Solutions-Linux 2005

Joe's j-chkmail

http://j-chkmail.ensmp.fr/papers

Jose-Marcio.Martins@ensmp.fr

Ecole des Mines de Paris - Centre de Calcul

Plan

- j-chkmail – Qu'est-ce ?
- Où et qui ?
- Filtrage viral
- Filtrage de SPAM
- Surveillance et commande
- Protection du serveur
- Conclusions

j-chkmail - Qu'est-ce ?

- Le complément des logiciels de serveurs de messagerie : sendmail fait du routage, j-chkmail fait du filtrage.
- Capable de traiter un trafic important
  - Organisations de taille moyenne et grande
- Faible consommation de ressources CPU
  - Même ordre de grandeur que le logiciel serveur
- Et surtout fiable - un serveur de mail est une cible privilégiée pour des attaques : il faut résister.

j-chkmail - Quoi ?

- Filtrage viral
- Filtrage de SPAM
- Protection du serveur de mail
- Surveillance en temps réel
- Commande du filtre

- Sendmail + API libmilter
- Environ 50000 lignes de code C

j-chkmail - Pour qui ?

- Établissements enseignement/recherche
- Population importante
  - Milliers d'utilisateurs
- Population hétérogène
  - Informaticien, infirmière, ...
- Trafic important
  - Centaines de milliers de messages par jour
  - Des gigaoctets par jour

j-chkmail - license

- Open source
- Gratuit
- Ce n'est pas exactement un logiciel libre
  - Distribué uniquement à des utilisateurs qui acceptent de s'identifier
  - Sans droit de redistribution ou réutilisation
- De la « Security by Obscurity » !!! Et alors ???
  - Temps de réaction plus important
  - Intervalle de mise à jour des serveurs moins critique

Quelques utilisateurs...

- Des institutions d'Enseignement/Recherche
  - Université de Jussieu – (75000 b.a.l.)
  - Université de Waterloo
- Des organismes internationaux
  - unon.org, icimog.org.np
- Des providers
  - pobox.sk (300000 b.a.l. – 20K messages/heure)
- L'administration (peu en France, en dehors des établissements Enseignement/Recherche)
- Banques
- Entreprises

j-chkmail – Filtrage viral

- Le code malveillant des virus sont dans des « fichiers exécutables »
- Détection des messages ayant des fichiers attachés susceptibles de contenir un code exécutable – « X-Files »

  http://support.microsoft.com/default.aspx?scid=kb;EN-US;q262631
- 50 à 1000 fois plus rapide qu'un scanneur antivirus classique
- Interface avec des scanneurs externes : ClamAv, Sophos, McAfee, F-prot, Trendmicro, ...

Le filtrage des X-Files - principe...

http://support.microsoft.com/default.aspx?scid=kb;EN-US;q262631
http://www.cknow.com/vtutor/vtextensions.htm

ade	**adp**	**bas**	**bat**	**bin**	**btm**	**chm**	**cmd**	**com**	**cpl**
crt	**dll**	**drv**	**exe**	**hlp**	**hta**	**inf**	**ini**	**ins**	**isp**
je	**js**	**jse**	**lnk**	**mdb**	**mde**	**msc**	**msi**	**msp**	**mst**
pcd	**pif**	**reg**	**scr**	**sct**	**shb**	**shs**	**sys**	**url**	**vb**
vbe	**vbs**	**vxd**	**wsc**	**wsf**	**wsh**				

j-chkmail - Filtrage SPAM

Un exemple de spam parmi d'autres...

```
<html>
<body bgcolor="white"> <div align="center">
<a href="http://xndish-aerial.com/agb/super.html?aa=agb&ab=martins&ac=ensmp.fr">
  <img src="http://xndishaerial.com/agb/images/agb.jpg?ba=agb&bb=martins&bc=ensmp.fr"
border="0"></a><br><br><br><br><br><br><br><br><br><br><br><br><br><br><br><br><br><br><b
r><br><br><br><br><br><br><br><br><br>  <a href="http://xndish-aerial.com/send-me-
out/index.html?ca=agb&cb=martins&cc=paris.ensmp.fr">
<img src="http://xndish-aerial.com/agb/images/outw.gif?da=agb&db=martins&dc=ensmp.fr"
border="0">
</a><br><br><br><br><br>
  <img src="http://xndish-aerial.com/agb/images/add.gif" border="0">
<br><br><br>
  <font size="1" color="#000099">southeastern tentatively musty baneful journeyings
jonquil definitive grunts may macroeconomics journeyed tau scurvy stretchers
loneliness  RzneXznegvafRzneXcnevf.rafzc.seRzneX kanji
coventry villainously primal granite yonkers harshness shepard gadgetry concepts
renee specifiable lowest astonishment osteopathic  volleyballs ecosystem definable
smartest portal rubies moneyed attributing handsomeness harmoniousness objectives
nester outputting chanter disproved  translucent lunar narragansett africanized
imagined staves housewife unblocks unneeded inexorable censure perverted overcame
allocatable anon </font></div></body>
</html>
```

Si on regarde bien...

- Le message publicitaire avec confirmation d'adresse, contenu dans un image

```
<img src="http://xndishaerial.com/agb/images/agb.jpg?ba=agb&bb=martins&bc=ensmp.fr">
```

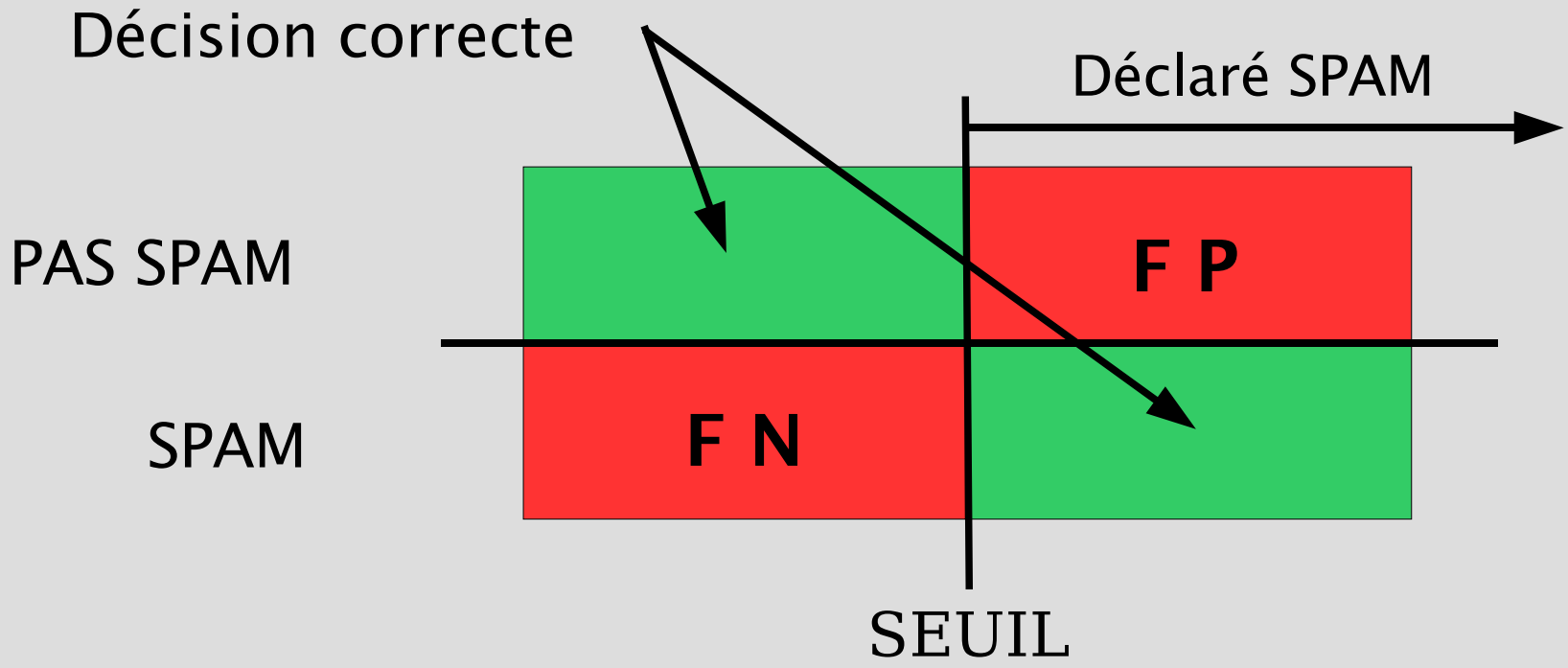
- Le moyen de contacter le spammeur

```
<a href="http://xndish-aerial.com/agb/super.html?aa=agb&ab=martins&ac=ensmp.fr">
```

- Un message textuel sans rapport avec le SPAM

<font size="1" color="#000099"> southeastern tentatively musty baneful journeyings jonquil definitive grunts may macroeconomics journeyed tau scurvy stretchers loneliness RzneXznegvaf RzneXcnevf.rafzc.seRzneX kanji coventry villainously primal granite yonkers harshness shepard gadgetry concepts renee specifiable lowest astonishment osteopathic volleyballs ecosystem definable smartest portal rubies moneyed attributing handsomeness harmoniousness objectives nester outputting chanter disproved translucent lunar narragansett africanized imagined staves housewife unblocks unneeded inexorable censure perverted overcame allocatable anon

(Parenthèse)
« Le » problème de tout filtrage...



- Si l'on augmente le seuil,
  - Le taux de faux positifs diminue
  - Le taux de détection diminue

L'approche j-chkmail

- L'objectif est la satisfaction de l'utilisateur (qualitatif et non pas quantitatif) – faciliter le classement
- Algorithmes rapides
  - Filtrage comportementale – « dégrossir » le trafic
  - Filtrage de contenu
- Seuils bas – plus de détection et plus de faux positifs
- Pas de critères « blanchissants » – les messages sont blanchis par l'utilisateur (les expéditeurs connus)
- Éviter dépendances externes

(Parenthèse – le protocole SMTP)

```
    martins@calloway:~> telnet paris smtp
    Trying 194.214.158.200...
    Connected to paris.
    Escape character is '^]'.
<-  220 paris.ensmp.fr ESMTP Sendmail 8.12.8/8.12.7/JMMC
->  helo calloway.ensmp.fr
<-  250 paris.ensmp.fr Hello calloway [194.214.158.171], pleased to meet you
->  mail from:joe@ensmp.fr
<-  250 2.1.0 joe@ensmp.fr... Sender ok
->  rcpt to:dupont
<-  250 2.1.5 dupont... Recipient ok
->  rcpt to:durand
<-  550 5.1.1 durand... User unknown
```

Enveloppe

```
->  data
<-  354 Enter mail, end with "." on a line by itself
->  From: Antoine
->  To: Sebastien
->  Subject: test telnet
->
->  C'est un test, je dis !
->  .
<-  250 2.0.0 h2QBmFBx017626 Message accepted for delivery
->  quit
<-  221 2.0.0 paris.ensmp.fr closing connection
    Connection to paris closed by foreign host.
    martins@calloway:~>
```

Corps du Message

Analyse du comportement

- Comment ? Entre autres choses,
  - Cadence de connexion
  - Pièges à SPAM
  - Détection des « erreurs d'adressage » (harvest)
  - Conformité des commandes SMTP, l'enveloppe et des en-têtes (RFCs)
- Résultats
  - Listes noires dynamiques
  - Rejet des connexions sans vérification de contenu
  - Ne détecte pas beaucoup, mais dégrossit bien le trafic

Analyse du contenu

- Comment ?
  - Recherche d'expressions régulières
  - Listes noires d'URLs : j-chkmail + SURBL.org
  - Heuristiques diverses (Oracle- 33 critères) : conformité du message, conformité et « richesse » code HTML, entropie, comparaison plain/html,...
- Résultats :
  - Score attribué aux messages, indiqué dans un en-tête
  - Message refusé si score important
  - Le client de messagerie du destinataire dirige le message vers une « boite à SPAM » si l'émetteur est ***inconnu*** et si le seuil est ***important***
  - Génération d'une blackliste locale

Filtrage par l'utilisateur

- Si l'expéditeur est connu
  - Alors met message dans la boite « Entrée »
- Si le score est supérieur à 3 (e.g.)
  - Alors met message dans boite « SPAM FORT »
- Si le score est supérieur à 0
  - Alors met message dans boite « SPAM FAIBLE »
- Met message dans boite « Entrée »

Résultats (BAL de l'auteur)

- Hypothèses
  - ~ 200 SPAMS par jour
  - Sur le poste utilisateur,
    - Pas de traitement pour les messages en provenance des correspondants courants (listes de diffusion, ...)
    - Deux boites à SPAM
- Résultats
  - Détection > 99 %
  - Faux positifs < 1 %
- Et la consommation de ressources ?

j-chkmail – Protection du serveur…

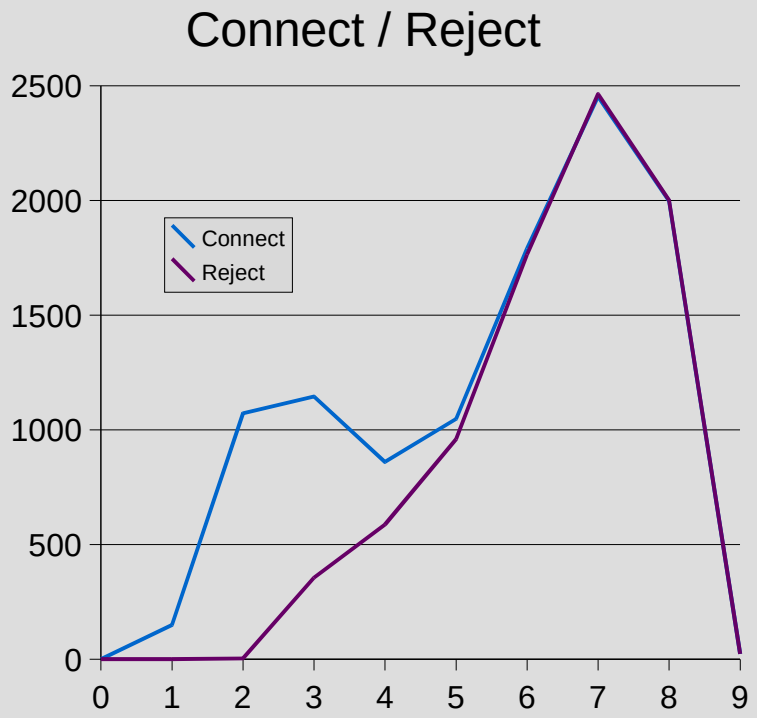
- Filtre auto-régénérant (« increvable »)
- Mesure de cadence de connexion (par client SMTP)
- Contrôle du nombre de connexions ouvertes (par client SMTP)
- Prise en compte des ressources disponibles (charge CPU, descripteurs de fichiers, …) et de la contribution de chaque client.

Résultats – mesure de cadence

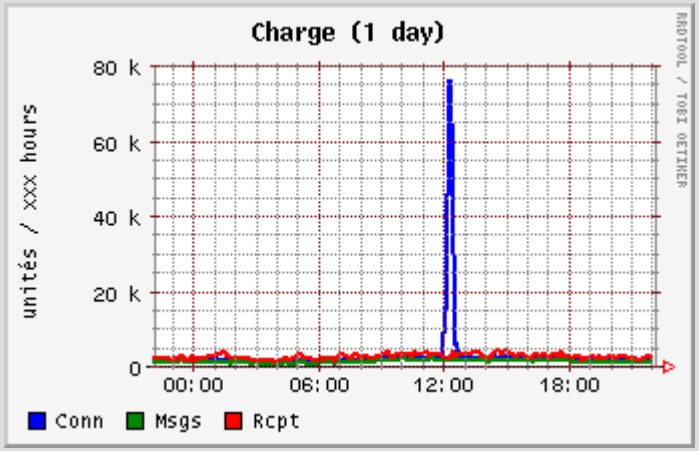
- 10536 connexions en 8 minutes
- 238 clients du réseau 66.216.119.0/24
- Connexions par client : [28 – 67]
- Pic : 86 connexions dans la même seconde
- 15 messages refusés par le contenu, dans les 3 premières minutes

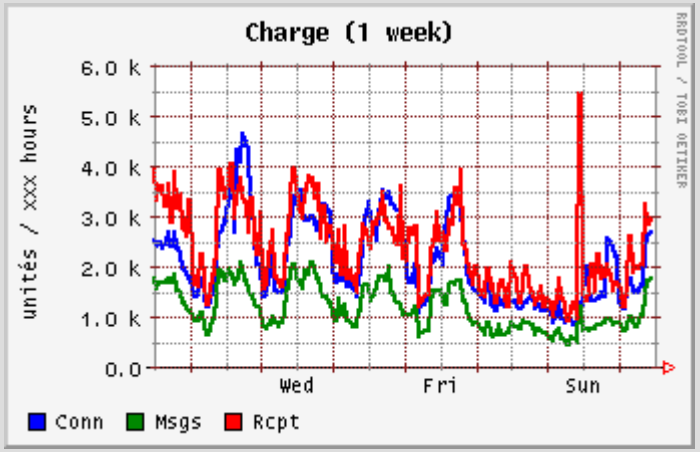
  `www.rapiddealsbyemail.com`
- 8156 connexions refusées par la mesure de cadence de connexion
- **Aucun message légitime perdu !**

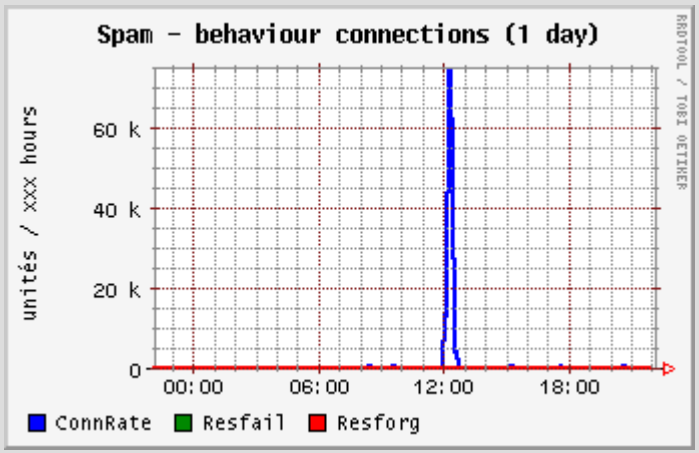
- Intégration dans sendmail depuis 2003 (contrib j-chkmail)

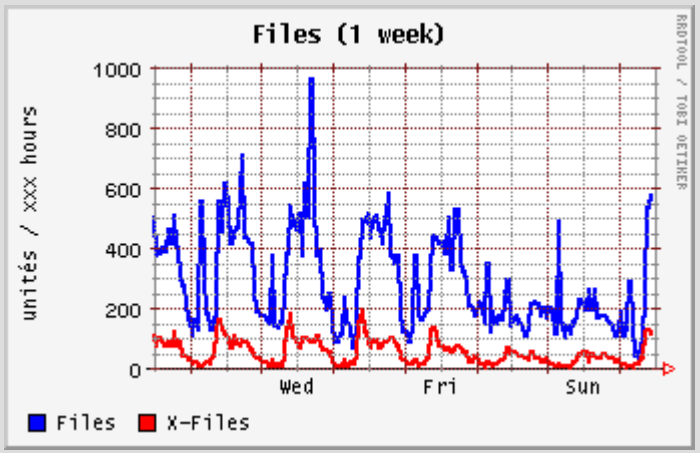


j-chkmail - Surveillance









j-chkmail - Surveillance

```
martins@paris:~> j-printstats -q -l 6h | more
Version                         : Joe's j-chkmail v1.7 - PreAlpha 6
*** Summary
*** TOTAL
 First Connection  : Sun Jun  6 17:33:11 2004
 Last Connection   : Sun Jun  6 23:33:09 2004
 Connections       :    9393
 Gateways          :    4258
 Throttle Max      :     445 / 10 min (for the server)
 Throttle Max      :     100 / 10 min (for a single gateway)
 Duration (sec)    :   0.005  16.931 7226.787 206.110 (min mean max std-dev)
 Work (sec)        :   0.001   0.028    1.803   0.150 (min mean max std-dev)
 Mean Throuput     :   0.647 KBytes/sec
Counts
 Messages          :    5471
 Volume            :  105393 KBytes
 Mean Volume       :   18.81 KBytes/msg
 Recipients        :    8256
 Bad Recipients    :    2145
 Yield             :    0.58 msgs/connection
 Yield             :    0.88 rcpt/connection
 Files             :     672
 X-Files           :     388
 Virus             :       0
```

Surveillance ...

```
martins@paris:~> j-printstats -q -l 6h -m c
                    CONN  MSGS REG.EXP  ORACLE
...
. 24.17.181.231  :     2     1       1       1 : c-24-17-181-231.client.comcast.net
. 24.17.211.202  :     1     1       1       1 : c-24-17-211-202.client.comcast.net
. 24.19.30.98    :     2     4       4       4 : c-24-19-30-98.client.comcast.net
. 24.19.33.250   :     1     1       1       1 : c-24-19-33-250.client.comcast.net
. 24.19.125.53   :     1     1       1       1 : c-24-19-125-53.client.comcast.net
. 24.19.212.106  :     2     1       1       1 : c-24-19-212-106.client.comcast.net
. 24.19.226.165  :     1     1       1       1 : c-24-19-226-165.client.comcast.net
. 24.20.26.47    :     1     1       1       1 : c-24-20-26-47.client.comcast.net
. 24.20.103.148  :     1     1       1       1 : c-24-20-103-148.client.comcast.net
. 24.20.172.173  :     2     1       1       1 : c-24-20-172-173.client.comcast.net
. 24.21.24.32    :     1     1       0       1 : c-24-21-24-32.client.comcast.net
. 24.21.143.163  :     6     1       1       1 : c-24-21-143-163.client.comcast.net
. 24.21.252.100  :     1     1       1       1 : c-24-21-252-100.client.comcast.net
. 24.24.107.127  :     2     1       1       1 : cpe-024-024-107-127.midsouth.rr.com
. 24.24.234.17   :     4     2       0       2 : cpe-24-24-234-17.socal.rr.com
. 24.25.4.97     :     1     1       0       1 : ncmx03.mgw.rr.com
. 24.25.37.146   :     3     2       1       2 : ilm25-37-146.ec.rr.com
. 24.26.180.239  :     1     1       1       1 : CPE-24-26-180-239.mn.rr.com
. 24.28.193.149  :     3     1       0       1 : vamx03.mgw.rr.com
. 24.29.47.228   :     1     1       1       1 : alb-24-29-47-228.nycap.rr.com
. 24.30.28.172   :     1     2       2       2 : c-24-30-28-172.mw.client2.attbi.com
. 24.30.84.80    :     1     1       1       1 : c-24-30-84-80.mw.client2.attbi.com
```

j-chkmail - Commande

```
martins@paris:~> telnet localhost 2010
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
200 OK - Waiting for commands !
reload databases
200 OK for RELOAD DATABASES !
200 URLBL : OK
200 RELOAD DATABASES done !

200 OK - Waiting for commands !
stats connopen
200 OK for STATS CONNOPEN !
*** Open connections :
  170.171.252.28    :   1 : silas.randomhouse.com
  193.251.37.182    :   2 : ASt-Lambert-109-2-4-182.w193-251.abo.wanadoo.fr
  193.49.22.101     :   1 : evry
  220.112.147.75    :   1 : [220.112.147.75]
  64.53.226.128     :   1 : d53-64-128-226.nap.wideopenwest.com
  83.130.129.80     :   1 : IGLD-83_130_129_80.inter.net.il
    6 entries on database
200 STATS CONNOPEN done !
Connection to localhost closed by foreign host.
martins@paris:~>
```

En cours/en étude...

- Encore des critères comportementaux (priorité...)
- Interface web
  - Interface d'administration
  - Gestion (partielle) des préférences des utilisateurs
  - Gestion semi-automatisée de la quarantaine
- Greylisting - ...
- Filtrage de contenu :
  - Des vérifications linguistiques plus évoluées
  - Disparition de l'Oracle ???
  - Filtrage morphologique (V1@gra, ...)
  - Bayesien ???
- Prise en compte de l'authentification de l'émetteur (DomainKeys, SPF) et gestion de listes blanches

Conclusions

- Filtre rapide et efficace
- Evolution en cours : améliorer la qualité de la détection et la facilité de mise en oeuvre.
- La principale difficulté pour le déploiement est la communication avec l'utilisateur : le « mode d'emploi »

Ce qui se passe sur un grand site...

SUR 440.000 MESSAGES

# 1	127535	URIBL_WS_SURBL
# 2	127101	URIBL_SBL
# 3	125917	URIBL_JP_SURBL
# 4	120728	URIBL_OB_SURBL
# 5	96849	BAYES_99
# 6	95827	RCVD_IN_BL_SPAMCOP_NET
# 7	90406	HTML_MESSAGE
# 8	71017	URIBL_SC_SURBL
# 9	46927	MIME_HTML_ONLY
#10	36806	URIBL_AB_SURBL
#11	33822	RCVD_IN_XBL
#12	30930	MIME_BOUND_DD_DIGITS
#13	30649	MPART_ALT_DIFF
#14	28472	URIBL_AH_DNSBL
#15	26638	RCVD_IN_SORBS_DUL
#16	26621	DRUGS_ERECTILE
#17	26394	MSGID_FROM_MTA_HEADER
#18	24615	RCVD_IN_DSBL
#19	23977	MSGID_FROM_MTA_ID
#20	23690	RCVD_IN_SORBS_SPAM
#21	22457	RCVD_IN_NJABL_DUL
#22	21115	RCVD_IN_NJABL_PROXY
#23	21013	RCVD_IN_SBL
#24	20262	X_MESSAGE_INFO
#25	18044	HTML_FONT_BIG

...

- Listes noires IP
- Listes noires URI
- Filtrage bayesien
- Heuristiques

Merci à Raymond Dijkxhoorn