An Empirical Study of Factors Affecting the Rate of Spam

Rodrigo Sanches Miani[1], **Danielle Oliveira**[1]
Kil Jin Brandini Park[2], **Bruno Bogaz Zarpelão**[3]

[1]Faculdade de Computação (FACOM)
Universidade Federal de Uberlândia (UFU) – Uberlândia, MG – Brasil

[2]Faculdade de Engenharia Elétrica (FEELT)
Universidade Federal de Uberlândia (UFU) – Uberlândia, MG – Brasil

[3]Departamento de Computação (DC)
Universidade Estadual de Londrina (UEL) – Londrina, PR – Brasil

{miani,kil}@ufu.br, danielle@si.ufu.br, brunozarpelao@uel.br

Abstract. *Several factors may influence the number of spam received by email users, from user's profile information such as age or nationality to the way the email account is exposed on the Web. We propose a replication study of an experiment conducted more than a decade ago to understand the changes in the dynamics of the business of spam. To that end, using real email addresses created and managed only for the experiment, we simulate four different behavior profiles: i) interaction on social networks, ii) purchase in e-commerce sites, iii) interaction on forums and message boards and iv) use of file sharing tools, and analyze the amount of spam received on those accounts. The results indicate that linking an email account to a social network is the most significant influence on the spam rate.*

1. Introduction

The increasing use of the Internet and the resulting popularization of email caused a significant impact on people's lives. A problem directly linked to the popularization of email involves receiving unsolicited electronic messages, also known as spam. The amount of spam generated is a direct consequence of the low financial cost of sending emails, especially when compared to the regular correspondence [Cerf 2005].

Hann et al. [Hann et al. 2006] suggested a study to understand the dynamics of spam and confirm whether spam follows some guidance or are randomly distributed. The authors conducted an experiment to evaluate whether some factors, such as email providers and interests declared in certain products or services, could determine the rate of spam. A secondary goal in [Hann et al. 2006] was to investigate the influence of other factors (age and geographic location of the owner's account) in the delivery of unsolicited messages.

To that end, 288 email accounts with distinct owner's profiles were created on *Hotmail*, *Lycos*, *Excite* e *Yahoo* providers. 192 were exposed in the *Yahoo Geocities*, currently available only in Japan, website hosting provider. A web page was constructed for each one of these accounts. The definition of profiles included characteristics such as interests declared, e.g., "computers and technology," age, gender and geographic location.

Over a 33 weeks period, spam sent to the email accounts were monitored and analyzed according to each feature. The conclusions gathered at the end of the experiment were:

- Spam is not random but targeted at consumer segments that are relatively more likely to make online purchases, who declare an interest in specific products or services, adults, and US residents (US);
- The most significant find was that the greatest influence on spam rate was the identity of email service provider, since accounts on *Hotmail* received significantly more spam.

The study proposed by Hann et al. [Hann et al. 2006] is particularly influential due to its design. The authors conducted a field study by establishing real email accounts and also proposed a way to other people interact with them using Geocities. The common approach discussed in related studies consisted in analyzing a snapshot of a spam dataset provided by some university or ISP [Clayton 2008], [Garg and Niliadeh 2013] and [Almaatouq et al. 2016] or using survey data [Siponen and Stucke 2006].

The aim of this study is to replicate many aspects of the experiment conducted by [Hann et al. 2006], using an updated database that reflects the behavior of a Web user with different characteristics, such as using social networks, buying at online shopping sites, sharing files and posting on online forums. This type of replication study is known as triangulation which we have the same study goal but different design. At the end we will conduct a comparison between the results obtained by [Hann et al. 2006] and those achieved in this study. Since the prior study was performed more than ten years ago and the Internet behavior is constantly changing, we believe that a new and renovated investigation is necessary to provide a better understanding of other factors influencing the distribution of spam.

Our idea is to verify which behavior profiles (online shopping, online social networks, file hosting and online forums) affects the rate of spam, how email providers treated unsolicited messages and also establish a dataset of email accounts that could be used in future work.

The rest of this paper is organized as follows: Section 2 outlines the main strategy adopted in the methodology applied. Section 3 presents the results obtained and the analysis of those results. Section 4 describes related work. Finally, Section 5 brings the conclusion and future work.

2. Methodology

This paper aims to replicate, extend and update the findings of the work proposed by Hann et al. [Hann et al. 2006]. The main difference in this study is that the email accounts will be exposed to certain classes of online services, called ”exposure groups.” The proposal consists of the following steps:

1. Build the user profile: the first step consists in creating profiles for emails of fictitious persons. Each profile is formed by name, age, gender, nationality, service provider and exposure group that the account is associated.
2. Create email accounts: create fictitious accounts in three free email providers. Each account will be created respecting the users' profiles defined in the previous step.

Table 1. Summarization of the characteristics considered in this study.

Age	Gender	Residence	Mail service provider	Exposure group
18, 35 and 60	Male and Female	Brazil and United States of Americca	GMX/Mail, India and Zoho	i) Online Social Network (Facebook); ii) Online Shopping (Amazon and Submarino), iii) File Hosting (4Shared) and iv) Discussion Forum (Reddit)

3. Association: with all the email accounts created, the next step, was to associate them with their respective exposure groups: online shopping, online social networking, file hosting, and discussion forums.
4. Data collection: the email accounts were accompanied by a period of four months. They were monitored periodically to facilitate the analysis of the amount of spam received at the end of the review period.
5. Analysis: gathered spam was analyzed according to previously defined characteristics (age, gender, nationality, service provider and exposure groups). A comparison of the results acquired here and those found in [Hann et al. 2006] should be performed.

The first step was to define the user profiles. The characteristics used to determine each profile were the same as seen in [Hann et al. 2006]. For the age attribute, the possible choices are 18, 35 and 60 years. We choose those ages because they represent the main groups: youth, adults, and seniors, respectively. For nationality, there are two possibilities, United States or Brazil. For gender, choices are male and female. To explore other factors that influence the receipt of spam, we considered the email service provider and the exposure of email addresses to particular Web services. We classified every service in exposure groups. Table 1 summarizes the characteristics of the email accounts created in this experiment.

The exposure groups were proposed based on a typical Web user's behavior. Our idea was to answer the following question: "What kind of activities a Web user could do using an email account?". Based on studies conducted by [Whang et al. 2003], we created the following exposure groups:

- Online Social Networks (OSN) - Online social networking sites are very popular, usually run by individual corporations (e.g., Google and Yahoo!), and are accessible via the Web. Participating users join a network, publish their profile and any content, and create links to any other users with whom they associate [Mislove et al. 2007]. In this work, we associated email accounts with a *Facebook* account;
- Online Shopping (OS) - This is a shopping activity performed by a consumer via a computer-based interface, where the consumer's computer is connected to and can interact with, a retailer's digital storefront [Häubl and Trifts 2000]. For Brazilian residents, email accounts were registered in *submarino.com*, while US residents were associated with *amazon.com*;
- Discussion Forum (DF) - A discussion forum is a web application that provides a virtual environment supporting discussion and debate among peers without tem-

poral or geographical barriers [Cheng et al. 2011]. *Reddit* accounts were created and linked to some of the email accounts;
- File Hosting (FH) - File Hosting (FH) services are used daily by thousands of people as a way of storing and sharing files. *4Shared* accounts were created and associated with some of the email accounts.

We would like to verify the incidence of spam as follows: i) in every isolated group, ii) in the case of an email account is associated with all groups at the same time or, iii) an email account is associated with OSN and OS or iv) an email account is associated with exposure groups DF and FH. The division between OSN/OS and DF/FH is related to the e-commerce potential of both groups. Thus the total number of cases to analyze is seven, four separate groups added to each of the three combination alternatives.

We created an email account for each combination of gender, nationality, age and exposure group. Based on the same method as proposed by Hann et al. [Hann et al. 2006], we will have the following number of accounts: 3 (age) x 2 (gender) x 2 (nationality) x 3 (email providers) x 7 (combinations of exposure groups), equal to 252 email accounts, 84 accounts in each of the providers and 36 email accounts in each of the seven combinations of exposure groups.

We developed a database containing 252 email accounts created in three different email providers following the profiles previously defined. Initially we intended to use *Yahoo*, *Gmail* and *Hotmail* providers to create those accounts. However, such providers demanded a mobile phone number where they would send a code as a form of identity confirmation. This two-factor authentication hindered the creation of email accounts in those providers. Therefore, the decision was to use email providers that in addition to being free, had no restrictions regarding the verification of the user identity through a mobile number. Based on the list suggested by [Listandyou 2013], the chosen service providers were *GMX/Mail*, *India* and *Zoho*.

The proposed work will test the following hypotheses:

- Hypothesis 1: Test the incidence of spam considering the characteristics age, gender, and residence. Here, we use the same hypotheses as described by the replicated work [Hann et al. 2006]: both age and residence affect the rate and spam while gender did not.
  - 1a) Spam rates would be higher for email accounts associated with individuals aged 18 to 35, compared to individuals with 60.
  - 1b) Spam rates did not differ for email accounts associated with men or women.
  - 1c) Spam rates would be higher for email accounts related to the United States than those associated with Brazil.

- Hypothesis 2: Verify the Spam incidence in each of the seven defined combinations of exposure groups.
  - 2a) Spam rates would be higher for email accounts associated with all exposure groups.
  - 2b) Spam rates would be higher for accounts concomitantly exposed to social network and e-commerce than for those concomitantly exposed to a discussion forum and file hosting due to the synergy created between an

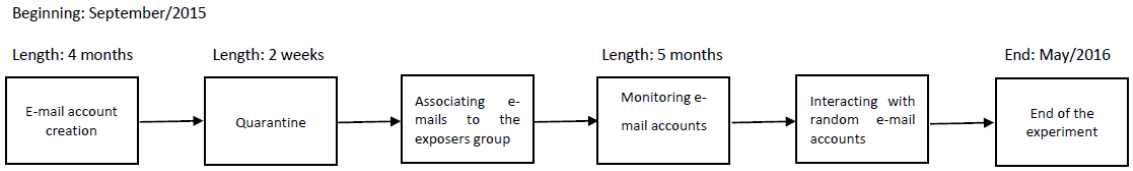

Figure 1. Flowchart of the proposed methodology.

email address with e-commerce profile, and it's owner's disclosed daily habits.

– 2c) For each separate exposure groups, spam rates would be higher for e-commerce and social networks exposed emails. According to [Jin et al. 2013], social networks are one of the most influential and important marketing tools, and the major goal of spam is to promote sales.

The experiment consists of monitoring the email accounts for 20 weeks. In the first two weeks, all email accounts go through a quarantine period in which no action related to those accounts will be conducted. In other words, none of the email accounts were associated with the exposure groups within these two weeks. Even in the quarantine period, we analyzed how many unsolicited messages were received. The results gathered can and will be used as a baseline for comparison between unexposed and exposed accounts. The accounts were monitored on a monthly basis to facilitate the analysis of the amount of received spam at the end of the experiment. Figure 1 shows each stage of the investigation.

The proposed methodology includes interacting with random email accounts two weeks before the end of the experiment. We carry out some activities within the scope of each exposure group, for instance, a friend request for a person with an email account associated with Online Social Network, upload a file for a person with an email account associated with File Hosting service and so on. The goal here is to understand the impact of some activities on the extends of spam received. The replicated work [Hann et al. 2006] suggested this direction for future research, in particular, to investigate accounts that engage in online transactions.

Table 2 describe the methodological differences between our work and the experiment proposed by Hann et al. [Hann et al. 2006] according to the email account characteristics, the length of the experiment, personal characteristics and how the email accounts were exposed. The main differences between the two works lie on the email service provider and in the way the accounts were exposed. While in Hann's work the email address was posted to the off-line Yahoo Geocities service, in our work, we link the email address to four different types of web services.

3. Results and Analysis

When creating the synthetic email accounts, we accepted the default type and level of anti-spam tools. All of the service providers chosen provided a simple spam guard that only directed suspected electronic messages into a bulk folder.

After creating the email accounts, they remained inactive for two weeks, a period named quarantine. During this time, none of the 252 email accounts received a single unsolicited email message. This behavior could be the result of some factors:

Table 2. Methodological differences

	Our work	Hann's work
Email account	252 email accounts for fictitious persons at GMX/Mail, India Mail, and Zoho. Default type and level of anti-spam tools.	288 email accounts for fictitious persons at Excite, Hotmail, Lycos and Yahoo. Default type and level of anti-spam tools.
Experiment length	20 week period (from September 2015 to May 2016). The association of the exposure groups was made after the first two weeks.	33 week period (from August 2003 to March 2004).
Person characteristics	i) Association to Web Services: Online Social Network (Facebook), Online Shopping (Amazon or Submarino), File Hosting (4Shared) and Discussion Forum (Reddit) ii) Age: 18, 35 and 60, iii) Gender: female and male and iv) Residence: Brazil or U.S.	i) Declared interests: computers and technology, travel, casino, or none; ii) Age: 15 or 30; iii) Gender: female or male and iv) Residence: Singapore or U.S.
Exposer accounts	Association of the person's email address to four types of Web Services: online social network, online shopping, file hosting and discussion forum (all of the email accounts were associated with at least one service).	Web page that included the person's email address and other personal details at Yahoo Geocities (192 out of 288 accounts).

- The chosen email service providers do not propagate user's data to third parties or;
- The quarantine period was not large enough to cover the propagation of user's data or;
- The chosen email server providers are not prime targets for spammers.

Continuing the experiment, after the first two weeks, we started the association of the email accounts with their respective exposure groups. Such accounts may be associated with one, two or all exposure groups. We opted out to receive notifications and marketing via email during their creation and their association to any of the four services. Nevertheless, those type of messages, when received, will still be deemed as spam.

We monitored the email addresses during the following period of 18 weeks and accounted for unsolicited messages both from the inbox and the bulk folder. The experiment was concluded in May 2016. Table 3 shows the descriptive statistics of the rate of spam.

Table 3. Spam rates for the 252 email accounts.

Provider	# of Spam	Mean	Std Dev
GMX/Mail	5208	62	7.87
Zoho	4185	49.82	7.05
India	4010	47.73	6.90
Total	13403	53.18	56.23

During the analysis period, the email accounts received an average of 53.18 spam messages. In the experiment conducted by Hann et al. [Hann et al. 2006], this number was 6.84 for the exposed email accounts. Even with the difference between the duration of the tests (20 and 33 weeks) and the use of different service providers, the average number of spam received in the new experiment is outstanding. It is also important to note that

Table 4. Regression results

Independent Variables	(a)	(b)	(c)	(d)
Constant	49.821** (6.121)	51.087*** (9.397)	35.012*** (7.8)	72.521*** (5.902)
India Mail	-2.083 (8.657)	-2.083 (8.7)	-2.083 (6.863)	-2.083 (5.008)
Mail.com (GMX)	12.179 (8.657)	12.179 (8.7)	12.179* (6.863)	12.179** (5.008)
Age 1 (18 or 35)	-	5.476 (8.7)	5.476 (6.863)	5.476 (5.008)
Age 2 (60)	-	-2.202 (8.7)	-2.202 (6.863)	-2.202 (5.008)
Residence	-	-5.817 (7.104)	-5.817 (5.604)	-5.817 (4.089)
Gender	-	1.103 (7.104)	1.103 (5.604)	1.103 (4.089)
Exposure group 1 (OSN+OS+DF+FH)	-	-	75.611*** (8.288)	-
Exposure group 2 (OSN+OS)	-	-	61.75*** (8.288)	-
Exposure group 3 (DF+FH)	-	-	-24.83*** (8.288)	-
Exposure group 4 (OSN)	-	-	-	57.991*** (6.246)
Exposure group 5 (OS)	-	-	-	-72.037*** (6.246)
Exposure group 6 (DF)	-	-	-	-69.148*** (6.246)
Exposure group 7 (FH)	-	-	-	-66.843*** (6.246)
R^2	0.013	0.019	0.397	0.68
	* p $<$ 0.1	** p $<$ 0.05	*** p $<$ 0.01	( ): error term

the accounts created on GMX/Mail received more spams than those created in Zoho and India.

Most of the spam is related to the email service provider itself, their marketing collaborators and all sort of notifications from the associated web service. The source is easily identified by subject or statements that marked their affiliation with the respective email service provider.

As in the experiment proposed by [Hann et al. 2006], a multiple linear regression model was used to verify the hypotheses. The amount of spam for each email account was the dependent variable (252 observations), and characteristics of each associated mail account such as gender (male, female), age (18, 35 or 60), and exposure group (seven combinations) were the independent variables. Four regression models were proposed to study the degree of significance of each independent variable.

The results were analyzed in agreement with the hypotheses and are described in Table 4. The significance levels adopted along this work were: * significance to the level of 99%, ** significance to the level of 95% and *** significance to the level of 90%.

In column (a), we report a regression with just a constant and two variables indicating the email service providers in which the accounts were created. We created a dummy variable to represent each service provider and, in this case, one variable (Zoho) has been removed from the model to avoid multicollinearity. The regression tool used (Minitab) removed one of the dummies automatically. Therefore, the coefficients of the two service provider variables were not significant, indicating that their accounts do not receive more or less spam than those registered with Zoho. However, when we consider the variables related to the exposure groups (columns (c) and (d)) the coefficient of the Mail.com (GMX) mail is positive and significant, indicating that such account received more spam than those registered with India Mail and Zoho.

Column (b) included additional variables (age, gender and residence) to test hy-

potheses 1 (1a, 1b and 1c). None of the personal characteristics associated with the email accounts were significant. This result is consistent only with hypothesis 1b) which states that men did not receive significantly more spam than women. Spam rates for Internet users aged between 18 and 35 were not higher than users aged 60. In other words, such characteristic does not influence the way spams are directed to users. Other feature studied that does not influence the rate of spam is the place where users live (0 = the USA; 1 = Brazil). Accounts of users who live in Brazil do not receive more emails than the USA-resident users.

Column (c) presents a regression model with three exposure groups: 1) email accounts registered in the four services, 2) email accounts registered only in social networks and e-commerce, and 3) email accounts registered in discussion forums and file hosting. The idea of this model is to evaluate hypotheses 2a) and 2b). First of all, the variable which represents exposure group 1, comprehending users registered in all services, is positive, significant and with the coefficient greater than the others. This implies that the rates of spam are higher for email accounts associated with all exposure groups, verifying hypothesis 1a).

The intuitive relationship that the greater the exposition of the user email account to different services, the higher the rate of spam was proved. Hypothesis 2b), which consists of verifying if email accounts associated with social networks and e-commerce tend to receive more emails than the email accounts associated with services with a less commercial appeal, such as discussion forums and file hosting, was duly proved. The coefficient of variables (OSN+OS) and (DF+FH) are significant and (OSN+OS) is positive while the coefficients for (DF+FH) is significant and negative, indicating that such accounts received less spam than those associated with Online Social Networks and Online Shopping.

Finally, column (d) includes the characteristics of the users and the individual association to each of the exposure groups. The coefficients of each variable show that the relationship of an email account to a social network is the most important evidence of receiving spams. This partially verifies hypothesis 2c), since the exposure group related to online shopping does not receive significantly more spams than the other groups. The negative values for OS, DF, and FH in column (d) mean that the number of spam received by the accounts associated with OS, DF and FH is lower than the accounts associated with OSN.

The comparison of each column (a-d) of Table 4 shows that the association of an email account to certain Web services (exposure groups) represents the greatest influence on the rate of spam. This can be seen by analyzing the explanatory power rates (R^2) for models that do not contain such variables (0.013 and 0.019, extremely low when compared to 0.397 and 0.68). Besides, all variables that represent the exposure groups were significant.

Hann et al. [Hann et al. 2006] suggested that users engaging in online transactions could affect the extent of spam received. To evaluate this hypothesis, we performed some interactions within the scope of each exposure group. We randomly choose five email accounts (one in each exposure group) during the last two weeks of the experiment (weeks 19 and 20). Interactions were carried out in each account according to the groups they

were associated with. The interactions performed for the accounts related to the social networks group were: make the email public by changing some privacy configurations in each Facebook account, "like" several pages and add friends randomly. For the online shopping group: purchase simulations, add an item to chart or create a bank payment slip without paying it. For discussion forums, comments were made on popular posts. Finally, we uploaded files, pictures, and others types of digital media for the email accounts associated to file hosting.

We employed a t-test in each exposer group combination to assess whether the mean number of spams received by the accounts with some interaction is statistically different from the accounts with no interactions. Table 5 shows the results. Our results show that the difference between the mean number of spam for accounts with interaction and accounts with no interaction is significant at the 0.05 level for the following groups: exposure group 1 (OSN+OS+DF+FH), exposure group 2 (OSN+OS), exposure group 4 (OSN) and exposure group 5 (OS). This result indicates that some actions performed by the users of Online Social Networks and Online Shopping services could lead to an increase or decrease of the spam rate, as suggested by the replicated work [Hann et al. 2006]. Such investigation should be extended to understand which sort of interaction might cause an increase or decrease in the spam rate for a certain email account. Knowing which activity for a web service is associated with an increase in the spam rate would be useful for providing new insights to fight against spam.

Table 5. T-test results

	Mean number of spam for accounts with interaction	Mean number of spam for accounts with no interaction	t-value	p-value*
Exposure group 1 (OSN+OS+DF+FH)	110.61	125.8	-1.76226	0.0435
Exposure group 2 (OSN+OS)	96.48	113.6	-2.24266	0.015774
Exposure group 3 (DF+FH)	12.94	8.2	0.70162	0.243847
Exposure group 4 (OSN)	127.77	162.6	-3.57541	0.000536
Exposure group 5 (OS)	2.16	5.2	-2.04572	0.024291
Exposure group 6 (DF)	6.13	2.4	0.73122	0.234903
Exposure group 7 (FH)	8.23	5	0.65091	0.259741

*p-values of one-tailed significance.

To improve comprehension regarding the impact of the study, it is important to compare the results found here with those discovered in the replicated work. Table 6 presents a summarization of the results found by both works.

Findings #1, #3, #5 and #6 are similar to the replicated work while findings #2 and #4 provide contradictory results when compared to the replicated work. Finding #1 of the replicated work states that spam rates are higher in email accounts with a declared interest in some product or service. Since this type of email account configuration is not available anymore, we decided to link a mail account to a Web Service, creating a declared interest

Table 6. Comparison between new results with the results being replicated

	Replicated work ([Hann et al. 2006])	Our work	Comparison
Finding #1	Spam rates are higher in email accounts with a declared interest in some product or service	Spam rates are higher in email accounts associated with Online Social Networks.	Similar result
Finding #2	Spam rates are higher in email accounts associated with individuals aged 30 than those aged 15.	Spam rates do not differ in email accounts associated with individuals aged 18, 35 or 60.	Contradictory result
Finding #3	Spam rates do not differ from email accounts associated with men about women.	Spam rates do not differ from email accounts associated with men about women.	Similar result
Finding #4	Spam rates are higher in email accounts associated with the U.S. than Singapore residents.	Spam rates do not differ in email accounts associated with U.S. and Brazil.	Contradictory result
Finding #5	Spam rates are higher among Hotmail accounts, followed in decreasing order by Lycos, Excite, and Yahoo! accounts.	Spam rates are higher among Mail.com (GMX) accounts when the association with exposure groups is taken into account.	Similar result
Finding #6	Spam rates are higher in email accounts exposed through Web pages.	Spam rates are higher in email accounts associated with any Web Service (exposure groups).	Similar result

for each mail account. Likewise, our work states that spam rates are higher in email accounts associated with Online Social Networks when compared to the other three types of exposure groups (Online Shopping, Discussion Forum, and File Hosting). At last, finding #6 is related to exposing an email account through Web pages. We did not create a Web page with person's email address at Yahoo Geocities. Instead, we used the person's email address to associate with some types of Web Services. This could also be seen as a way to expose or share a certain email account, since, in our case, all of the email accounts were kept in secret. Our quarantine period showed that unsolicited emails were only received after the association with a Web Service, which means that spam rates are higher in email accounts exposed through Web pages and associated with Web Services.

Findings #2 and #4 presents some contradictory results regarding the link between spam rate and two factors: age and residence. Hann et al. [Hann et al. 2006] found that spam rates are higher in email accounts associated with individuals aged 30 than those aged 15. They used the following argument to justify their hypothesis: historically, the 30–49 age group exhibited the highest rate of online purchases, but by December 2002, the 18–29 group had caught up, and both groups exhibited the same 63% rate. Our idea was to extend the age group to 18, 35 and 60 years old. However, no influence was found regarding the proposed age groups. Using a similar argument, we found that the percentage of U.S. adults in each age group (18–29, 30–49, 50–64 and 65+) whoever buy something online is 90%, 87%, 72% and 59% [Smith and Anderson 2016].

Since every group has a rate of online purchase higher or very similar to that 63%, we might infer that that age groups above 18 would not be an influence of spam rate. Finding #4 is related to the relationship between spam rate and the person's country of residence. Since in 2003 the e-commerce participation rate was 22.7% among Singaporeans with Internet access as compared with 61% of Americans, the residence would be a potential influence on spam rate. Hann et al. found that, in this case, the residence is a factor that influences the distribution of spam. However, in 2015, the e-commerce participation rate was 51% among Brazilians with Internet access as compared with 75,6%

among Americans[1]. This difference indicates that both countries have a strong digital buyer penetration which helps explain why the residence was not a relevant factor in the spam rate in this study.

Another difference between the two studies is that the main conclusion of the experiment conducted by Hann et al. is that the most significant influence on the spam rate was the identity of the email service provider. Hotmail accounts, in particular, received significantly more spam than accounts set up with other email service providers. In our case, the most significant influence on the spam rate was the association of an email account to an online social network (*Facebook*).

4. Related Work

Most of the spam literature is focused on providing new ways of detecting and filtering unsolicited messages [Androutsopoulos et al. 2000, Blanzieri and Bryl 2008]. Some empirical studies were performed to discuss relevant questions such as how unsolicited emails sent with commercial purpose affects the consumer privacy [Jacobsson and Carlsson 2004], quantitative analysis of spam traffic and the use of DNS blacklists [Jung and Sit 2004], and characterization of types of vulnerable email accounts [Dhinakaran et al. 2007].

Two empirical studies follows a similar methodology as that described in [Hann et al. 2006]: [Clayton 2008] and [Garg and Niliadeh 2013]. Both of them aim to understand the underlying characteristics of the factors that influence the distribution of spam. During eight weeks, Clayton [Clayton 2008] collected records through the UK Internet Service Provider (ISP) Demon Internet and found out that the first character of email addresses affects the proportion of spam received. Garga and Nilizadeh [Garg and Niliadeh 2013] present a study that examines whether economic, structural, and cultural characteristics of a community explain the incidence of Craigslist-based scams. The authors find that scams are targeted and influenced by community characteristics. Communities with a higher proportion of educated white males specifically are most exposed to online fraud most likely due to the purchasing behavior of this community.

The exposure groups were proposed based on a typical behavior of Internet users who own an email account. The following studies helped us to define and to understand the characteristics of each proposed group (online shopping, online social networks, discussion forum, and file hosting): [Hart 2008], [Schryen 2007], [Almaatouq et al. 2016] and [Ezpeleta et al. 2016].

Hart [Hart 2008] examines a large sample of web users to study their online behavior especially their online purchasing behavior, their perceptions of web personalization and their concerns about privacy. The author highlights that once people have commenced shopping online, few decide not to continue. The type of online purchasing shows significant variation across age groups due to lack of earning power or to less Internet experience or access. This could be linked to the rate of spam received by each group.

In a study conducted by Schryen [Schryen 2007], he acknowledges that Web pages

[1]http://www.statista.com/statistics/420391/spam-email-traffic-share and http://www.statista.com/statistics/252404/digital-buyer-penetration-in-brazil/

and Usenet groups belong to the most vulnerable Internet spots regarding email address harvesting. However, this scenario is changing due to the dissemination of Online Social Networks (OSNs), that is also one of the exposers group studied in our work.

Almaatouq et al. [Almaatouq et al. 2016] present an analysis of spam accounts in OSNs. The authors analyzed over 100 million messages collected from Twitter over the course of 1 month and conclude that there exist two behaviorally distinct categories of spammers (profile and social interactions) and that they employ different spamming strategies.

Ezpeleta et al. [Ezpeleta et al. 2016] propose a study to evaluate the consequences of displaying information publicly in OSNs. The authors demonstrated that 19% of the collected email addresses have a corresponding Facebook account. Since basic public information can be extracted from those users, creating personalized email subject and bodies it is also possible. The authors show that these emails can have a click-through rate higher than 7.62%, more than 1,000 times higher than typical spam campaign rates.

5. Conclusion

We presented a study of factors that influence the spam rate by replicating a study conducted in 2004 [Hann et al. 2006]. We showed that the sending of spam is still oriented to a given segment and also that an email account associated with an online social network (*Facebook*) receives significantly more emails than email accounts associated with other types of Web services such as online shopping, discussion forum or file hosting.

This paper investigated factors that influence the spam rate, or the number of unsolicited messages received by an email account, among them: age, nationality, and gender. Besides, the identity of the email provider and the exposition of email addresses in different types of Web services, here called exposure groups, were also researched. Just like in the experiment carried out by Hann et al. [Hann et al. 2006], this study aimed to confirm that the sending of spam is not random, but oriented to a given segment. We also showed that the factor with greater influence in the receipt of unsolicited electronic messages is the exposure of the email address to the online social network exposure group (*Facebook*).

As a future work, we expect to analyze the characteristics of received spams in each email account to know the incidence of malicious spams related to cyber crimes, as *phishings*, and explore which exposure groups are more likely to receive malicious spams. Another important direction for future work is to extend this paper by creating new exposure groups, for example, including the email address in personal sites or providing the email address in registers in physical stores.

Acknowledgment

The authors would like to thank FACOM/UFU and PROPP/UFU for supporting this work.

References

Almaatouq, A., Shmueli, E., Nouh, M., Alabdulkareem, A., Singh, V. K., Alsaleh, M., Alarifi, A., Alfaris, A., and Pentland, A. (2016). If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security*, (February):1–17.

Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., and Stamatopoulos, P. (2000). Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach. *Proceedings of the workshop Machine Learning and Textual Information Access*, (September 2000):1–12.

Blanzieri, E. and Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92.

Cerf, V. G. (2005). Spam, spim, and spit. *Communications of the ACM*, 48(4):39.

Cheng, C. K., Paré, D. E., Collimore, L. M., and Joordens, S. (2011). Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers and Education*, 56(1):253–261.

Clayton, R. (2008). Do Zebras get more Spam than Aardvarks? In *Proceedings of the Fifth Conference on Email and Anti-Spam*.

Dhinakaran, C., Lee, J. K., Nagamalai, D., and Chae, C. J. (2007). An empirical study of spam and spam vulnerable email accounts. *Proceedings of Future Generation Communication and Networking, Main Conference Papers, Vol 1*, 1:407–412.

Ezpeleta, E., Zurutuza, U., and Hidalgo, J. M. G. (2016). A study of the personalization of spam content using facebook public information. *Logic Journal of the IGPL*, 25(1):30–41.

Garg, V. and Niliadeh, S. (2013). Craigslist scams and community composition: Investigating online fraud victimization. In *Security and Privacy Workshops (SPW), 2013 IEEE*, pages 123–126. IEEE.

Hann, I.-H., Hui, K.-L., Lai, Y.-L., Lee, S.-Y. T., and Png, I. P. (2006). Who gets spammed? *Communications of the ACM*, 49(10):83–87.

Hart, M. (2008). Do online buying behaviour and attitudes to web personalization vary by age group? In *Proceedings of the 2008 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries: Riding the Wave of Technology*, SAICSIT '08, pages 86–93, New York, NY, USA. ACM.

Häubl, G. and Trifts, V. (2000). Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids. *Marketing Science*, 19(1):421.

Jacobsson, A. and Carlsson, B. (2004). Privacy and Spam: Empirical Studies of Unsolicited Commercial E-Mail. In *Proceedings of IFIP Summer School on Risks & Challenges of the Network Society*, pages 241–251.

Jin, L., Chen, Y., Wang, T., Hui, P., and Vasilakos, A. V. (2013). Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, 51(9):144–150.

Jung, J. and Sit, E. (2004). An empirical study of spam traffic and the use of DNS black lists. *4th ACM SIGCOMM conference on Internet measurement*, pages 370–375.

Listandyou (2013). 10 services to create free email accounts.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement - IMC '07*, pages 29–42.

Schryen, G. (2007). The impact that placing email addresses on the Internet has on the receipt of spam: An empirical analysis. *Computers and Security*, 26(5):361–372.

Siponen, M. and Stucke, C. (2006). Effective anti-spam strategies in companies: An international study. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 6, pages 127c–127c.

Smith, A. and Anderson, M. (2016). Online Shopping and E-Commerce. Technical report, Pew Research Center.

Whang, L. S.-M., Lee, S., and Chang, G. (2003). Internet over-users' psychological profiles: a behavior sampling analysis on internet addiction. *Cyberpsychology & behavior*, 6(2):143–150.