promised outcome: for example, content in the post (e.g., preview image shown) does not match the content the user receives upon clicking. To identify spam, Facebook relies on a dynamic set of machine learning classifiers to automatically identify and remove spam from the platform [47]. Once identified, spam is immediately removed from users' News Feed.

In this paper, we examine only spam containing URLs, rather than spam, for example, that requests people call the spammer, as we can observe when people on Facebook have clicked, not when they have contacted a spammer offline. Facebook spam containing URLs exhibits similar features to other social media spam: the URLs are often links to blacklisted websites, the spam is produced primarily by compromised accounts, has multiple vectors of delivery (e.g., through posts, comments on posts, etc.), and propagates through the social graph, spreading by country and region [19, 20, 47]. Facebook is a widely used social platform, with over 2 billion active users [10] who account for over 50% of the worldwide Internet population [56], making it a valuable source of data for generalizable findings.

In this paper we present the following key findings. First, higher activity level on Facebook reduces an individual's likelihood of clicking on spam. Activity level is also weakly correlated with skill using Facebook and the Internet more broadly, suggesting that skill-improvement may help drive down clicks on spam. Second, we find that people in communities with more spam are less likely to click. This suggests support for interventions that expose people to spam or phishing content as part of a training process, as exposure to this content may help them discern trustworthy content from spam. Third, we find that clicking norms within a peoples' community influence their individual consumption behavior. This suggests that interventions targeted to communities with high rates of spam consumption may be particularly effective and that leveraging social influence [11] may be effective for shaping better spam behaviors. Fourth, we find that people treat reshared content differently than original content, and their behavior suggests that they are using reshares as a heuristic to assess content credibility. Thus, we suggest that surfacing more information about the number of reshares and who has reshared content may be help to create an even more useful robust signal for people to evaluate content. This also raises the possibility of identifying other features that may be surfaced or made accessible to help them discriminate potentially spammy content.

Fifth, and finally, we observe a gender effect in spam clicking. After segmenting a random sample of the spam in our dataset by topic, we find that spam topics fall broadly into three categories: sales oriented (e.g., clothes for sale, modeling opportunities), media (e.g, videos, pictures), and interactives (e.g., quizzes, games). The sales-oriented content was viewed more by women and also had a click through rate that was 2 times that of media spam, which was viewed more often by men. Thus, we suspect that gender does not drive clicking but rather serves as an index for the different prevalence and effectiveness of different spam topics. Additionally, the similarity between the most consumed spam (sales oriented spam) and organic sponsored stories underscores the importance of distinguishing advertisements from general content more clearly, in order to provide people with a way to verify the authenticity of an offer.

In the remainder of the paper, we present related work, summarize our methods and detail our findings, and conclude with suggestions for new interventions to reduce the consumption of spam and disrupt the spam value chain.

## RELATED WORK

Below, we review prior work on social media spam, the demand for spam, and the demographics of spam consumers.

### *Social Media Spam*

The majority of prior work on social media spam has focused on Twitter and MySpace spam. Researchers have examined what type of spam is on these platforms [20], how to detect it [39, 60, 25, 52, 61], how it spreads throughout the social network [45, 33, 21, 27, 59], and by whom it is produced [33, 55, 54, 38, 36]. Examinations of spam on Facebook have focused either on characterizing Facebook spam [19], investigating malicious fake accounts [30, 8], including those aimed at generating fake likes [12, 37], or identifying malicious Facebook apps [44, 58]. Most relevant to our work, Gao et al. analyzed 3.5 million wall posts from Facebook in 2010 [19] and used URL classification based on URL blacklists to find that approximately 0.05% of these posts were spam. 70% of the spam they identified was phishing spam, and 97% of it was produced by real, compromised accounts rather than what the authors identified as fake accounts. In this work, we explore user interactions with Facebook spam, rather than characterizing the spam or the spammer.

### *Demand for Spam*

Prior work has attempted to characterize spam and the spam value chain in order to disrupt spammers profit models [28, 35, 9]. Examination of pharmaceutical spam has specifically illustrated consumers' demand for spam, underlining the importance and challenge of steering people away from enticing offers. Chachra et al. found that consumers sought out spam websites on which to buy counterfeit pharmaceuticals [9] either through Google searches or by hunting through their spam inboxes, circumventing the blacklisting designed to protect them. Relatedly, McCoy et al. found that three pharmaceutical spam websites had over 500k new consumers per month, suggesting significant demand for the spam goods. In our work we expand beyond pharmaceutical and email spam to explore the factors that drive the consumption of social media spam more broadly.

### *Spam Consumer Demographics*

Sheng et al. examined the demographics of those who fall for phishing emails [49]. They found that women and younger (ages 18-25) people were more likely to fall for phishing attacks. Jagactic et al. also found that women were more likely to fall for phishing attacks in their work [26]. While controlling for Internet experience and technical skill has been shown [49] to reduce these effects, prior work has still found an unexplained gender effect in spam clicking. More generally, there have been mixed results regarding whether higher Internet familiarity or Internet skill leads to lower phishing success rates [57, 24, 13, 49]. On the other hand, McCoy et al. found that men were more likely to engage with pharma email spam due to the type of pharmaceuticals typically advertised (erectile dysfunction, baldness) [40], suggesting that gender may serve as a proxy for spam topics rather than skill or any intrinsic propensity to click more frequently. In our work, we explore the relationship between Internet skill, gender, and spam topics to spam consumption in a larger, ecologically valid sample, and in a new domain: social media spam.

Focusing on peoples' relationship to spam senders and the amount of spam they receive, Jagactic et al. found that emails from friends' accounts tend to produce a higher victimization rate [26] and Vishwanath et al. found that people who receive more emails are more likely to fall for email spam [57]. The latter work, however did not control for the fact that people who receive more email also receive more spam, so it is unclear whether email load is truly the important factor. Prior work has also found that habitual media use was associated with higher likelihood of falling for phishing attacks due to inattention. In our work, we also consider the person's relationship to the content, as well as their friend count and activity level on Facebook to re-examine these prior findings in the social media context. Finally, we draw a new comparison, evaluating whether the factors that lead people to consume spam on Facebook differ from the factors related to whether they click on regular content.

## METHODOLOGY

We use aggregated, de-identified Facebook log data to explore the relationship between user and content factors and spam consumption. In this section we describe our datasets, sampling method, and modeling approach, and detail two supplemental data analyses.

### Consumption Behavior Modeling

**Data.** We used two log datasets in our analysis: one for spam and one for non-spam, which we will refer to as *ham*. The vast majority of attempts to spam Facebook are caught when an account attempts to publish spam. Thus, the spam dataset used in our study focused on the small subset of spam that was viewed by people on Facebook and later identified as spam by Facebook's machine learning classifier.

Both datasets contained only content shown in people's Facebook News Feeds [17] and could be any type of Facebook content including posts, videos, pictures, and Sponsored Stories [15]. This content was either posted by individual Facebook users or by Facebook Pages [14], and could be either original content (produced by the account making the post) or reshared content, where one account shares a piece of original content produced by another account [16]. We randomly sampled 300,000 viewer-content pairs from each dataset over 20 days on Facebook (2017/07/05 - 2017/07/25) resulting in our final analysis dataset of 600,000 viewer-content pairs. In our analysis datasets we include a parsimonious set of features of the account that viewed the content (but no uniquely identifying information about the account itself), features of the content, and features describing the relationship between the viewer's account and the account that produced the content. We selected features that we hypothesized – based on prior work – were relevant to spam consumption behavior [49, 26, 57, 24, 13]. Table 1 summarizes these features.

**Analysis.** To examine which of the features in Table 1 relate to spam consumption behavior, we utilized two binary logistic regression models: one to model spam consumption and one to model ham consumption. The spam model and the ham model included these viewer-content pair features as inputs and whether the viewer in the viewer-content pair had clicked on the content in the pair as the output. To validate our model fit we split each dataset 80-20 into a training set and a test set. Using AUC - a standard metric for evaluating the fit of logistic regression models [34] - as our metric of model fit, we found the AUC for the spam model is 0.72 and for the ham model is 0.80, which suggests that our models are reasonably well fit [46]. For each model, we present the outcome variable, including factors, log-adjusted regression coefficients (odds ratios), 95% confidence intervals, and p-values. Finally, to determine whether the explanatory features had different effects in our two models, we compare overlap in the confidence intervals for the ORs. Variables with non-overlapping confidence intervals between the two models are considered to be unique explanatory factors for that model.

### Supplemental Data Collection & Analyses

As is true of most analyses of log data at a single point in time, we cannot explain the decision-making that leads to the importance of these features nor can we determine causality. To partially mitigate the explanatory limitations of our work, we conducted supplemental analyses, described in more detail below, to evaluate hypotheses we established based on our log data findings; we suggest that other findings from our analysis should be similarly explored in future work.

#### *Qualitative Coding of Spam Topics*

We qualitatively examined a sample of the spam content in our dataset. We randomly sampled pieces of spam from our spam dataset and then qualitatively coded the topic of the spam into one of three groups that emerged during our codebook generation. We continued coding until new themes stopped emerging [41]: resulting in 250 coded pieces of spam. Two researchers (one male and one female) independently coded the content. Intercoder agreement was computed with Krippendorf's Alpha [18]; resulting in an alpha = 0.91, which is above the recommended threshold of 0.80 [29]. Further, after calculating this inter-coder agreement score, the researchers met to iterate on the codes until they reached 100% agreement on final codes for each piece of content. With our sample of 250, our maximum margin of error [4] is 6%, with a per item margin of error of 3% for two of our spam topics and 2% for the third.

We examined the click-through rate for the different content groups by gender, in order to more deeply understand whether the relationship we found between gender and spam clicking was potentially due to spam topics, as suggested in prior work [40].

#### *Survey: Activity Level and Skill*

Second, to explore whether activity level on Facebook, a significant feature in our spam model, was related to Internet skill and skill on Facebook, we surveyed 1784 Facebook users who were using Facebook with an English locale, to assess

| **Feature** | **Type** | **Description** | **Attribute of** |
|---|---|---|---|
| Gender | Categorical | Self-reported gender in the viewer's Facebook profile. | Viewer |
| Age | Numeric | Self-reported age in years in Facebook profile. | Viewer |
| Friend Count | Numeric | Number of friends of the viewer's Facebook profile. | Viewer |
| L28 | Numeric | Number of days out of the last 28 days that the viewer went to the Facebook News Feed. | Viewer |
| Country: Spam Prev. | Numeric | Volume of spam on the platform, calculated by weighted sampling of a subset of viewed content on Facebook, which is then human-labeled as spam or non-spam in order to obtain a non-biased estimate of the amount of spam on the platform, in a given country, etc. [47]. | Viewer |
| Country: Spam CTR | Numeric | Spam vs. ham clicking norms measured as a normalized spam click rate. This rate is calculated by taking the click through rate (number of clicks / number of views) of spam in a given country and normalizing this number by the general content click-through rate. This provides context for the rate of spam clicking in the country, contextualizing whether or not it is on par with general content click rates. | Viewer |
| Reshare | Categorical | Original content (produced by the account making the post) or reshared content, where the account owner shares a piece of original content produced by another account [16]. | Content |
| Content Origin | Categorical | Relationship between the viewer and the spammer. We consider only the three most prevalent relationships in our dataset: if the content came from a friend of the viewer (friend), from a friend of friend (fof), or from a Page they follow (Page). | Content |

**Table 1. Features included in the model: feature name, feature type (categorical or numeric), description of the feature, and whether the feature is an attribute of the viewer or the content.**

each type of skill. We deployed a two-item survey on Facebook, which contained a standardized, pre-validated measure of Internet skill [22] and a measure of Facebook skill directly derived from that measure (in supplementary material). The order in which the two measures were administered, as well as the ordering of terms, was randomized; the scales of understanding were also flipped, to reduce priming and bias. Additionally, our sample oversampled spam viewers, to ensure that our findings were generalizable to these users; our survey sample consisted of 918 users who had seen spam in the past month. We used Pearson correlation to examine the relationship between L28 (our measure of Facebook activity level), Internet skill, and Facebook skill.

### Ethics and the Use of Facebook Data

In this work we analyze anonymous, aggregated Facebook log data, anonymized content posted on Facebook, and data from a survey completed voluntarily by users with their Facebook locale set to English. There was no manipulation of any Facebook user's experience and no personal identifying information was used in this work. All spam was immediately removed from users' News Feed as soon as it was identified.

## RESULTS

Based on the results of our binary logistic regression models of spam click behavior (Table 2) and ham click behavior (Table 3) we find that age, gender, friend count, activity level on Facebook, country features (e.g., prevalence of spam in that country, clicking norms in the country), whether the content is reshared, and the relationship between the viewer and the content are all related to spam clicking. However, age, friend count, and whether the content is reshared are all also significantly related to ham clicking, in the same direction and with overlapping CIs. Thus, we consider only the features shown in bold in Table 2 to be uniquely related to spam clicking. We discuss these features, and their interpretation in detail below.

| Feature | O.R. | C.I. | p-value |
|---|---|---|---|
| **Days Active** | **0.98** | **[0.97, 0.99]** | **<0.001*** |
| Age | 1.01 | [1.01, 1.01] | <0.001* |
| **Male** | **0.81** | **[0.77, 0.85]** | **<0.001*** |
| Friend Count | 1.01 | [1.01, 1.01] | <0.001* |
| **Country Spam Prev.** | **0.59** | **[0.5, 0.7]** | **<0.001*** |
| **Country Spam CTR** | **1.01** | **[1, 1.01]** | **<0.001*** |
| Reshare | 2.70 | [2.37, 3.06] | <0.001* |
| **Content from FoF** | **1.75** | **[1.57, 1.94]** | **<0.001*** |
| **Content from Page** | **5.89** | **[5.44, 6.38]** | **<0.001*** |
| **Reshare:Content from FoF** | **0.35** | **[0.29, 0.41]** | **<0.001*** |
| Reshare:Content from Page | 0.37 | [0.32, 0.44] | <0.001* |

**Table 2. Spam model results. Friend count is in 100s of friends; the baseline for categorical "Content from" variable is Content from Friend. OR is the odds ratio between the given factor and the baseline; CI is the 95% confidence interval. Features in bold are unique to spam clicking as compared to ham clicking.**

### Activity Level

We find that users who are more active on Facebook, as measured by the number of days out of the last 28 that they were active on the site (L28), are less likely to click on spam. That is, our results show that, in this sample, a daily active user is 58% as likely to click on spam as a user who was active only once during the same 28 day period. We hypothesized, along the lines of prior work [49, 57], that this may be due to more active users having higher Internet and Facebook skill levels and thus being more savvy at identifying spam. To test the relationship between L28 and skill, we collected survey data for Internet skill and Facebook skill (see Section 3.2.2 for a detailed description of the survey instrument). We sampled 1784 Facebook users who had their Facebook set to English, resulting in a sample of responses from 54 countries. 46% of our respondents were Female and the sample had a mean age of 34.3 years. Based on the results of this survey, we find that L28 is significantly, weakly correlated to Facebook skill

| Feature | O.R. | C.I. | p-value |
|---|---|---|---|
| Days Active | 0.98 | [0.94, 1.02] | 0.324 |
| Age | 1.03 | [1.02, 1.03] | <0.001* |
| Male | 1.10 | [0.92, 1.31] | 0.303 |
| Friend Count | 1.01 | [1, 1.02] | 0.03* |
| Country Spam Prev. | 0.33 | [0.09, 1.23] | 0.098 |
| Country Spam CTR | 0.93 | [0.91, 0.95] | <0.001* |
| Reshare | 2.15 | [1.58, 2.92] | <0.001* |
| Content from FoF | 0.24 | [0.03, 1.72] | 0.155 |
| Content from Page | 9.47 | [7.33, 12.22] | <0.001* |
| Reshare:Content from FoF | 2.64 | [0.27, 25.94] | 0.405 |
| Reshare:Content from Page | 0.24 | [0.16, 0.37] | <0.001* |

**Table 3. Ham model results. See Table 2 for detailed legend.**

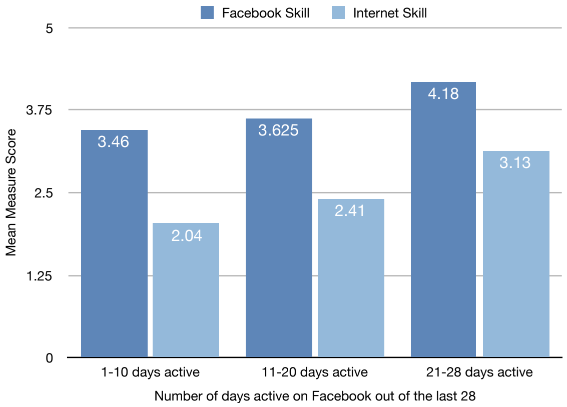

**Figure 1. Mean Facebook and Internet skill for users with different activity levels on Facebook, as measured by self report skill measures and number of days out of the last 28 that the user was active. Skill scores range from 1 to 5.**

($r = 0.16$, $p < 0.001$) and Internet skill ($r = 0.13$, $p < 0.001$) as shown in Figure 1, suggesting partial support for our hypothesis that activity level is an index for skill, which consequently helps users identify and avoid clicking on spam. However, this correlation is weak, suggesting that other causes may also factor into the relationship between L28 and spam. In Section 6, we discuss potential directions for future work to explore this relationship further.

### Gender

We find that women are more likely to click on spam, but there is no gender relationship for ham. This finding parallels findings in prior work [26, 49]. Similar to the results of Sheng et al., we find that controlling for activity level on the platform (weakly correlated with Internet skill) does not mitigate this effect [49]. Based on the findings of McCoy et al. who noted that pharma spam presenting different types of pharmaceuticals (e.g., erectile dysfunction medication) resulted in significantly different click-through-rates by gender [40], we hypothesized that our gender findings might be explained by a relationship between the topic of the spam and different click rates between the genders.

To evaluate this hypothesis, we qualitatively coded the topics of 250 pieces of spam content randomly sampled from the content viewed by users in our sample. We find that spam topics fall broadly into three categories: sales oriented (e.g., clothes for sale, modeling opportunities; 38%), media (e.g., videos, photos; 42%), and interactives (e.g., quizzes, games; 18%); with 2% of content falling in the "other" category. It is of note that these topics are fairly different than those identified in 2010 by Gao et al. in their characterization of Facebook spam, with the exception of quizzes and pharmaceutical spam, likely because spam evolves significantly over time, and because our dataset includes spam removed by Facebook whereas Gao et al. had access only to spam posts that remained undetected at the time they collected their dataset [19].

Of the topics in our dataset, we find that the sales-oriented content was viewed more by women (66%) - potentially because much of this content featured female products (e.g., beauty products for sale) or was explicitly gender targeted (e.g., "take this quiz to find out if he likes you") . On the other hand, more of the media spam was viewed by men (75%) - potentially because the majority of this spam was porn or violent content, which prior work has found is more appealing to men [3]. Interactive content was clicked slightly more often by men (55%) but this difference was not significant. Overall, regardless of the gender of the viewer, sales spam had a click-through-rate 2 times higher than that of media spam. We hypothesize that this is because sales spam is more similar to typical Facebook content than media spam, the majority of which was porn and violent content (82%). Sales spam that looks like regular content may consequently be more likely to go unnoticed and thus is more likely to be clicked, as compared to porn and violent content, which may appear clearly out of place or even be offensive to users. Thus, we suspect that gender, in and of itself, is unlikely to necessarily be related to spam clicking, but rather serves as an index for spam topics, which have differing effects on click behavior.

### Country Features

We find that prevalence of spam in a country, as well as country clicking norms are both related to an individual users' likelihood to click on spam. Users who reside in countries where spam is more prevalent are 59% less likely to click on it. On the other hand, in countries in which people proportionally frequently click on spam as compared to ham, individual users are more likely to click on spam.

This suggests that exposure to spam may help users learn to avoid engaging with spam; while country/communal norms around clicking may have the opposite effect, with individual users tending to behave in-line with country tendencies. This is true even when controlling for other features often correlated with country (such as activity level on Facebook, which strongly relates to Internet and Facebook skill level). These findings suggest that educational campaigns around spam may be especially influential when targeted to countries with high proportional rates of spam clicking. Future work should also consider more deeply exploring the decision-making patterns around clicking in countries at both ends of this scale - those with high spam to ham clicking ratios and visa versa - to better understand how to guide users toward safer clicking decisions.

### Viewer Relationship to Content & Resharing

We find a nuanced relationship between spam (and ham) clicking, resharing, and viewers' relationship to the content (e.g. whom or what shared or reshared the content). People on Facebook are less likely to click on spam content from friends than from friends-of-friends, with spam viewers being 1.75 times more likely to click on spam from friends-of-friends. We hypothesize that this is because people on Facebook have established expectations for content that should come from friends, and spam often does not fit these expectations. This relationship is the opposite for ham, with ham viewers being 0.24 times as likely to click on content from friends-of-friends. We hypothesize that this is the case because ham content from friends is more likely to align with their interests and thus they are more likely to click on it. Finally, people on Facebook are generally most likely to click on content, either ham or spam, from Pages. The OR for this relationship is greater for ham than for spam, with people being 9.47 times more likely to click on ham from Pages than from friends vs. 5.89 times more likely to click on spam from Pages than from friends-of-friends.

However, when the content is reshared the relationship between content-viewer connections and click behavior becomes more complex. We find that, generally, reshared content is more likely to be clicked whether it is ham (2.15 times as likely), or spam (2.7 times as likely). However, when considering the interaction between resharing and content-viewer relationships, we find a number of more nuanced relationships. Spam from friends that is reshared is more likely to be clicked - suggesting that resharing of spam from friends may serve to vouch for otherwise un-expected content. However, this effect does not hold for reshared content created by Pages and friends-of-friends. People on Facebook are 0.35 times as likely to click on reshared spam created by friends-of-friends, and are 0.37 times as likely to click on reshared ham or spam created by Pages. This suggests that people on Facebook may have nuanced heuristics for assessing spamminess - where reshares of unusual content posted by friends lend credibility while reshares of content posted by unknown sources adds suspicion. On the other hand, for ham, people are 2.64 times more likely to click on reshared content created by friends-of-friends, suggesting that reshares may proxy for content that is more aligned with their interests. On the other hand, just like for spam, people are 0.24 times as likely to click on reshared ham produced by Pages; perhaps suggesting that reshared content from Pages, even if it is ham, may begin to make people on Facebook think it is spam.

In summary, we hypothesize that people on Facebook click on spam from friends less often because they have stronger hypotheses for what content from friends should look like and thus are suspicious of unusual content from known sources. Relatedly, they may have weak assumptions for what content from friends-of-friends should look like and may even expect that it will differ somewhat from their own interests - thus making it more likely that they will fall for spam from these sources. Finally, they may anticipate that Pages will post promotional content, which looks similar to spam, and thus may be most likely to click Page spam.

Regarding reshares, we hypothesize that resharing of friend content serves to double vouch for the content: a known source has posted it and, as the majority of re-shares come from friends or Pages, a known or at least followed source re-shared it, supporting the likelihood that this is "safe" content. On the other hand, unknown content from an unknown source being reshared appears to trigger people on Facebook's sense of "spamminess" and thus they tend to click less on reshared spam from friends-of-friends or from Pages. This idea that reshares of unknown content makes people more suspicious is supported by prior work on email spam [48], and may explain why even reshared ham that originates from Pages is less likely to be clicked than if it was not reshared. These relationships between content and viewer on social media are relatively unexplored in prior work. Future work on spam clicking, and potentially on social media clicking in general, may wish to more deeply explore the reasoning behind users choices to click on content with which they have different relationships, to confirm the volitionality of users choices and to better understand how we should surface features such as the number of content reshares [42] to help users avoid clicking on spam.

## LIMITATIONS

We address first the limitations of our log data analysis, and second the limitations of our supplemental analyses. Our data come from only one platform, Facebook. While Facebook is the largest social media site in use today, and thus we feel a representative platform on which to conduct our work, we can only hypothesize about the generalizability of our findings to other platforms. Similarly, our datasets contain viewer-content pairs over a period of 20 days in July 2017; it is possible that events during this time of which we are unaware may have negatively impacted the generalizability of our findings. Additionally, our two demographic features (age, gender) rely on users' self-reports. It is possible that users' may have misreported; we do not, however have any reason to suspect that such mis-reporting is systematic or widespread enough to influence our results [6]. Further, Facebook's spam classifiers identify and remove the vast majority of spam before it is viewed. As such, this study examines spam that takes Facebook longer to catch. This population of spam may be somewhat unique to Facebook, as it is driven both by spammers and by Facebook's unique processes for detecting and removing spam. Finally, given that this work was conducted with proprietary log data it may be difficult to reproduce the results.

Our supplemental qualitative coding analysis suffers from a number of limitations common to qualitative analyses: there may have been coder bias that influenced coding, to mitigate this, we intentionally double coded and reached high agreement. Additionally, we coded only 250 of 300,000 pieces of content from each of our samples; while these pieces of content were randomly sampled, it is possible that the content we coded was not representative and thus resulted in biased findings. Finally, our survey data collection and analysis may suffer from response bias of those who were willing to voluntarily take a survey while on Facebook during the time period of survey collection (2017/08/11 - 2017/08/14). Second, and finally, the survey was conducted only with people using Facebook in English and thus the results are not representative of all users in our datasets; to partially mitigate this, we conducted the survey at varying times of the day, ensuring that users in a wide spread of geographies saw and completed the survey.

## DISCUSSION

Below we distill suggestions for the design and targeting of interventions to reduce consumption of spam and help users stay secure.

### *Training to Increase Skills and Awareness*

Our findings confirm many of the results of smaller scale studies on email phishing [49, 13]: we find that users who are more active on the platform are less likely to click on spam. We find a weak correlation between Facebook and Internet skill and activity level, suggesting that skill may partially explain this effect. Additionally, we find that exposure to spam reduces likelihood of clicking. This suggests that content awareness, in combination with general skills may help users avoid spam. Thus, we recommend evaluating the use of educational interventions that raise users' Internet and Facebook skill levels such as platform-specific walkthroughs, as well as continuing to use phishing-awareness-style trainings [32] that safely expose users to spam.

### *Customize Training Materials*

Relatedly, we find that spam related to different topics may be easier or harder to detect. Thus, we hypothesize, in line with theories presented in prior work [53], that training materials, especially for content awareness, should be personalized to show users the spam they are most likely to encounter. For example, future academic research may explore which types of spam are most likely to be encountered in particular countries or communities in order to personalize training materials more granularly. Such segmentation may also allow for integration of topical features into automated spam classifiers.

### *Provide Content Heuristics*

Our results also suggest that surfacing additional heuristics about the content itself may help users to gauge authenticity. We find that users leverage reshares as a way to determine if they should click on a piece of spam. We suggest, in line with recommendations from other work on email spam [42], that future research should explore whether providing additional context to users about resharing or about the authenticity of content may help them avoid spam.

For example, given our findings that sales-oriented spam is harder for users to detect, we suggest future work examine whether small UI indicators to show that a piece of content is a validated sponsored advertisement may be effective for reducing spam clicking and helping users evaluate the authenticity of a sales offer. Additionally, platforms may consider studying whether down-ranking (e.g., displaying lower in the Facebook or Twitter feed) content that contains a promotion and an URL but is not an officially sponsored advertisement is useful, at least for those users who are especially at risk of falling for spam.

### *Leverage Social Influence*

We find that users from countries with higher spam clicking tendencies click on spam *more* often. This may suggest that social norms around clicking are influencing users' security behaviors. Prior work by Das et al. found that telling users which of their Facebook friends had enabled a particular account security feature increased adoption of that feature [11]. Along the same lines, we suggest that, in addition to the above suggestions for content heuristics, social media platforms may wish to surface the number of friends who have *reported* a particular piece of content as spam, similar to Facebook's current approach to addressing fake news [1] and to how torrent platforms display the number of people who recommend a particular file.

### *Who to Target?*

Older users are more likely to click on both ham and spam. While they are not particularly at risk for spam, their "clicky" tendencies put them at greater risk than other users. Thus, they may be good candidates for a personalized intervention increasing skills or awareness of spam topics, especially if based on future work to better understand what spam topics may be most viewed by older users. Building on the relationship we find between individual spam consumption and country clicking norms we also recommend targeting users in high spam-CTR countries with similar personalized interventions — as targeting a relatively small, high-CTR set of users may have wide reaching effects. Finally, users who see a particularly low volume of spam but who have other risk factors (e.g., age, high friend count, high CTR social network) may be especially at risk, since they will not have exposure to spam content from which to learn heuristics for discerning legitimate content. Such personalized targeting of security education has been suggested in prior work [23, 53] but not evaluated, suggesting an area of fruitful future work.

## SUMMARY

In this work we contribute one of the largest scale studies of spam clicking behavior to date, and the first study to examine spam consumption behavior on social media. We find that more active users are less likely to consume spam, and find a weak positive correlation between activity level and Internet and Facebook skills – suggesting that higher skill may, in part, help users avoid spam. We also find that the volume of spam in a users' social network as well as clicking norms in that network influence their behavior. We identify a new, nuanced relationship between resharing, content-viewer relationships and click behavior. We find that resharing increases clicking on spam from friends, but decreases clicking on spam posted by friends-of-friends or Pages, suggesting that resharing serves as different heuristic roles depending on the type of content.

Additionally, we echo prior results finding that women are more likely to click on spam [49, 26], but find based on the results of open-coding of 250 pieces of spam content, that this relationship between gender and clicking is likely due to differences in the type of spam seen on Facebook, rather than anything intrinsically gender-related. Our results suggest that spam topics are likely a significant feature driving click behavior. In summary, we provide some of the first insights on users' consumption of spam on social media and find that spam clicking behavior on Facebook is affected by factors unique from those that affect clicking on general content. Based on these results, we suggest new directions for the design and targeting of spam interventions to disrupt the spam value chain and keep users secure.

## ACKNOWLEDGEMENTS

The authors would like to thank Mark Handel, Josh Krivoshein, Michelle M. Mazurek, Jay Nancarrow, David Radburn-Smith, Josh Rosenbaum, and Sudhir Venkatesh for their invaluable assistance on this project.

## REFERENCES

1. Davey Alba. 2016. Facebook finally gets real about fighting fake news. (2016).

2. Abdullah Alnajim and Malcolm Munro. 2009. An anti-phishing approach that uses training intervention for phishing websites detection. In *Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on*. IEEE, 405–410.

3. Feona Attwood. 2005. 'Tits and ass and porn and fighting': Male heterosexuality in magazines for men. *International Journal of Cultural Studies* 8, 1 (2005), 83–100.

4. James E Barlett, Joe W Kotrlik, and Chadwick C Higgins. 2001. Organizational research: Determining appropriate sample size in survey research. *Information technology, learning, and performance journal* 19, 1 (2001), 43.

5. Adam Beautement, M Angela Sasse, and Mike Wonham. 2009. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 workshop on New security paradigms*. ACM, 47–58.

6. Jeremy Birnholtz, Moira Burke, and Annie Steele. 2017. Untagging on social media: Who untags, what do they untag, and why? *Computers in Human Behavior* 69 (2017), 166–173.

7. Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Roland Borza. 2014. NoPhish: an anti-phishing education app. In *International Workshop on Security and Trust Management*. Springer, 188–192.

8. Qiang Cao and Xiaowei Yang. 2013. SybilFence: Improving social-graph-based sybil defenses with user negative feedback. *arXiv preprint arXiv:1304.3819* (2013).

9. Neha Chachra, Damon McCoy, Stefan Savage, and Geoffrey M Voelker. 2014. Empirically characterizing domain abuse and the revenue impact of blacklisting. In *Proceedings of the Workshop on the Economics of Information Security (WEIS)*. 4.

10. Josh Constine. 2017. Facebook now has 2 billion monthly users and responsibility. (2017). https://techcrunch.com/2017/06/27/facebook-2-billion-users/

11. Sauvik Das, Adam DI Kramer, Laura A Dabbish, and Jason I Hong. 2015. The role of social influence in security feature adoption. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1416–1426.

12. Emiliano De Cristofaro, Arik Friedman, Guillaume Jourjon, Mohamed Ali Kaafar, and M Zubair Shafiq. 2014. Paying for likes?: Understanding facebook like fraud using honeypots. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM, 129–136.

13. Julie S Downs, Mandy Holbrook, and Lorrie Faith Cranor. 2007. Behavioral response to phishing risk. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*. ACM, 37–44.

14. Facebook. 2017a. About Pages. (2017). https://www.facebook.com/help/282489752085908/?helpref=hc_fnav

15. Facebook. 2017b. About placements. (2017). https://www.facebook.com/business/help/407108559393196

16. Facebook. 2017c. How Do I Share a Post I See On My News Feed? (2017). https://www.facebook.com/help/163779957017799?helpref=search&sr=2&query=reshare

17. Facebook. 2017d. Your Home Page. (2017). https://www.facebook.com/help/753701661398957/?helpref=hc_fnav

18. Deen G Freelon. 2010. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science* 5, 1 (2010), 20–33.

19. Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. 2010. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 35–47.

20. Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*. ACM, 27–37.

21. Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 729–736.

22. Eszter Hargittai and Yuli Patrick Hsieh. 2012. Succinct survey measures of web-use skills. *Social Science Computer Review* 30, 1 (2012), 95–107.

23. Mariana Hentea, Harpal Dhillon, and Manpreet Dhillon. 2006. Towards changes in information security education. *Journal of Information Technology Education: Research* 5, 1 (2006), 221–233.

24. Jason Hong. 2012. The state of phishing attacks. *Commun. ACM* 55, 1 (2012), 74–81.

25. Danesh Irani, Steve Webb, and Calton Pu. 2010. Study of Static Classification of Social Spam Profiles in MySpace.. In *ICWSM*.

26. Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. *Commun. ACM* 50, 10 (2007), 94–100.

27. Jing Jiang, Christo Wilson, Xiao Wang, Wenpeng Sha, Peng Huang, Yafei Dai, and Ben Y Zhao. 2013. Understanding latent interactions in online social networks. *ACM Transactions on the Web (TWEB)* 7, 4 (2013), 18.

28. Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M Voelker, Vern Paxson, and Stefan Savage. 2008. Spamalytics: An empirical analysis of spam marketing conversion. In *Proceedings of the 15th ACM conference on Computer and communications security*. ACM, 3–14.

29. Klaus Krippendorff. 2004. Reliability in content analysis. *Human communication research* 30, 3 (2004), 411–433.

30. Katharina Krombholz, Dieter Merkl, and Edgar Weippl. 2012. Fake identities in social media: A case study on the sustainability of the facebook business model. *Journal of Service Science Research* 4, 2 (2012), 175.

31. Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. 2009. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 3.

32. Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 905–914.

33. Kyumin Lee, James Caverlee, and Steve Webb. 2010. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 435–442.

34. Stanley Lemeshow and David W Hosmer Jr. 1982. A review of goodness of fit statistics for use in the development of logistic regression models. *American journal of epidemiology* 115, 1 (1982), 92–106.

35. Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, and others. 2011. Click trajectories: End-to-end analysis of the spam value chain. In *Security and Privacy (SP), 2011 IEEE Symposium on*. IEEE, 431–446.

36. Chengfeng Lin, Jianhua He, Yi Zhou, Xiaokang Yang, Kai Chen, and Li Song. 2013. Analysis and identification of spamming behaviors in sina weibo microblog. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 5.

37. Xinye Lin, Mingyuan Xia, and Xue Liu. 2015. Does" Like" Really Mean Like? A Study of the Facebook Fake Like Phenomenon and an Efficient Countermeasure. *arXiv preprint arXiv:1503.05414* (2015).

38. Benjamin Markines, Ciro Cattuto, and Filippo Menczer. 2009. Social spam detection. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*. ACM, 41–48.

39. Michael Mccord and M Chuah. 2011. Spam detection on twitter using traditional classifiers. In *international conference on Autonomic and trusted computing*. Springer, 175–186.

40. Damon McCoy, Andreas Pitsillidis, Grant Jordan, Nicholas Weaver, Christian Kreibich, Brian Krebs, Geoffrey M Voelker, Stefan Savage, and Kirill Levchenko. 2012. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proceedings of the 21st USENIX conference on Security symposium*. USENIX Association, 1–1.

41. Janice M Morse. 1994. Emerging from the data: The cognitive processes of analysis in qualitative inquiry. *Critical issues in qualitative research methods* 346 (1994), 350–351.

42. James Nicholson, Lynne Coventry, and Pam Briggs. 2017. Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, Santa Clara, CA, 285–298.

43. Patrick G Nyeste and Christopher B Mayhorn. 2010. Training Users to Counteract Phishing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 1956–1960.

44. Md Sazzadur Rahman, Ting-Kai Huang, Harsha V Madhyastha, and Michalis Faloutsos. 2012. Frappe: detecting malicious facebook applications. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies*. ACM, 313–324.

45. Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 249–252.

46. Marnie E Rice and Grant T Harris. 2005. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and human behavior* 29, 5 (2005), 615.

47. Hervé Robert. 2016. Spam Fighting at Scale. (2016). `https://code.facebook.com/posts/894756093957171/spam-fighting-scale-2016/`

48. Scott Ruoti, Tyler Monson, Justin Wu, Daniel Zappala, and Kent Seamons. 2017. Weighing Context and Trade-offs: How Suburban Adults Selected Their Online Security Posture. In *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*. USENIX Association, 211–228.

49. Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 373–382.

50. Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security*. ACM, 88–99.

51. Sukamol Srikwan and Markus Jakobsson. 2008. Using cartoons to teach internet security. *Cryptologia* 32, 2 (2008), 137–154.

52. Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference*. ACM, 1–9.

53. Shuhaili Talib, Nathan L Clarke, and Steven M Furnell. 2013. Establishing a personalized information security culture. In *Contemporary Challenges and Solutions for Mobile and Multimedia Technologies*. IGI Global, 53–69.

54. Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 243–258.

55. Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. 2013. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse.. In *USENIX Security Symposium*. 195–210.

56. International Telecommunication Union. 2016. Measuring the Information Society Report. (2016). `https://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2016/MISR2016-w4.pdf`

57. Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H Raghav Rao. 2011. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems* 51, 3 (2011), 576–586.

58. Kaze Wong, Angus Wong, Alan Yeung, Wei Fan, and Su-Kit Tang. 2014. Trust and privacy exploitation in online social networks. *IT Professional* 16, 5 (2014), 28–33.

59. Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 71–80.

60. Sarita Yardi, Daniel Romero, Grant Schoenebeck, and others. 2009. Detecting spam in a twitter network. *First Monday* 15, 1 (2009).

61. Xianchao Zhang, Shaoping Zhu, and Wenxin Liang. 2012. Detecting spam and promoting campaigns in the twitter social network. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 1194–1199.