Filtrage de messages sur des gros serveurs

José-Marcio Martins da Cruz

Ecole des Mines de Paris

JRES 2005 ± Marseille

Jose-Marcio.Martins@ensmp.fr - http://j-chkmail.ensmp.fr

Plan

- La vie sur les gros serveurs
- Un peu de stats et les filtres statistiques
- Filtrage de contenu versus comportement
- Le filtrage sur des gros serveurs
- j-chkmail – les nouveautés

Les gros serveurs

- Passerelle de campus d'université
  - Beaucoup de : utilisateurs, messages et volume
  - Des profils d'utilisateurs très variés : sciences exactes, sociales, marketing, finance, ... -> difficile de définir le profil typique
  - Contraintes fortes : sécurité, fiabilité, disponibilité, ...
  - Facteurs humains :
    - l'administrateur de la messagerie ne connais pas les utilisateurs –> relations tendues
    - Peu de retour utile -> la mise au point des filtres est plus difficile
    - Pression des utilisateurs, pas droit à l'erreur... -> position souvent conflictuelle
  - Dynamique de trafic : des pics importants
  - Limitations informatiques : matériel et logiciel
  - Objectifs : ration efficacité/ressources

La vie sur un gros serveur

# 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

...

- Liste noire IP
- Liste noire URL
- Filtrage Bayesien
- Filtrage Heuristique
- Expressions régulières

Domaine prolocation.net
6 (??) heures – 440K messages
Janvier 2005
Source : Raymond Dijkxhoorn

Ce que l'on remarque :

- L'efficacité decroit rapidement -> peu de critères détectent beaucoup de spam – «les 20/80»

- Les critères apparemment les plus «efficaces» sont des critères externes puis le filtrage bayesien

- Quelques critères heuristiques parmi les premiers, et celui en tète n'est pas fiable (HTML_MESSAGE)

A propos des listes noires

- Des résultats de mail-abuse.com sur une semaine sur un MX de ensmp.fr...

mail-abuse.com		: 144505			
-- 127.1.0.1	:	889	127.1.0.2	:	140416
-- 127.1.0.3	:	1082	127.1.0.4	:	556
-- 127.1.0.6	:	271	127.1.0.8	:	924
-- 127.1.0.9	:	11	127.1.0.10	:	356

- Conclusion :
  - Sert surtout à bloquer les adresses terminaux (dynamiques, disent-ils)
  - 20K à 30K connexions bloquées par jour -> intéressant !
  - Peu d'erreurs, mais...
  - Pas de problème si vous êtes un client, gênant si vous êtes un fournisseur (le client est toujours en position de force).

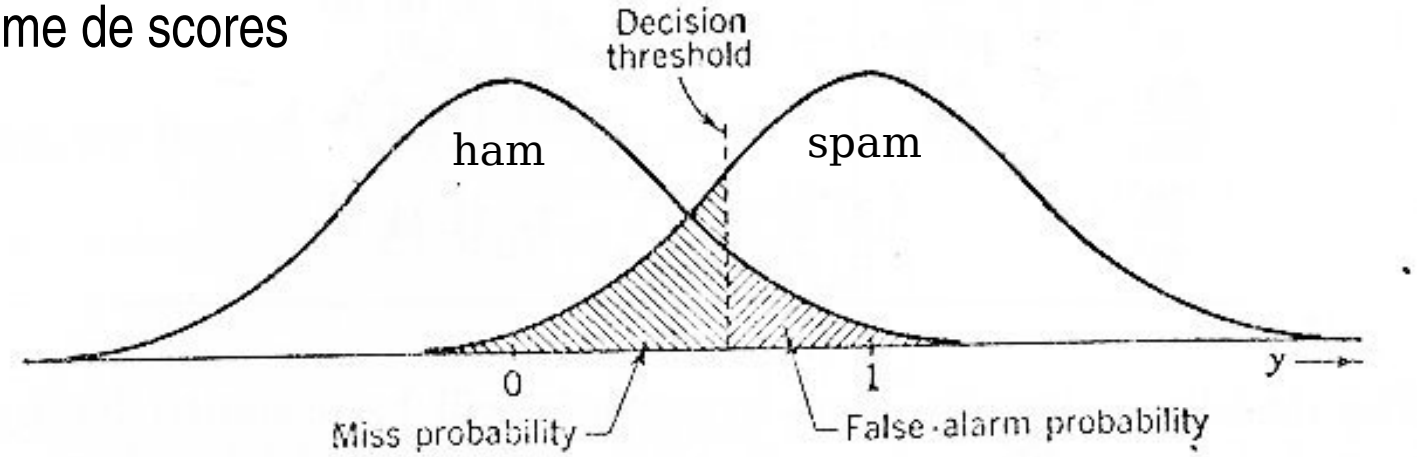
Filtrage statistique et statistiques de filtrage...

Filtrage statistique

- Dans un filtre statistique le réglage des paramètres de fonctionnement est basé sur des mesures faites sur un échantillon
  - Une série de tests pondérés par la probabilité de de leur apparition sur un corpus de ham/spam.
  - La probabilité d'erreur n'est pas forcément bonne mais une estimation est possible.
- Un filtre statistique mesure la *distance* entre un message arrivant et les messages typiques utilisés dans l'apprentissage.
- Exemples :
  - Filtrage heuristique
    - Un nombre important (mais fini) de tests différents définis par l'auteur du filtre sur des caractéristiques précises des messages
  - Filtrage bayesien
    - Un nombre important (limité par les ressources disponibles) de tests définis par le protocole d'apprentissage sur les mots (tokens)

Filtrage statistique ± les probabilités

- Histogramme de scores



- Moins les courbes se chevauchent, moins le taux d'erreur est important
- Plus les courbes sont «ouvertes», plus le taux d'erreur est important
- L' «ouverture» des courbes est une mesure de la dispersion des critères (diversité, variance, ...)
- La forme des courbes dépend **aussi** des critères de filtrage choisis et des poids
- L©idéal serait : avoir un point de passage par zero

Dessin : W. Harman – Principles of Statistical Theory of Communication

Filtrage bayesien dans une communauté universitaire

	ADM ASU	ADM INF0	ENS ANG	ENS INF0	ENS MARK	RESP ASPAM	TOTAL
0.10	97.21	99.03	71.43	97.06	50.86	78.12	82.76
0.20	0.00	0.00	1.39	0.00	2.15	0.54	0.52
0.30	0.00	0.00	0.55	0.59	2.58	0.45	0.38
0.40	0.51	0.00	1.66	0.20	3.65	0.29	0.56
0.50	1.27	0.08	13.73	0.78	16.74	2.03	3.17
0.60	0.00	0.08	3.88	0.20	1.50	0.12	0.38
0.70	0.00	0.00	0.28	0.00	0.21	0.07	0.10
0.80	0.00	0.00	0.42	0.00	0.21	0.04	0.10
0.90	0.00	0.00	0.97	0.00	0.21	0.10	0.09
1.00	1.02	0.81	5.69	1.18	21.89	18.24	11.94
HAM	97.21	99.03	72.82	97.06	53.00	78.66	83.28
UNSURE	1.78	0.16	20.94	1.77	25.11	3.08	4.78
SPAM	1.02	0.81	6.24	1.18	21.89	18.26	11.94
FP	0	0	5	2	10	3	
FN			2			10	
Messages	394	1239	721	510	466	8213	194549

- OBS : filtrage bayesien appliqué après greylisting
- Source : Fabrice Prigent – Université de Toulouse

Ce que l©on peut dire :

- Le corpus de *ham* varie de personne à personne, mais pas le corpus de *spam*
- Quand la dynamique des profils est faible, l'erreur existante entre le profil d'un utilisateur et le profil moyen est faible, elle aussi.
- Les utilisateurs dont le profil est plus éloigné du profil type sont ceux dont l'erreur est la plus importante
- Plus vous avez des «unsure», plus les histogrammes des scores se chevauchent -> plus vous avez des erreurs

Filtrage bayesien dans une autre communauté

	JOE- ENSMP		ENS ANG	ENS INF0	ENS MARK	RESP ASPAM
0.10	56.37		71.43	97.06	50.86	78.12
0.20	4.07		1.39	0.00	2.15	0.54
0.30	3.86		0.55	0.59	2.58	0.45
0.40	4.38		1.66	0.20	3.65	0.29
0.50	17.22		13.73	0.78	16.74	2.03
0.60	2.40		3.88	0.20	1.50	0.12
0.70	0.31		0.28	0.00	0.21	0.07
0.80	0.31		0.42	0.00	0.21	0.04
0.90	0.94		0.97	0.00	0.21	0.10
1.00	10.13		5.69	1.18	21.89	18.24
HAM	60.44		72.82	97.06	53.00	78.66
UNSURE	28.50		21.50	1.77	25.11	3.11
SPAM	11.06		5.69	1.18	21.89	18.24
FP	10		5	2	10	3
FN	14		2			10
Messages	958		721	510	466	8213

Filtrage déterministe (presque, en fait...)

- Pas d'erreur de filtrage (ou erreur de filtrage négligeable)
  - OBS : Aucun filtrage n'est infaillible
- Types :
  - Filtre noir – bloque les messages dont on est «sur» d'etre un spam
    - Ex : listes noires «fiables» – surbl.org : *«If it appears in HAM, don't list it !»*
  - Filtre blanc – laisse passer, sans aucun filtrage, les messages légitimes
    - Messages signés, SPF (?), DKIM (?)

Filtrage de contenu et comportement

Filtrage de contenu

- Seul le contenu effectif du message en cours entre en compte
- Pas de corrélation entre messages et connexions SMTP
- Exemples :
  - «Pattern matching» – recherche de mots ou expressions régulières
    - maintenance très difficile, gourmand, peu efficace
  - Filtrage d'URLs
    - peu gourmand, efficace, maintenance presque automatisable (extraction de URLs avec validation manuelle)
  - Filtrage heuristique
    - gourmand, besoin de beaucoup de tests pour être effectif
  - Filtrage bayesien – divise le message en unités élémentaires (mots, par exemple) et vérifie si ceux les plus significatifs se trouvent plus souvent dans les *hams* ou *spams*.
  - ...

Filtrage de comportement

- Comportement :
  - Corrélation, dans le temps, de transactions (connexions ou messages) successifs
  - Fenêtre temporelle de taille variable
  - Le comportement est, en général, associé au client SMTP
  - Ex : Cadence de connexions, taux d'erreurs divers, listes noires d'adresses IP.
- Le but : utiliser les résultats antérieurs pour biaiser les résultats futurs (dépenser de la mémoire au lieu de cycles CPU).
- Le filtrage de contenu donne des résultats plus fins, tandis que le filtrage de comportement permet surtout de bloquer les abus et les envois en masse.

Le filtrage de trafic important

Filtrage efficace sur des gros serveurs

- Filtrage comportemental – rapide, dégrossit le trafic et protège contre des attaques.
- Privilégier des critères objectifs : taux d'erreur faible et peu/pas de filtrage statistique.
- Prendre des décisions dès que possible : arrêter d'exécuter des tests si l'on sait déjà ce que sera fait du message en cours d'analyse.
- Éviter les vérifications coûteuses en temps, même s'il s'agit de temps d'attente
- Éviter les traitements dont la complexité algorithmique est important ou dont le temps de traitement ne soit pas déterministe.
  - Expressions régulières – nombre x f(complexité) x f(longueur message)
- Éviter les dépendances externes
  - DNS, bases de signatures
- Compromis entre efficacité et rapidité – l'objectif sur le serveur est la réduction du trafic inutile. La «perfection de filtrage» se trouve plutôt sur le poste de l'utilisateur.
- Communiquer !

j-chkmail

j-chkmail

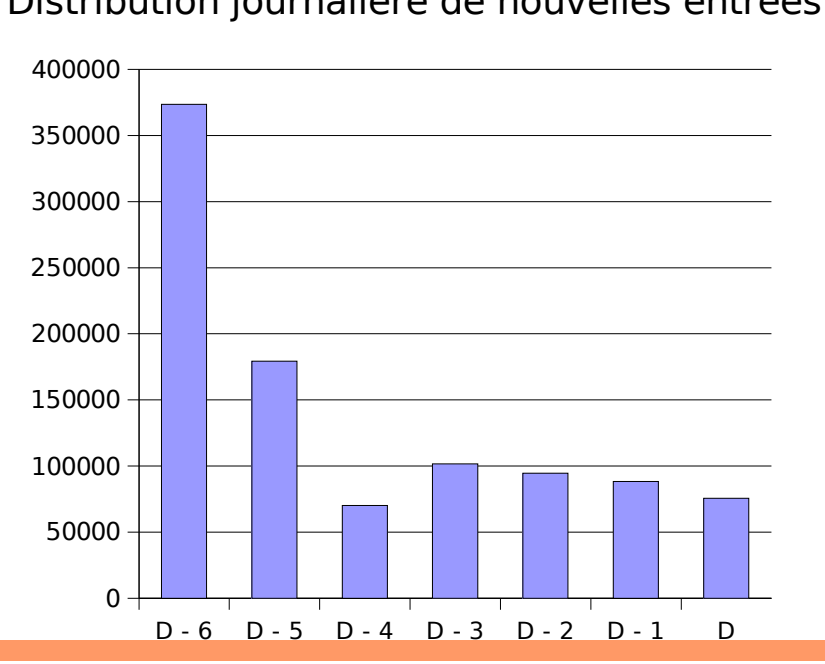
- Filtrage virale
  - présence de fichiers attachés susceptible de contenir du code exécutable (exe, pif, ...). Pas d'analyse du fichier et pas besoin de base de signatures
  - Interface scanneur externe (ClamAV, ...)
- Anti-spam
  - Cadences (connexion, bounce, temps de traitement, score, ...)
  - Greylisting
  - Expression régulières et filtrage d'URLs
  - «Petits testes entre amis» : BadMX, argument de la commande EHLO, pièges à spam, erreurs de destinataires, ...
  - Filtrage heuristique (actuellement 30 critères) – en cours d'extinction
- Protection du serveur
  - Consommation individuelle de ressources (cadences, connexions ouvertes, ...)
  - Charge du serveur

j-chkmail ± filtrage comportemental

- Dépense de la mémoire au lieu de la CPU
- Trois niveaux de historique possibles :
  - Court – besoin de réaction rapide
    - 10 minutes (20 min enregistrées)
    - un résumé de l'activité de chaque client SMTP – nombre de connexions, messages, bounces, virus, X-Files, somme des scores, temps CPU, volume, ...
  - Moyen – liste noire
    - 5 heures
    - les comportements douteux confirmés : erreurs de destinataire, messages vers des pots de miel, ...
  - Long – traitement externe des fichiers de log ou des résultats de filtrage

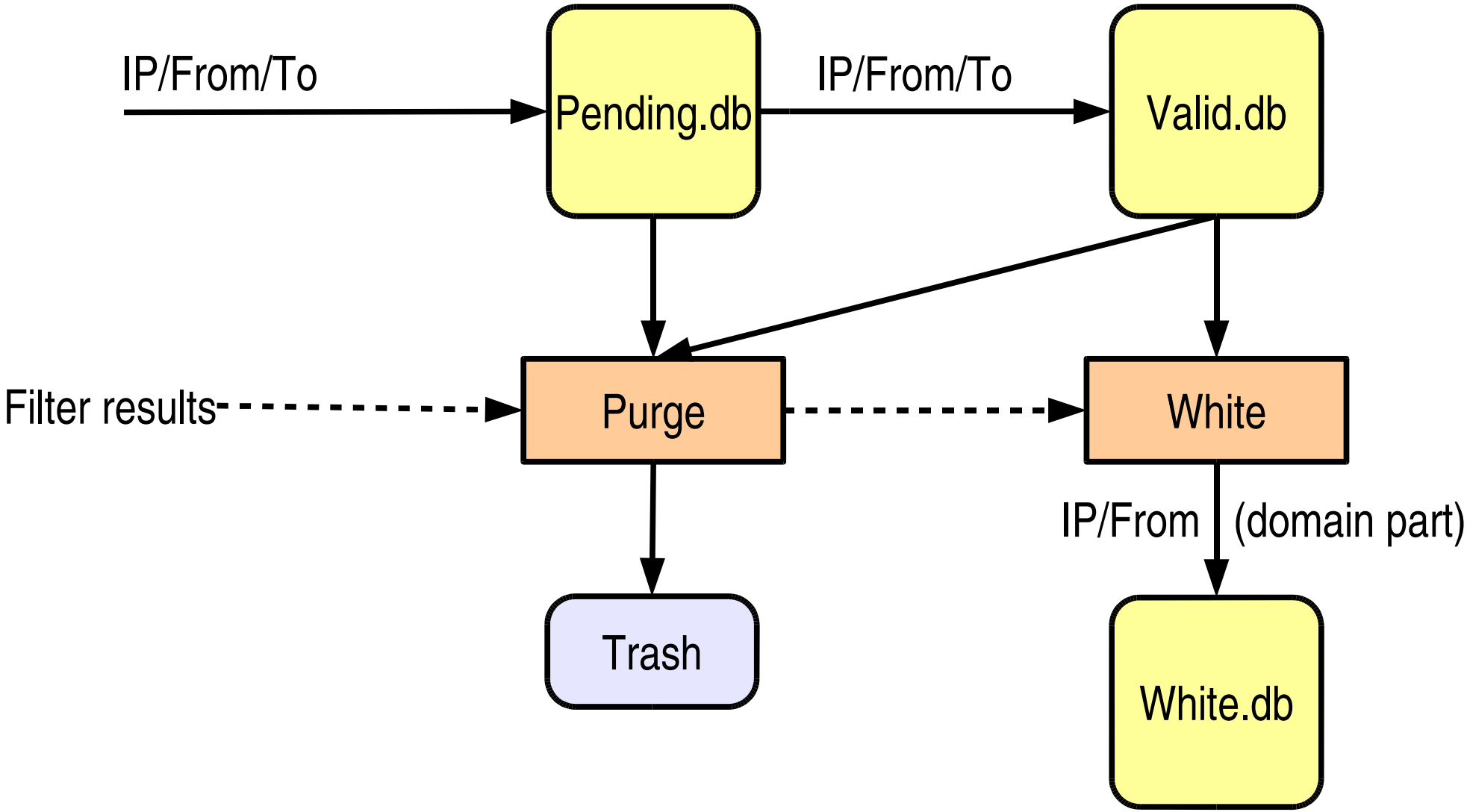
Greylisting

- Problèmes avec Greylisting classique
  - Taille des bases de données – 1M enregistrements - maintenance difficile
  - Enregistrements inutiles : moins de 5 % des entrées en attente sont validées après 12 heures
  - Faux négatifs : ne pas laisser en attente des enregistrements non confirmés...
  - Empoisonnement de la base – répéter un même message, changeant uniquement l'expéditeur.
  - Scalabilité – le nombre d'enregistrements croit avec le nombre de destinataires et non pas avec le nombre de connexions (pas tout à fait vrai, mais...)
- Ex :
  - Un client SMTP envoi un message à N utilisateurs -> N entrées crées dans la base des entrées en attente
  - Il change d'expéditeur et répète l'opération
  - A chaque fois, N nouvelles entrées sont créées



Greylisting avec paramètres temporels variables

- Les paramètres temporels ne sont pas fixes mais dépendent de la «qualité» de l'enregistrement – moins la «qualité» de l'enregistrement est bonne, plus courte est sa durée de vie.
  - Mail from Jose-Marcio.Martins@ensmp.fr venant de paris.ensmp.fr ou smtp.hotmail.com -> Pénalité
  - Durée de vie pour des clients dont le score moyen est élevé -> Pénalité
  - Détection des clients SMTP habituels (plusieurs triplets similaires) -> Bonus (IP/domaine)
- Limitation du nombre d'enregistrements en attente, par client SMTP
- Trois types d'enregistrements au lieu de deux : en attente, validées ou couples blanchis


IP/From/To
Pending.db
IP/From/To
Valid.db
Filter results
Purge
White
IP/From
(domain part)
Trash
White.db

Filtrage d'URLs

- Extraction d'URLs puis recherche dans une liste – bien plus rapide que le filtrage d'expressions regulières
- Deux formats : liste noire DNS ou base de données (hash BerkeleyDB)
- Actuellement
  - surbl.org
    - Efficacité annoncée ~ 70 à 80 % des spams reçus.
    - «If it appears in ham, don't list it»
  - j-chkmail - spams reçus en France
  - spamanti.net – liste noire française

L©avenir de j-chkmail

- Validation de greylisting partagé sur ferme de serveurs de mail (bientôt)
- Filtrage bayesien
- De la doc (contrib ???)
- Des scripts pour traitement par lot
  - Gestion de la quarantaine
  - Gestion des notifications
  - ...
- Mise à plat de certaines parties du filtre - abstractions

Conclusions

- Projet commencé début 1992
- Seul filtrage utilisé sur les MXs du domaine ensmp.fr
  - Temps moyen de traitement des connexions de l'ordre de 50 ms (Sun E280R)
- Le plus gros utilisateur connu : pobox.sk
  - trafic habituel : 20K msgs/h pics de 40K msgs/h
- Des méthodes originales : filtrage des exécutables (pas vraiment), cadence de connexion (depuis 2002) et greylisting modifié
- Seul filtre concerné par la protection du serveur et limitation des ressources
- Filtre toujours en évolution