A visual history of spam (and virus) email

devblogs.microsoft.com/oldnewthing/20040916-00

September 16, 2004

Raymond Chen

I have kept every single piece of spam and virus email since mid-1997. Occasionally, it comes in handy, for example, to add naïve Bayesian spam filter to my custom-written email filter. And occasionally I use it to build a chart of spam and virus email.

The following chart plots every single piece of spam and virus email that arrived at my work email address since April 1997. Blue dots are spam and red dots are email viruses. The horizontal axis is time, and the vertical axis is size of mail (on a **logarithmic** scale). Darker dots represent more messages. (Messages larger than 1MB have been treated as if they were 1MB.)

Note that this chart is not scientific. Only mail which makes it past the corporate spam and virus filters show up on the chart.

Why does so much spam and virus mail get through the filters? Because corporate mail filters cannot take the risk of accidentally classifying valid business email as spam. Consequently, the filters have to make sure to remove something only if they has extremely high confidence that the message is unwanted.

Okay, enough dawdling. Let's see the chart.

Overall statistics and extrema:

- First message in chart: April 22, 1997.
- Last message in chart: September 10, 2004.
- Smallest message: 372 bytes, received March 11, 1998.

```
From: 15841.
To: 15841.
Subject: About your account…
Content-Type: text/plain; charset=ISO-8859-1
Content-Transfer-Encoding: 7bit


P
```

- Largest message: 1,406,967 bytes, received January 8, 2004. HTML mail with a lot of text including 41 large images. A slightly smaller version was received the previous day. (I guess they figured that their first version wasn't big enough, so they sent out an updated version the next day.)
- Single worst spam day by volume: January 8, 2004. That one monster message sealed the deal.
- Single worst spam day by number of messages: August 22, 2002. 67 pieces of spam. The vertical blue line.
- Single worst virus day: August 24, 2003. This is the winner both by volume (1.7MB) and by number (49). The red splotch.
- Totals: 227.6MB of spam in roughly 19,000 messages. 61.8MB of viruses in roughly 3500 messages.

Things you can see on the chart:

- Spam went ballistic starting in 2002. You could see it growing in 2001, but 2002 was when it really took off.
- Vertical blue lines are "bad spam days". Vertical red lines are "bad virus days".
- Horizontal red lines let you watch the lifetime of a particular email virus. (This works only for viruses with a fixed-size payload. Viruses with variable-size payload are smeared vertically.)
- The big red splotch in August 2003 around the 100K mark is the Sobig virus.
- The horizontal line in 2004 that wanders around the 2K mark is the Netsky virus.
- For most of this time, the company policy on spam filtering was not to filter it out at all, because all the filters they tried had too high a false-positive rate. (I.e., they were rejecting too many valid messages as spam.) You can see that in late 2003, the blue dot density diminished considerably. That's when mail administrators found a filter whose false-positive rate was low enough to be acceptable.

As a comparison, here's the same chart based on email received at one of my inactive personal email addresses.

This particular email address has been inactive since 1995; all the mail it gets is therefore from harvesting done prior to 1995. (That's why you don't see any red dots: None of my friends have this address in their address book since it is inactive.) The graph doesn't go back as far because I didn't start saving spam from this address until late 2000.

Overall statistics and extrema:

- First message in chart: September 2, 2000.
- Last message in chart: September 10, 2004.

- Smallest message: 256 bytes, received July 24, 2004.

```
Received: from dhcp065-025-005-032.neo.rr.com ([65.25.5.32]) by …
         Sat, 24 Jul 2004 12:30:35 -0700
X-Message-Info: 10
```

- Largest message: 3,661,900 bytes, received April 11, 2003. Mail with four large bitmap attachments, each of which is a Windows screenshot of Word with a document open, each bitmap showing a different page of the document. Perhaps one of the most inefficient ways of distributing a four-page document.
- Single worst spam day by volume: April 11, 2003. Again, the monster message drowns out the competition.
- Single worst spam day by number of messages: October 3, 2003. 74 pieces of spam.
- Totals: 237MB of spam in roughly 35,000 messages.

I cannot explain the mysterious "quiet period" at the beginning of 2004. Perhaps my ISP instituted a filter for a while? Perhaps I didn't log on often enough to pick up my spam and it expired on the server? I don't know.

One theory is that the lull was due to uncertainty created by the CAN-SPAM Act, which took effect on January 1, 2004. I don't buy this theory since there was no significant corresponding lull at my other email account, and follow-up reports indicate that CAN-SPAM was widely disregarded. Even in its heyday, compliance was only 3%.

Curiously, the trend in spam size for this particular account is that it has been going **down** since 2002. In the previous chart, you could see a clear upward trend since 1997. My theory is that since this second dataset is more focused on current trends, it missed out on the growth trend in the late 1990's and instead is seeing the shift in spam from text to <IMG> tags.

Raymond Chen