Why Am I Getting All This Spam?

Unsolicited Commercial E-mail Research
Six Month Report

Center for Democracy & Technology
March 2003

Summary

Every day, millions of people receive dozens of unsolicited commercial e-mails (UCE), known popularly as "spam." Some users see spam as a minor annoyance, while others are so overwhelmed with spam that they are forced to switch e-mail addresses. This has led many Internet users to wonder: *How did these people get my e-mail address?*

In the summer of 2002, CDT embarked on a project to attempt to determine the source of spam. To do so, we set up hundreds of different e-mail addresses, used them for a single purpose, and then waited six months to see what kind of mail those addresses were receiving. It should come as no surprise to most e-mail users that many of the addresses CDT created for this study attracted spam, but it is very interesting to see the different ways that e-mail addresses attracted spam -- and the different volumes -- depending on where the e-mail addresses were used.

The results offer Internet users insights about what online behavior results in the most spam. The results also debunk some of the myths about spam.

Major Findings

- Our analysis indicated that e-mail addresses posted on Web sites or in newsgroups attract the most spam.

    - Web Sites — CDT received the most e-mails when an address was placed visibly on a public Web site. Spammers use software harvesting programs such as “robots” or “spiders” to record e-mail addresses listed on Web sites, including both personal Web pages and institutional (corporate or non-profit) Web pages.

      CDT tested two methods of obstructing address harvesting:
        - Replacing characters in an e-mail address with human-readable equivalents, e.g. "example@domain.com" was written "example at domain dot com;" and

        - Replacing characters in an e-mail address with HTML equivalents.

      E-mail addresses posted to Web sites using these conventions did not receive any spam.

    - USENET newsgroups -- Newsgroups can expose to spammers the e-mail address of every person who posts to the newsgroup. Newsgroup postings, on average, generated less spam than posting an e-mail address on a high-traffic web site. In our study, we discovered that most newsgroup-related spam is sent to the address in the message header, even if other e-mail addresses are included in the text of the posting.

- For the most part, companies that offered users a choice about receiving commercial e-mails respected that choice. Most of the major Web sites to which we provided e-mail addresses respected the privacy choices we made -- when a choice was made available to us.

- Some spam is generated through attacks on mail servers, methods that don't rely on the collection of e-mail addresses at all. In "brute force" attacks and "dictionary" attacks, spam programs send spam to every possible combination of letters at a domain, or to common names and words. While these attacks can be blocked, some spam is likely to get through. In many cases, spam generated by these attacks will be directed to shorter e-mail address (like bob@domain.com) before it is directed to longer addresses (like bobwilliams@domain.com).

Tips for Avoiding Spam

Currently there is no foolproof way to prevent spam. Based on our research, we recommend that Internet users try the following methods to prevent spam:

- **Disguise e-mail addresses posted in a public electronic place.**
  CDT received the most spam just by placing an e-mail address at the bottom of a webpage. Spammers "harvest" these addresses with computer programs that collect and process addresses and add them to spam mailing lists. If a user must post his/her e-mail address in a public place, it is useful to disguise the address through simple means such as replacing "example@domain.com" with "example at domain dot com" or other variations such as the HTML numeric equivalent, in which "example@domain.com" could be written "example@d omain.com."

  Opt out of member directories that may place your e-mail address online. If your employer places your e-mail address online, ask the Webmaster to make sure it is disguised in some way.

- **Read carefully when filling out online forms requesting your e-mail address, and exercise your choice.**

  If you don't want to receive e-mail from a Web site operator, don't give them your e-mail address unless they offer the option of declining to receive e-mail and you exercise that option. If you are asked for your e-mail address in an online setting such as a form, make sure you pay attention to any options discussing how the address will be used. Pay attention to check boxes that request the right to send you e-mails or share your e-mail address with partners. Read the privacy policies of Web sites. If you suspect that a Web site has violated its privacy policy, you can report it to your state attorney general or the Federal Trade Commission.

- **Use multiple e-mail addresses**

  When using an unfamiliar Web site or posting to a newsgroup, establish an e-mail address for that specific purpose. Alternatively, instead of just using one or two e-mail addresses, you can use "disposable e-mail addresses," which consolidate e-mail in a single location but allow you to immediately shut off any address that is attracting spam. By recording which disposable address was used at which web site, one can track what sites are causing spam. Many Web sites are now providing free e-mail accounts. A search in Google Directory for "disposable e-mail addresses" provides a list of e-mail providers designed for one-time use e-mails.

- **Use a filter.**

  Many ISPs and free e-mail services now provide spam filtering. While filters are not perfect, they can cut down tremendously the amount of spam a user receives.

- **Short e-mail addresses are easy to guess, and may receive more spam.**

  At least one spammer tried to guess the e-mail addresses used in this study by sending mail to short and common addresses. E-mail addresses composed of short names and initials like bob@ or tse@, or basic combinations like smithj@ or toms@ will probably receive more spam. E-mail addresses need not be incomprehensible, but a user with a common or short name may want to modify or add to it in some way in his or her e-mail address.

For further information, please contact Ari Schwartz at the Center for Democracy & Technology, 202-637-9800, ari@cdt.org.

Why Am I Getting All This Spam?
Unsolicited Commercial E-Mail Research
Six Month Report

Introduction

Junk e-mail, a.k.a. spam, inconveniences tens of millions of Internet users and imposes huge costs on ISPs. Armed with lists of e-mail addresses, "spammers" send billions of e-mail messages every day -- messages that most users don't want.

It is often difficult or impossible to tell how a spammer acquired a user's e-mail address. Was it a result of some activity the user engaged in? Did the user give his/her e-mail address to the wrong person? Was the user randomly targeted? Are there steps the user could take to avoid such spam in the future?

This study attempts to answer some of these questions by analyzing common activities of Internet users and looking for evidence of some activities that resulted in one e-mail address receiving more spam than others. We do not believe that this report answers every question about spam, where it comes from, or how to stop it. However, by illuminating some of the ways that an e-mail address can be added to a spam list, the study provides users and policymakers with a better understanding of the problem and some guidance about how to better avoid spam in the future.

Methodology

The goal of this study was to understand whether certain kinds of Internet activities make a user an easy target for spam.

To determine how a person's e-mail address finds its way onto spam lists, CDT created hundreds of e-mail accounts and seeded the addresses in dozens of popular Internet locations.

Each e-mail address was used or posted in only one place; Table 1 summarizes the ways in which the addresses were used or posted. The addresses themselves were randomized, making it unlikely that a spam sender could guess them[1] -- one sample address was "m45k5e@egovtoolkit.org."[2]

---

[1] During the course of this project CDT's mail system suffered a "dictionary attack," in which a would-be spam sender attempted to guess every e-mail address on our system.

[2] We used the *egovtoolkit.org* domain for all addresses in this project. The domain is owned and operated by CDT, but is not presently used except internally. This was done to avoid the small chance that a spam-sender might recognize the *cdt.org* domain and treat those addresses differently from all others.

Table 1 - Usage Categories

Type of online activity	Control addresses	Experimental addresses
Public Web posting: * *www.cdt.org* * *www.getnetwise.org* * *www.consumerprivacyguide.org*	Addresses were posted on a publicly accessible Web page and left online for six months.	1. Address removed from Web two weeks after posting. 2. Address posted in "human-readable" form 3. Address posted in HTML-obscured form.
USENET: • *alt.internet.commerce* • *alt.health* • *alt.kids-talk* • *alt.news-media* • *alt.sex.erotica* • *alt.showbiz.gossip* • *misc.consumers.house* • *misc.industry.insurance* • *rec.gambling.misc* • *rec.humor* • *rec.travel.misc* • *soc.senior.issues* • *us.jobs*	Addresses were used in the headers of posted messages.	1. Address included in text in "plaintext" form 2. Address included in text in "human-readable" form 3. Address included in text in HTML-obscured form.
Web services: *Appendix 1 lists the Web-based companies and organizations to which e-mail addresses were provided.*	Addresses were provided to Web sites offering various online services using default and/or "opt-in" privacy preferences.	1. After two weeks, changed personal preferences to "opt-out" of future e-mail communication. 2. Upon receiving e-mail, unsubscribe request was submitted (where available).
Web-based postings • *amazon.com* • *careerbuilder.com* • *ebay.com* • *intelihealth.com* • *joehollywood.com* • *monster.com* • *popbitch.com* • *seniornet.org* • *thirdage.com* • *webmd.om*	Provided an e-mail address as part of posting to a Web-based job, auction, or discussion board.	1. Address included in text in "plaintext" form 2. Address included in text in "human-readable" form 3. Address included in text in HTML-obscured form.
WHOIS database • *netsol.com* • *npsis.com*	Provided an e-mail address as part of registering a ".com" or ".org" domain.	None.

The project examined five basic ways of posting or otherwise disclosing an e-mail address, and how each could affect the amount of spam a user received. The activities examined were: 1) posting an e-mail address on a public Web site; 2) posting an e-mail address on a USENET newsgroup; 3) providing an e-mail address to a popular Web site in connection with some service; 4) providing an e-mail address to a popular Web site in order to post on a job, auction, or

discussion board; and 5) posting the address in the WHOIS database of information about domain name registrants.

In each area of online activity, we created a set of "control" addresses, provided in a straightforward manner with no attempt to avoid future spam, and one or more "experimental" addresses, each of which utilized a particular anti-spam measure.

Experimental Anti-Spam Measures

1. Removal from public accessibility. A number of e-mail addresses were posted on publicly accessible Web sites for two weeks, then removed. The goal was to determine whether removing the address from public view would have an effect on the overall amount of spam received.

2. Posting in "human-readable" form. Some Internet users posting their addresses in public places have altered the form of their e-mail address in such a way that another user can still easily reach them, but an automated tool would not recognize them. For example, a user with e-mail address *example@domain.com* could post his address as "example at domain dot com." We tested the effectiveness of this practice by posting addresses on the Web and on USENET newsgroups in this "human-readable" form.

3. Posting in HTML-obscured form. Tech-savvy Internet users have sometimes used special codes in HTML -- Hypertext Markup Language, used to construct Web pages -- to post their addresses in a way that Web browsers can interpret, but that is an obstacle to automated spam tools. In HTML, the letter "e" can be written "e" and the "@" symbol "@." So, the address "example@domain.com" could be written "example@domain.com."[3] We tested the effectiveness of this practice by posting addresses on the Web and on USENET in this HTML-obscured form.

4. Changing personal preferences on a Web site. Many Web sites provide users with the opportunity to alter their personal preference so that they no longer receive e-mail communication from that site. Some Internet users, however, have been concerned that changing those preferences will have little effect on the amount of spam received, believing that once an address is "out," there is little they can do about it. We tested the effectiveness of changing one's personal preferences by returning to Web sites to which we'd submitted e-mail addresses and changing the addresses' associated preferences to request no further e-mail communication. We tried this in two separate ways. For certain addresses, we would "opt-in" to certain kinds of communication, then log back in and

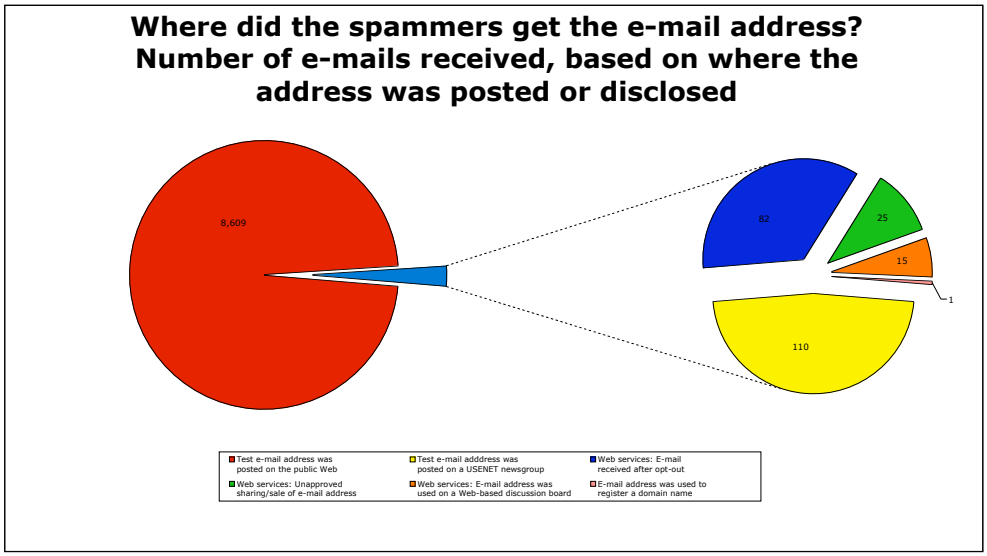
[3] If you'd like to obscure your e-mail address, or any other piece of text, try the free E-mail Address Encoder at <http://www.wbwip.com/wbw/emailencoder.html>.

immediately change our preferences to "opt-out." For another set of addresses, we allowed at least two weeks to elapse before changing preferences. In both cases, we allowed a two-week "grace period" for our changes to take effect before classifying received e-mails as spam.

Results

In six months of operation, our project received over 10,000 e-mail messages to the more than 250 single-use e-mail addresses we created. About 1,600 of these were legitimate e-mail communications that we'd requested from various online services. Another 62 were unclassifiable due to incomplete e-mail headers or other missing data. And 16 messages were received after we'd opted-out of future communications from a business we'd given an e-mail address to, but were received within a two-week "grace period" that our methodology allowed. We classified the remaining 8,842 as unsolicited, a.k.a. spam, e-mail.

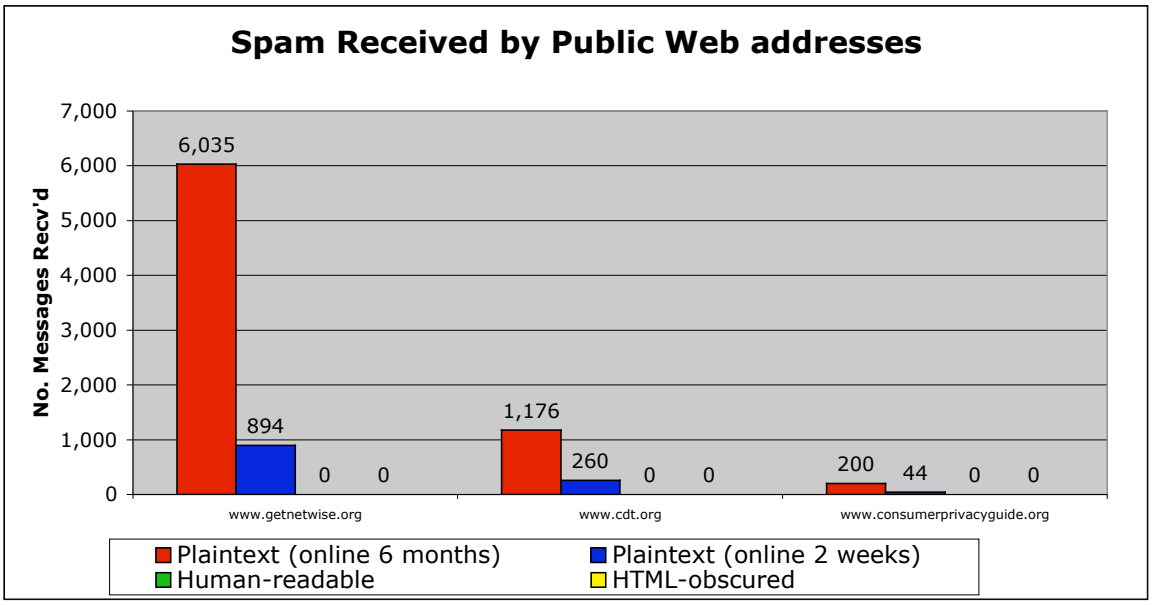
Figure 1 - Sources of addresses used by spammers



1. Addresses Posted on the Public Web

The vast majority of the spam we received -- over 97% of it -- was delivered to addresses that had been posted on the public Web.

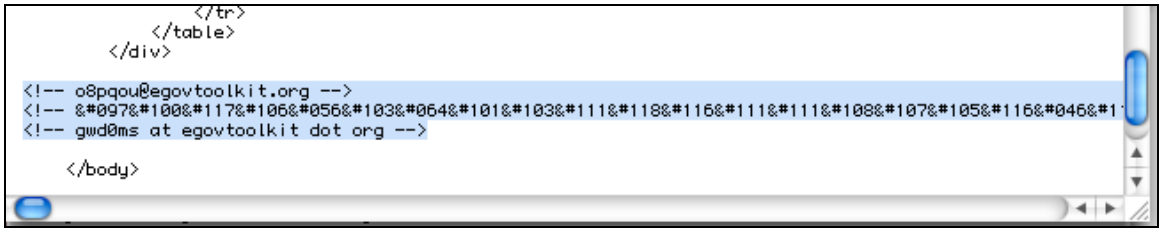
Figure 2 - Messages received by addresses on the public Web



All the plaintext e-mail addresses we placed on the public web received some spam. The number of messages received seems to be related to the popularity of the web site. GetNetWise.org is a well-known online safety site that is linked to by major portals like AOL and Yahoo!, and the addresses posted there received a lot of spam, while ConsumerPrivacyGuide.org is a relatively new site, and addresses posted there received much less spam.

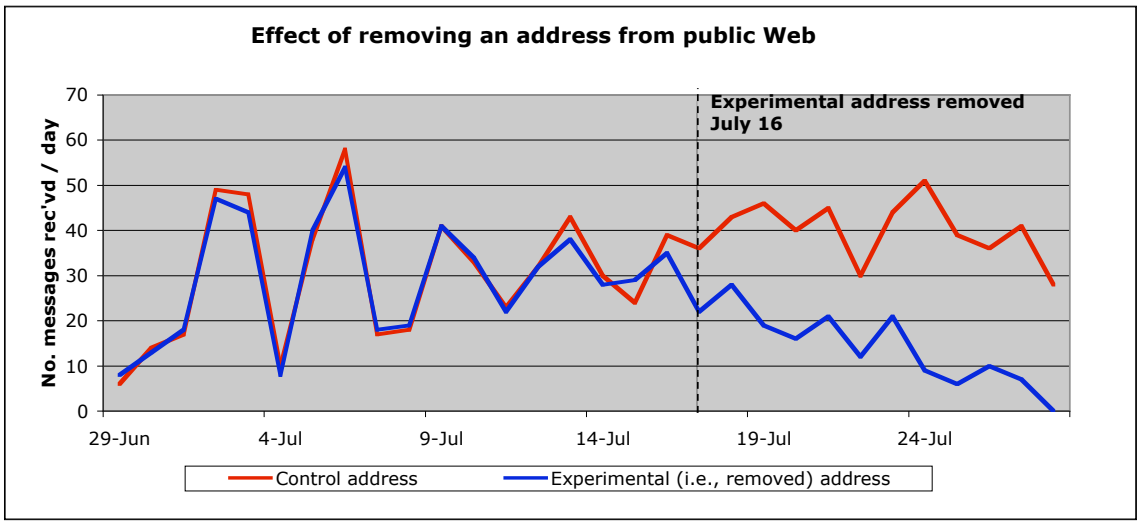
But none of the addresses that were obscured, whether in "human-readable" or "HTML-obscured" form, received a single piece of spam, leading us to conclude that e-mail address "harvesters" are not presently capable of collecting such addresses. While this may change as time passes and technology develops, for the time being it appears that obscuring an e-mail address is an effective means of avoiding spam.

Figure 3 - Sample HTML code from GetNetWise.org/index.html



Two weeks after placing our test addresses on the public Web, we removed some of them in order to determine how long an e-mail address, once placed on the public Web, would continue to receive spam after its removal. The effect was significant for all three Web sites tested.

Figure 4 - Effect of removing an address from the public Web



Over the remainder of the study, the address that had been removed from the public Web received significantly less spam than the address still on the Web.

2. Public Postings to USENET Newsgroups

The second-greatest amount of spam we received was from public postings to USENET newsgroups. Once again, we posted addresses in plaintext, "human-readable," and "HTML-obscured" form.

Figure 5 - Sample USENET posting with e-mail addresses in plaintext, human-readable, and HTML-obscured form

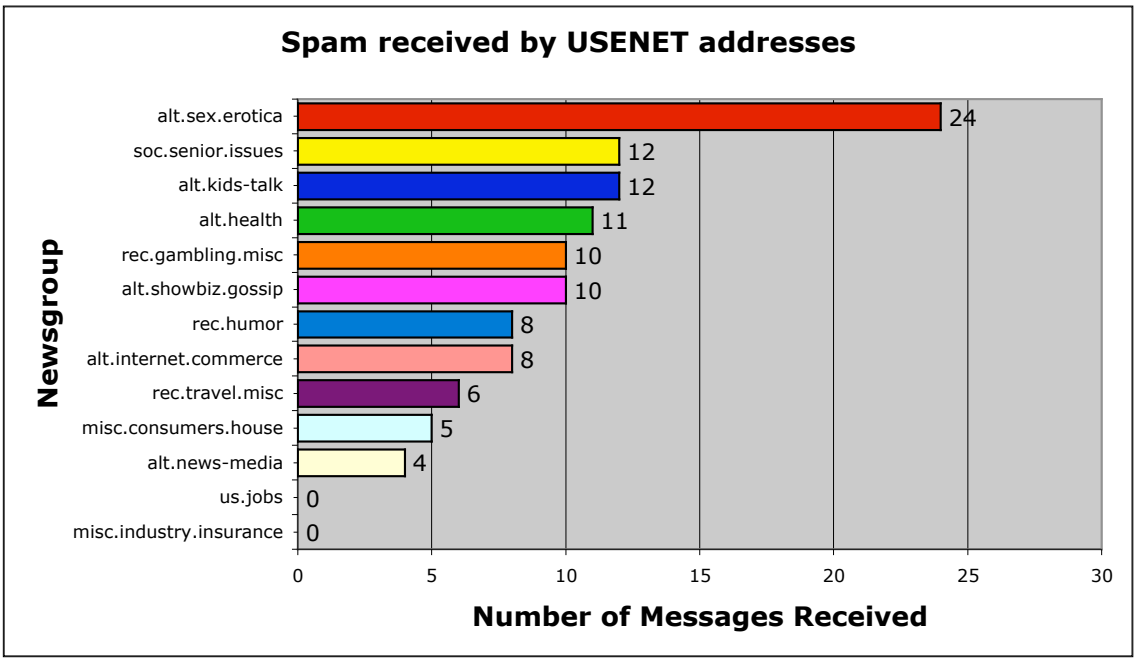
```
From: CDT Test <house.email@egovtoolkit.org>
Newsgroups: misc.consumers.house
Subject: Please Ignore, Test Message
Date: Sun, 23 Jun 2002 17:39:03 -0400
Lines: 8
Message-ID: <cvfchu08h1rts4n3ej8hh3coi26qqnvuae@4ax.com>
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-Trace: UmFuZG9tSVbBOp1UAr9/m5nxxKc2CdzWINfT778C/iz9M5iCzS9SRVk+hr0URMK6
X-Complaints-To: abuse@rcn.com
NNTP-Posting-Date: 23 Jun 2002 21:39:08 GMT
X-Newsreader:  Forte Agent 1.91/32.564

This posting is part of a research project at the
Center for Democracy and Technology, please
ignore.

Justin Cohen
house.plain@egovtoolkit.org
house.english at egovtoolkit dot org
&#104&#111&#117&#115&#101&#046&#104&#101&#120&#064&#101&#103&#111&#118&#116&#111&#111&#108&#107&#10
```

Once again, neither the "human-readable" nor the "HTML-obscured" e-mail addresses received any spam. And while not every message posted to USENET generated spam to the plaintext address we provided, most (85%) did.

Figure 6 - Messages received by addresses on USENET newsgroups



For the vast majority of the spam we received due to USENET postings, messages were sent to addresses referenced in the message header, not to addresses referenced in the text of the message. In a very few cases (<1% of all USENET-related spam we received), messages were sent to addresses referenced in the message text. In all cases, spam was sent to addresses that were included in plaintext, not obscured in any way.

The chart above indicates the distribution of spam we received relative to the newsgroups to which we posted. While "alt.sex.erotica" generated twice as much spam as the next newsgroup, we do not believe that this data supports any strong conclusion regarding which newsgroups are the most susceptible to spam.

3. Consumer Preferences

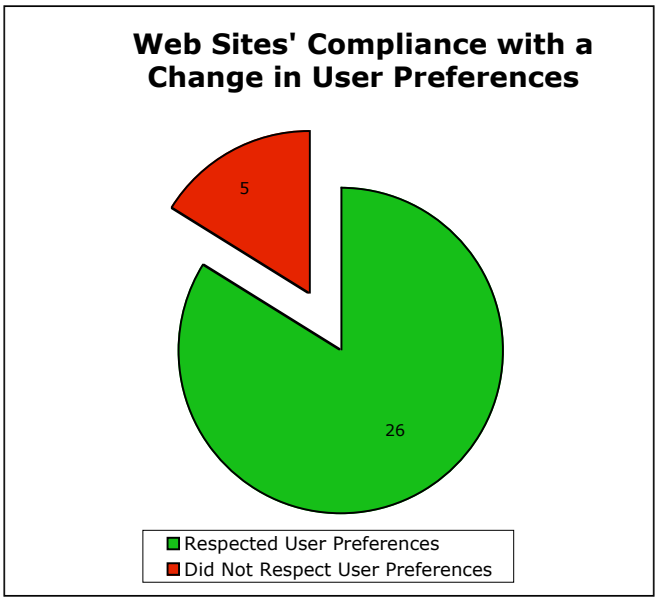
The third area we tested was the degree to which Web companies respected consumer attempts to opt out of receiving commercial e-mail.

First of all, in all of the cases where we disclosed an e-mail address and asked not to receive commercial e-mail, the Web site operator respected that request -- we received no spam when we opted out when first giving our e-mail address. In a variation on this test, we changed our preference from permitting commercial e-mail to opting out of it. Any e-mail we received more than two weeks after an attempt to "opt-out" was classified as spam. We tested two different kinds of opt-out: first, opt-out immediately after opting-in (simulating a consumer changing his/her mind immediately about his/her privacy preferences), and second, opt-out

two or more weeks after the initial opt-in (simulating a consumer changing his/her mind after some time).

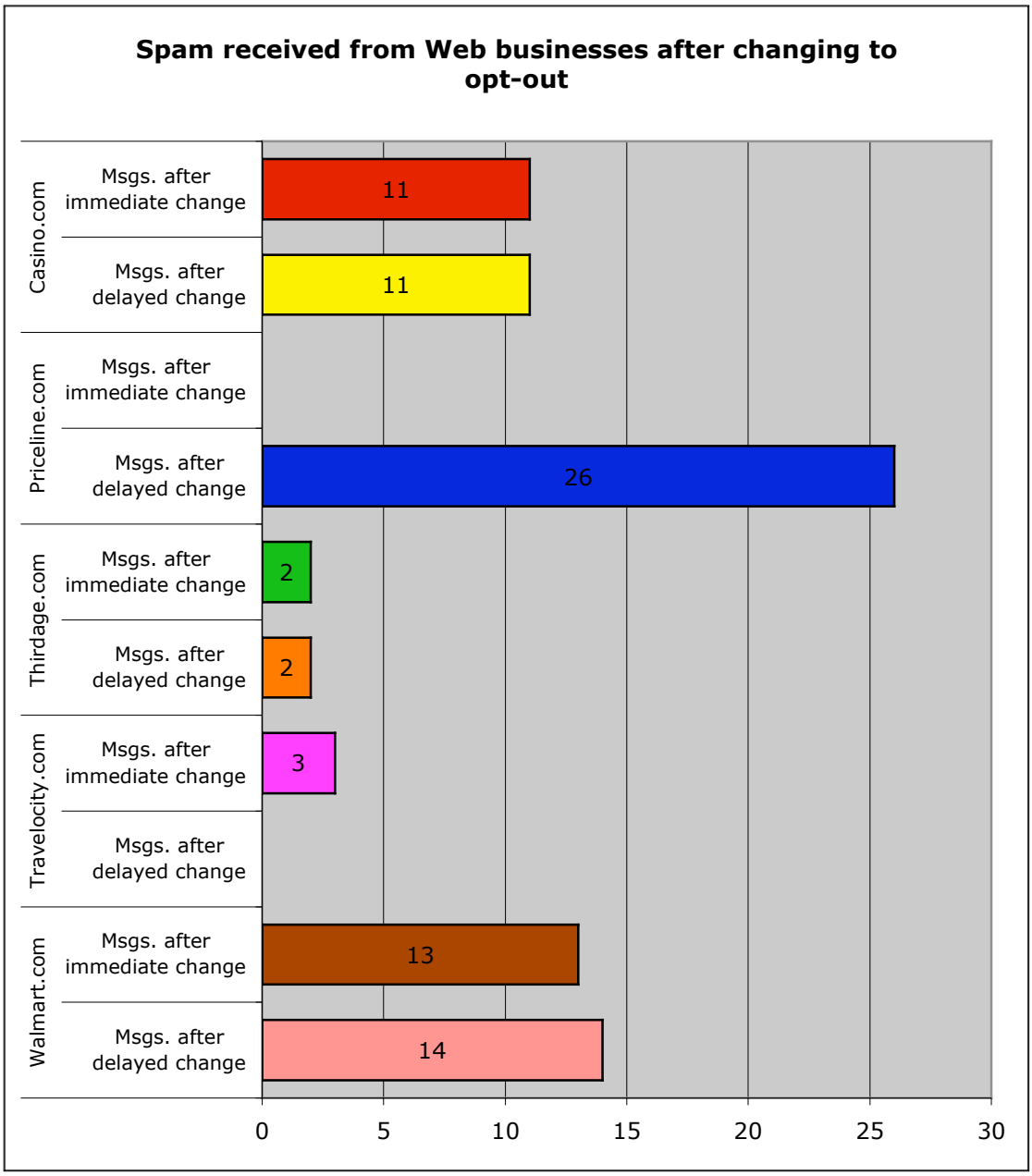
We pursued this methodology with thirty-one Web businesses and other organizations with myriad offerings.

Figure 7 - Web sites' respect for a change in user privacy preferences



For the majority of Web sites we encountered no difficulty and found that "opt-outs" were respected within the two-week grace period our methodology provided. In five cases, however, opt-outs were not respected, and a total of eighty-two "spam" messages were received from the companies well after an opt-out request had been submitted.

Figure 8 - Messages received after changing to opt-out from further communication



Our study also tested whether Web companies and other organizations shared or sold e-mail addresses in inappropriate ways. For this study, "inappropriate" was defined as sharing/sale (i) without notice to the consumer (in the form of a Web site privacy policy, or some other notice), (ii) in contradiction to the terms of the stated privacy policy, or (iii) in contradiction to the personal preferences we entered.

In general, we found inappropriate sharing/sale of e-mail addresses to be limited. We received just twenty-five such messages, mostly from gambling- and adult-content related websites.

4. Web Discussions

We also reviewed how much spam might result from a user's participation in a Web-based discussion board. In most cases, no spam was received. The only exception was an e-mail provided to *Intelihealth.com*, which generated fifteen spam messages.

5. Domain Name Registration

When a user registers a domain name in one of the Internet's seven global Top-Level Domains or certain country-code Top-Level Domains, his or her contact information is entered into a publicly accessible database known as the WHOIS database. We tested how much spam would be received to an address provided in the WHOIS database. Contrary to our expectations, just one spam e-mail was generated in the six months that our project was operational. Since WHOIS records are permanent, however, it is possible that additional spam could still be generated. Additionally, it should be noted that in the six months that this project was operational, none of the domains in question were up for renewal -- anecdotal reporting from many Internet users describes a significant increase in spam when renewals draw near.

6. Mail Server Attacks

Finally, at one point in the project our mail system began receiving spam messages to addresses that had never been used for any purpose, had been submitted to no one and, in many cases, did not even exist. By reviewing the server logs, we determined that our system had been the victim of a "brute force attack" in which a spammer had attempted to send e-mails to every possible combination of letters that could form an e-mail address.

Figure 9 - Example addresses used in a brute-force attack

a@egovtoolkit.org	aa@egovtoolkit.org	aaa@egovtoolkit.org	aaaa@egovtoolkit.org
b@…	ab@…	aab@…	aaab@…
c@…	ac@…	aac@…	aaac@…
d@…	ad@…	aad@…	aaad@…

The strain of so many e-mails severely impaired our mail server, and our team decided to install a block that would prevent any more messages from the responsible network (in this case, *h8h.com*) from entering our server. Our system received 8,506 "brute force" e-mails before the block was installed. Few, if any, of these e-mails actually made their way to existing e-mail addresses. In order to maintain the integrity of our conclusions, we did not include these 8,506 messages in the data above.

Conclusions

1. E-mail addresses harvested from the public Web are frequently used by spammers. By an overwhelming margin, the greatest amount of spam we received was to addresses posted on the public Web.

When an address has been posted on the public Web, it can potentially be viewed by hundreds of millions of users. People who develop spam lists exploit this feature by using address-harvesting programs to surf across thousands of web sites, collecting any e-mail addresses that they encounter. Most users have no idea that their addresses have been harvested until they begin receiving spam.

2. The amount of spam received by an address posted on the public Web is directly related to the amount of traffic that Web site receives. The more visitors a Web site has in a given period of time, the greater the likelihood that an address-harvesting program used to send spam will scour it. As a result, addresses posted on high-traffic Web sites are likely to receive a greater amount of spam than address posted on smaller sites -- popular Web sites are more frequently "harvested," and addresses posted on those Web sites are added to a greater number of spam lists.

3.E-mail addresses harvested from the public Web appear to have a relatively short "shelf life." When e-mail addresses we posted on the public Web were removed, there was a pronounced drop in the amount of spam they received each day. The change was not absolute -- on a given day, an address might receive a few spam messages even months after it had been removed from the public Web. But such spam was on the order of 2 or 3 messages per day, compared to the thirty or more messages received by addresses still on the public Web.

4. Addresses posted in the headers of USENET messages can receive significant spam, though less than a posting on the public Web. Like most Web sites, USENET postings are publicly accessible and may be targeted by e-mail address-harvesting programs. When a user includes his or her address in the heading of a USENET message, that address can be harvested and used to send spam. Our preliminary data indicates that some USENET newsgroups are more frequently harvested for e-mail addresses than others.

5. Obscuring an e-mail address is an effective way to avoid spam from harvesters on the Web or on USENET newsgroups. Even when posted in publicly accessible areas, none of the addresses we obscured -- whether in English ("example at domain dot com") or in HTML -- received a single piece of spam. Users who want to avoid spam should consider obscuring their addresses when possible.

6. Sites that publish their policies and make choice available to users generally respected those policies. A major element of the CDT project was to submit e-mail addresses to a number of popular businesses and other organizations on the Web. Many of these sites had privacy policies describing how they handle e-mail addresses and other potentially sensitive pieces of information. While the terms of these policies varied, we found that almost all sites followed their policies. In addition, when consumers were offered choices about how their personal information would be handled, those choices were respected.

7. Domain name registration does not seem to be a major source of spam. Despite the fact that the WHOIS database is publicly accessible, our project received just a single spam message to an address that was in WHOIS for six months. This leads us to believe that, at least for some people registering new domain names, listings in the WHOIS database may not be a major source of spam. However, because our project had a relatively short duration, we were not able to examine whether additional spam would be received as a domain name approached its renewal date.

8. Even when an e-mail address has not been posted or shared in any way, it is still possible to receive spam through various "attacks" on a mail server. In our study, a "brute force" attack on the mail server generated a tremendous amount of spam, even to addresses that hadn't been shared anywhere. Anecdotal evidence from network operators indicates that such attacks are not uncommon, and that while alert network administrators can sometimes block them, a significant amount of spam can still result. Sometimes, these attacks take the form of "dictionary attacks," in which the attacker sends e-mail to all the words in the dictionary, or attacks in which e-mail is sent to common surnames and first initials (such as "jsmith" or "bjones"). For individual Internet users, there is little that can be done to avoid the spam that may result from such attacks.

For further information, contact Ari Schwartz at the Center for Democracy & Technology, 202-637-9800, ari@cdt.org.

Appendix 1: Service Providers to Whom E-Mail Addresses Were Provided

Web services:

a-bad-credit-loans.com
amazon.com
careerbuilder.com
casino.com
cnn.com
democrats.org
easylaugh.com
ebay.com
expedia.com
gambling.net
intelihealth.com
joehollywood.com
joker.org
libertymutual.com
lp.org
macys.com
monster.com
moving.com
msnbc.com
nakedmail.com
popbitch.com
pornmail.org
priceline.com
reformparty.org
rnc.org
seniornet.org
statefarm.com
thirdage.com
travelocity.com
walmart.com
webmd.com