

Un filtre anti-spam bayésien :

Bogofilter

Solutions Linux 2006 - 1er février 2006



Document sous licence libre Creative Commons by-nc-sa
<http://creativecommons.org/licenses/by-nc-sa/2.0/fr/>
Certains droits réservés



R. Marichez
raphael.marichez@polytechnique.org
<http://falco.bz>

Un filtre anti-spam bayésien : Bogofilter

- Aperçu de la théorie bayésienne
- Implémentation à Polytechnique.org
- Généralisations et limites

Solutions Linux 2006 - 1er février 2006

Un aperçu de la théorie bayésienne (1/5)

Processus d'apprentissage



Thomas Bayes
1702 - 1761

Un aperçu de la théorie bayésienne (2/5)

Tableaux de mots (tokens ou words) rencontrés associés à :



Paul Graham

$s(\mathbf{w})$ taux d'occurrence dans les spams :
nombre de spams contenant w / nombre total de spams

$h(\mathbf{w})$ taux d'occurrence dans les hams :
nombre de hams contenant w / nombre total de hams

Un aperçu de la théorie bayésienne (3/5)

“Un email constitué uniquement de 'w' est-il un spam ?”

$$p(\mathbf{w}) = s(\mathbf{w}) / (s(\mathbf{w}) + h(\mathbf{w}))$$

Optimisation de l'algorithme :

$$p(\mathbf{w}) = \frac{\text{spams contenant } \underline{\mathbf{w}} * \text{hams}}{(\text{spams contenant } \underline{\mathbf{w}} * \text{hams}) + (\text{hams contenant } \underline{\mathbf{w}} * \text{spams})}$$

Un aperçu de la théorie bayésienne (4/5)

Mots rares ou encore inconnus

x : probabilité pour un mot nouveau d'apparaître dans un spam
(= connaissance générale)

s : pondération de la connaissance générale

$p(\mathbf{w})$: probabilité pour un mot connu d'apparaître dans un spam

n : nombre de mails référençant le mot w

note de w :

$$f(w) = \frac{(s * x) + (n * p(w))}{s + n}$$

Un aperçu de la théorie bayésienne (5/5)

Agrégation des scores

$$H = C^{-1}(-2 \ln \prod_w f(w), 2n)$$

Cf.

Paul Graham :

<http://www.paulgraham.com/spam.html>

Gary Robinson :

<http://www.linuxjournal.com/article.php?sid=6467>

Wikipedia (en)

Bogofilter à Polytechnique.org (1/4)

Bogofilter :
langage C



Eric S. Raymond

nombreuses plates-formes :
Linux, BSD, Solaris, Mac OS X, HP-UX, AIX, ...

S'interface sur Postfix, Sendmail, ...

R. Marichez
raphael.marichez@polytechnique.org
<http://falco.bz>

Bogofilter à Polytechnique.org (2/4)

Postfix :

/etc/postfix/master.cf :

```
-o content_filter=bogofilter:
```

```
bogofilter: unix      -      n      n      -      10      pipe  
      flags=R user=bogofilter argv=/etc/postfix/bin/script_shell.sh -f \  
      ${sender} -- ${recipient}
```

Attention aux codes de retour d'erreur

```
/usr/bin/bogofilter -p -e -O msg.$$ || exit $EX_TEMPFAIL
```

ou : Proxy SMTP (clamsmtp, proxsmtp...)

Bogofilter à Polytechnique.org (3/4)

/etc/bogofilter.cf

➤ Ajout d'un entête

➤ X-Spam-Flag: {No|Unsure|Yes}
scores (spamicité) :

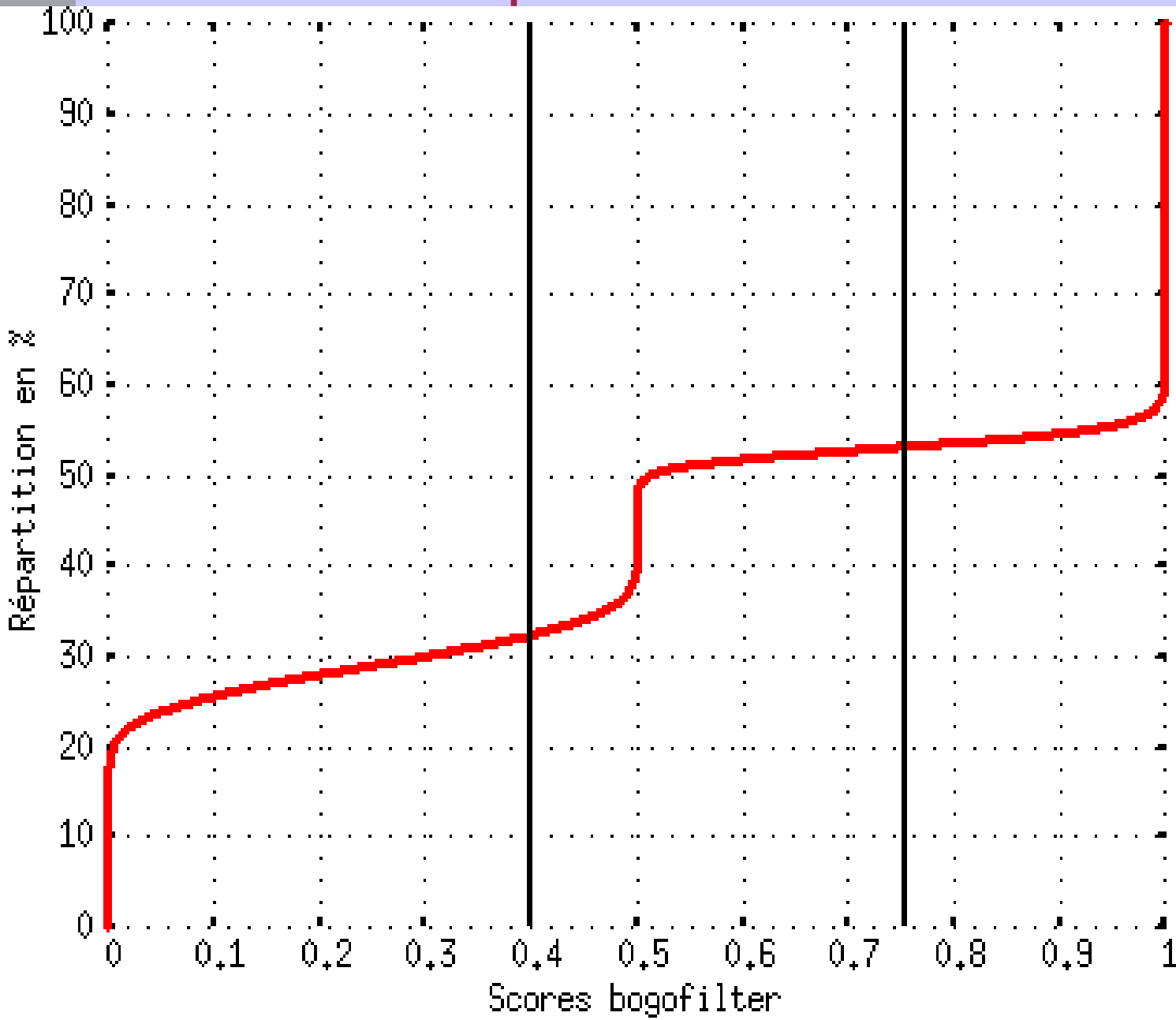
0 -> 0.40 ; 0.40 -> 0.75 ; 0.75 -> 1

➤ Format de l'entête :

```
header_format = %h: %c, tests=bogofilter, spamicity=%p,\  
queueID=%Q, IP=%A
```

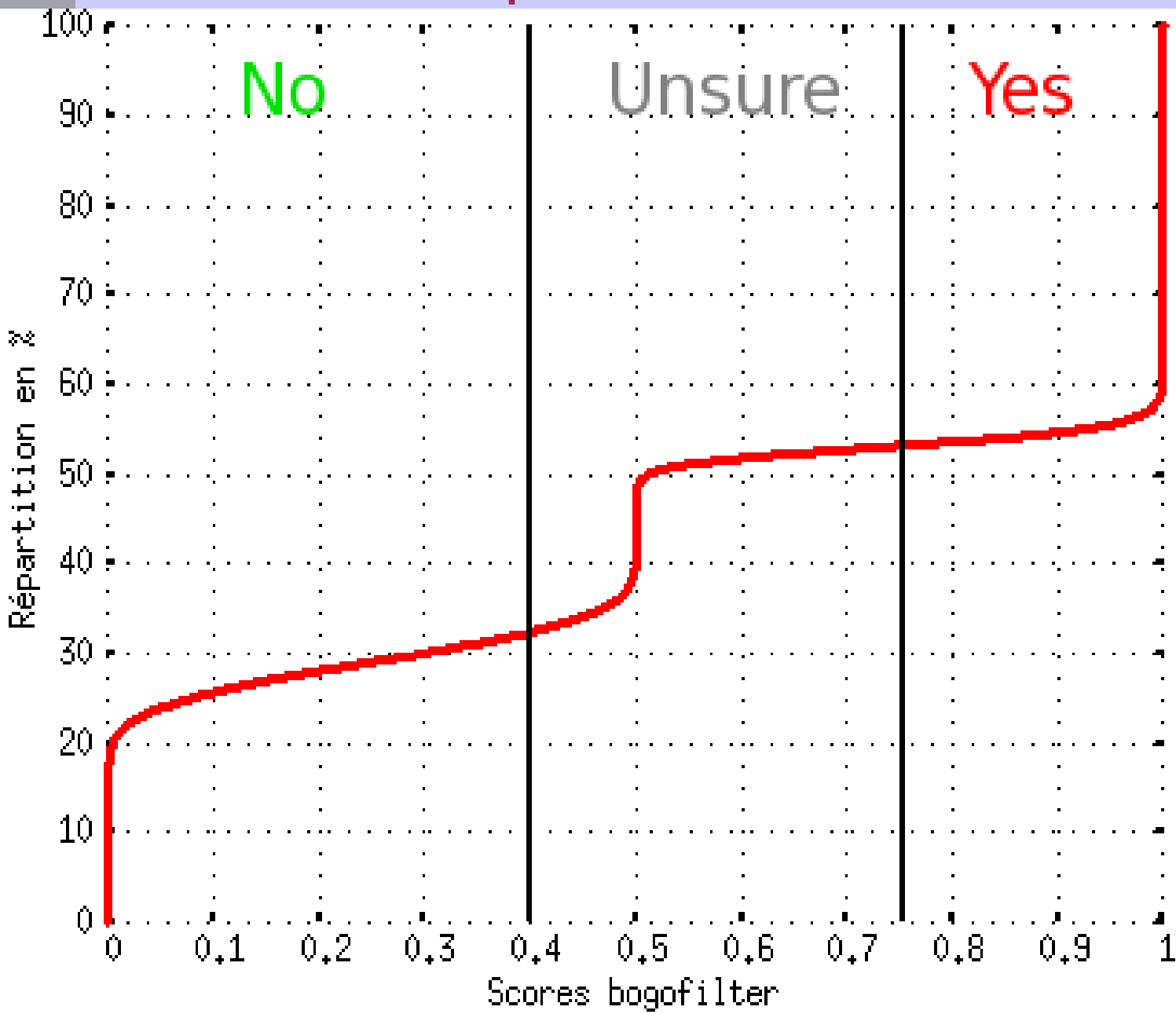
Bogofilter à Polytechnique.org (4/4)

Fonction de répartition



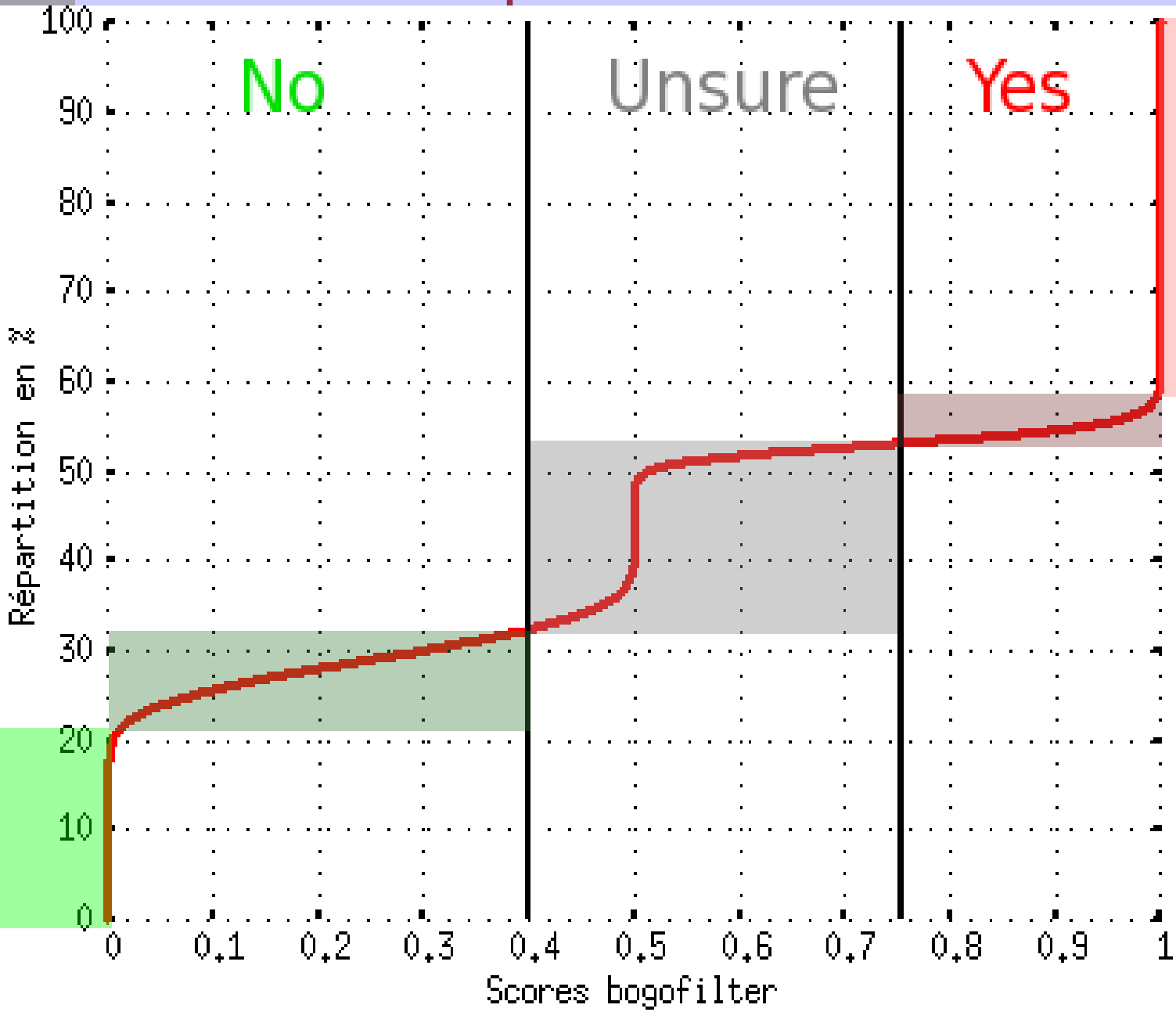
Bogofilter à Polytechnique.org (4/4)

Fonction de répartition



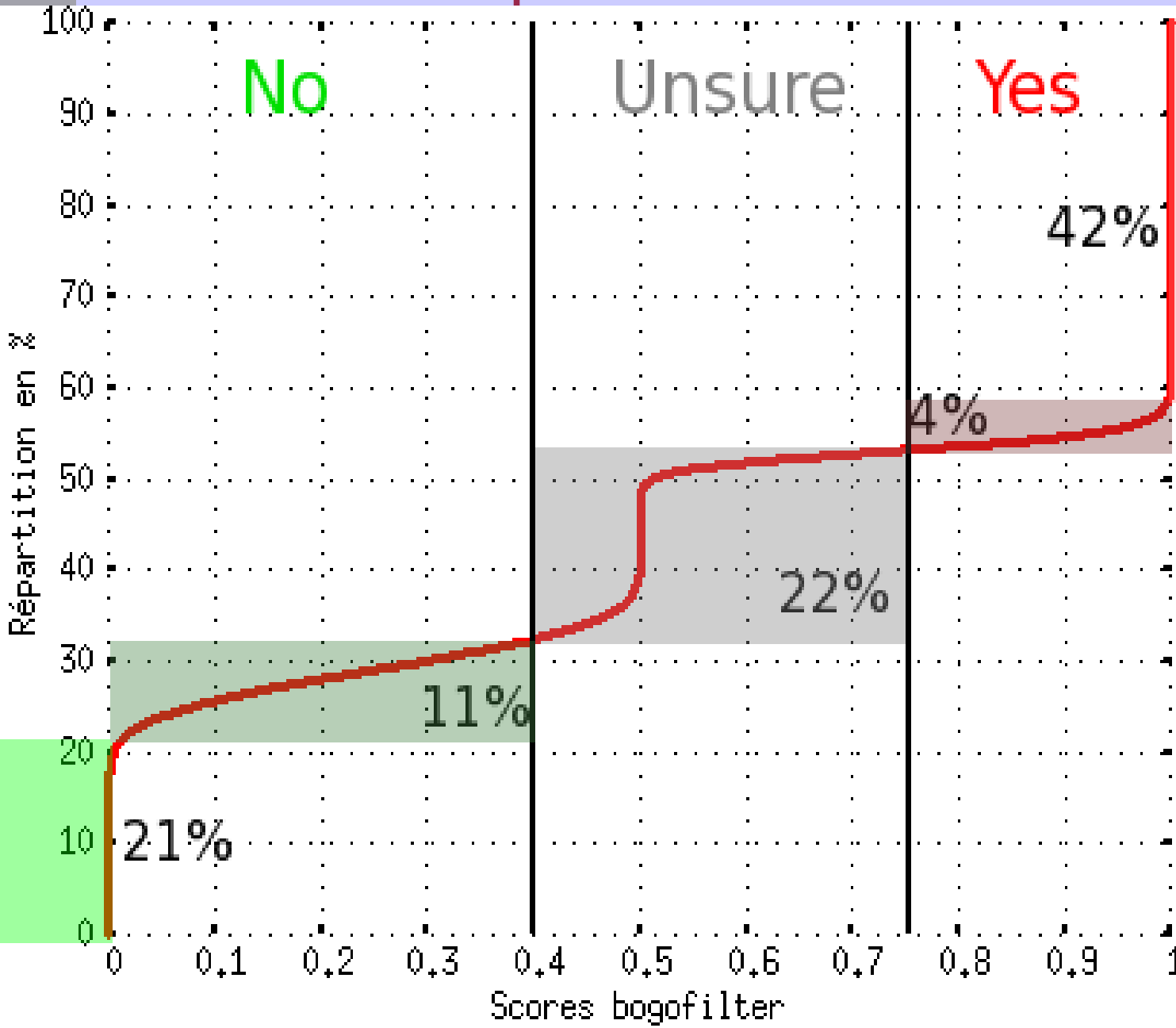
Bogofilter à Polytechnique.org (4/4)

Fonction de répartition



Bogofilter à Polytechnique.org (4/4)

Fonction de répartition



34.0%
 $s = 1.000000$

13.0%
 $s = 0.49 \rightarrow 0.51$

13.0%
 $s = 0.000000$

Généralisations et limites (1/5)

Erreurs rapportées par les utilisateurs (0.4% de rapporteurs réguliers)

Boîtes spam@ et nospam@

Sur 100 messages rapportés en ~10 jours :

spam@ (faux-négatifs) :

contient 85 messages (dont 45 “Unsure”; 40 “No”)

Taux global d'erreur (non détection) : 4%

nospam@ (faux-positifs) :

contient 15 messages (dont 14 “Unsure”; 1 “Yes”)

Taux global d'erreur (fausse détection) : 0,05 %

Généralisations et limites (2/5)

Mots rares ou encore inconnus

note de w :

$$f(w) = \frac{(s * x) + (n * p(w))}{s + n}$$

Bogofilter (compile-time par défaut ou run-time) :

s=0.0178 et x=0.520 actuellement

s=0.0100 et x=0.415 anciennes versions

Généralisations et limites (3/5)

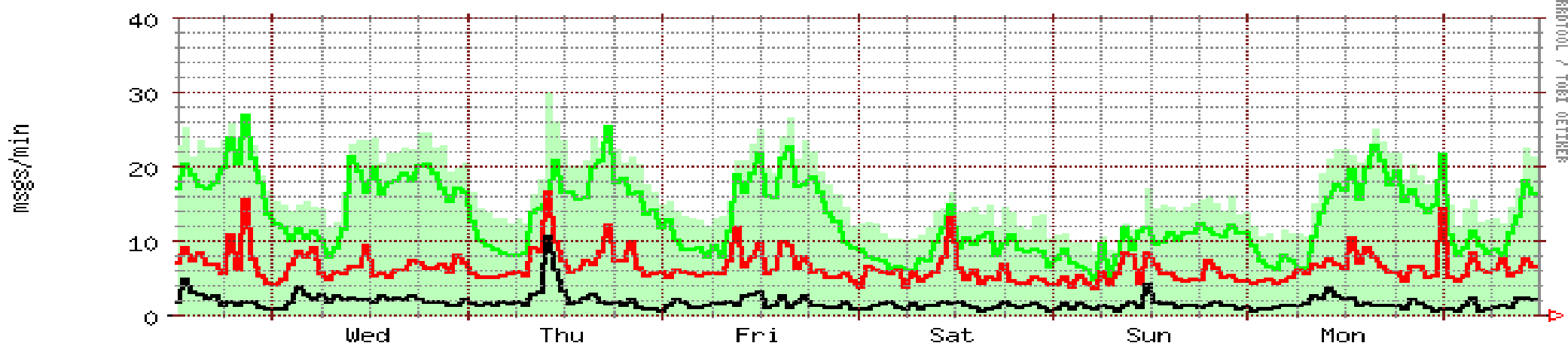
Conditions particulières :

- + : Rapidité
- + : Ensemble homogène d'utilisateurs
- : Langues étrangères
- : Contournements
- : Peu de configuration possible

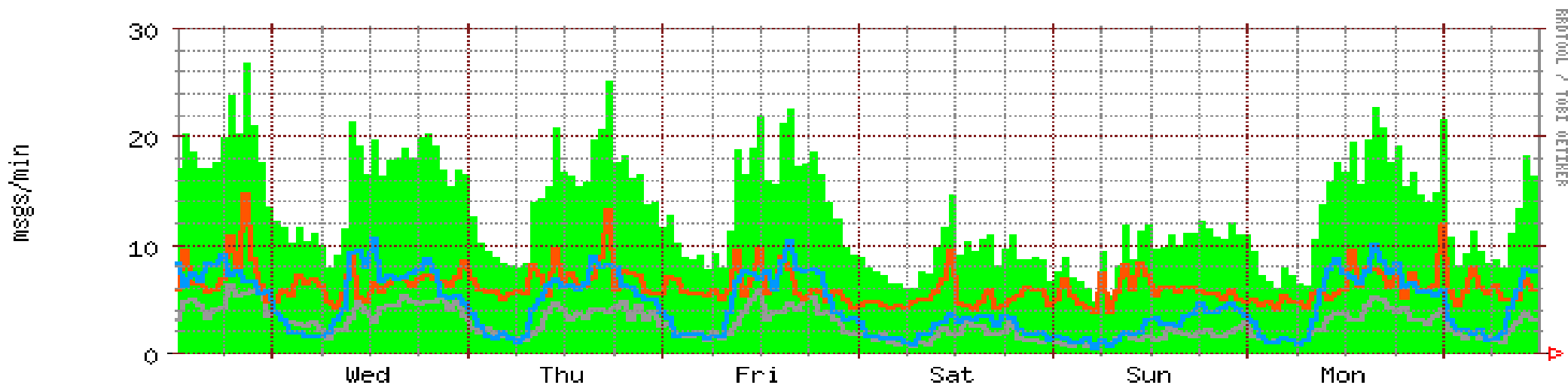
Une base par utilisateur ?

Couplage à d'autres méthodes ?

Généralisations et limites (4/5)

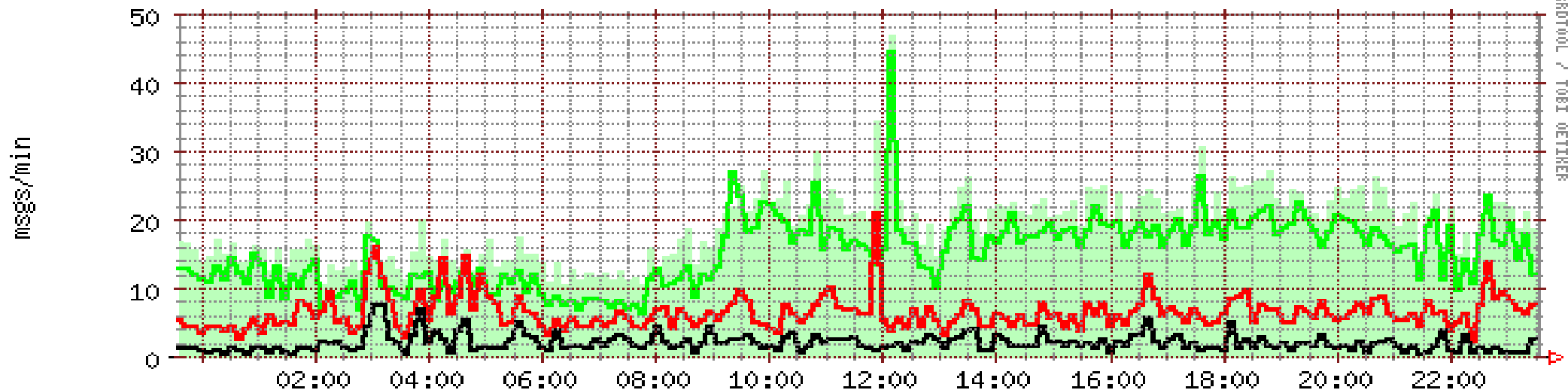


Incoming connections	total:	170677 msgs	avg:	16.81 msgs/min	max:	139 msgs/min
Accepted mails	total:	132135 msgs	avg:	13.01 msgs/min	max:	138 msgs/min
Rejected mails	total:	64324 msgs	avg:	6.34 msgs/min	max:	88 msgs/min
Virus (included in rej.)	total:	16344 msgs	avg:	1.61 msgs/min	max:	32 msgs/min



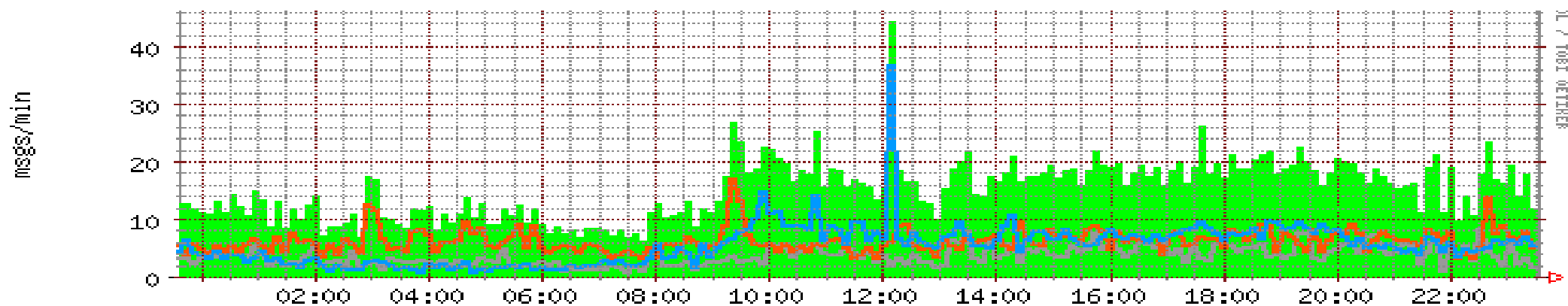
Accepted	total:	132135 msgs	avg:	13.01 msgs/min	max:	138 msgs/min
unsure	total:	26900 msgs	avg:	2.65 msgs/min	max:	39 msgs/min
spam	total:	61398 msgs	avg:	6.06 msgs/min	max:	74 msgs/min
ham	total:	43737 msgs	avg:	4.29 msgs/min	max:	128 msgs/min

Généralisations et limites (5/5)



Incoming connections	total:	27715 msgs	avg:	19.15 msgs/min	max:	139 msgs/min
Accepted mails	total:	21912 msgs	avg:	15.15 msgs/min	max:	138 msgs/min
Rejected mails	total:	9242 msgs	avg:	6.39 msgs/min	max:	49 msgs/min
Virus (included in rej.)	total:	2820 msgs	avg:	1.95 msgs/min	max:	22 msgs/min

[Wed Jan 25 23:35:48 2006]



Accepted	total:	21912 msgs	avg:	15.15 msgs/min	max:	138 msgs/min
unsure	total:	5162 msgs	avg:	3.57 msgs/min	max:	35 msgs/min
spam	total:	8872 msgs	avg:	6.13 msgs/min	max:	29 msgs/min
ham	total:	7806 msgs	avg:	5.40 msgs/min	max:	128 msgs/min

[Wed Jan 25 23:35:48 2006]

Fin

R. Marichez
raphael.marichez@polytechnique.org
<http://falco.bz>