WHITE PAPER
Algunos datos estadísticos sobre el spam

Sergio Alejandro Hernando Westerheide

Consultor de Seguridad

sergio@sahw.com

Resumen

Cuando se hace seguimiento y análisis del correo basura, suelen aparecer algunos datos colaterales, que habitualmente no son consignados en los documentos de investigación que se generan a raíz de los experimentos. El objetivo de este artículo es dar a conocer algunas de estas cifras reales, pero no citadas, sobre el spam. Los datos que contiene este documento derivan de analizar más de 10.000 mensajes de correo basura, y en su gran mayoría, son cifras poco significativas. En ningún caso deben tomarse estas cifras como definitivas, ni extensibles a cualquier receptor de correo. Son, simple y llanamente, mis cifras de spam, que quiero compartir con los interesados. Bajo ninguna circunstancia, pese a la apariencia de *white paper*, se trata de un documento de investigación profunda sobre tendencias o características avanzadas sobre el correo no deseado. Este texto debe tomarse únicamente como una recopilación de datos de carácter anecdótico.

1. Datos globales

Las estadísticas se han creado con mboxstats[1], una aplicación liberada según GPL. Aprovechando que gestiono el correo con un cliente que trabaja de manera nativa en mbox[2], en este caso, Mozilla Thunderbird[3], es factible aplicar mboxstats a las cuentas del perfil, almacenadas, en el caso que nos ocupa, en el directorio local *./mozilla-thunderbird* de una máquina Linux con kernel 2.6.15-26-386. Una vez analizados los ficheros correspondientes, el programa permite generar estadísticas numéricas en texto plano, relacionadas con el correo residente en una ubicación determinada dentro del perfil analizado.

Nótese que se ha respetado la notación original de mboxstats, por lo que los datos estadísticos aparecerán principalmente en lengua inglesa:

- Statistics created on: Wed Aug 16 07:12:12 2006
- First message was written at: Sat Dec 14 04:45:51 1901
- Last message was written at: Mon Jan 18 20:09:09 2038
- Total number of messages: 10450
- Number of messages that is a reply: 120 (1.15%)
- Total size: 146830825
- Average size: 14050
- Total number of attachments: 200
- Total size of attachments: 21402218 (14.58% of total)
- Number of PGP signed messages: 2 (0.02%)
- Most busy day in number of msgs: 2038-01-19 (339 msgs, 1174164 bytes)
- Most busy day in number of bytes: 2006-08-11 (191 msgs, 7782631 bytes)
- Total number of writers: 8536

[1] http://www.vanheusden.com/mboxstats/
[2] http://www.qmail.org/man/man5/mbox.html
[3] http://www.mozilla.com/thunderbird/

- Number of people who wrote more than one message: 570 (6.68%)
- Total number of lines: 2163486
- Average lines per message: 207
- Total header length (lines): 335979
- Average header length (lines): 32 (15.53%)
- Average line length: 61
- The header is 9.47% bytes in size of the total.
- Total number of unique user-agents: 187
- Total number of unique organisations: 38
- Total number of unique top-level domains: 116
- Average spam score: 26.62
- Spammiest writer: qznegpg887353525@ms53.url.com.tw (86.47)
- Most spam by: William (john@e-zone-defense.biz) (48)

1.1. Conclusiones

1. Los datos relativos a las fechas no pueden ser considerados jamás como fuente de información, puesto que la mayoría de las veces, no se corresponden con la realidad.

2. El spam es, por lo general, ligero. El tamaño medio de 14050 bytes indica que no suelen ser mensajes con un peso elevado, lo que sin duda está orientado a poder enviar grandes cantidades de correo basura sin castigar excesivamente los recursos del emisor.

3. Tal y como era de esperar, el correo basura no suele entregarse firmado vía PGP y similares. Sólo 2 mensajes emplearon esta fórmula.

4. Sólo un 6,68% de los remitentes envió más de un mensaje de correo, con lo que en un total de 10450 mensajes hay 8536 remitentes distintos. Las listas negras de remitentes, por sí mismas, no son útiles para filtrar el spam.

5. La puntuación asignada por Spamassassin[4] a esta colección es de 26.79 de valor medio, bastante superior a los 5 puntos habituales para marcar las muestras como basura. Por término medio, y afortunadamente, el spam es fácil de "cazar". El caso más flagrante lo supone un mensaje con una puntuación de 86.47.

2. Clasificación según importancia

El siguiente punto analizado en las estadísticas es el marcado de importancia de los mensajes. Al contrario de lo que podría pensarse en primera instancia, la gran mayoría de los mensajes no incluye información sobre importancia. Sólo un 1.34% de los mensajes analizados contuvo datos relativos a esos parámetros:

- Low : 0.00
- Normal: 1.32
- High : 0.02
- (the rest is unspecified)

2.1. Conclusiones

1. El factor importancia no es válido para clasificar muestras de spam, ya que no es utilizado en la gram mayoría de las ocasiones.

[4] `http://spamassassin.apache.org`

3. Los remitentes más insistentes

En este punto, tratamos de ver hasta qué puunto los remitentes del correo basura se repiten o no. Tal y como se demuestra, no suele ser así. Recordemos que sólo el 6.68% de los remitentes emite más de un mensaje. En nuestras estadísticas, estos son los diez remitentes más insistentes a la hora de enviar spam a mi buzón:

- 51 mensajes de necojp@citiz.net
- 48 mensajes de john@e-zone-defense.biz
- 47 mensajes de john@englishforum.biz
- 46 mensajes de robert@scandinavian-seed.biz
- 46 mensajes de geoffrey@psychologen.biz
- 46 mensajes de robert@scandinavian-seed.biz
- 44 mensajes de simon@repairnet.biz
- 41 mensajes de reginald@pradella.biz
- 41 mensajes de richard@guitarra.biz
- 41 mensajes de william@pellicano.biz

3.1. Conclusiones

1. Los remitentes no suelen repetirse, para evitar las posibles filtraciones en función de los buzones de origen. Los spammers más efectivos son siempre los menos abundantes.
2. Abundan remitentes con cuentas con TLD .biz, con la idea de aparentar ser negocios en la red, y generar confianza en el receptor.
3. Los remitentes emplean, por norma general, nombres comunes (John, Robert, Simon, William, etc) como identificadores de usuario@dominio, lo que crea confianza en el receptor, dificultando su marcado como basura.

4. Los asuntos más utilizados

Uno de los ganchos más importantes que tiene el correo basura es el asunto. Descartando un total de 154 mensajes marcados con la cadena ***spam***, personalizada en el servidor y no generada por los remitentes, y otros 146 mensajes con el asunto en blanco (generalmente spam de prueba o fallido), los diez asuntos más empleados son los siguientes:

- 122 mensajes con asunto she wants a better sex? all you need's here!
- 121 mensajes con asunto best love dr@gs at best store!
- 121 mensajes con asunto all products for your health!
- 113 mensajes con asunto full of health? then don't click!
- 113 mensajes con asunto we cure any desease!
- 111 mensajes con asunto need medicine? all here!
- 110 mensajes con asunto all love enhancers on one portal!
- 110 mensajes con asunto why seek? choose any love pi11 you want!
- 107 mensajes con asunto any med for your girl to be happy!
- 104 mensajes con asunto our store is your cureall!

4.1. Conclusiones

1. Según el asunto del mensaje, los correos relacionados con la salud y la farmacia ilegal son los más abundantes.
2. Entre los diez asuntos más empleados no aparece ningún asunto relacionado con el software ilegal, otro de los reclamos más empleados entre los spammers. Tampoco aparecen estafas nigerianas, ni temáticas relacionadas con las imitaciones o el robo de identidad.
3. Se contabilizaron un total de 6442 asuntos distintos.

5. Los dominios de origen más habituales

Según el dominio del buzón de origen, estos son los diez dominios de origen más frecuentes:

- 5711 mensajes procedían de dominios .com
- 1375 mensajes procedían de dominios .biz
- 879 mensajes procedían de dominios .net
- 446 mensajes procedían de dominios .jp
- 310 mensajes procedían de dominios .ru
- 212 mensajes procedían de dominios .org
- 145 mensajes procedían de dominios .tw
- 142 mensajes procedían de dominios .uk
- 112 mensajes procedían de dominios .ar
- 100 mensajes procedían de dominios .es

5.1. Conclusiones

1. El TLD más popular, .com, es también el más popular a la hora de remitir spam, buscando con ello los remitentes el factor familiaridad.
2. Los remitentes .biz destacan, de un modo cada vez más creciente, por su presunta vinculación con negocios.
3. Se detectaron 116 TLDs distintos en el total de muestras recibidas, muchos de ellos ficticios (%anydomain, .or, websitesource, etc.)
4. El spam oriental, con remites de Japón y Taiwán, es notorio y creciente.
5. El dominio .es aparece como origen en 100 mensajes, y ocupa plaza en el *top ten* de dominios de origen.

6. Clasificación según franjas horarias

Ua vez analizadas las franjas horarias asignadas a cada mensaje, podemos establecer las siguientes como las diez más frecuentes:

- 1621 mensajes en la franja +0100
- 1515 mensajes en la franja -0700
- 1149 mensajes en la franja +0800
- 951 mensajes en la franja +0200
- 458 mensajes en la franja -0400
- 427 mensajes en la franja -0500
- 407 mensajes en la franja +0300
- 400 mensajes en la franja -0060
- 347 mensajes en la franja -0600
- 317 mensajes en la franja -0800

6.1. Conclusiones

1. La franja más empleada, -0700, pertenece principalmente al oeste de EEUU.
2. La franja menos utilizada es +1200, usual de Nueva Zelanda.
3. Numerosos mensajes provienen de franjas ficticias o erróneas, como -01200, +1300, +0720, etc. Una de estas franjas ficticias, -0060, alcanza 400 mensajes.
4. No está claro el porqué de las franjas ficticias, ya que las franjas de origen no suelen ser un factor discriminante del correo basura, ni suele ser un factor de ordenado habitual en los receptores.

7. Clasificaciones según organizaciones de origen

Es un dato muy poco relevante. Como curiosidad, se citan a continuación las cinco organizaciones de origen más populares:

- 35 mensajes sin organización especificada
- 6 mensajes con organización productos varios2
- 4 mensajes con organización qualcomm windows eudora version 5.1
- 4 mensajes con organización internet mail service (5.5.2650.21)
- 3 mensajes con organización microsoft outlook express 5.50.4522.1200

7.1. Conclusiones

1. Muy esporádicamente, aparecen mensajes procedentes de alguna organización real, empleada fraudulentamente, como por ejemplo (aui) asociacion de usuarios de internet, instituto secretariado europeo o grupo raac empresarial.

2. Las muestras analizadas procedían de 38 presuntas organizaciones, en su mayoría, ficticias. No se puede descartar que algunas organizaciones reales posean máquinas comprometidas desde las que se ha enviado correo basura.

8. Los agentes de usuario más utilizados

Los diez *user-agents* más empleados son:

- microsoft, con 4426 mensajes
- the, con 777 mensajes
- foxmail, con 106 mensajes
- volleymail, con 75 mensajes
- mozilla, con 72 mensajes
- squirrelmail/1.4.3a, con 76 mensajes
- writely, con 63 mensajes
- mime-tools, con 58 mensajes
- mozilla/5.0, con mensajes
- qualcomm, con 58 mensajes

8.1. Conclusiones

1. La gran mayoría de user-agents detectados sólo proporcionan la información "microsoft" con lo que, unido a que no es un parámetro crucial, le resta importancia a esta categoría.

2. Como curiosidad, citar un mensaje apareció presuntamente enviado desde un agente microsoft-outlook-express-macintosh-edition/5.02.2022. Otro agente nada habitual, aparecido entre las muestras, es microsoft-entourage/11.0.0.040405.

3. El cómputo asciende a un total de 187 agentes de usuario, en su mayoría ficticios, como motor de origen de las muestras.

9. Número de mensajes por día de la semana

El siguiente dato analizado es el día de la semana en el que fueron recibidas las muestras. El patrón no es claro, no existiendo un comportamiento tendencial, como cabría esperar, como sí acontece en los casos de spam orientado (phishing, spam segmentado, etc.):

- 1323 mensajes recibidos en lunes
- 1683 mensajes recibidos en martes
- 1243 mensajes recibidos en miércoles
- 1418 mensajes recibidos en jueves
- 1549 mensajes recibidos en viernes
- 1389 mensajes recibidos en sábados
- 1234 mensajes recibidos en domingos

9.1. Conclusiones

1. El día con más actividad es el martes y el menos activo, el domingo. La cantidad de spam emitido el viernes es principalmente phishing, con el objeto de maximizar su duración. Estos patrones son muy oscilantes, cambiando los días máximos y mínimos prácticamente cada vez que se toman muestras actualizadas.

10. Número de mensajes por mes, día del mes y hora

El dato de número de mensajes por mes no es relevante, ya que las muestras no abarcan un año completo de recepción. En cuanto a los días del mes, el más activo es el día 19 de cada mes registrado, con 597 mensajes. En cuanto a las horas, la más relevante es las 11 a.m, con 786 mensajes.

10.1. Conclusiones

1. El único patrón destacable es la hora del día a la que se suele recibir más correo basura, las 11 de la mañana, si bien no existen causas claras y concluyentes del porqué de esta hora como más frecuente.

11. URLs más utilizadas

Las URLs más utilizadas suelen corresponder a imágenes de productos en servicios públicos de comercio electrónico. Son las siguientes:

- 886 mensajes con la URL `http://g-images.amazon.com/images/g/01/detail/add-to-cart-midsize.gif`
- 877 mensajes con la URL `http://img.shopzilla.com/shopzilla/rating_5_star_104x19.gif`
- 453 mensajes con la URL `http://companiespharma.info/cs/?cid`
- 294 mensajes con la URL `http://image.shopzilla.com/resize?sq=3d100&uid=3d321652686`
- 294 mensajes con la URL `http://image.shopzilla.com/resize?sq=3d100&uid=3d6260970`
- 294 mensajes con la URL `http://image.shopzilla.com/resize?sq=3d100&uid=3d8778190`
- 244 mensajes con la URL `http://www.w3.org/tr/rec-html40`
- 180 mensajes con la URL `http://zaicheg.com/`
- 151 mensajes con la URL `http://capsulesworld.info/cs/?cid`
- 132 mensajes con la URL `http://g-images.amazon.com/images/g/01/promotions/sticker/newest_versio`

11.1. Conclusiones

1. Se capturaron un total de 11299 URLs distintas en los mensajes.
2. El servicio con más imágenes robadas es Shopzilla.
3. Muchos mensajes hacen alusión a una publicación de la World Wide Web Consortium, la especificación HTML 4.01

12. Referencias

[1] *Sergio Hernando. Correo no deseado. Causas, tipologías y medidas de prevención.* http://www.sahw.com/wp/archivos/2006/03/10/correo-no-deseado-causas-tipologias-y-medidas-de-prevencion/

[2] *Sergio Hernando. Filtrar el correo basura procedente de Spamassassin con Mozilla Thunderbird.* http://www.sahw.com/wp/archivos/2006/07/21/filtrar-el-correo-basura-procedente-de-spamassassin-con-mozilla-thunderbird/

[3] *El spam financiero y su impacto en los mercados de valores.* http://www.hispasec.com/unaaldia/2762

[4] *El spam sigue creciendo inmune a cualquier medida en su contra.* http://www.hispasec.com/unaaldia/2763

[5] *Un día cualquiera en la vida de un receptor de spam.* http://www.hispasec.com/unaaldia/2801

[6] McWilliams, Brian "Spam Kings. The Real Story behind the High-Rolling Hucksters Pushing Porn, Pills, and ....". O'Reilly, 2005. ISBN: 0596007329

[7] Posluns, Jeffrey "Inside the SPAM Cartel". Syngress, 2004. ISBN: 1932266860

13. Licencia

[5] http://creativecommons.org/licenses/by-nc-sa/2.5/es/
[6] http://www.latex-project.org