

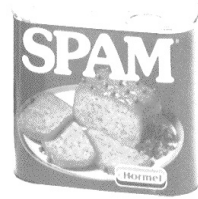
# Tutorial on Junk Mail Filtering

Geoff Hulten

*MSN Safety Team*

Joshua Goodman

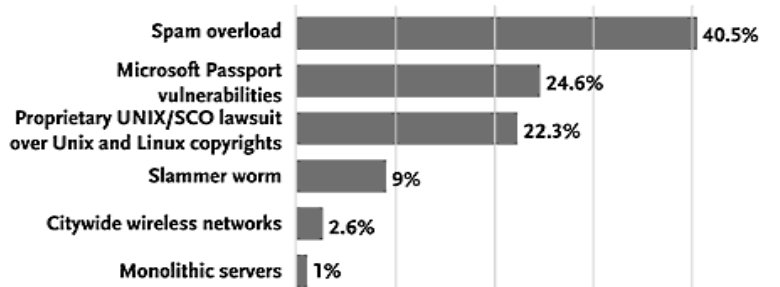
*Microsoft Research*



## InfoWorld Poll July 25, 2003

### WORST DISASTER

When it comes to bogging down readers, even security issues and vendor squabbling barely register compared to spam. As one reader put it, "The volume of e-mail spam is threatening to overwhelm the usefulness of e-mail."



## Statistics

- ◆ Over 50% of all email now, up from 8% in 2001
- ◆ Cost to business \$10B/yr in US (2002)



## Pew Internet Study Numbers

- ◆ 25% of email users say spam has reduced their overall email use
- ◆ 76% of email users are bothered by offensive or obscene content of spam.
  - 24% like obscene or offensive content?
- ◆ 40% of email users spend fewer than 5 minutes a day dealing with spam; 12% spend a half hour or more.
- ◆ Economics favor spam
  - 7% of email users report that they have ordered a product advertised in unsolicited email.
  - Cost of sending spam is only about .01 cents/message!
    - If 1 in 100,000 people buy, and you earn \$11, you make a profit



## Overview Overview

- ◆ Introduction to spam
- ◆ Techniques spammers use
- ◆ Kinds of spam
- ◆ Solutions to spam
- ◆ Machine Learning solutions
- ◆ Analyzing all that data
- ◆ Conclusion



## Introduction to Spam Overview

- ◆ The Pew study and other numbers
- ◆ What is Spam?
- ◆ Isn't this text categorization? Isn't it solved already?
  - No!



## Techniques Spammers Use Overview

- ◆ Obscuring spam
  - All sorts of fun tricks to confuse filters, etc.
- ◆ Getting email addresses
  - Dictionary attacks
  - Web Crawling
- ◆ Sending spam
  - Open proxies
  - Open relays
  - Zombies



## Kinds of Spam Overview

- ◆ Email spam (you already know about that)
- ◆ Chat rooms
- ◆ Instant Messenger
- ◆ Popups (lots of kinds of popups)
- ◆ Search engine spam
  - Tricking search engines into returning your page
- ◆ Conclusion: If you can advertise for free, someone will



## Solutions to Spam Overview

- ◆ Filtering
  - Machine Learning
  - Matching/Fuzzy Hashing
  - Blackhole Lists (IP addresses)
- ◆ Postage
  - Turing Tests, Money, Computation
- ◆ SmartProof
  - Combines the best of Filtering with all of Postage
- ◆ Safe sender lists (lists of known good senders)
- ◆ Disposable Email Addresses
- ◆ Stopping Outbound Spam
- ◆ Deployment issues (client vs. server, etc.)



## Machine Learning Solutions Overview

- ◆ Evaluating Spam Filters
  - Hard to get good corpora for evaluation
  - People take shortcuts, get overoptimistic results
- ◆ Building Models of Spam
  - “Bayesian” Techniques (Naïve Bayes)
  - Support Vector Machines
  - Rules/Trees
- ◆ Miscellaneous (e.g. Beating machine learning solutions)
- ◆ Personalization (Threshold drift problem)



## DATA Overview

- ◆ How we get our data at Microsoft
  - How we got millions of hand labeled messages for free!
- ◆ Analysis of data from the feedback loop
  - Where spam is from?
  - Can legal approaches work?



## Conclusion Overview

- ◆ Spam and Email as a new field of study
  - Related to, but different than text classification
- ◆ Spam is a huge problem,
  - It will be with us in one form or another for the foreseeable future (email, SPIM, web, etc.)
- ◆ We'll need a variety of solutions
  - Filtering will be part of the solution
- ◆ Email and Spam present tons of great new problems!



## Introduction to Spam Overview

- ◆ The Pew study and other numbers
- ◆ What is Spam?
- ◆ Isn't this text categorization? Isn't it solved already?
  - No!
  - Why spam is hard



## What is Spam?

- ◆ Pew study numbers
  - 92%: unsolicited mail containing adult content
  - 89%: unsolicited financial offers, etc.
  - 76%: unsolicited political or religious mail
  - 32% consider unsolicited commercial email to be spam, *even if it came from a sender with whom they've "already done business."*
- ◆ Typical legal definition: unsolicited commercial email from someone without a pre-existing business relationship.
- ◆ Our definition: whatever our users tell us (when we are showing you numbers from one of our studies, that's the definition we will use)



Isn't this just text categorization?  
Isn't it solved already?

- ◆ For a machine learning person, natural inclination is to use standard text classification techniques.
- ◆ Simple text-categorization techniques, e.g. Naive Bayes, have reported 99% accuracy.



Isn't this just text categorization?  
Isn't it solved already?

**NO!**



## Why most people don't get 99% spam filtering

- ◆ Spammers only try to engineer around commonly used techniques
  - You can try out a technique with a few people and it will work great
  - You ship it to 100 million people and spammers immediately try to engineer around it.
- ◆ Reports with Naive Bayes are typically for personalized filters
  - Reports assume users correct every mistake
    - User has to examine all mail marked as junk and check if good
  - Most people don't want to have to correct their mistakes
  - Possible catastrophic filter degradation if all mistakes not corrected or user misunderstands what spam is (e.g. off topic posts to newsgroups.)
- ◆ Later we'll describe other issues – hard to get good corpora to evaluate on, and many shortcuts lead to overoptimistic results



## Text classification alone is not enough

- ◆ Spammers now often try to obscure text (examples coming up)
  - Inherently adversarial problem
- ◆ Even text features are special
  - Treat Subject lines as different than body text
- ◆ Use special features
  - Example: Mail in the middle of the night is more likely to be spam than mail in the middle of the day
- ◆ Can combine text classification/learning with other techniques, e.g. postage techniques
- ◆ All sorts of interesting problems and specialized solutions



## Techniques Spammers Use Overview

- ◆ Obscuring mail to avoid filters
  - Content in Images, Good word Chaff, Content Chaff, URL Spamming, Hidden Text, Character Encoding, etc...
- ◆ Getting email addresses
  - Dictionary attacks
  - Web Crawling
- ◆ Sending spam
  - Open proxies
  - Open relays
  - Zombies



## Obscuring mail Overview

- ◆ Modify spam to avoid filters
  - Content in Images
  - Good word Chaff
  - Content Chaff
  - URL Spamming
  - Hidden Text
  - Character Encoding
  - Etc...





# Weather Report Guy

- ◆ Content in Image
- ◆ Good Word Chaff

Weather, Sunny, High 82, Low 81, Favorite...

**100's of Lenders Compete for your Loan to get you the *Lowest Rate!***

- Refinancing
- New Home Loans
- Debt Consolidation
- Debt Consultation
- Auto Loans
- Credit Cards
- Student Loans
- Second Mortgage
- Home Equity

**Good Credit - Bad Credit  
Bankruptcy - Foreclosure**



Interest Rates are at their lowest point in 40 years! We help you find the best rate for your situation by matching your needs with hundreds of lenders!

**100% Free Service!**

**Click Here To Begin**

henny4 info  
procto ja kigs.

## Weather

NA, NA - Sunny

High: 82 , Low: 81 degrees

## Favorites



# The Hitchhiker Chaffer

- ◆ Content Chaff
  - Random passages from the Hitchhiker's Guide
  - Footers from valid mail

"This must be Thursday," said Arthur to himself, sinking low over his beer, "I never could get the hang of Thursdays."

Express yourself with MSN Messenger 6.0...

We are offering a 7 day / 6 night vacation at a huge discount and we will also add 4 / 3 night vacation to any of our 17 destinations.

Many hotels pay for most of your accommodations in exchange for you to view their resorts, hoping you will spend money on their services.

Click below to claim your trip:

<http://www.nickeltype.com/ttl/>

To update your list preference: [nickeltype.com/re](http://www.nickeltype.com/re)

"Does your health insurance cover pets?," "This must be Thursday," said Arthur to himself, sinking low over his beer, "I never could get the hang of Thursdays."

Express yourself with MSN Messenger 6.0 -- download now!  
[http://www.msnmessenger-download.com/tracking/teach\\_general](http://www.msnmessenger-download.com/tracking/teach_general)



# Hitchhiker Chaffer's Later Work

- ◆ There is nothing fancy about this spam

What would you consider a good deal? How about two free airline tickets to the destination of your choice? If you think THAT is a good deal then [click here to keep reading!](#)

- "A spam filter will catch that in its sleep" – anonymous

To change your list options: 1-800-2offer.biz/re

- ◆ Or maybe not...



# Hitchhiker Chaffer's Later Work

- ◆ Hidden Text
- ◆ Content Chaff
- ◆ URL Spamming

Dear adena,

What would you consider a good deal? How about two free airline tickets to the destination of your choice? If you think THAT is a good deal then [click here to keep reading!](#)

To change your list options: 1-800-2offer.biz/re

Also included a number of unusual statements made by candidates during, 'On display? I eventually had to go down to the cellar to find them.'

Also included are a number of unusual statement made by candidates during, 'On display? I eventually had to go down to the cellar to find them.'

<http://join.msn.com/?Page=features/es>

Get McAfee virus scanning and cleaning of incoming attachments. Get Hotmail  
<http://join.msn.com/?PAGE=features/es>



# Secret Decoder Ring Dude

- ◆ Another spam that looks easy

**Online Pharmacy - 24/7 Customer Care**

Hundreds of products for dozens of ailments. We carry everything from Pain Relief to Skin Care products. Our most popular include:

- Viagra - Proven sexual aid to enhance performance
- Soma - The best in muscle relaxation available
- Phentermine - Safe, proven way to reduce weight
- ... and more!

**Visit the Online Pharmacy for your medical needs**

- ◆ Is it?

Please unsubscribe me



# Secret Decoder Ring Dude

- ◆ Character Encoding
- ◆ HTML word breaking

Phar&#109;acy	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; padding: 2px;">Prod&amp;#117;c&lt;!LZJ&gt;t&lt;!LG&gt;s</td> <td style="padding: 2px;"> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; padding: 2px;">Phar&amp;#109;acy</td> <td style="padding: 2px;">Prod&amp;#117;c&lt;!LZJ&gt;t&lt;!LG&gt;s</td> </tr> </table> </td> </tr> </table>	Prod&#117;c<!LZJ>t<!LG>s	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; padding: 2px;">Phar&amp;#109;acy</td> <td style="padding: 2px;">Prod&amp;#117;c&lt;!LZJ&gt;t&lt;!LG&gt;s</td> </tr> </table>	Phar&#109;acy	Prod&#117;c<!LZJ>t<!LG>s
Prod&#117;c<!LZJ>t<!LG>s		<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; padding: 2px;">Phar&amp;#109;acy</td> <td style="padding: 2px;">Prod&amp;#117;c&lt;!LZJ&gt;t&lt;!LG&gt;s</td> </tr> </table>	Phar&#109;acy	Prod&#117;c<!LZJ>t<!LG>s	
Phar&#109;acy	Prod&#117;c<!LZJ>t<!LG>s				



# Diploma Guy

## ◆ Word Obscuring

Dplmoia Pragorm  
Caerte a mroe prosoeprus

### Dlpmoia Pragorm

Caerte a mroe prosoeprus fituae for yorsuelf

Recieve a ful doplma from non arecredited  
unversities baesd upon your rael life exmeriepc

You will not be tseted, or rteeviewnd  
Receive a Msetar's, Bachelor's or Dtctarooe

Clal 24 horus a day 7 days a week

1 - 2 7 0 - 8 1 7 - 8 2 4 7



# Diploma Guy

## ◆ Word Obscuring

Dipmloa Paogrrm  
Cterae a more presporous

### Dipmloa Paogrrm

Cterae a mroe presporous future for yrloseuf

Rvceiee a full dipmola form non arecredited  
universities based upon yuor real life eneriepxce

You will not be tteesd, or intrievewed  
Reecieve a Msaetr's, Bochaler's or Dtctoraoe

Call 24 hrous a day 7 dyas a week

1 - 2 7 0 - 8 1 7 - 8 2 4 7





# Diploma Guy

## ◆ Word Obscuring

Dimlpoa Pgorram  
 Cearte a more poosperrus

### **Dimlpoa Pgorram**

Cearte a more poosperrus fituae for yeulsof

Recveie a flul dlpmoia from non accredited  
 uviterinies beasd uopn your rael life exeprience

You will not be teetd, or intvweiered  
 Rveciee a Mtsear's, Bheaclor's or Dttcoraoc

Call 24 hrous a day 7 days a week

1 - 2 7 0 - 8 1 7 - 8 2 4 7



# Diploma Guy

## ◆ Word Obscuring

Dpmlpia Pragorm  
 Caetre a more prorpeosus

### **Dpmlpia Pragorm**

Caetre a more prorpeosus fituae for ysueolf

Rcievee a flul dpiloma from non arccedietd  
 univereitnis based upon yuor rael life exeeprince

You will not be teetd, or inveritwed  
 Reecieve a Mtsear's, Bloheacr's or Dootcrate

Call 24 huors a day 7 dyas a week

1 - 2 7 0 - 8 1 7 - 8 2 4 7





# Diploma Guy

## ◆ Word Obscuring

Dplmoia Pragorm  
Carete a mroe propseous

### Dlpmoia Pgorarm

Carete a mroe propseous firute for yourself

Receive a flul dlpmoia from non atciedred unvneristeis besad upon your real lfe eeprincee

You will not be tetsed, or intewvired Rceveise a Meatsr's, Beohalcr's or Dotcoarte

Clal 24 huros a day 7 days a week

1-270-817-8247



# More of Diploma Guy

## ◆ Diploma Guy is good at what he does

lzqqbdhp

```

'U' 'D' 'M' '1
'n' 'i' 'a'
'i' 'p' 's' '2
'v' 'l' 't' '7
'e' 'o' 'e' '0
'r' 'm' 'r'
's' 'a' 's' '8
'i' 's' '1
't' 'P' '7
'y' 'H'
'D' '8
'B' '4
'a' '7

```

vvsmll

# Trends in Spam Exploits

- Based on 1,200 spam messages sent to Hotmail

Exploit	2003 Spam	2004 Spam	Delta (Absolute %)	Description
Word Obscuring	4%	20%	16%	Misspelling words, putting words into images, etc.
URL Spamming	0%	10%	10%	Adding URLs to non-spam sites (e.g. msn.com).
Domain Spoofing	41%	50%	9%	Using an invalid or fake domain in the from line.
Token Breaking	7%	15%	8%	Breaking words with punctuation, space, etc.
MIME Attacks	5%	11%	6%	Putting non-spam content in one body part and spam content in another.
Text Chaff	52%	56%	4%	Random strings of characters, random series of words, or unrelated sentences.
URL Obscuring	22%	17%	-5%	Encoding a URL in hexadecimal, hiding the true URL with an @ sign, etc.
Character Encoding	5%	0%	-5%	Phar&#109;acy renders into Pharmacy.

## Obscuring mail Conclusion

- ◆ Tons of ways to obscure mail
- ◆ Sometimes we can detect obscuring method (leads to something strange and detectable)
  - Example: character encoding
  - When detection software becomes common, spammers change method
- ◆ Always something new to look for
  - Not just adding new features
  - Change the space that needs to be modeled



## Techniques Spammers Use Overview

- ✓ Obscuring mail to avoid filters
  - ✓ Content in Images, Good word Chaff, Content Chaff, URL Spamming, Hidden Text, Character Encoding, etc...
- ◆ Getting email addresses
  - Dictionary attacks
  - Web Crawling
- ◆ Sending spam
  - Open proxies
  - Open relays
  - Zombies



## Getting Email Addresses

- ◆ Dictionary Attacks:
  - Try millions or billions of possible email addresses
    - Put together first-name and last-name, or first-name + number, etc.
    - See if the mail “bounces”: if not, you have a live address
    - Use “Web beacons” to check if mail is being read
- ◆ Web crawling:
  - Look for email addresses on web pages
- ◆ Send spam to these addresses or sell them to other spammers



## Sending Spam: Open Relays

- ◆ In the old days, before the fully connected internet, it was “nice” to forward mail meant for someone else.
- ◆ A mail server is an “open relay” if it will forward on behalf of anyone.
- ◆ Spammers love open relays
  - A little harder to trace them
  - Shifts bandwidth and other burdens to someone else
- ◆ But open relays are added to blackhole lists, and quickly cannot send legitimate mail



## Sending Spam: Open Proxies

- ◆ These are web-page proxy servers
  - Used for getting web-pages past firewalls
  - Should have nothing to do with email
- ◆ Spammers exploit holes
  - Exploit a hole that you can use some proxies to send email
  - Exploit another hole that anyone can access the proxy-server
    - Should be behind firewall
  - Both holes must be present to use an open proxy
- ◆ Spammers really love these
  - Almost impossible to trace spammer
  - Less incentive for owner to close the proxy than to close open mail relays: they don't care if their web proxy is on an email blackhole list
- ◆ Almost impossible to have a Microsoft open proxy
  - Default settings are correct and it's difficult to change them



## Sending Spam: Zombies

- ◆ Consumer computers taken over by viruses or trojans
  - Spammer tells them what to send
  - Very difficult to trace
  - Very cheap for spammer
  - Rent a zombie for about \$3/month!
- ◆ As much as 40-60% of spam may originate from zombies now!



## Techniques Spammers Use Conclusion

- ◆ Spammers are incredibly clever
  - Always finding new ways to obscure mail or send spam
- ◆ Spammers are evil
  - Willing to resort to blatantly illegal activity
- ◆ Always a new challenge



## Kinds of Spam: Overview

### “Advertising Wants to be Free”

- ◆ Email spam (you already know about that)
- ◆ Newsgroup spam
- ◆ Chat rooms
- ◆ Instant Messenger
- ◆ Popups
  - Web pages
  - Spyware
  - Windows Messenger (not IM)
- ◆ Search engine spam
  - Link spam
  - Word spam
  - Blog spam
- ◆ Conclusion: If you can advertise for free, someone will.



## Newsgroup Spam

- ◆ Just like email spam
- ◆ Cheaper for spammers – post once  
download many times
- ◆ Not big issue because ... no one uses  
newsgroups
  - That's because they are all full of spam
  - Could the same thing happen to email?!



## Chat Room Spam

- ◆ Spambots come in and pretend to chat
  - But really just advertising porn sites
  - Some spambots trivial
    - Don't talk at all, but take up space
    - Link to porn spam in their profile
  - Some spambots very sophisticated
    - You can have a short conversation with them before they try to convince you to go to their website
- ◆ MSN closed its free chat rooms



## Instant Messenger Spam "SPIM"

- ◆ Send messages to people via IM
- ◆ Microsoft solved this by requiring to get permission before IMing
- ◆ Spammers put spam in their "name" – so permission request message now has spam!



# Popup Spam

- ◆ Web page popups
  - You go to a web page, and get a popup
  - May be a “pop under” that appears under all other windows, so you don’t even know where it came from
- ◆ Spyware (e.g. Gator)
  - Software installed on your computer either without your permission, or where permission is hidden deep in license agreement.
  - Creates popups all the time
- ◆ Messenger Spam (not IM)
  - Method meant to deliver notices like “Printer is out of paper”
  - Spammers exploit it to create notices like “Buy a diploma”



# Search Engine Spam

- ◆ Link spam
  - Search engines use number of links to determine rankings
  - Spammers create millions of pages that link to their site
  - Fake pages may be realistic and may be returned as search results, too.
- ◆ Word spam
  - Spammers put misleading words on their page, e.g. celebrity names or technical terms
  - Page is actual porn
- ◆ Blog spam
  - Some web pages let anyone post comments
  - Spammers automate comment posting, add links to their pages



## Kinds of Spam

### Conclusion

- ◆ Anywhere you can place free advertising, someone will
- ◆ Even after we solve email spam, there will be lots more to work on
- ◆ Some techniques from email spam transfer, some don't



## Solutions to Spam

### Overview

- ◆ Filtering
  - Machine Learning (quick version)
  - Matching/Fuzzy Hashing
  - Blackhole Lists (IP addresses)
- ◆ Postage
  - Turing Tests, Money, Computation
- ◆ SmartProof
  - Combines the best of Filtering with all of Postage
- ◆ Safe sender lists (Bonded Sender)
- ◆ Disposable Email Addresses
- ◆ Stopping Outbound Spam
- ◆ Deployment Constraints
  - Client versus server, issues for large organization



## Filtering Technique Machine Learning

- ◆ Learn spam versus good
- ◆ Problem: need source of training data
  - Get users to volunteer GOOD and SPAM
  - Will talk more about this later
- ◆ Should generalize well
- ◆ But spammers are adapting to machine learning too
  - Images, different words, misspellings, etc.
- ◆ We use machine learning – details later



## Filtering Technique Matching/Fuzzy Hashing

- ◆ Automatically get examples of known spam
  - Use “Honeypots” – addresses that should never get mail
    - All mail sent to them is spam
  - Use “Report Junk” button data
- ◆ Look for similar messages that arrive in real mailboxes
  - Exact match easily defeated
  - Use fuzzy hashes
    - How effective?
- ◆ The Chinese menu (madlibs) attack will defeat any exact match filters or fuzzy hashing
  - Make thousands of dollars working at home !!!**
  - Earn lots of money in the comfort of your own house .**
- ◆ Spammers already doing this



## Blackhole Lists



### MSN blocks e-mail from rival ISPs

By Stefanie Olsen  
Staff Writer, CNET News.com  
February 28, 2003, 2:34 PM PT

- ◆ Lists of IP addresses that send spam
  - Open relays, Open proxies, DSL/Cable lines, etc...
- ◆ Easy to make mistakes
  - Open relays, DSL, Cable send good and spam...
- ◆ Who makes the lists?
  - Some list-makers very aggressive
  - Some list-makers too slow

**Microsoft's MSN said its e-mail services had blocked some incoming messages from rival Internet service providers earlier this week, after their networks were mistakenly banned as sources of junk mail.**

The Redmond, Wash., company, which has nearly 120 million e-mail customers through its Hotmail and MSN Internet services, confirmed Friday it had wrongly placed a group of Internet protocol

## Solutions to Spam Overview

- ✓ Filtering
  - ✓ Machine Learning (quick version)
  - ✓ Matching/Fuzzy Hashing
  - ✓ Blackhole Lists (IP addresses)
- ◆ Postage
  - Turing Tests, Money, Computation
- ◆ SmartProof
  - Combines the best of Filtering with all of Postage
- ◆ Safe sender lists (Bonded Sender)
- ◆ Disposable Email Addresses
- ◆ Stopping Outbound Spam
- ◆ Deployment Constraints
  - Client versus server, issues for large ISPs



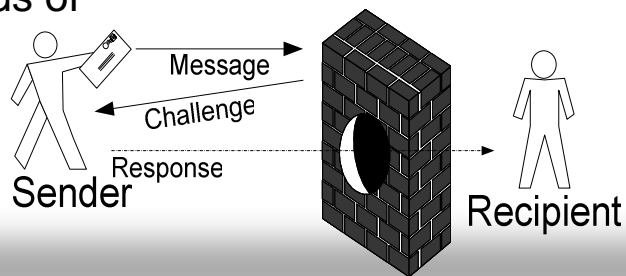
## Postage

- ◆ Basic problem with email is that it is free
  - Force everyone to pay (especially spammers) and spam goes away
  - Send payment pre-emptively, with each outbound message, or wait for challenge

- ◆ Multiple kinds of

payment:

Turing Test,  
Computation,  
Money



## Turing Tests (Naor '96)

- ◆ You send me mail; I don't know you
- ◆ I send you a challenge: type these letters

F \ X L T 6 H E B



- ◆ Your response is sent to my computer
- ◆ Your message is moved to my inbox, where I read it



## Computational Puzzle (Dwork and Naor '92)

- ◆ Sender must perform time consuming computation
- ◆ Example: find a hash collision
  - Easy for recipient to verify, hard for sender to find collision
- ◆ Requires say 10 seconds (or 5 minutes?) of sender CPU time (in background)
- ◆ Can be done preemptively, or in response to challenge



## Money

- ◆ Pay actual money (1 cent?) to send a message
- ◆ My favorite variation: take money only when user hits "Report Spam" button
  - Otherwise, refund to sender
  - Free for non-spammers to send mail, but expensive for spammers
- ◆ Requires multiple monetary transactions for every message sent – expensive
- ◆ Who pays for infrastructure?



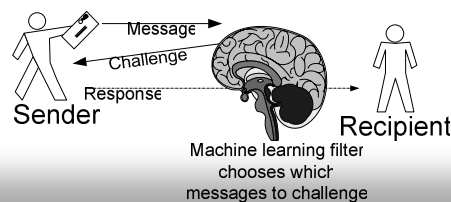
## The SmartProof Approach Overview

- ◆ Combines best aspects of several previous techniques:
  - Machine learning
  - Challenge response
  - Postage (multiple techniques)



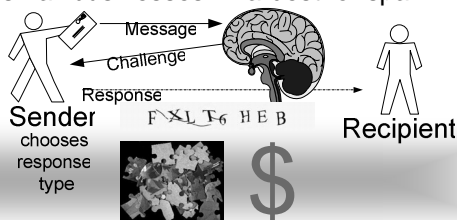
## SmartProof: Selective Challenging

- ◆ Most challenge-response approaches challenge every message
- ◆ We use machine learning to challenge only some messages
  - Definite spam deleted (saves processing costs)
  - Definite good passed through to inbox (avoids annoying challenges, and avoids many challenges that will not be answered.)
  - Only possible spam, possible good is challenged



## SmartProof: Sender Chooses Type of Proof

- ◆ Can auto-respond with computation
  - Least annoying to sender – he may never see the challenge
  - Usable by people with disabilities
- ◆ Can respond by solving a Turing Test
  - Works for people with old computers or incompatible computers or who do not want to download code
- ◆ Future?: Can respond with micro-payment
  - Works for small businesses. Hardest for spammers to work around.



## Solutions to Spam Overview

- ✓ Filtering
  - ✓ Machine Learning (quick version)
  - ✓ Matching/Fuzzy Hashing
  - ✓ Blackhole Lists (IP addresses)
- ✓ Postage
  - ✓ Turing Tests, Money, Computation
- ✓ SmartProof
  - ✓ Combines the best of Filtering with all of Postage
- ◆ Safe sender lists (Bonded Sender)
- ◆ Disposable Email Addresses
- ◆ Stopping Outbound Spam
- ◆ Deployment Constraints
  - Client versus server, issues for large ISPS



## Safe Sender Lists

- ◆ Block lists are hard
  - People get very angry when you call them a spammer
  - There are  $2^{32}$  IP addresses
  - Hard to know when to stop blocking
- ◆ Instead, make lists of known good senders
  - Amazon pays someone, e.g. Bonded Sender, to be added to list of known good senders
  - Users download or check Bonded Sender list
  - No one can get angry
  - Relatively small number (millions instead of billions of IPs)
  - Sender can be given incentives to tell you when he changes IP addresses or use non-IP based methods



## Bonded Sender

- ◆ Program developed by IronPort systems, now working with TrustE
- ◆ List safe IP addresses
- ◆ Senders must “post a bond” – deposit money with Bonded Sender
- ◆ Portion of bond is confiscated based on complaints
  - However, some users make mistakes, so you get some complaints “for free.”
- ◆ Need to monitor volume rate (to compute complaint rate) and so that a spammer cannot post a small bond and then send billions of messages quickly
  - Record number of queries about each sender



## Disposable Email Addresses

- ◆ Also called Ephemeral Addresses
- ◆ You have one address for each sender
  - JOSHUAGO1895422@microsoft.com
  - All go to same mailbox
- ◆ If I give you my address, and you send me spam, I just delete the address
- ◆ How do new senders get an address?
- ◆ If I send mail to 3 people, which address is it From?
- ◆ Hard to remember!



## My Favorite Solution

- ◆ If we could get everyone at Hotmail to never answer any spam, spammers would just give up sending to Hotmail.
- ◆ So, when new Hotmail users sign up, send them 100 really tempting ads



## My Favorite Solution

- ◆ If we could get everyone at Hotmail to never answer any spam, spammers would just give up sending to Hotmail.
- ◆ So, when new Hotmail users sign up, send them 100 really tempting ads
- ◆ If they answer *any* of them, terminate account



## My Favorite Solution

- ◆ If we could get everyone at Hotmail to never answer any spam, spammers would just give up sending to Hotmail.
- ◆ So, when new Hotmail users sign up, send them 100 really tempting ads
- ◆ If they answer *any* of them, terminate account
- ◆ Hotmail management refuses to consider this.



## Solutions to Spam Overview

- ✓ Filtering
  - ✓ Machine Learning (quick version)
  - ✓ Matching/Fuzzy Hashing
  - ✓ Blackhole Lists (IP addresses)
- ✓ Postage
  - ✓ Turing Tests, Money, Computation
- ✓ SmartProof
  - ✓ Combines the best of Filtering with all of Postage
- ✓ Safe sender lists (Bonded Sender)
- ✓ Disposable Email Addresses
- ◆ Stopping Outbound Spam
- ◆ Deployment Constraints
  - Client versus server, issues for large ISPS



## Stopping Outbound Spam Overview

- ◆ Free email services like Yahoo, Hotmail, Gmail, etc. need to make sure that spammers do not send spam from them
- ◆ Spammers automate account creation, send tons of spam
- ◆ Solution: prevent automatic account creation – use a HIP
- ◆ Cost analysis shows this is insufficient
  - Costs about 2 cents per HIP (domestic) to pay people to solve, 0.2 cents using foreign labor
  - Can send maybe 1000 messages before account shutdown
  - Cost per message = .002 or .0002. Spammers charge/earn .01



## Stopping Outbound Spam Repeated HIPs

- ◆ Ask people to solve HIP once per every 100 messages (recipients)
  - CPM = .02 or .002 (versus .01 spammer profit)
- ◆ Annoying for users
- ◆ Interesting analysis
  - Ask people to solve one HIP for every 100 messages, up to 30 HIPs total.
  - After that, never ask them again (until a complaint)
  - Almost as effective!
  - If first 3000 messages are all spam, someone would have complained and shut down account.
  - If first 3000 are good, and next 3000 are spam, they solved 30 HIPs and sent at most 3000 spam
  - Nice proof by induction that optimal spammer strategy is to spam immediately



## Stopping Outbound Spam Variations

- ◆ Charge more for some messages than others
  - E.g. use a machine learning spam filter
- ◆ Allow money or computation instead of HIP
  - Money might come from paid accounts
    - If you are already paying money for your account, e.g. \$10 per month, and you don't get too many complaints, you are not a spammer
- ◆ Work to increase complaint rate (critical factor in math) ( $\text{Cost per message} = \text{cost per HIP} * \text{number of HIPs} / \# \text{ messages until complaint}$ )
  - Get better industry standards for complaints



## Deployment Constraints Overview

- ◆ Client, Server, Large Orgs all have advantages and disadvantages



## Client Pros and Cons

- ◆ Client Disadvantages
  - Delivery cost already imposed on email system
  - Annoying if over a slow connection (must download spam before discarding it.)
  - Difficult to determine IP address of sender
    - IP address of sender is very valuable
    - Used by safe sender lists, many anti-spoofing proposals
    - There are many received from lines in the headers showing the various IP addresses the mail went through
- ◆ Client Advantages
  - More storage available, for, e.g. personalized filters
  - UI advantage: Can easily introduce features like safe list that require UI changes or user feedback to update machine learning filter
  - More CPU available for e.g. complex HTML analysis, more complex machine learning algorithms, etc.



## Server Pros and Cons

### ◆ Server Disadvantages

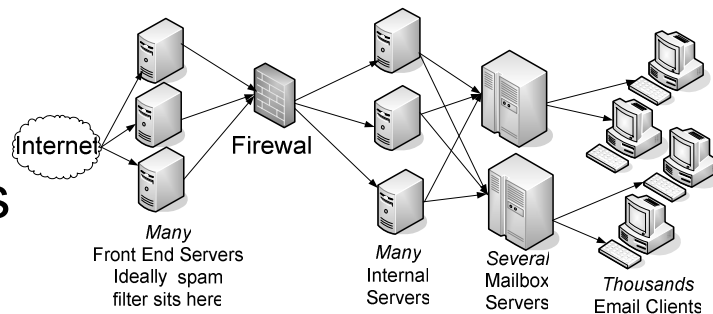
- Servers must be fast, i.e. 1 msec per message, 100 msec max

### ◆ Server Advantages

- IP address
- Most cost effective (stop spam early)
- Other information available (recipients including BCCed at same company)



## Large Org Issues



- ◆ Disadvantage: large orgs will typically have very many “front end” servers to receive mail, channel to back end, user specific storage
  - Difficult/impossible to put per user information on the front end servers, e.g. safe lists, personalization, preferences, etc.
- ◆ Advantage: lots of information about senders, e.g. historical complaint rate data, etc.

## Solutions to Spam Conclusion

- ◆ Lots of different solutions
- ◆ Often, they work best in combination
  - Example: combine machine learning filters with postage or use machine learning filters and HIPs to help stop outbound spam
  - No single discipline can solve this
- ◆ Complex issues
  - Client versus server versus large org affect which filtering technique can be used
- ◆ Final solution will combine approaches
  - Safe sender lists and postage help ensure all good mail gets through
  - Filters, blackhole lists stop mail from unknown or known bad senders



## Machine Learning Solutions Overview

- ◆ Evaluating Spam Filters
  - Lots of overoptimistic numbers in the literature
  - Hard to get good corpora for evaluation (bad ones too easy)
  - ROC curves and why they are important
- ◆ Building Models of Spam
  - Data sets used in spam studies
  - Details of important spam filtering papers
  - Synthesis
  - Other spam filtering papers
- ◆ Personalization (Threshold drift problem)



## Evaluating Spam Filters Overview

- ◆ Lots of overoptimistic numbers in the literature – you won't get 99% in real life
- ◆ Hard to get good corpora to evaluate with so people take shortcuts
  - But shortcuts typically make problem easier
- ◆ ROC curves and why they are important



## Why you won't get 99% in real life

- ◆ Academic/researcher email often not typical
  - Many academics get mostly plain text mail
  - HTML spamming tricks are especially hard
  - Images in mail are especially hard
- ◆ Some test corpora are unrealistically easy
  - Hard to get good mail, so spam and good may come from different corpora
  - A personalized filter might learn that mail with your name or one of your interests is always good
  - Real spam is sometimes targeted: includes your name or your interests – but if spam is from a different corpus, it's untargeted
- ◆ Evaluating spam filters is very hard (details in two minutes)  
Some very good reported results are partially because of test shortcuts
  - Reports may use filter judgment as part of test labeling process
  - Many tests done with cross validation



## Evaluation Shortcuts (and why they are bad)

- ◆ Shortcut: use filter judgment as part of test labeling process
  - Experimenter may not notice filter mistake
  - Experimenter may not disagree with filter on marginal or difficult cases
- ◆ Collected good data may be after some filter (e.g. on the server) has already run
  - Hardest false positives have already been discarded
- ◆ Shortcut: get spam and good from different corpora
  - There may be detectable differences between the corpora
    - e.g. they are from different times, so spam is Christmas oriented while good is about summer vacation
    - Date lines might be different! (December is spam, June is good)
  - Sent to different people so have different names in them



## Why not to do Cross Validation

- ◆ Use of cross-validation in test protocols leads to unrealistic results
  - Imagine you get mail on a new topic, e.g. your "spam filtering tutorial"
  - "Spam" is normally a very spammy word
  - Messages on Spam Filtering Tutorial get filtered until you correct the error
  - In a cross validation scenario, the training and test will both always have some mail on this topic, so no mistakes
  - Cross validation assumes exchangeable data but email is inherently temporal: non-exchangeable.
- ◆ But it's so darn hard to get data that most people need to cross validate.



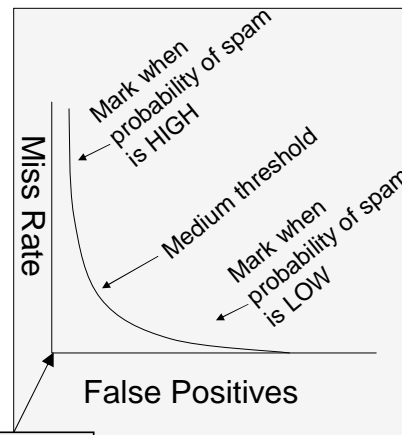
## ROC Curves Numbers Aren't Enough

- ◆ It's easy to catch 100% of spam – mark everything as spam.
  - Except that this marks an awful lot of good mail
- ◆ It's easy to make no mistakes on good mail (don't mark anything)
  - Except this misses all the spam.
- ◆ Need two numbers – *Miss Rate* (percent spam missed), and *False Positive Rate* (percent good caught as spam)
  - Except you can get incomparable numbers



## Probability Tradeoffs

- ◆ Can use different probability thresholds to trade off False Positive Rate and Miss Rate
- ◆ Mark only when probability of spam high
  - Lots of missed spam, low false positive rate
- ◆ Mark even when probability of spam low
  - Lots of false positives, low miss rate

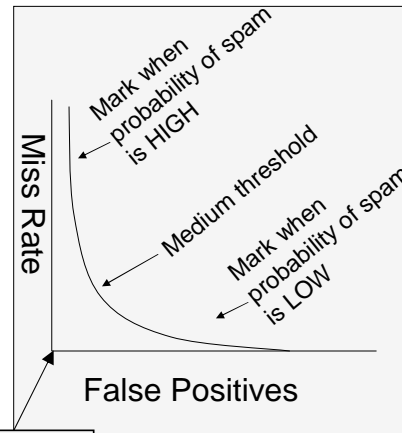


Zero Miss, Zero  
False Positive  
Ideal Location



## What Probability to Use

- ◆ If you're deleting, only delete when probability of spam is very high
- ◆ If you're marking with "\*SPAM\*", you can use low probability
- ◆ If you're putting spam in a special folder, use a high probability, so that people never have to look in the folder.

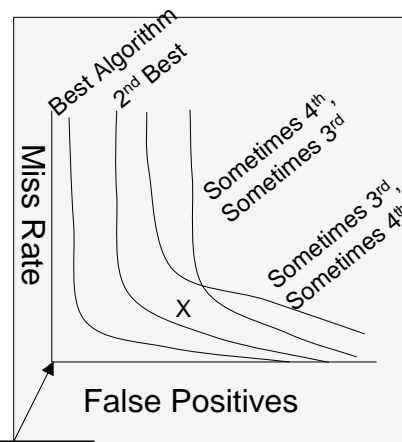


Zero Miss, Zero  
False Positive  
Ideal Location



## ROC Curves

- ◆ 5 algorithms
- ◆ One is clearly best, one is clearly 2<sup>nd</sup>, but two of them cross
- ◆ One of them only has a single point
  - Typical of matching algorithms like Brightmail



Zero Miss, Zero  
False Positive  
Ideal Location



## Evaluation Conclusion

- ◆ You won't get 99% in real life
- ◆ Careful evaluation is especially hard for spam filters
  - Hard to get good corpora and most shortcuts make the problem easier
- ◆ Important to show range of results (ROC curve)
  - Otherwise, it may be difficult to compare different filters or understand advantages of each



## Machine Learning Solutions Overview

- ✓ Evaluating Spam Filters
  - ✓ Lots of overoptimistic numbers in the literature
  - ✓ Hard to get good corpora for evaluation (bad ones too easy)
  - ✓ ROC curves and why they are important
- ◆ Building Models of Spam
  - Data sets used in spam studies
  - Details of important spam filtering papers
  - Synthesis
  - Other spam filtering papers
- ◆ Personalization (Threshold drift problem)



## Detail on Data Sets

- ◆ Ling-Spam – <http://www.iit.demokritos.gr/~ionandr/publications/>
  - Good mail taken from a moderated linguistics mailing list
  - Spam from volunteers
- ◆ PU1 – <http://www.iit.demokritos.gr/~ionandr/publications/>
  - Good mail
    - All good mail the author received and saved over 36 months
    - Deleted 6+th occurrence of mail from any user to simulate the effects of safelists
  - Spam
    - All spam the author received over 22 months
    - Non-English and Same-day-repeats removed
  - Encoded into token IDs to preserve privacy



## Data Sets from Spam Studies

Name	Year Introduced	Times Used	Num Messages	Spam %
Ling-Spam	2000	5	3k	16%
PU1	2000	3	1k	45%
PU2 / PU3	2004	1	721 / 4k	20% / 44%
Non-Public	Various	4	1.7k – 11k	28% – 88%

- Other Data sets
  - Spam Assassin corpus
  - Spambase (from the UCI repository)



## Why is it Hard to get Good Spam Data Sets?

- ◆ No one is yet will share large amounts of un-encoded good mail
- ◆ Sharing encoded mail doesn't allow researchers to explore solutions to current spammer attacks
- ◆ Time changing adversarial nature
  - Results on old spam may not generalize to new spam



## Naïve Bayes

- ◆ Sometimes called “Bayesian techniques”
  - Misnomer, now popularly misused in the open source community and popular press
- ◆ Proper name is “Naïve Bayes”
  - Called Naïve because makes assumptions that aren't true
- ◆ Want  $P(\text{spam}|\text{words})$
- ◆ Use Bayes Rule:  $P(\text{spam} | \text{words}) = \frac{P(\text{words} | \text{spam}) \times P(\text{spam})}{P(\text{words})}$
- $$P(\text{words}) = P(\text{words} | \text{spam}) \times P(\text{spam}) + P(\text{words} | \text{good}) \times P(\text{good})$$
- ◆ Assume independence: probability of each word independent of others (wrong assumption)

$$P(\text{words} | \text{spam}) \approx P(\text{word1} | \text{spam}) \times P(\text{word2} | \text{spam}) \times \dots \times P(\text{wordn} | \text{spam})$$

## A Bayesian Approach to Filtering Junk E-Mail

1998 - Sahami, Dumais, Heckerman, Horvitz

- ◆ One of the first papers on using machine learning to combat spam
- ◆ Used Naïve Bayes
- ◆ Feature Space: Words, Phrases, Domain-Specific Features
- ◆ Feature Selection: 500 by mutual information
- ◆ Evaluation Data: ~1700 Messages, ~88% Spam, from volunteer's private e-mail
- ◆ Results: Spam precision 100%, Spam recall 98.3%
- ◆ Other Highlights: Decision theoretic decisions – cost of false positive much higher than false negative



## A Bayesian Approach to Filtering Junk E-Mail

1998 - Sahami, Dumais, Heckerman, Horvitz

- ◆ Hand Crafted Features
  - 35 Phrases
    - 'Free Money'
    - 'Only \$'
    - 'be over 21'
  - 20 Domain Specific Features
    - Domain type of sender (.edu, .com, etc)
    - Sender name resolutions (internal mail)
    - Has attachments
    - Time received
    - Percent of non-alphanumeric characters in subject
- ◆ Best collection of heuristics discussed in literature
  - Without them: precision 97.1% recall 94.3%
  - With them: precision 100% recall 98.3%



## Digression: Cost Sensitive Method

- ◆ False positives worse than false negatives
- ◆ Naïve Bayes easy to adjust
  - Raise the threshold for marking a message as spam
- ◆ Other learners adapted with standard techniques
  - Weighted re-sampling of training data
  - Metacost
  - Etc.



## A Plan for Spam

2002 – P. Graham

- ◆ Widely cited in the open source community
- ◆ Uses a heavily tuned version of Naïve Bayes
  - Tokens in good mail get double weight
  - At test time only use the 15 tokens with most extreme probabilities
  - Several other tweaks designed to avoid false positives
- ◆ Feature Space: Words in header and body
- ◆ Feature Selection: ~23,000 features
  - all that appeared more than 5 times
- ◆ Evaluation Data: ~8000 messages from author; ~50% spam
- ◆ Results: Spam precision 100%, Spam recall 99.5%



## Digression: Naïve Bayes pros and cons

- ◆ Pros: Naïve Bayes uses “sufficient statistics”:
  - Don’t need to store actual messages
  - Just need counts of how often each word occurred as good or spam
  - Data storage size and privacy advantage for personalized filters
- ◆ Cons: Performance may be worse because of Naïve assumptions
  - Consider “Click Here to Unsubscribe”
  - Phrase occurs 10 times as often in spam as good
  - $P(\text{click}|\text{spam})$  is 10 times higher,  $P(\text{here}|\text{spam})$  is 10 times higher,  $P(\text{unsubscribe}|\text{spam})$  is 10 times higher
  - Multiply together, get factor of 1000
- ◆ See “Tackling the poor assumptions of Naïve Bayes Text Classifiers”, Rennie et al., ICML 2003
  - Variations on Naïve Bayes that help, including term frequency transformations, document length normalization, etc.

## Support Vector Machines for Spam Categorization

1999 – Drucker, Wu, Vapnik

- ◆ Algorithms: Linear SVMs, Ripper, Rocchio, Boosted C4.5
- ◆ Feature Space: Words, binary, TF, TF-IDF, Stop List
- ◆ Feature Selection: 1000 or 7000 by Mutual Info
- ◆ Evaluation Data: 3000 messages, 28% spam; from a volunteer
- ◆ Results: Boosted decision tree with TF best in error
  - Boosted Tree FP 5%, FN 1.2%
  - SVM FP 5%, FN 1.5%
  - Rocchio FP 5%, FN 4.7%
  - Ripper FP 5%, FN 7.9%
- ◆ Other Highlights: SVMs with binary features better error dispersion; Stoplists don’t seem to help



## Digression: non-linear SVMs

- ◆ Non-linear SVMs might be ok for personal filters but cannot be used for general filtering
  - They are too slow (require computing distance from all support vectors, may be many thousands of supports)
  - Support vectors are actual email messages
    - Can't ship people's email!
    - Need good mail as well as spam



## Learning to Filter Unsolicited Commercial E-Mail

2004 – I. Androutsopoulos, G. Paliouras, and E. Michelakis

- ◆ Algorithms: Naïve Bayes, Flexible Bayes, SVMs, LogitBoost with decision stumps
- ◆ Feature Space: Word n-grams for n between 1-3
- ◆ Feature Selection: 1000 or 3000 by Mutual Info
- ◆ Evaluation Data
  - LingSpam; PU1-3
- ◆ Results
  - LogitBoost was best
  - 1-grams usually best



## Learning to Filter Unsolicited Commercial E-Mail

2004 – I. Androutsopoulos, G. Paliouras, and E. Michelakis

- ◆ Real world workload: the Filtron system
  - SVM with 520 1-gram features built on PU3
- ◆ Run for 212 days on authors mail
  - 6732 messages, 24% spam
  - Spam precision 96.54%
  - Spam recall 89.34%
- ◆ False negatives
  - 1/3 due to spammer tricks
  - 1/3 due to message encoding
  - 1/4 due to non-English mail
- ◆ False positives
  - 1/2 auto generated or commercial newsletters
  - 1/5 3-5 words plus link or attachment
  - 1/4 1-2 lines



## Boosting Trees for Anti-Spam Email Filtering

2001 – Carreras and Marquez

- ◆ Algorithms: AdaBoost with trees and Naïve Bayes
- ◆ Feature Space: Words, binary
- ◆ Feature Selection: 26,000 (all available features)
- ◆ Evaluation Data: PU1 Corpus
- ◆ Results: Spam Prec. 98.7% Recall 97.1%
- ◆ Other Highlights: Found that predictions far from the margin were more reliable
- ◆ Results summary slide: one of the best algorithms for spam filtering



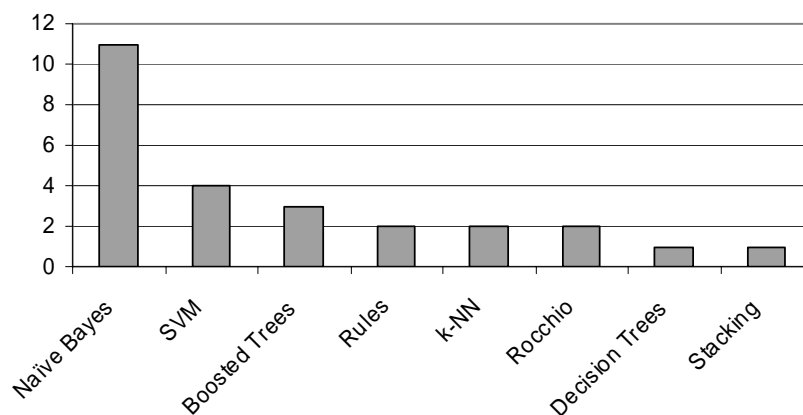
## How to Beat an Adaptive Spam Filter

2004 – J. Graham-Cumming

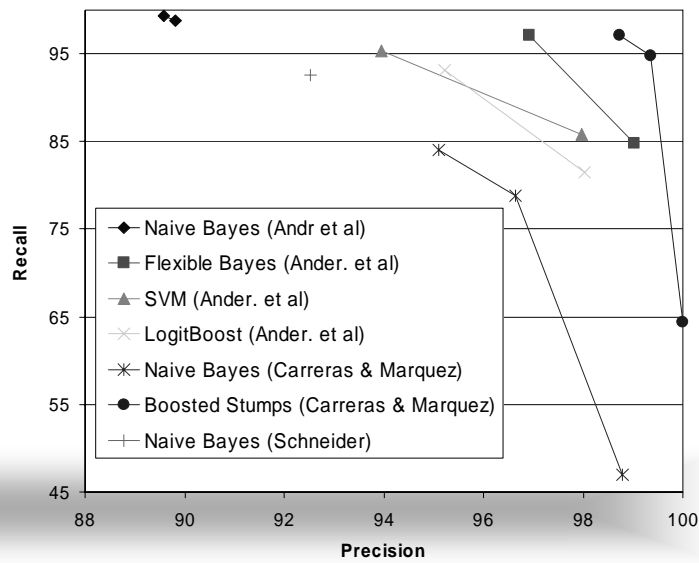
- ◆ Use machine learning to discover words that beat an adaptive filter
  - Take a message that is near spam threshold
  - Send it to the target filter 10,000 times
    - Each time adding 5 random words and a unique web beacon
  - Messages that beat the filter have beacons activated (when the user views them)
  - Train an 'evil' filter to learn which messages beat the target filter
  - Use 'evil' filter to modify new spam messages
- ◆ Found single word additions to get new spam by the filter



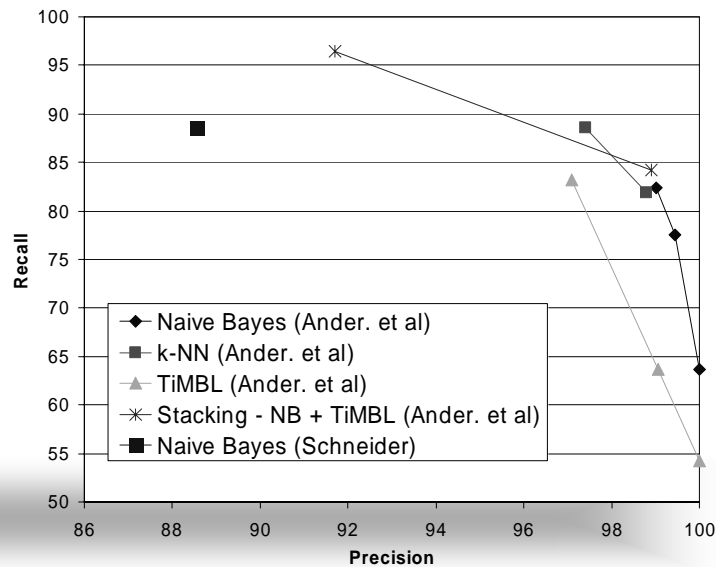
## Algorithms Used in Spam Filtering



## Performance on PU1



## Performance on LingSpam



## Which Algorithm is Best?

- ◆ Very difficult to tell
  - Not enough information on precision/recall tradeoffs
  - Most consistently-used data set is spam from volunteers, good from Linguistics mailing list
- ◆ Lessons learned
  - Naïve Bayes reported to do very well
  - More complex algorithms have some gain



## Summaries of Other Spam Papers



## Directions for Machine Learning in Spam

- ◆ “In Vivo” spam filtering: A challenge problem for data mining  
2004 – Tom Fawcett
  - Skewed and changing class distributions
  - Unequal and uncertain error costs
  - Complex text patterns requiring sophisticated parsing
  - A disjunctive target concept comprising superimposed phenomena with complex temporal characteristics
  - Intelligent, adaptive adversaries



## SpamCop: A Spam Classification & Organization Program

1998 – P. Pantel and D. Lin

- ◆ Algorithms: Naïve Bayes
- ◆ Feature Space: Words, stemming
- ◆ Feature Selection: ~3,800
  - Token must show up at least 4 times
  - Token must be somewhat informative
- ◆ Evaluation Data: ~1,200 messages; ~35% spam; mail from author & spam from the Internet
- ◆ Results: Spam Perc. 98.8% Recall 91.7%
- ◆ **Other Highlights: Compared to RIPPER and found Naïve Bayes makes half the errors**



## An Evaluation of Naïve Bayesian Anti-Spam Filtering

2000 – Androutsopoulos, Konstantinos, Chandrinos, Paliouras, Spyropolous

- ◆ **Algorithm:** Naïve Bayes (Multi-Variate Bernoulli Model)
- ◆ **Feature Space:** Words, Stemming, Stop List
- ◆ **Feature Selection:** Varied 50 – 700, best  $\leq$  200
- ◆ **Evaluation Data:** Ling-Spam corpus; ~2900 Messages, ~16% spam; good from linguistics mailing lists, spam from volunteers.
- ◆ **Results:** Spam precision 99.5%, Spam recall 92.8%
- ◆ **Other Highlights:** Varied cost of misclassification, introduced the Ling-Spam corpus.



## An Experimental Comparison of Naïve Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-Mail Messages

2000 – Androutsopoulos, Konstantinos, Chandrinos, Spyropolous

- ◆ **Algorithm:** Naïve Bayes, Keywords (Outlook 2000)
- ◆ **Feature Space:** Words, Stemming, Stop List
- ◆ **Feature Selection:** Varied 50 - 700
- ◆ **Evaluation Data:** PU1 Corpus; ~1100 messages, ~45% spam; trivially encrypted versions of author's mail
- ◆ **Results:** Spam precision 98%, Spam recall 76%
- ◆ **Other Highlights:** Varied cost of misclassification, introduced the PU1 corpus.



Stacking Classifiers for anti-spam filtering of e-mail  
2001 – Sakkis, Androutsopoulos, Paliouras, Karkaletsis, Spyropoulos,  
Stamatopoulos

- ◆ **Algorithm: Stacked Naïve Bayes and k-NN**
  - Stacking used k-NN augmented with the predictions of the base classifiers.
- ◆ **Feature Space: Words, Stemming, Stop List**
- ◆ **Feature Selection: 100 NB, 600 – 700 k-NN**
- ◆ **Evaluation Data: Ling-Spam**
- ◆ **Results: Spam precision ~98.8%, Spam recall 85-90%**
- ◆ **Other Highlights: NB and k-NN make relatively uncorrelated errors**



Learning to Filter Spam E-Mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach  
2001 – Androutsopoulos, Paliouras, Karkaletsis, Sakkis, Spyropoulos,  
Stamatopoulos

- ◆ **Algorithm: Naïve Bayes and k-NN**
  - k-NN distance metric weighted by feature value
  - Made cost sensitive by weighting votes
- ◆ **Feature Space: Words, Stemming, Stop List**
- ◆ **Feature Selection: 50 – 700 by Mutual Info**
- ◆ **Evaluation Data: Ling-Spam**
- ◆ **Results for k-NN: Spam precision 98.9%, Spam recall 74%**
- ◆ **Other Highlights:**
  - Naïve Bayes generally better than k-NN
  - Fewer neighbors usually better (1-2)



## SVM-based Filtering of E-mail Spam with Content-specific Misclassification Costs

2001 – Kolcz and Alspector

- ◆ Algorithms: Cost Sensitive Linear SVMs
- ◆ Feature Space: Words, hand-crafted features, binary and normalized instances
- ◆ Feature Selection: 10,000 by Mutual Info
- ◆ Evaluation Data: 11,000 Messages, 47% spam; from a set of volunteers
- ◆ Results: FP 2.2%, FN 0.6%
- ◆ Other Highlights: Normalizing input vectors helps, Instance dependent costs, Discusses ESP (Email Service Provider) needs



## A Comparison of Event Models for Naïve Bayes Anti-Spam E-mail Filtering

2003 - Schneider

- ◆ Algorithm: Naïve Bayes (Multi-Variate Bernoulli and Multinomial)
- ◆ Feature Space: Words, Stemming, Stop List
- ◆ Feature Selection: Varied < 5000 ; Mutual Information
  - Feature rank adjusted by differences between in class term-frequency and global term-frequency
- ◆ Evaluation Data: Ling-Spam corpus; PU1 corpus
- ◆ Results: Multinomial better tradeoffs between classes
  - Ling-Spam: Spam prec 99.3%, Spam recall 96.05%
  - PU1: Spam prec 97.7%, Spam recall 97%



## Machine Learning Solutions Overview

- ✓ Evaluating Spam Filters
  - ✓ Lots of overoptimistic numbers in the literature
  - ✓ Hard to get good corpora for evaluation (bad ones too easy)
  - ✓ ROC curves and why they are important
- ✓ Building Models of Spam
  - ✓ “Bayesian” Techniques
  - ✓ Support Vector Machines
  - ✓ Rules/Trees
- ✓ Miscellaneous
  - ✓ Cost sensitive learning
  - ✓ Beating machine learning solutions
- ◆ Personalization (Threshold drift problem)



## Machine Learning Solutions Overview

- ✓ Evaluating Spam Filters
  - ✓ Lots of overoptimistic numbers in the literature
  - ✓ Hard to get good corpora for evaluation (bad ones too easy)
  - ✓ ROC curves and why they are important
- ✓ Building Models of Spam
  - ✓ Data sets used in spam studies
  - ✓ Details of important spam filtering papers
  - ✓ Synthesis
  - ✓ Other spam filtering papers
- ◆ Personalization (Threshold drift problem)



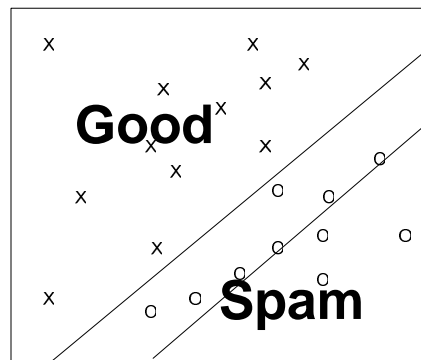
# Personalization

- ◆ Doing work recently on personalization
  - You label your mail as spam or good
  - We adapt the filter as you go
- ◆ Nasty problem: Threshold drift
- ◆ Online training



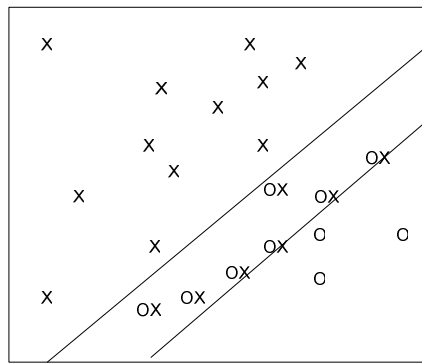
## Threshold Drift Conservative Threshold Setting

We are conservative in our filtering. For instance, maybe we need to be 96% certain that mail is spam before we classify as spam



# Threshold Drift

## Lots of Spam Classified as Good



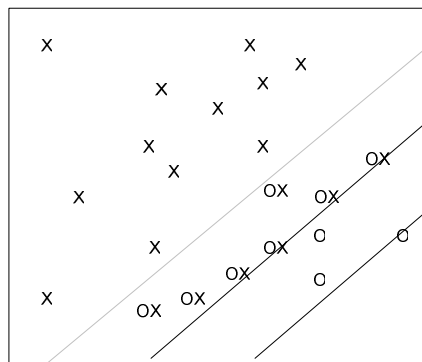
← Separator:  
50/50 mark

← Conservative Threshold:  
96% sure



# Threshold Drift

## New Separator Parallel to Old



← Old Separator:  
50/50 mark

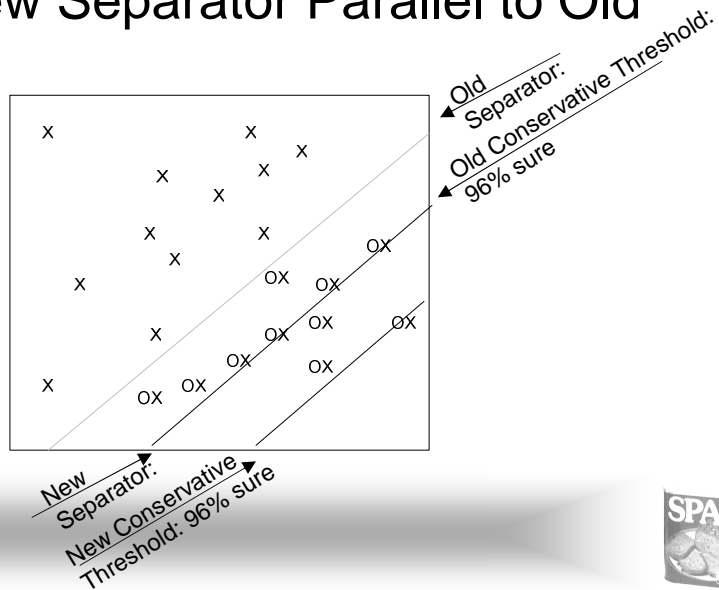
← Old Conservative Threshold:  
96% sure

← New Separator:  
50/50 mark

← New Conservative Threshold:  
96% sure



## Threshold Drift New Separator Parallel to Old



## More Personalization Problems

- ◆ Users may correct all errors, or only all spam, all good, 50% spam, 10% spam, no errors, etc.
- ◆ Need to work no matter what the user correction rate is
- ◆ Still working on this problem
  - Current solution: Cross Validate, throw away new filter if not better than old



## Online Training

- ◆ Need online training
  - Need to throw away data for privacy reasons
  - Want fast updates, after each new message
- ◆ Online algorithms haven't gotten that much attention
- ◆ Our threshold drift solution, etc. depend on cross-validation. How do we do that online?



## DATA Overview

- ◆ How we get our data at Microsoft
  - The Feedback Loop
  - Noise in the feedback loop
- ◆ Analysis of data from the feedback loop
  - Where spam is from
  - Can legal approaches work?



## Feedback Loop

- ◆ We asked Hotmail users to help us fight spam
- ◆ Each day, we send volunteers a random message they would have received and ask them “Spam or Good”
- ◆ We get tens of thousands of classified messages per day
  - We get more on a single day than was used in any published study we are aware of!
- ◆ Now have over ten million training messages



## Advantages of Feedback Loop

- ◆ Alternative to feedback loop: spam in unused accounts; or users reports of spam (and maybe mistakes on good)
- ◆ We have Good and Spam
  - Some senders are mixed
  - Many users make mistakes
  - We can learn anyway
- ◆ We select messages *before* they reach spam filter, so we don't have bias
  - Some methods exclude all filtered messages from further training
    - Can't learn about mail they delete
    - Can't learn about mail already filtered, so may “drift” back to allowing it



## Feedback Loop: User Errors



**Patti Robles**  
Art Director

"Every day, I get these annoying spams that are nothing but cookie recipes and articles about the benefits of Vitamin C. Wait-those are from my mom."



**Andrew Reed**  
Forklift Operator

"Gee, you sign up for one little mailing on how to enlarge yourself, and you're swamped for the rest of your life."

- ◆ About 3% of labels are errors
- ◆ User error rate is sometimes thought of as upper bound on computer performance, but you can do better
- ◆ Example: classify based on sender email address
  - If 97% mark as spam, then we classify 100% as spam, make no errors
- ◆ We have so much training data, we are robust to user error

## Data Analysis

- ◆ Feedback loop is incredible source of data
  - Where does Spam come from?
  - Where does Good come from?
- ◆ Legal approaches
  - What does CAN-SPAM actually say?
  - Will legal approaches work?
  - Analyze based on products spam sells, difficulty of offshoring each product



## Mapping Messages to Countries

- ◆ Messages tagged with the IP that relayed them to Hotmail
- ◆ Use data from Internet bodies
  - ARIN, APCNIC, LACNIC, and RIPENCC
  - Allocate IP ranges, each tagged with country
- ◆ Mapped over 99% of the IPs in our data
- ◆ Some Caveats
  - Open relays
  - Sub-allocations

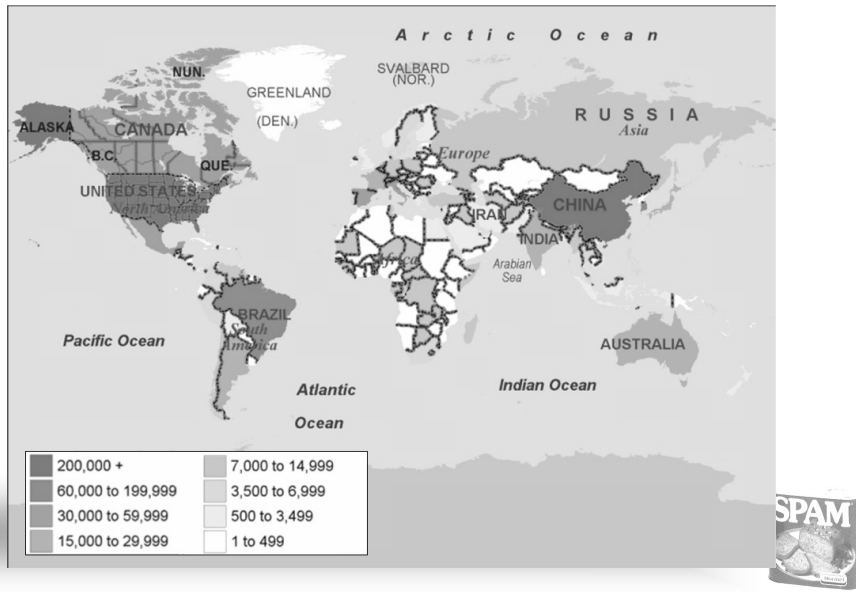


## Analyzed Two Million Messages

- ◆ Arrived via the Feedback Loop between April and June of 2003
- ◆ Came from 214,000 distinct IP addresses
- ◆ Mapped to 157 different countries
  - 42% sent less than 100 messages
  - 33% sent 100 - 1000 messages
  - 25% sent 1000+ messages



## Where is the Mail from?

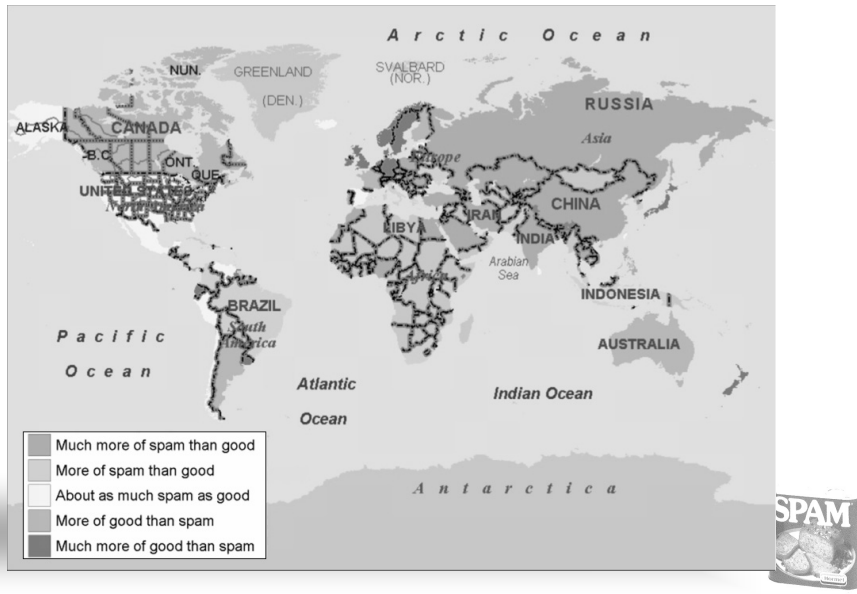


## Where is the Mail from?

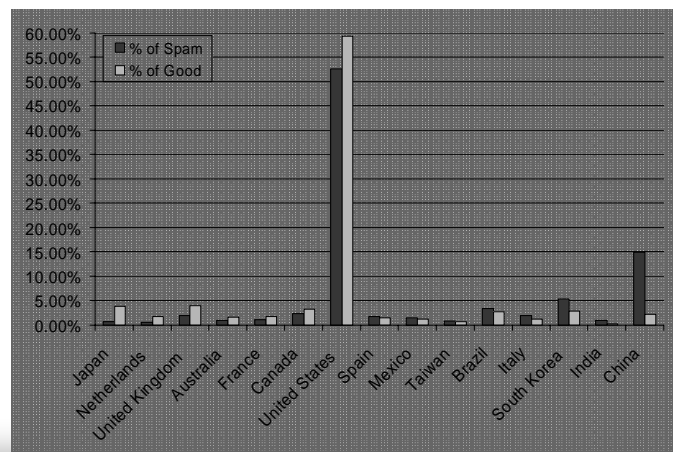
Country	% of Mail
United States	53.50%
China	13.19%
South Korea	5.04%
Brazil	3.38%
Canada	2.32%
United Kingdom	2.17%



## Diversity of Mail by Country



## Diversity of Mail by Country



## Who is the Mail to?

Character Set	% of Good	% of Spam
ISO Latin	1.41%	0.31%
ANSI/OEM Korean	2.33%	0.45%
ANSI/OEM Japanese	4.16%	0.18%
ANSI Latin	29.18%	7.52%
US-ASCII	60.31%	90.80%
Other	2.61%	0.74%

- ◆ Character set: a proxy for language
- ◆ Vast majority of spam is in US-ASCII
- ◆ Good mail more diverse than spam



## CAN-SPAM Act

- ◆ Does not ban spamming
  - Requires opt-out and a postal address
  - Preempts State laws, including California's law
- ◆ FTC must consider a "Do not spam" list
- ◆ Any hope for legal solutions?



## What Businesses does it Support?

### ◆ **Domestic**

Require a domestic presence:

- financial services (credit card, mortgage)
- Insurance
- Government grant programs

### ◆ **Semi-domestic**

Require shipping, cheap to ship:

- Viagra and other medical products
- College diplomas
- Magazines

### ◆ **International**

Do not require shipping or domestic presence:

- Porn sites
- Software
- Scams



## Businesses Supported by Spam

### ◆ Examined a sample of spam by hand

Type of Business	% of Sample
Domestic	30%
Semi-Domestic	32%
International	38%

### ◆ Currently well more than 30% our spam comes from domestic sources

### ◆ Much spam could migrate internationally!



## More Detail on the Products in Spam

Product	2003 Spam	2004 Spam	Delta (Absolute %)	Description
Porn/Sex Non-graphic	17%	34%	17%	Enhancers with sexual connotation, links to porn.
Insurance	1%	4%	3%	Health, dental, life, home, auto insurance.
Rx / Herbal	8%	10%	2%	Cheap drugs or herbal supplements.
Financial	12%	13%	1%	Refinancing, get out of debt, financial advice.
Travel / Casino	2%	3%	1%	Selling airlines tickets, hotel reservations, rental car. Internet casino sites. Other gaming sites.
Scams	8%	6%	-1%	Get rich quick, Phisher scams, etc.
Newsletters	9%	6%	-3%	Any newsletter that isn't selling something.
Other Spam	13%	8%	-5%	Everything else that appears to be spam.
Porn/Sex Graphic	13%	7%	-5%	Anything that contains pornographic images.
Dubious Products	20%	10%	-10%	Pirated software, diplomas, etc.



## Data Analysis Summary

- ◆ Feedback Loop very valuable data source
- ◆ About half of the spam and half of the good mail comes from foreign sources
- ◆ Countries are very diverse in the ratio of spam mail to good mail that they send
- ◆ Most of the spam is in English
- ◆ 70% of the businesses that send spam could be run with no domestic presence



## Why Email and Spam need Their Own Field/Conference

- ◆ Email is one of top two applications
  - Search is the other (TREC, SIGIR)
  - Email is why my grandfather and my wife's grandmother bought computers
- ◆ Compare to databases, operating systems, speech recognition, natural language processing, graphics, ...
- ◆ Historically, email was simple and not that important
  - Complex, formatted, key to work, key to e-commerce



## Examples: Text and From lines

- ◆ Email is not Text Classification
  - Example: On-line learning needed for privacy reasons
  - Example: On-line user adaptation
  - Example: Special information like From header information
- ◆ Example: From header field
  - 20 Newsgroups data deleted From: information
    - TOO Predictive!
  - Chris Meek found that Sender Most-Recently Used was best way to do auto-folding
    - Really Boring machine learning result
    - Really Great email result!
  - Andrew McCallum noticed that one person may have different From: lines



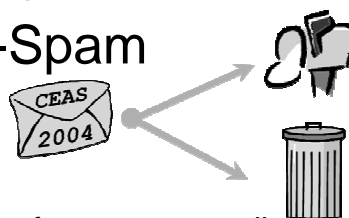
## Example: Anti-Spoofing

- ◆ Cryptographic approaches
  - S/MIME, PGP
  - Small adoption because of problems distributing keys – need solutions that work for email
- ◆ Systems/networking approaches
  - DNS/IP address-based approaches (e.g. Caller-ID, SPF)
- ◆ Combination approaches
  - Put key in DNS entry (e.g. Yahoo's DomainKeys)
- ◆ Need a conference where the machine learning people and crypto people and systems people and email people and spam people all come together to compare and learn



## Conference on Email and Anti-Spam

- ◆ [www.ceas.cc](http://www.ceas.cc)
- ◆ How it's different:
  - First academic-style research conference on email or spam
    - Plenty of informal conference, industrial conferences
- ◆ Key Details:
  - Mountain View (Near San Francisco)
  - Immediately after AAAI
  - July 30 & 31



## Email and Spam: New Field Conclusion

- ◆ Email and spam are a great new field
- ◆ Have a conference.
- ◆ A Textbook? A Journal? A Course? NSF Grants?



## Conclusion Overview

- ◆ The spam end game
- ◆ The future of spam fighting in general
- ◆ The future of machine learning filtering in particular
- ◆ The future of spam (no more email, lots more other places)
- ◆ Spam is a great problem



## Conclusion

### The Spam End-Game

- ◆ Most large senders use reputations services, like Bonded Sender, etc. to prove they are “good.”
- ◆ Many individual users piggyback on a large sender, e.g. Verizon.com certifies its users, and rate limits them
  - Outbound spam techniques and prompt responses to complaints are critical
- ◆ Remaining individuals treated as suspicious
  - Mail is aggressively filtered (machine learning) unless on recipient's safe list
  - Challenges (money, HIP or computation) sent, then added to safe list
  - Some spammers willing to pay the price, will get through, but mail must be targeted to be profitable
- ◆ Almost no spam gets through. All good mail gets through by one path or another.



## Conclusion

### Future of spam fighting for email

- ◆ Legal
  - One part of the solution, but at most 30-50%
- ◆ Identity (Caller-ID, DomainKeys)
  - Helps make safe-listing work
- ◆ Safe Sender Lists (Bonded Sender, etc.)
  - Helps, especially large senders. Lets filters be more aggressive
- ◆ Payment (Money or HIP or computation)
  - Essential component for smaller senders
- ◆ Filters (machine learning)
  - Key component for all senders not on some kind of safe list
  - Combine with payment approaches (e.g. SmartProof)



## Conclusion

### Future of Machine Learning Filtering

- ◆ Lots to do
- ◆ Spammers are always adapting (remember tricks, like images, good word chaff, “How to Beat an Adaptive Spam Filter”)
- ◆ Lots of interesting work for personalization
- ◆ Apply email filtering techniques to other kinds of spam (e.g. SPIM)



## Conclusion

### The Future of Spam

- ◆ Most people will get very little spam
- ◆ Spam is not just email
  - Advertising Wants to be Free
  - Chat room, SPIM, Web spam, blog...
- ◆ Different kinds of spam need different techniques but often similar ideas can be used



## Conclusion Conclusion

- ◆ Spam is a huge problem,
  - It will be with us in one form or another for the foreseeable future (email, SPIM, web, etc.)
- ◆ We'll need a variety of solutions
  - Filtering (Machine Learning)
  - Postage
  - Safe sender lists
- ◆ Email and Spam present tons of great new problems, and we're making progress!



## Bibliography of Spam

- ◆ Thanks to Tom Fawcett, Trevor Stone, Loder et al. and others from whom this bibliography was assembled.
- ◆ Don't forget to look at the <http://www.ceas.cc> website for new papers
- ◆ Very rough chronological order



# Bibliography of Spam

## 0-1999

- ◆ C. Dwork and M. Naor, "Pricing via Processing or Combatting Junk Mail", *Lecture Notes in Computer Science 740* (Proceedings of CRYPTO'92), 1993, pp. 137--147.
- ◆ W. Cohen, Learning Rules that classify Email, *Advances in Inductive Logic Programming* 1996. <http://citeseer.ist.psu.edu/cohen96learning.html>
- ◆ Moni Naor, "Verification of a human in the loop or Identification via the Turing Test", unpublished manuscript.
- ◆ M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian Approach to Filtering Junk E-mail. *AAAI'98 Workshop on Learning for Text Categorization*, July 27, 1998, Madison, Wisconsin.
- ◆ L. F. Cranor and B. A. LaMacchia. Spam! *CACM*, 41(8):74-83, August 1998.
- ◆ Pantel, P., Lin, D., SpamCop: a spam classification and organization program". In *AAAI'98 Workshop on Learning for Text Categorization*, 1998.
- ◆ J. Provost. Naive-bayes vs. rule-learning in classification of email. *Technical Report AI-TR-99-284*, University of Texas at Austin, Artificial Intelligence Lab, 1999.
- ◆ H. Drucker, Donghui Wu, and V.N. Vapnik. Support vector machine for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048-1054, September 1999.
- ◆ Robert Hall. A countermeasure to duplicate-detecting anti-spam techniques., *Technical Report 99.9.1*, AT&T Research Labs, 1999.
- ◆ H. Katirai. Filtering junk e-mail: A performance comparison between genetic programming naive bayes. Available: <http://members.rogers.com/hoomank/papers/katirai99filtering.pdf>, 1999.



# Bibliography of Spam

## 2000

- ◆ Rennie, J., Ifile: An application of machine learning to mail filtering, *Proceedings of the KDD-2000 Workshop on Text Mining*, 2000 <http://citeseer.ist.psu.edu/rennie00ifile.html>
- ◆ I. Androutsopoulos, J. Koutsias, K. Chandrinou, G. Paliouras, and C. Spyropoulos. An evaluation of naive bayesian anti-spam filtering. In G. Potamias, V. Moustakis, and M. van Someren, editors, *Proceedings of the Workshop on Machine Learning in the New Information Age*, pages 9-17, 2000.
- ◆ Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis G., Spyropoulos, C. and P. Stamatopoulos, Learning to filter spam-email: A comparison of a naive Bayesian and memory-based approach. In *workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practices of KDD*. 2000 <http://citeseer.ist.psu.edu/androutsopoulos00learning.html>
- ◆ Yanlei Diao, Hongjun Lu, and Dekai Wu. A comparative study of classification-based personal e-mail filtering. In Takao Terano, Huan Liu, and Arbee L. P. Chen, editors, *Proceedings of PAKDD-00, 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 408-419, Kyoto, JP, 2000.
- ◆ I. Androutsopoulos, J. Koutsias, K. V. Chandrinou, and C. D. Spyropoulos. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of SIGIR-2000*, pages 160-167. ACM, 2000.



# Bibliography of Spam 2001

- ◆ G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. Spyropoulos, and P. Stamatopoulos. *A memory-based approach to anti-spam filtering*. TechReport DEMO 2001, National Centre for Scientific Research "Demokritos", 2001.
- ◆ Georgios Sakkis, Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos, Panagiotis Stamatopoulos. Stacking Classifiers for Anti-Spam Filtering of E-mail. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pp. 44–50, Carnegie Mellon University, Pittsburgh, PA, USA, 2001.
- ◆ A. Kolcz and J. Alsdpector. SVM-based filtering of email spam with content-specific misclassification costs. In *Proceedings of the TextDM'01 Workshop on Text Mining* - held at the 2001 IEEE International Conference on Data Mining, 2001.
- ◆ X. Carreras and L. Marquez. Boosting trees for antispam email filtering. In *Proceedings of RANLP-2001, 4th International Conference on Recent Advances in Natural Language Processing*, 2001.
- ◆ Crawford, E, J Kay and E McCreath, (2001) Automatic Induction of Rules for e-mail Classification, Proceedings of ADCS'2001, Australian Document Computing Symposium, 13-20. [online proceedings](#)



# Bibliography of Spam 2002

- ◆ J. M. Gómez Hidalgo. Evaluating cost-sensitive unsolicited bulk email categorization. In Proceedings of SAC-02, 17th ACM Symposium on Applied Computing, pages 615-620, Madrid, ES, 2002.
- ◆ A. Back, "HashCash -- A Denial of Service Counter-Measure" (5 years on), *Tech Report*, 2002.
- ◆ S. E. Fahlman. Selling interrupt rights: A way to control unwanted email and telephone calls. *IBM Systems Journal*, 41(4):759-766, 2002
- ◆ S.Hird. *Technical Solutions for Controlling Spam* In the proceedings of AUUG2002, Melbourne, 4-6 September, 2002
- ◆ Hidalgo, J., (2002) Evaluating cost-sensitive Unsolicited Bulk Email categorization, *Proceedings of the 2002 ACM symposium on Applied Computing* <http://portal.acm.org/citation.cfm?id=508911&dl=ACM&coll=portal>



## Bibliography of Spam 2003 part 1

- ◆ K. Schneider. *A comparison of event models for naive bayes anti-spam e-mail filtering*. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), 2003.
- ◆ K. Schneider. Fighting spam in real time. In Proceedings of the 2003 Spam Conference, Jan 2003. Available: [http://www.brightmail.com/press/2003\\_MIT\\_Spam\\_Conference/](http://www.brightmail.com/press/2003_MIT_Spam_Conference/).
- ◆ L. Weinstein. *Inside risks: Spam wars*. CACM, 46(8):136, Aug 2003.
- ◆ C. Dwork, A. Goldberg, and M. Naor, "On Memory-Bound Functions for Fighting Spam", *Proceedings of the 23rd Annual International Cryptology Conference (CRYPTO 2003)*, August 2003.
- ◆ John Ioannidis, *Fighting Spam by Encapsulating Policy in Email Addresses*, *Proceedings of Network and Distributed Systems Security Conference (NDSS)*, 2003.
- ◆ M. Abadi, M. Burrows, M. Manasse, and T. Wobber, "Moderately Hard, Memory-bound Functions", *Proceedings of the 10th Annual Network and Distributed System Security Symposium*, February 2003.
- ◆ Center for Democracy and Technology. Why am I getting all this spam? Unsolicited commercial email six month report. Available: <http://www.cdt.org/speech/spam/030319spamreport.shtml>, March 2003.
- ◆ P. Cunningham, N. Nowlan, S. J. Delany, and M. Haahr. A case-based approach to spam filtering that can track concept drift. In *The ICCBR'03 Workshop on Long-Lived CBR Systems*, Trondheim, Norway, June 2003. Available from: <http://www.cs.tcd.ie/publications/tech-reports/reports.03/TCD-CS-2003-16.pdf>



## Bibliography of Spam 2003 part 2

- ◆ K. Eide. Winning the war on spam: Comparison of bayesian spam filters. Available: <http://home.dataparty.no/kristian/reviews/bayesian/>, August 2003.
- ◆ P. Graham. Better bayesian filtering. Available: <http://www.paulgraham.com/better.html>, Jan 2003.
- ◆ R. Kraut, S. Sunder, J. Morris, M. Cronin, and D. Filer. Markets for attention: Will postage for email help? In *ACM Conference on CSCW*, pages 206-215, 2003.
- ◆ Tom Fawcett, " 'In vivo' spam filtering: A Challenge Problem for Data Mining", *KDD Explorations* vol.5 no.2, December 2003
- ◆ J. Graham-Cumming. Field Guide to Spam. Sophos, 2003. Available from [http://www.activestate.com/Products/PureMessage/Field\\_Guide\\_to\\_Spam/](http://www.activestate.com/Products/PureMessage/Field_Guide_to_Spam/) and updated periodically.
- ◆ Kevin R. Gee. Using latent semantic indexing to filter spam, 2003. *ACM Symposium on Applied Computing, Data Mining Track*, 2003.
- ◆ Trevor Stone, *Parameterization of Naive Bayes for Spam Filters*, *Masters Comphrehensive Exam*, University of Colorado at Boulder, 2003.
- ◆ T. Tompkins and D. Handley. Giving e-mail back to the users: Using digital signatures to solve the spam problem. *First Monday*, 8(9), September 2003. [http://firstmonday.org/issues/issue8\\_9/tompkins/index.html](http://firstmonday.org/issues/issue8_9/tompkins/index.html)
- ◆ David Madigan, *Statistics and the War on Spam* in *Statistics, A Guide to the Unknown*, 2003.
- ◆ M. Abadi, A. Birrell, M. Burrows, F. Dabek, and T. Wobber, "Bankable Postage for Network Services", *Proceedings of the 8th Asian Computing Science Conference*, Mumbai, India, December 2003..



# Bibliography of Spam 2004 and To Appear

- ◆ Thede Loder, Marshall Van Alstyne, and Rick Wash. An Economic Solution to the Spam Problem. *ACM E-Commerce*, 2004
  - ◆ Hulten, Geoff, Goodman, Joshua and Rounthwaite, Robert. Filtering Spam E-mail on a Global Scale. World Wide Web Conference, 2004.
  - ◆ Goodman, Joshua and Rounthwaite, Robert. Stopping Outgoing Spam, *ACM Conference on E-Commerce*, May 2004.
  - ◆ Goodman, Joshua and Heckerman, David. Stopping Spam with Statistics. *Significance Magazine*, to appear.
  - ◆ Goodman, Joshua and Rounthwaite, Robert. SmartProof. *To appear*.
  - ◆ Goodman, Joshua. IP Addresses and Clients. *To appear*.
  - ◆ T. V. Zandt. Information overload in a network of targeted communication. RAND (forthcoming).
  - ◆ Hulten, G, Penta, A, Seshadrinathan, G. and Mishra, M. Trends in Spam Products and Methods (to appear.)
- ◆ Don't forget <http://www.ceas.cc> for new papers

