Receiving spam depends on the occurrence of the e-mail address in Hyper Text and on the nature of the link to the file containing the e-mail address.

Robert B. Mellor
Faculty of Computing, Information Systems & Mathematics
Kingston University
London KT1 2EE

Keywords
Spam, e-mail, marketing, HTML, JavaScript.

Abstract

Using a case study, spam levels were recorded for e-mail addresses embedded in the companies web site (HTML files) as "mailto" links, either with or without the address being shown in the Hyper Text. Spam was received at addresses not stated in the Hyper Text. However including the address in the Hyper Text resulted in approx. 17-fold more spam being received, including malicious (virus-containing) spam. Addresses on HTML files linked to using JavaScript did not receive spam. This led to the hypothesis that "ripper" software (software that extracts e-mail addresses from HTML files) can follow conventional HTML anchor tags, but cannot follow JavaScript links. This hypothesis was tested using a dummy web site before being confirmed on the case web site.

Introduction.

Clearly everyone who uses the Internet as a marketing channel has an interest in knowing how to avoid receiving time-consuming spam. Companies are realizing that spam takes up a large percentage of their contact resources, and are starting to incur both direct expenses by e.g. buying anti-spam software, or indirect expenses by removing customer-contact e-mail addresses from their web sites. This study investigates the roots of spam by looking at which e-mail addresses receive spam, using both a case study commercial web site as well as a contrived model web site.

Bulk e-mailing is a marketing tool supposed to increase sales (Tomasula, 2002). The point of bulk mailing, and bulk e-mailing, is to offer a genuine and useful service to those who have asked for it and thus harvest customer gratitude and goodwill. Sending mail or e-mail to people who have not asked for it, who are not interested in the product and who have never heard of the producer, will probably result in the exact opposite. Despite this, many millions of e-mails are sent daily describing unwanted products to uninterested receivers, and this problem has been significant for

some time (Cranor & LaMacchia, 1998, and Denning, 1982). Such e-mails are given the name "spam". Spam is the inappropriate use of a mailing list, Usenet or other networked system to send the same message to a large number of people who didn't ask for it. The term arose from a famous Monty Python sketch that features the word "spam" over and over and reflected their low opinion of the food product with the same name, which they perceived as a content-free waste of resources. Spam, the processed meat product, is a registered trademark of Hormel Corporation.

E-mail is sent by various protocols. E-mails routed between servers (i.e. between domains) travels by SMTP (Simple Mail Transfer Protocol), whereas e-mail sent inside a domain, e.g. between work colleagues, can be sent by various other protocols, e.g. IMAP (Internet Mail Application Protocol). In 1999, Wood (Wood, 1999) estimated that "spam" accounted for around 15% of all e-mail messages sent. In 2003 the Center for Democracy and Technology estimated that 55% of all e-mails sent are spam (CDT, 2003) and Mellor (2003) estimated that for SMTP, the figure had risen to 90%. The difference between these latter two estimates lies in that CDT took all types of e-mail (including both IMAP and SMTP), which led to a dilution effect, because e-mails sent between members of one domain are rarely spam (or rarely perceived as spam). However the scale of spam is enormous, the British Broadcasting Corporation (BBC, 2003) report that AOL (America On Line) are deflecting 1 thousand million spam messages each day sent by SMTP to their subscribers.

Often, spam originates with unscrupulous persons who typically use a type of software called a "ripper" or "spambot". This software crawls the Internet in a random fashion and extracts any e-mail addresses found in HTML (Hyper Text Markup Language) files. Lists of e-mail addresses found are subsequently compiled and used either to send out marketing offers from third persons, or are sold to the naive and unwary, who then use it to send their own marketing offers. Such lists may contain up to 500000 e-mail addresses.

How successful is spam? The answer can be no more than a guess. Mellor (2005) described one successful company who sent highly crafted e-mail messages, containing details of special- and last minute offers, on a quarterly basis to those who had expressly subscribed to the service. Despite this high degree of precision, the company revealed that on average over four years the sales response was only one per 21870 successful e-mails (Mellor, 2004). The sales rate connected to spam advertising may be 100 times smaller, i.e. roughly one sale per 2.2 million e-mails and permission-based business-to-consumer e-mail lists had an average price of $170/M in April 2004 (Courtenay Communications Corporation, 2004). This may thus explain why spam is so prevalent; the sales rate resulting from spam advertising is so tiny that the lists have to be used again and again in order to get any return on investment. Indeed, with an optimistic profit of 10 dollars per sale, and the above sales rate, spam senders must send 7.3 million e-mails per day to make 1000 dollars a month before tax (and before covering outlay).

The sheer volume of spam leads to further complications. Most spam senders forge message headers in order to suggest that the message originated from a different, normally fictitious, address. This is to avoid millions of irate replies. Similarly "remove" links in these e-mails seldom work, because, if they did, then the server would crash under the weight of the remove messages.

Methods.

Data Sources.

All e-mail addresses and domain names have, where appropriate, been made anonymous for the purpose of this study. Of the e-mail addresses used, none had an active anti-spam filter applied, and none of the e-mail addresses have ever been knowingly published on Usenet or other networked system. All addresses were published on the Internet in HTML files on UNIX Apache platform software and Robots.txt specified that all files are to be crawled.

Addresses were derived from two web sites. The first group of e-mail addresses were all derived from the same commercial UK-based company, (this case company is here called company C, using the domain here called xx.org.uk) which has operated with these addresses since February 2000. All these e-mail addresses are based on the same provider. The usage of the various e-mail addresses is shown in the following table.

Usage	**c@xx.org.uk**	**d@xx.org.uk**	**e@xx.org.uk**	**f@xx.org.uk**
Technical address (not published as a HTML link)	Used mostly for submissions, but also help desk for customers.			
Administrative address (not published as a HTML link)		Administrative mail to customers		
Mail to customers and as link on one HTML page			Link only in the Markup Language e-mail	
General mail to customers and as link on one HTML page				Link in both Hyper Text and in Markup Language f@xx.org.uk

Table 1: *Overview of addresses under dot.org.uk Top Level Domain, where the HTML files were first published in February 2000.*

The second web site group was derived from a dot.org.uk domain (here called yy.org.uk), which was registered for the purposes of this study, and a corresponding web site that was opened on a virtual server ("web hotel") in May 2003. This "dummy" web site consisted of 7 HTML files. Index.html contained only five hyperlinks to five different HTML files, called 1.html, 2.html, 3.html, 4.html and

5.html (the links occurred in the file in reverse order, i.e. from 5 to 1). A sixth file, 6.html, was an "orphan URL", in that index.html did not link to it using a HTML anchor tag. These six files contained nothing apart from an e-mail address link in a different format (for general background see standard DHTML textbooks, e.g. Mellor, 2002).

Address	File	Code	Remarks
1@yy.org.uk	1.html	`<a href="mailto:1@yy.org.uk">1@yy.org.uk</a>`	The address written in both Hyper Text and in Markup Language
2@yy.org.uk	2.html	`<a href="mailto:2@yy.org.uk">e-mail</a>`	The address written only in Markup Language
3@yy.org.uk	3.html	`<p>3@yy.org.uk</p>`	The address written only in Hyper Text

4@yy.org.uk	4.html	`<form action="../cgi-bin/mail.pl" method="post">`
		`<input type="hidden" recipient="4@yy.org.uk">`
		`<input type="text">`
		`<input type="submit" value="send"></form>`
			The address written only in Markup Language, but in a Form Field and not as a "mailto"

	5.html	`<script language="JavaScript">`
		`function popup(URL)`
		`{window.open(URL,'popup')}`
		`</script>`
		`<A href="javascript:popup('6.html)">e-mail</A>`
			Click the link and 6.html pops up in a new window.

6@yy.org.uk	6.html	`<a href="mailto:6@yy.org.uk">6@yy.org.uk</a>`	The same as in 1.html, but at an orphan URL
7@yy.org.uk	Used in submitting the web site to various search engines		

Table 2: *Overview of addresses delegated in May 2003 under dot.org.uk Top Level Domain.*

Data gathering.

Data was gathered on xx.org.uk from the 01 January 2003 to the 21 July 2003 and on yy.org.uk from 01 July 2003 to 01 January 2004. Confirmation of results on xx.org.uk was from 01 February 2004 to 01 April 2004. All spam arriving was copied to a special folder for each of the various mailboxes. The number of mails containing virus were subtracted, and the remainder counted and the daily average calculated. Statistical variations are not given, because not all e-mail accounts were checked every day.

Other methods.

Automated keyword submissions to search engines used Submit Wolf version 4 (www.trellian.com) starting on the 01 January 2003 for xx.org.uk and on the 01 July

2003 for yy.org.uk and every week thereafter. Computers were protected by continually updated versions of Norton Anti-Virus (www.norton.com).

Results.

The amounts of spam received by the various addresses were as follows.

Address	Spam (av. number of mails/day)	Total number of virus attacks
1@yy.org.uk	6.77	17
2@yy.org.uk	0.40	0
3@yy.org.uk	5.82	5
4@yy.org.uk	0.41	0
6@yy.org.uk	0	0
7@yy.org.uk	0.73	0
c@xx.org.uk	0.62	0
d@xx.org.uk	0	0
e@xx.org.uk	0.32	0
f@xx.org.uk	5.52	23

Table 3: *Amount of spam received at addresses contained in HTML files*

It can be seen that:

- An address, which was not published on the web site, did not result in spam, likewise simply sending e-mail (as d@xx.org.uk) did not result in spam.
- Writing the address in Hyper Text and in Markup Language resulted in relatively high amounts of spam (1@yy.org.uk/6.77 and f@xx.org.uk/5.52)
- Writing the address in Hyper Text only, without the corresponding address in Markup Language, also attracted large amounts of spam (3@yy.org.uk/5.82).
- However writing in only Markup Language, without Hyper Text, attracted lower, but still significant, amounts of spam (2@yy.org.uk/0.40 and e@xx.org.uk/0.32)

In order to check these results, the Danish Veterinary and Food Administration was approached. This was because they present two e-mail addresses on every HTML file of the ca. 8000 HTML files on their web site (www.fdir.dk), where one is in Hyper Text and Markup Language, whilst the other one is directly adjacent to it and only in Markup Language.

```
<a href="mailto:fdir@fdir.dk">fdir@fdir.dk</a>  
 <a href="mailto:web@fdir.dk">Webmaster</a>
```

Because it is a government ministry, neither of these e-mail addresses could possibly have been used in Usegroups or in connection with advertising. Despite this web@fdir.dk received 1.64 spam e-mails daily and fdir@fdir.dk received 28.21 spam mails daily during the time period 01 January 2003 to the 21 July 2003.

It must therefore be presumed that the results presented above, despite being derived from a small number of addresses, some of which have been artificially made for this study, probably do reflect the real situation.

Discussion.

Types of spam

Spam can be divided into four categories. These four categories ignore a "lowest" category where the sender is often the producer of the goods or services involved. This is because mailing is by means of a "home made" mailing list where the customer has expressly subscribed to receive marketing information on the goods or services involved and "remove" options are honoured. Such a service cannot be described as spam. These four categories likewise do not include malicious virus attacks caused by e.g. e-mail worms, because such attacks do not have a root in marketing, and are therefore likewise not spam. The four categories of spam are:

Category	Characteristics
1. Benevolent spam	The sender is often the producer of the goods or services involved. Mailing is by means of a commercially acquired mailing list. "Remove" options are normally honoured.
2. Submission spam	Bulk keyword submissions to e.g. search engines by means of automated software may often result in simultaneous subscription to e-zines, newsletters and suchlike. The submitter has thus subscribed, albeit unknowingly. "Remove" options may be honoured.
3. Hopeless spam	The sender is a semi-professional "middle man" sending descriptions of products or services derived from other sources and where mailing is by means of a mailing list acquired either commercially or directly by "ripping". "Remove" options are seldom honoured.
4. Malicious spam	The sender is a malicious prankster sending a virus or worm via mail where the mailing list has been acquired directly by "ripping" or by randomly combining letters to form an e-mail address ("brute force" attack). "Remove" options do not exist.

Table 4: *Four categories of spam*

In this study type 4 spam could not be estimated from the web mail addresses, since anti-virus scanning at the mail server level automatically protected these. Only

c@xx.org.uk and 7@yy.org.uk were exposed to Type 2 spam. The nature of Types 1 and 3 spam means that all addresses were exposed to these types.

A new European law (Eur-Lex, 2002) on spam came into force on the 31 October 2003. This law states that bulk e-mails that include a "Remove" link are legally acceptable, although disguising or falsifying the sender address is illegal. This law is primarily directed at Type 1 and Type 2 spam, and future research will be aimed at monitoring its effect on spam, especially Type 3 spam (it is clearly not applicable to Type 4 spam).

In the USA, the Senate Commerce Committee on June 19 2003 approved a revised version of the Burns-Wyden CAN-SPAM Act (CAN-SPAM, 2003). The bill would require spam senders to offer Internet users the opportunity to "opt-out" of further commercial e-mail (Types 1, 2 and perhaps type 3 spam). The bill also would impose penalties for sending commercial email with false header information or misleading subject lines and the harvesting ("ripping") of addresses from the Web (Type 3 spam), and for using other computers to relay spam without authorization (Type 3, overlapping with Type 4, spam). The final outcome of this lawmaking process remains uncertain.

Clearly sending e-mail to customers and business associates does not provoke spam responses (d@xx.org.uk). Indeed even using an address in bulk keyword submissions (c@xx.org.uk and 7@yy.org.uk) to a wide spectrum of search engines provoked surprisingly little spam (Type 2 spam).

Clearly writing the e-mail address in Hyper Text (with or without Markup Language) attracted the most spam (Types 1 and 3 spam) and the most virus attacks (Type 4 spam). This was shown by 1@yy.org.uk, 3@yy.org.uk, f@xx.org.uk and is in agreement with the CDT report (CDT, 2003), which also recommended that hypertext renderings of e-mail addresses should be in human readable form (i.e. "f at xx dot co dot uk") only.

However "ripper" or "spambot" software could still detect e-mail addresses where they occurred in Markup Language only (2@yy. org.uk and e@xx.org.uk), although addresses without Hyper Text received around 17 times less spam than those with Hyper Text. This relationship was confirmed by the figures from the Danish Veterinary and Food Administration, despite the differences in Top Level Domain (in this case, dot.dk) and in web site size. The ability to extract e-mail addresses from Markup Language was not connected with the occurrence of the "mailto" code, since 4@yy.org.uk attracted similar amounts of spam, even though this address was encapsulated in a Form Field. In fact, it is perhaps remarkable that “ripper” software managed to extract it at all.

Most encouraging is that 6@yy. org.uk received no spam, despite 6.html being similar to 1.html. 6.html was an "orphan URL", not linked by a conventional HTML Anchor tag (<a href="URL/path">) to any other file. Thus it may be hidden from "ripper" software, which only can crawl HTML hyperlinks, and not JavaScript links. Opening the file 6.html in a new popup window from 5.html thus preserves its isolation.

In an effort to research this technique further, company C was persuaded to publish its administration e-mail address on an orphan URL (i.e. d@xx.org.uk in a 6@yy. org.uk situation), and still has not received any spam, three months later. Thus this technique seems to hold promise for web site owners unwilling to be deluged by spam and exposed to virus attacks. The generic code used was:

```
<script language="JavaScript">
Function box(URL)
{
window.open(URL,'box','menubar=no,toolbar=no,status=no,directories=no
,resizable=yes,scrollbars=no,width=300,height=20')
}
//where the width attribute reflects the length of the mail address
</script>
<A href="javascript:box('6.htm')">Click here for e-mail</A>
```

Literature

BBC (2003). Spam Attack. http://www.bbcworld.com/content/clickonline_archive_17_2003.asp?pageid=666&co_pageid=2

CAN-SPAM, (2003). http://www.cdt.org/legislation/108th/junkemail/s877sub4.pdf

CDT, (2003). Why am I getting all this spam? http://www.cdt.org/speech/spam/030319spamreport.pdf

Courtenay Communications Corporation. (2004). Worldata Index: BTB E-Mail List Prices Rise, BTC Slips. http://www.dmnews.com/cgi-bin/artprevbot.cgi?article_id=27142

Cranor, L. F. & LaMacchia, B. A. (1998). Spam! Communications of the ACM, *41*, 74-83.

Denning, P. (1982). Electronic Junk. Communications of the ACM, *3*, 163-165.

Eur-Lex, (2002). Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector. http://europa.eu.int/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=en&numdoc=32002L0058&model=guichett

Mellor, R. B., (2002). DHTML, Learning by Example. Wilsonville, Franklin Beedle.

Mellor, R. B., (2003). The Web Managers Handbook. Copenhagen, Globe.

Mellor, R. B., (2005). The Correlation between Visits and Product Sales on Three Business-to-Customer Internet Web Sites. KURIR, *1*, 17-30.

Mellor - Avoiding unwanted e-mail.

Tomasula, D. (2002). DMA: Use of Email Increases Sales. iMarketing News. http://www.dmnews.com/cgi-bin/publogin.cgi?article_id=19909

Wood, D., (1999). Programming Internet email. Sebastopol, O'Reilly