By Il-Horn Hann, Kai-Lung Hui, Yee-Lin Lai, S.Y.T. Lee, and I.P.L. Png

WHO GETS SPAMMED?

Spam is no random event, but specifically targets those with purchasing power.

EMAIL USERS all over the world are being swamped by unsolicited commercial email ("spam"). Even the world's richest person has not been spared:

"Like almost everyone who uses email, I receive a ton of spam every day. ...But spam is worse than irritating. It is a drain on business productivity, an increasingly costly waste of time and resources that clogs corporate networks and distracts workers."

—*Bill Gates* [2]

Policymakers, Internet service providers, software vendors, and scholars are struggling to devise technological, regulatory, and social solutions [6]. However, a major obstacle for policymakers is that scientific research into the spam industry has been very limited. Even the most basic question—whether spam is sent randomly or targeted—remains open. U.S. Federal Trade Commission Chairman Timothy Muris [7] has asserted:

"Unlike phone calls or mail solicitations, sending additional spam is essentially costless. ...Because email technology allows spammers to shift the costs almost entirely to third parties, there is no incentive for the spammers to reduce the volume. ...At our Spam Forum, a bulk emailer testified that he could profit even if his response rate was less than 0.0001%."

If spam is essentially costless to send, spammers should broadcast solicitations repeatedly to all available email addresses. As Bill Gates remarked:

"Knowing that only a small percentage of their output will get past today's filters, spammers have responded by significantly cranking up the volume of emails they send" [3].

We conducted a field experiment to learn more about spam. Our first objective was to confirm whether spam is randomly distributed or targeted. If spam is not randomly broadcast, what factors determine the rate of spam? It is already well known that

We found that spam was highest among Hotmail accounts, followed in decreasing order by Lycos, Excite, and Yahoo! accounts.

email addresses posted on Web sites or in newsgroups, [1, 9] as well as those that do not opt out of receiving marketing communications [5], attract relatively more spam. Accordingly, Internet users have been advised to disguise or conceal their email addresses to avoid them being harvested by spammers, and to opt out of receiving communications. Our second objective was to investigate what other factors influence the distribution of spam.

To jumpstart our experiment, we established email accounts at various Web-based email services for fictitious persons with various demographic characteristics (declared interest in particular products, age, and nationality). Over a period of 33 weeks, we monitored the resulting spam and analyzed the spam according to the personal characteristics.

Persons who declared interest in particular products received more spam than those who did not; those aged 30 received more spam than those aged 15; and U.S. residents received more spam than Singapore residents. Spam rates, however, did not differ across email accounts that were associated with men versus women. All of these findings support the hypothesis that spam is targeted at segments that are relatively more likely to make online purchases.[1]

Among the other factors that influenced spam rates, we found that spam was highest among Hotmail accounts, followed in decreasing order by Lycos, Excite, and Yahoo! accounts. Indeed, the identity of the email provider was the most important determinant of the spam rate. Consistent with previous studies, we also found that email addresses exposed through Web pages received more spam.

Our experimental procedure involved proceeding from the basis of several hypotheses, detailed as follows:

Hypothesis #1. Spam rates would be higher for persons with declared interest in some product or service than for those with no declared interest. The objective of spam is to promote sales. Hence, if spammers target their email messages, they should target the consumer segments more likely to purchase the item being promoted. However, if spam is randomly distributed, then consumers who are more likely to make online purchases should not receive any more spam than others.

We noted consumers as being relatively more likely to make online purchases in two ways. One way was for the person to explicitly state his or her interest in some product or service at the point of registration for the email account. Our other approach was to manipulate consumers' demographic characteristics, such as age, gender, and nationality. We relied on the Pew Internet and American Life Project [8], which provides a comprehensive picture of U.S. consumer behavior online.

Hypothesis #2. Spam rates would be higher in email accounts associated with individuals aged 30 than those aged 15. 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 [8].[2] The Pew Project did not even consider individuals aged below 18 in its e-commerce sample. An obvious reason is that they would not be eligible for credit cards.

Hypothesis #3. Spam rates would not differ for email accounts associated with men relative to women. The Pew Project found no significant difference in online consumer behavior by gender: "On any given day between March 2000 and December 2002, one would find roughly the same portion [sic] of men and women buying products online" [8].

Hypothesis #4. Spam rates would be higher in email accounts associated with U.S. than Singapore residents. With regard to nationality, the e-commerce participation rate is 22.7% among Singaporeans with Internet access [4] as compared with 61% among Americans [8].

Finally, to explore other factors that influence the spam rate, we considered the identity of the email service provider in addition to a known determinant: publication of the email address on a Web page.

In early August 2003, we created a total of 288 Web-based email accounts for fictitious persons at Excite, Hotmail, Lycos, and Yahoo. The persons were distinguished on the following dimensions:

[1]Men and women are equally likely to make online purchases; hence, spam should be targeted at them equally.

[2]Another reason could be income—people with higher income are more likely to shop online. According to the Pew Project [8], 49% of users in households earning $30,000 or less had tried shopping online, compared to 74% of those living with incomes of $75,000 or more.

Controls						
Age	**Gender**	**Residence**	**Mean**	**Standard Deviation**	**Maximum**	**Minimum**
15	Female	U.S.	4.583	5.728	13	0
15	Male	U.S.	4.750	5.956	13	0
30	Female	U.S.	4.667	5.867	14	0
30	Male	U.S.	4.833	6.177	16	0
15	Female	Singapore	4.417	5.551	13	0
15	Male	Singapore	4.583	5.728	13	0
30	Female	Singapore	4.500	5.697	14	0
30	Male	Singapore	4.500	5.697	14	0
Exposers						
Age	**Gender**	**Residence**	**Mean**	**Standard Deviation**	**Maximum**	**Minimum**
15	Female	U.S.	5.500	5.993	13	0
15	Male	U.S.	5.417	5.934	14	0
30	Female	U.S.	11.333	8.437	25	0
30	Male	U.S.	11.500	8.475	26	0
15	Female	Singapore	5.083	5.524	13	0
15	Male	Singapore	5.500	5.993	13	0
30	Female	Singapore	5.167	5.654	14	0
30	Male	Singapore	5.250	5.597	14	0

Table 1. Spam rates.

- Declared interests: computers and technology, travel, casino, or none,
- Age: 15 or 30,
- Gender: female or male,
- Residence: Singapore or U.S.

We created three accounts in each unique demographic combination, or a total of 4 (interests) x 2 (age) x 2 (gender) x 2 (nationality) x 3 (accounts) = 96 email accounts at Lycos and Excite. We created only 72 accounts in Yahoo as it did not offer casino gambling on its list of interests, and created only 24 accounts in Hotmail as it did not allow for the indication of interests at the point of registration. Hence, the total number of accounts created was 288.

For 192 "exposer" accounts (two in each demographic combination), we created a Web page that included the person's email address and other personal details at Yahoo Geocities.[3] In order to maintain an appearance of activity, we regularly sent email from these accounts. For the remaining 96 "control" accounts (one in each demographic combination), we did not construct a GeoCities Web page.

When establishing the synthetic email accounts, we accepted the default type and level of anti-spam tools. Of the email service providers that we used for the experiment, all except Excite provided a basic spam guard that directed suspected spam into a bulk folder.

Over the subsequent 33-week period, we monitored the number of unsolicited commercial email messages received ("spam rate") at each email address. The experiment concluded in March 2004.

Results and Analysis

Over the experimental period, the control accounts received an average of 4.60 (standard deviation 5.59) spam email messages, while the exposer accounts received an average of 6.84 (standard deviation 6.96). Table 1 reports descriptive statistics of the spam rates among the control and exposer accounts in the various demographic segments.

Most of the spam originated from the email service providers and their marketing collaborators. The source could be identified from statements or illustrations that marked their affiliation to the respective email service provider.

Our reported spam rates seemed low. However, they are reasonable given that the email accounts opted out of receiving special offers and other marketing communications. Further, the email accounts were not used to engage in any online transactions. If the accounts had not opted out, any subsequent commercial email would not have been "unsolicited." In August 2001, Jamal et al. [5] registered 200 email addresses in 69 top commercial Web sites. They opted out in 100 registrations. Over the subsequent 26 weeks, the 100 opt-out addresses received an average of 5.01 spam email messages. This is strikingly similar to the spam rate in our experiment. Their other 100 addresses received an average of 151.43 spam email messages.[4]

Jamal's experiment shows that email accounts that do not opt out will receive substantially more spam. Further, we conjecture that accounts used to engage in online transactions would also receive substantially more spam. These two factors probably account for most of the difference in spam rates between our synthetic accounts and those of real people.

We performed ordinary least squares regressions to test our hypotheses. For each email account, the quantity of spam was the dependent variable, and the various account characteristics were the independent variables. Table 2 reports results of the least squares

[3]U.S. Federal Trade Commission investigators seeded 250 email addresses across the Internet and observed the following rates of spam: 86% of addresses posted to newsgroups; half of addresses posted on free personal Web pages; 27% of addresses posted to message boards; and 9% of addresses listed in membership directories [9].

[4]The U.S. Federal Trade Commission's experiment attracted 3,349 spam emails to 250 email accounts or an average of 13.4 per account in six weeks [9].

Table 2. Regressions (Dependent variable: quantity of spam).

Independent variables	(a)	(b)	(c)
Constant	15.50*** (0.591)	13.35*** (0.570)	11.85*** (0.548)
LYCOS	-3.125*** [-0.223]*** (0.661)	-3.564*** [-0.255]*** (0.627)	-3.564*** [-0.255]*** (0.567)
EXCITE	-13.48*** [-0.963]*** (0.661)	-13.92*** [-0.994]*** (0.627)	-13.92*** [-0.994]*** (0.567)
YAHOO	-15.47*** [-1.015]*** (0.683)	-15.94*** [-1.046]*** (0.636)	-15.94*** [-1.046]*** (0.575)
GENDER	-	0.1389 [0.011] (0.294)	0.1389 [0.011] (0.266)
AGE	-	1.972*** [0.149]*** (0.294)	1.972*** [0.149]*** (0.266)
RESIDENCE	-	2.194*** [0.166]*** (0.294)	2.194*** [0.166]*** (0.266)
TRAVEL	-	0.6667 [0.044] (0.416)	0.6667* [0.044]* (0.376)
COMPTECH	-	0.7222* [0.047]* (0.416)	0.7222* [0.047]* (0.376)
CASINO	-	0.3657 [0.021] (0.480)	0.3657 [0.021] (0.434)
WEB PAGE (EXPOSER)	-	-	2.240*** [0.160]*** (0.282)
No. of observations	288	288	288
Adjusted R2	0.8078	0.8573	0.8833
F-statistic	403.09	192.62	218.32

Standardized coefficients in brackets [ ]; standard errors in parentheses ( ).
* significant at 99% level
** significant at 95% level
*** significant at 90% level

regressions. In column (a), we report a regression with just a constant and three variables indicating the various email service providers in which the accounts were created. Relative to Hotmail (the default email service provider), the coefficients of the three service provider variables were all negative and significant, indicating that their accounts received less spam than those registered with Hotmail.

Column (b) included additional variables characterizing differences among the persons associated with the email accounts. The results were partly consistent with Hypothesis 1. Accounts that declared interest in travel or computing and technology received significantly more spam than those that did not declare such interest.[5] However, accounts that declared interest in casino gambling did not receive significantly more spam than those that did not declare such interest. This result is consistent with Hypothesis 1 because U.S. law prohibits online gambling.

The empirical results were consistent with our second, third, and fourth hypotheses. The coefficient of AGE (0 = 15 years old; 1 = 30 years old) was positive and significant. The coefficient of GENDER (0 = female; 1 = male) was not significantly different from zero. The coefficient of NATIONALITY (0 = Singapore, 1 = U.S.) was positive and significant. We infer that Internet users aged 30 and U.S. residents received significantly more spam than those aged 15 and Singapore residents respectively, and men did not receive significantly more spam than women.

Finally, we investigated other factors that influenced spam rates. Table 2, column (c), includes an additional variable, EXPOSER (0 = no Web page; 1 = published a Web page on Yahoo GeoCities). The coefficient was positive and significant, which result is consistent with prior studies [1, 9].

Comparing Table 2, columns (a)–(c), the identity of the email service provider is evidently the most important influence on the spam rate. The model in column (a) accounts for over 80% of the variance in spam rates. Each of the variables representing a particular email service provider is significant at far above the conventional levels. The other explanatory variables—declaration of interest in product or service, age, gender, nationality, and exposure on a Web page—added less than 8% additional explanatory power.

Conclusion

We found that spam is not random, but quite systematically targeted at consumer segments that are relatively more likely to make online purchases—those who declare interest in particular products or services, adults, and U.S. residents.

Our most surprising finding was that, by far, the most important influence on the spam rate was the identity of the email service provider. Specifically, Hotmail accounts received significantly more spam than accounts set up with other email service providers. This effect was more important than declaration of interest, demographic factors, and whether the email address had been published on a Web page.

We should caution that this finding arose in a context where spam was truly unsolicited and almost all the spam originated from email service providers and their marketing collaborators. Further, the email accounts we created were not used to engage in any online transactions. Subject to these provisos, our results imply that consumers should take care in choosing email service providers and declaring inter-

[5]The coefficient of TRAVEL was just marginally short of significant in Table 2, regression (b).

ests when registering for an email account.

An important direction for future research is to extend our experiments by using some of the email accounts to engage in online transactions, and specifically, make online purchases. It would be important to observe the impact of these activities on the extent of spam received.

Policymakers, Internet service providers, software vendors, and scholars all over the world are struggling to devise technological, regulatory, and social solutions to spam. Our results contribute to these efforts by providing a better understanding of the business of spam. C

References

1. Center for Democracy and Technology. "*Why Am I Getting All This Spam? Unsolicited Commercial E-mail Research Six Month Report*", Washington, D.C., Mar. 2003; www.cdt.org/speech/spam/030319spamreport.shtml.
2. Gates, B. Why I hate spam. *Wall Street Journal*, (June 23, 2003).
3. Gates, B. "Preserving and enhancing the benefits of email—A progress report." Executive E-mail, June 28, 2004; www.microsoft.com/mscorp/execmail/2004/06-28antispam.asp.
4. Infocommunications Development Authority (IDA) of Singapore. *Annual Survey on Infocomm Usage in Households and by Individuals for 2002* (Aug. 5, 2003).
5. Jamal, K., Michael, M., and Shyam, S. Privacy in e-commerce: Development of reporting standards, disclosure and assurance services in an unregulated market. *Journal of Accounting Research 41*, 2 (May 2003), 285–309.
6. Loder, T.C., Van Alstyne, M.W., and Wash, R.L. An economic response to unsolicited communication. *Advances in Economic Analysis & Policy 6,* 1, Article 2.
7. Muris, T.J. The Federal Trade Commission and the future development of U.S. consumer protection policy. Aspen Summit: Cyberspace and the American Dream, The Progress and Freedom Foundation (Aug. 19, 2003).
8. Pew Internet & American Life Project. America's online pursuits: The changing picture of who's online and what they do. Washington, D.C. (Dec. 22, 2003).
9. U.S. Federal Trade Commission. Email address harvesting: How spammers reap what you sow. FTC Consumer Alert (Nov. 2002).

Il-Horn Hann (hann@marshall.usc.edu) is an assistant professor at the Marshall School of Business, University of Southern California.
Kai-Lung Hui (is_lung@cityu.edu.hk) is an associate professor in the Department of Information Systems, City University of Hong Kong.
Yee-Lin Lai (laiyeeli@comp.nus.edu.sg) is a research assistant in the School of Computing, National University of Singapore.
Sang-Yong Tom Lee (tomlee@hanyang.ac.kr) is an associate professor in the College of Information and Communications, Hanyang University, Seoul, Korea.
Ivan Png (ipng@comp.nus.edu.sg) is Kwan Im Thong Hood Cho Temple Professor in the School of Computing, National University of Singapore. He was a Nominated MP (10th Parliament of Singapore, Second Session), 2005–06, and is a member of the Trustworthy Computing Academic Advisory Board of Microsoft Corporation.