Improving Spam Detection Based on Structural Similarity

Jussara Almeida

**Luiz Henrique Gomes, Fernando Castro,
Rodrigo Almeida, Luis Bettencourt, Virgílio Almeida**

Federal University of Minas Gerais - Brazil

Los Alamos National Laboratory - US

Motivation e Goals

- Volume of spam traffic is increasing at very fast rate
  - 83% of all incoming e-mails in 2005
- Current detection techniques are not fully successful
  - Spammers escape by frequently changing e-mail characteristics traditionally used for detection/filtering
    - E-mail content, sender domain, sender IP address
  - False positives: high "cost" to end-users
- Our goals:
  - Improve spam detection by reducing the number of false positives

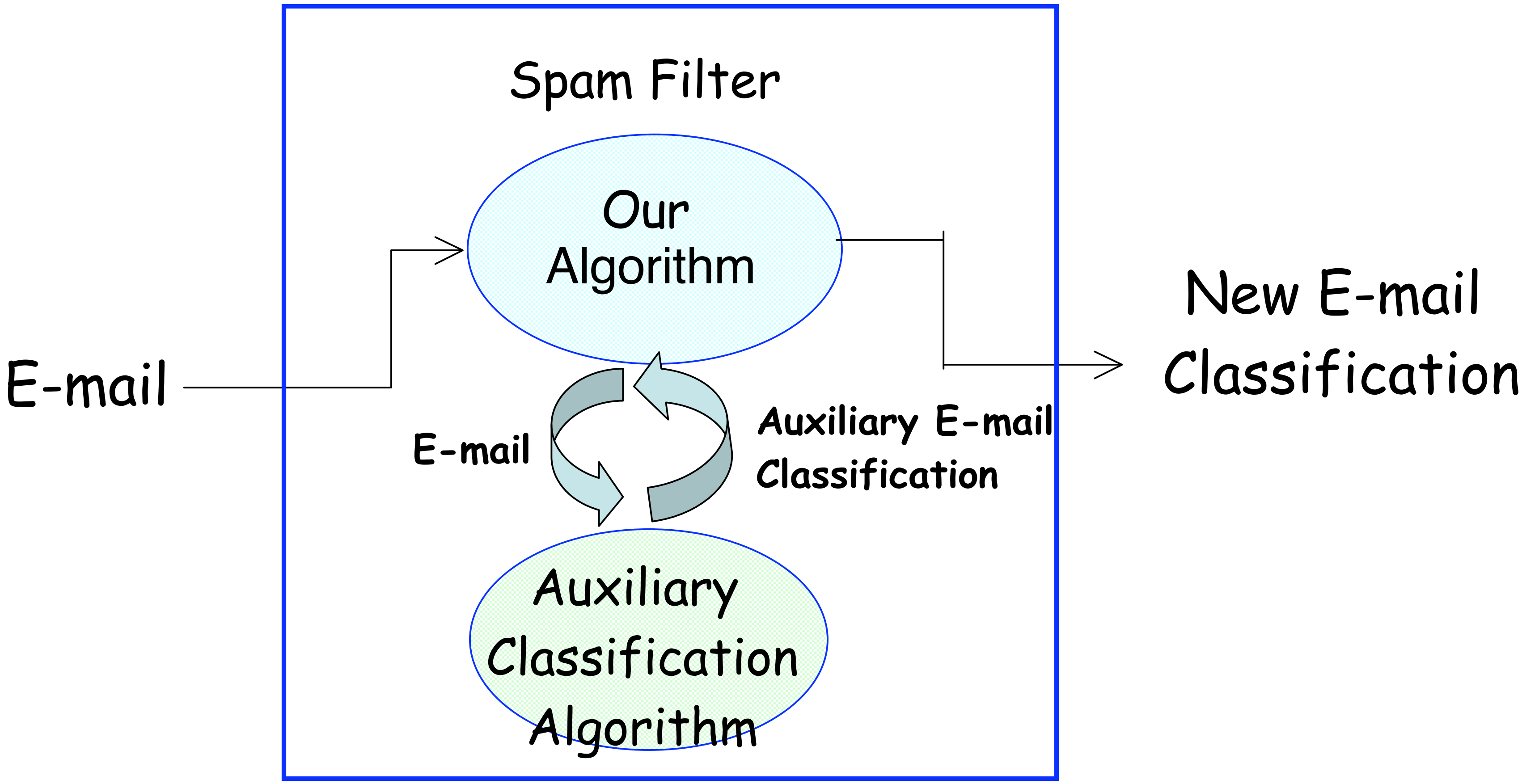
Key Question

What are the e-mail characteristics that are most costly to change from the point of view of the spammer?

Fundamentals of our Algorithm

- Exploit structural relationships between senders and recipients: sender/recipient contact list
- Assumption: contact lists change less frequently than other characteristics
  - Set of recipients targeted by a sender tends to remain stable for longer periods than e-mail content, sender domain or IP address
- Senders / recipients are clustered based on similarity of their contact lists
- Historical information on spam activity from/to a

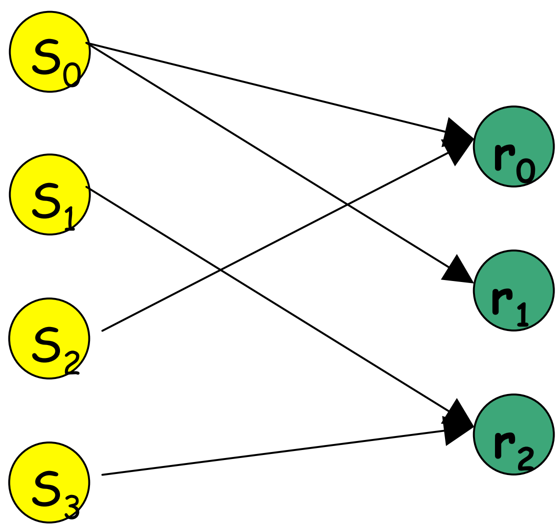
Proposed Architecture



Representing Users and Clusters

- Vectorial representation of an e-mail sender:

$$\vec{s_i}[n] = \begin{cases} 1, & if\ s_i\ sent\ at\ least\ one\ e\text{-}mail\ to\ r_n \\ 0, & otherwise \end{cases}$$



$\vec{S_0} = (1,1,0)$

$\vec{S_1} = (0,0,1)$

$\vec{S_2} = (1,0,0)$

$\vec{S_3} = (0,0,1)$

$\vec{r_0} = (1,0,1,0)$

$\vec{r_1} = (1,0,0,0)$

$\vec{r_2} = (0,1,0,1)$

Representing Users and Clusters

- Vectorial representation of an e-mail sender:

$$\vec{s_i}[n] = \begin{cases} 1, & \textit{if } s_i \textit{ sent at least one e-mail to } r_n \\ 0, & \textit{otherwise} \end{cases}$$

- Vectorial representation of a sender cluster:

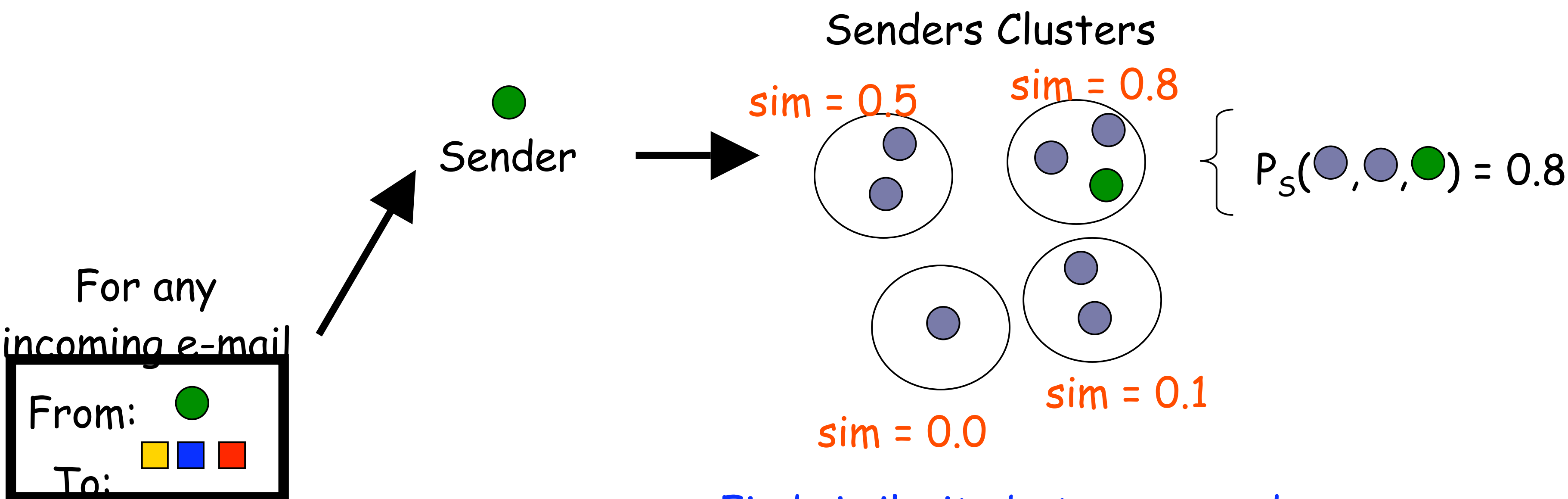
$$\overrightarrow{sc_i} = \sum_{s_i \in sc_i} \vec{s_i}$$

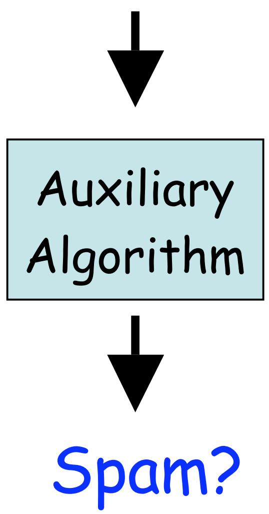
- Similarity between a sender and a sender cluster:

$$sim(\overrightarrow{sc_i}, \vec{s_i},) = \begin{cases} \cos(\overrightarrow{sc_i} - \vec{s_i},\ \vec{s_i}), & \textit{if } \vec{s_i} \in \overrightarrow{sc_i} \\ \cos(\overrightarrow{sc_i},\ \vec{s_i}), & \textit{otherwise} \end{cases}$$

Similar representations for recipients

Our Algorithm





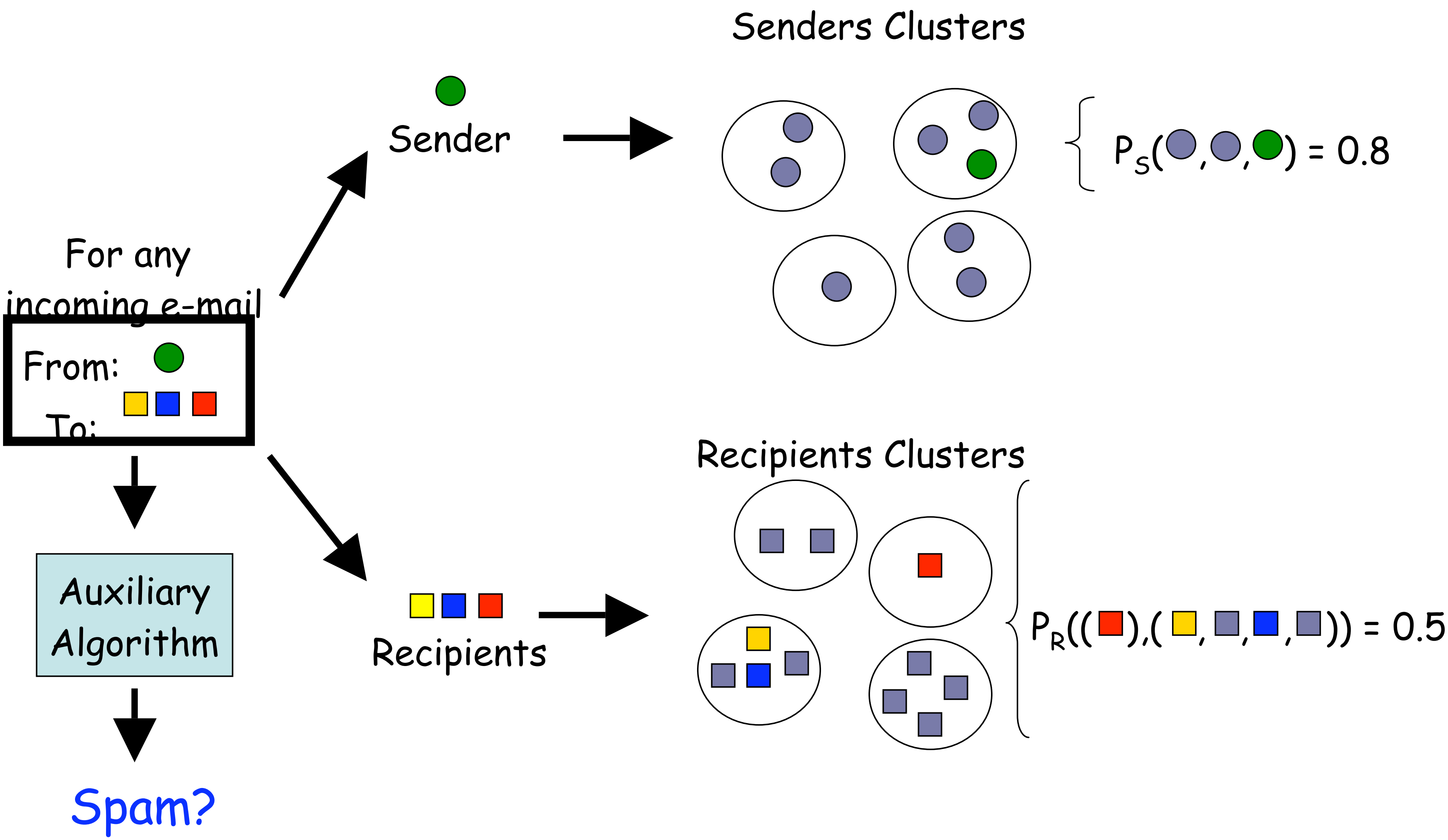
Find similarity between sender and each sender cluster

Add sender to cluster that is most similar to it as long as similarity > τ

Find the probability of cluster sending spam (auxiliary classification of this e-mail and previous e-mails sent by cluster)

Repeat process with the recipients

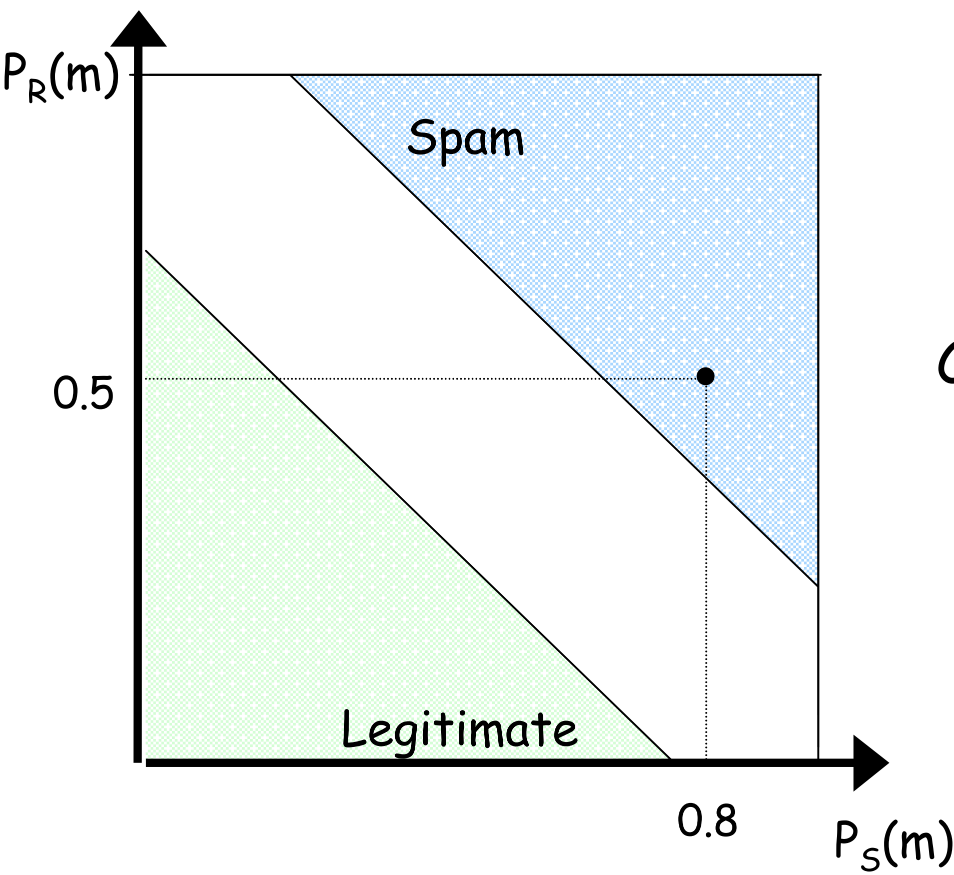
Our Algorithm



Our Algorithm

$P_S(\bullet,\bullet,\bullet) = 0.8$

$P_R((\blacksquare),(\blacksquare,\blacksquare,\blacksquare,\blacksquare)) = 0.5$



Key Ideas

Classify the e-mail as spam if the point (P_S, P_R) falls in the blue area

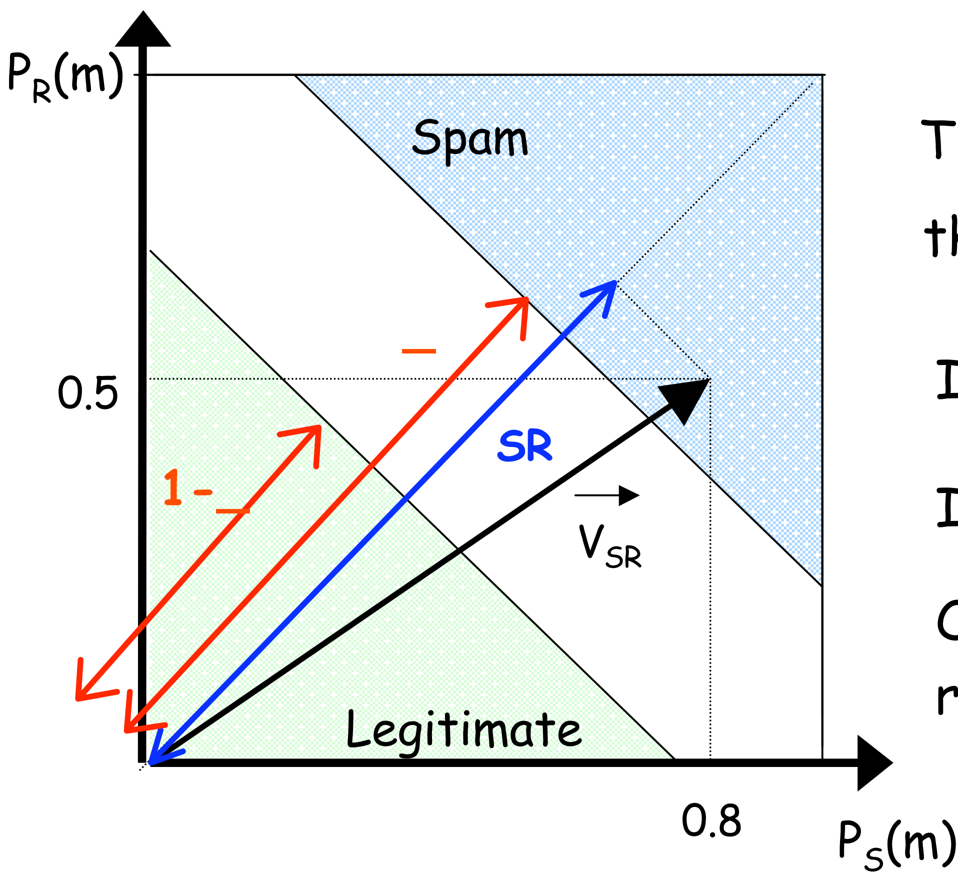
Classify the e-mail as legitimate if the point (P_S, P_R) falls in the green area

Compute a Spam Rank

Our Algorithm

P_S(●,●,●) = 0.8

P_R((■),(■,■,■,■)) = 0.5



Spam Rank Computation:

The Spam Rank vector is:

$\vec{V}_{SR}$(e-mail) = (P_S , P_R) = (0.8, 0.5)

The Spam Rank (SR) is the norm of the projection of $\vec{V}_{SR}$ over diagonal

If SR > _ : classify e-mail as spam

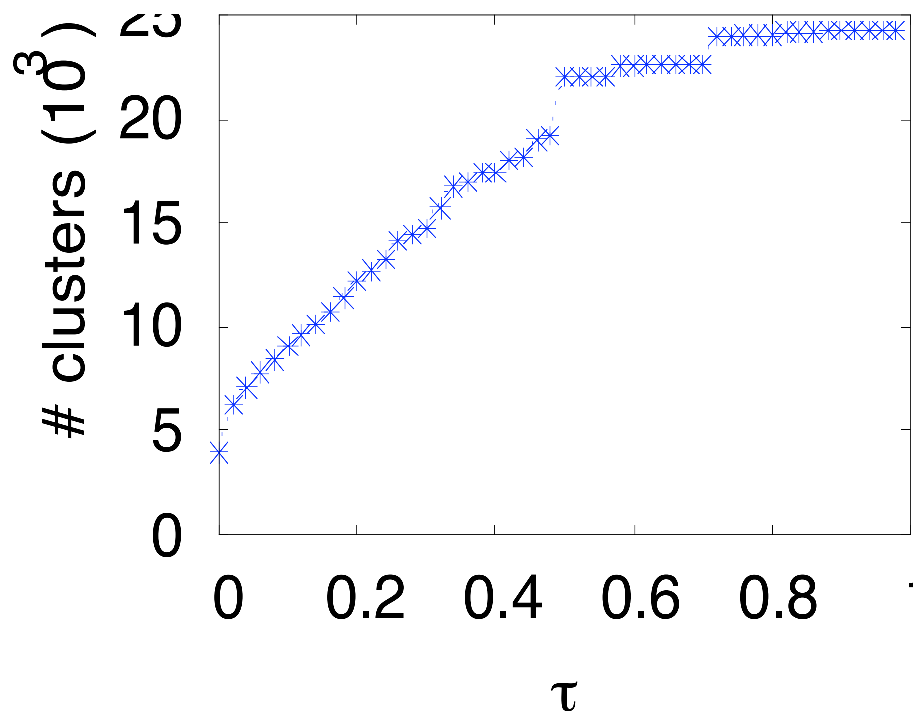
If SR < 1-_: classify it as legitimate

Otherwise, use classification reported by auxiliary algorithm

Preliminary Evaluation

- Eight-day SMTP log of incoming e-mails to UFMG
  - 321K e-mails, 8.3 GB of data
  - 23K distinct sender domain names
  - 34K distinct recipients
- E-mails originally classified by Spam Assassin
  - 154K spams, 0.8 GB
- In our experiments:
  - Auxiliary algorithm = Spam Assassin
  - Sender = sender domain name

Selecting the Similarity Threshold τ

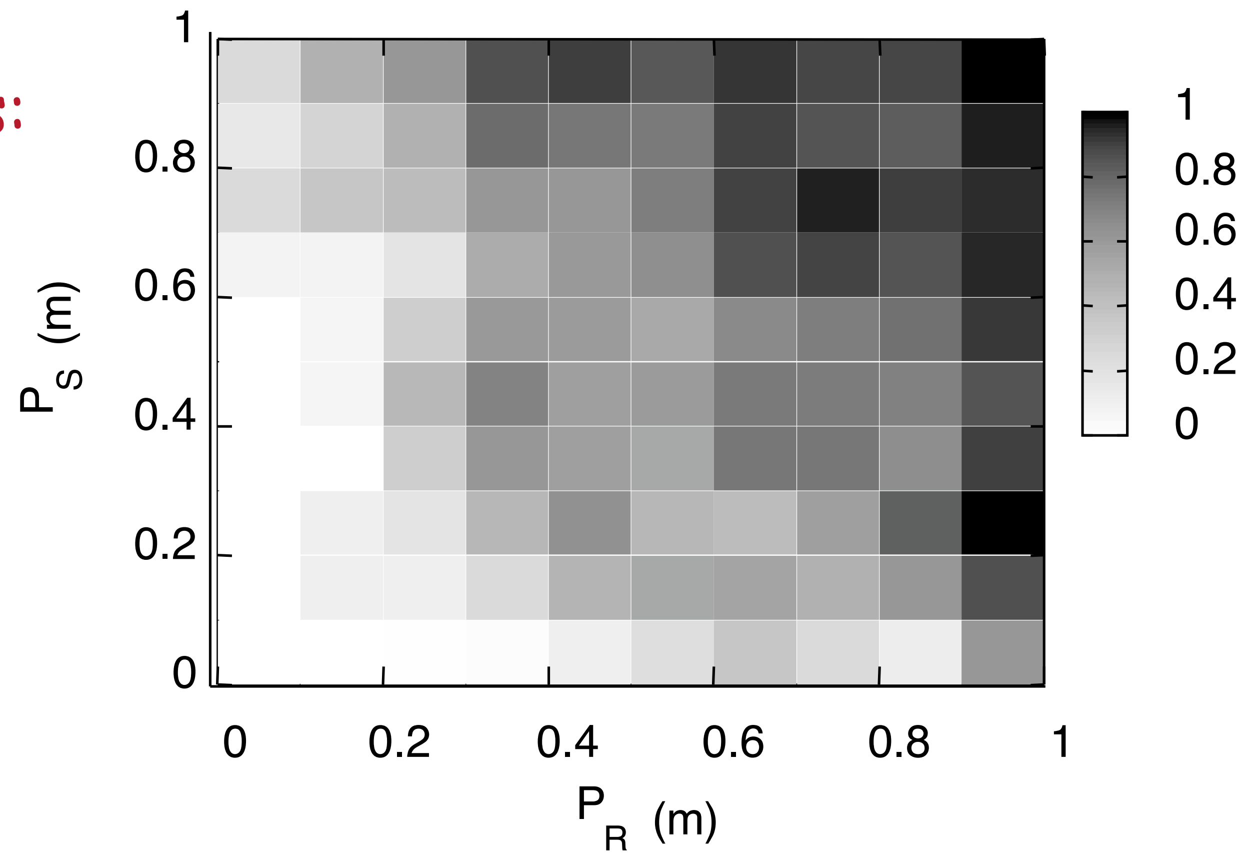


- Number of sender/recipient clusters is roughly stable for $\tau \geq 0.5 \Rightarrow$ use $\tau = 0.5$ in experiments

Effectiveness of Spam Rank

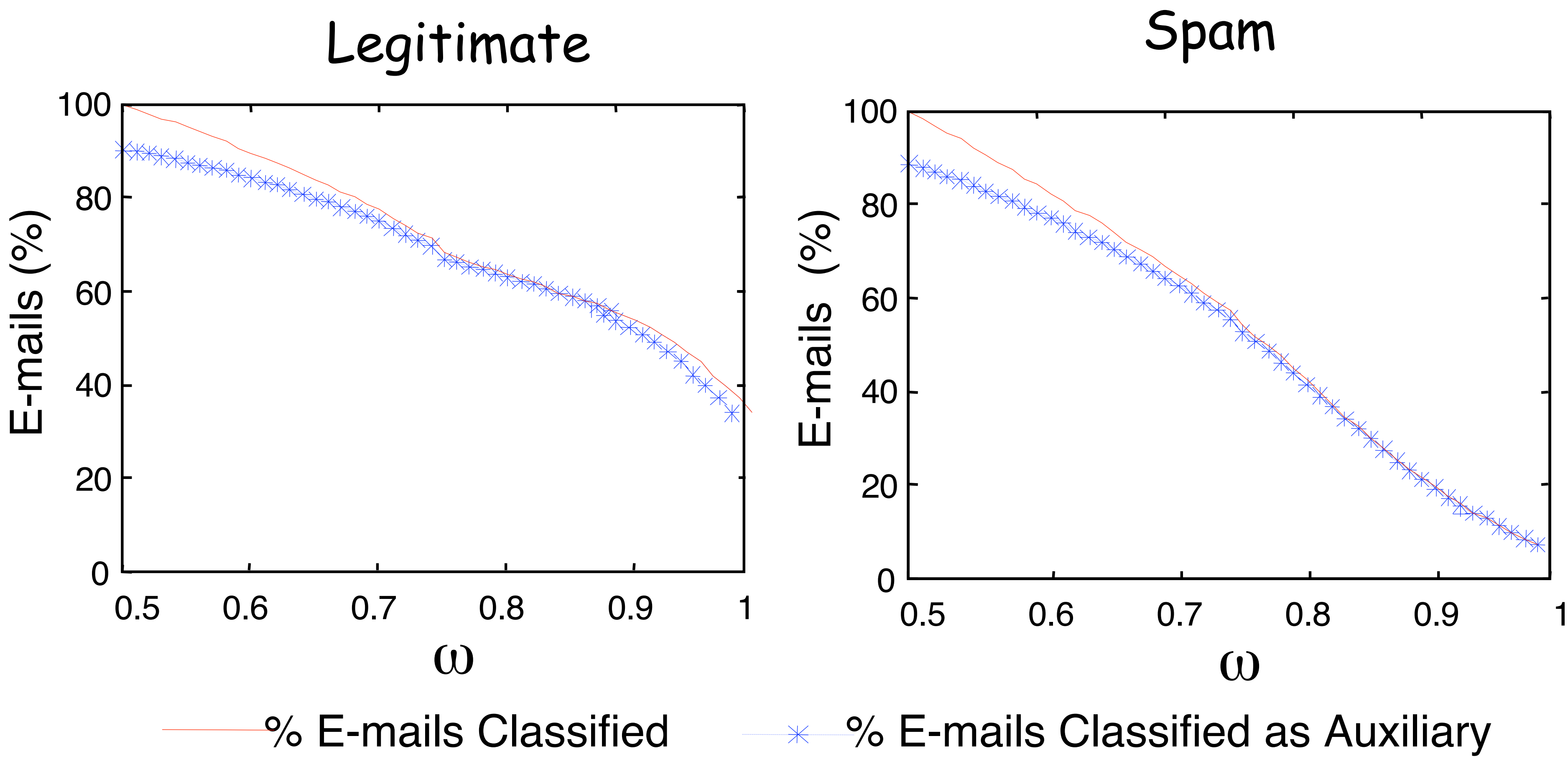
Fraction of spams:

$\tau = 0.5$



- Clusters with high P_S / P_R send/receive large # of spams
  - There are sender/recipients clusters that are predominantly spam/legitimate clusters

E-mail Classification



- Higher _ → smaller # of e-mails can be classified
- For fixed _, we are able to classify more legitimate e-mails than spams

Accuracy of our Classification

$\tau = 0.5$, _ = 0.85:

Classification		% e-mails	Accuracy of our Algorithm
Auxiliary	Our Algorithm		
Spam	Legitimate	0.27% (879 emails)	60%
Spam	Spam	15% (48,277 emails)	99.99%
Legitimate	Spam	0.11% (352 emails)	????

- Our algorithm avoids filtering 528 legitimate e-mails in 8 days
- It moves 352 e-mails originally classified as legitimate to the spam category (unable to verify correctness)

Conclusions and Future Work

- New e-mail classification algorithm that exploits structural similarities of senders and recipients
  - Clustering senders/recipients based on contact lists
- Using historical information of each cluster can improve accuracy of existing detection algorithms
  - Reduction of a non-negligible number of false positives caused by Spam Assassin

- Future Work
  - Several extensions to our algorithm:
    - Take traffic between sender/recipient into account
    - Consider spam probability of a sender-recipient pair