The Past

The Development of Spam

The Development of Spam Filters

The Present

Challenges

Web Spam

The Future?

Challenges in Spam Filtering Research

Christian Siefkes

Freie Universität Berlin
Berlin-Brandenburg Graduate School in Distributed Information Systems
(DFG grant no. GRK 316)

16th February 2007

The Past
The Development of Spam
The Development of Spam Filters
The Present
Challenges
Web Spam
The Future?

Overview

1. The Past
   - The Development of Spam
   - The Development of Spam Filters
2. The Present
   - Challenges
   - Web Spam
3. The Future?

The Past

The Development of Spam

The Development of Spam Filters

The Present

Challenges

Web Spam

The Future?

The Past

One of the Oldest Successful E-Commerce Application

- Google: founded in 1998
- eBay: founded in 1995
- Amazon: launched in 1995
- Lycos (remember?): online since 1994
- Spam: bringing e-commerce right into your inbox since 1994! (or 1978, but that one can be debated)

The Business Model of Spammers

- Increase sale of their products, esp. in highly competitive, dubious, or illegal areas (weight loss, pharmaceuticals, gambling, pornography, “sexual enhancement”, real estate, loans, counterfeit products, fake diplomas . . .)
- “Pump and dump” stock spam: gives spammer an average return of 5.8%; average buyer will lose 5.5% [Fri06]
- Trick users to make bad deals (Nigeria Connection and other scams) or to give away their credentials (phishing)

Early Spam

1 May 1978 DIGITAL WILL BE GIVING A PRODUCT PRESENTATION

Sent by Gary Thuerk, an aggressive DEC marketer, to *all* known Arpanet users (593)

18 Jan 1994 Global Alert for All: Jesus is Coming Soon

Posted to *all* ($\approx$ 6000) Usenet newsgroups
Claimed that "this world's history is coming to a climax"—but actually, this was just the beginning...

12 Apr 1994 Green Card Lottery- Final One?

Two lawyers (Canter and Siegel) advertising services for upcoming U.S. Green Card lottery—first large-scale *commercial* Usenet spam

Rule-based Filters (1997–)

- Perl script filter.plx developed by Mark Jeftovic since August 1997
- Series of patches by Justin Mason finally turned into SpamAssassin—first official release in April 2001:
  - Combines any number of hand-written rules to detect suspicious mails
  - Rules are assigned manually defined scores
  - Mail is marked as spam if summed score surpasses an adjustable threshold
- As spam evolves, such filters tend to become slow and hard to maintain

"A Plan for Spam" (2002)

- Paul Graham proposal for statistical spam filtering, published in August 2002
- Probability of mail being spam is estimated based on conditional probabilities of words to occur in spam:

$$P(S|w_n) = \frac{P(w_n|S) \times P(S)}{P(w_n)}$$

$P(S)$: prior probability of mail being spam
$P(w_n|S)$: conditional probability of word w_n occurring in spam mail
$P(w_n)$: probability of word w_n occurring in any mail

- Final probability is estimated using Bayes' theorem which only holds if features are *independent*
  $\longrightarrow$ Naive Bayes classification

From Words to Phrases (2003)

- Naive Bayes treats text as “bag of words”
- Phrases can be taken account by using e.g. *bigrams* or *trigrams* as features
- Most refined variant used in CRM114 (http://crm114.sourceforge.net)—first alpha release in Jan 2003

From Words to Phrases (2003) (2)

- CRM114 uses all possible combinations of a token (word) with any of the 4 preceding tokens as features:

Number		Sparse Binary Polynomial Hashing (SBPH)				
1	(1)					today?
3	(11)				lucky	today?
5	(101)			feel	<*skip*>	today?
7	(111)			feel	lucky	today?
9	(1001)		you	<*skip*>	<*skip*>	today?
11	(1011)		you	<*skip*>	lucky	today?
13	(1101)		you	feel	<*skip*>	today?
15	(1111)		you	feel	lucky	today?
17	(10001)	Do	<*skip*>	<*skip*>	<*skip*>	today?
19	(10011)	Do	<*skip*>	<*skip*>	lucky	today?
21	(10101)	Do	<*skip*>	feel	<*skip*>	today?
23	(10111)	Do	<*skip*>	feel	lucky	today?
25	(11001)	Do	you	<*skip*>	<*skip*>	today?
27	(11011)	Do	you	<*skip*>	lucky	today?
29	(11101)	Do	you	feel	<*skip*>	today?
31	(11111)	Do	you	feel	lucky	today?

Sparse Bigrams (2004)

- Introduced in [Sie04] (ECML/PKDD 2004)
- Uses only *word pairs* (sparse bigrams) of a token (word) with each one of the 4 preceding tokens as features

Number		Sparse Binary Polynomial Hashing (SBPH)					Orthogonal Sparse Bigrams (OSB)				
1	(1)					today?					
3	(11)				lucky	today?				lucky	today?
5	(101)			feel	<*skip*>	today?			feel	<*skip*>	today?
7	(111)			feel	lucky	today?					
9	(1001)		you	<*skip*>	<*skip*>	today?		you	<*skip*>	<*skip*>	today?
11	(1011)		you	<*skip*>	lucky	today?					
13	(1101)		you	feel	<*skip*>	today?					
15	(1111)		you	feel	lucky	today?					
17	(10001)	Do	<*skip*>	<*skip*>	<*skip*>	today?	Do	<*skip*>	<*skip*>	<*skip*>	today?
19	(10011)	Do	<*skip*>	<*skip*>	lucky	today?					
21	(10101)	Do	<*skip*>	feel	<*skip*>	today?					
23	(10111)	Do	<*skip*>	feel	lucky	today?					
25	(11001)	Do	you	<*skip*>	<*skip*>	today?					
27	(11011)	Do	you	<*skip*>	lucky	today?					
29	(11101)	Do	you	feel	<*skip*>	today?					
31	(11111)	Do	you	feel	lucky	today?					

Sparse Bigrams (2004) (2)

- OSB (sparse bigrams) improves classification speed without degrading results compared to SBPH
- Further improvement in [Sie04]: Uses the Winnow algorithm instead of Naive Bayes or a variation
  - Avoids the untenable *independence assumption*
  - Still supports *incremental training* (no need to know all training data in advance) and is very fast for both training and application

Sparse Bigrams (2004) (3)

Proposed improvements turned out to be highly successful:

- Winnow+OSB was one of the two best (if not the best) of the spam filters evaluated in the Spam Filtering Task of the *Text REtrieval Conference (TREC) 2005*
- OSB features are the default in newer versions of CRM114
- Strato has chosen Winnow+OSB for their server-side spam filter, filtering more than 15 million mails per day
- Fidelis Assis' OSBF-Lua (http://osbf-lua.luaforge.net/) won the Spam Filtering Task of *TREC 2006*

The Past

The Development of Spam

The Development of Spam Filters

The Present

Challenges

Web Spam

The Future?

The Present

The Co-evolution Problem

Spammers react to spam filters becoming more effective and more widespread:

- They increase their reach to make up for it:

Mid 2003	50% of all e-mail is spam
Early 2004	70% is spam
Dec 2006	Spam percentage estimated to be $> 90\%$!

  In absolute numbers, spam tripled from 30 billion mails per day in June 2005 to 85 billion per day in December 2006

- They turn to business areas where they need to reach fewer users to be successful, such as phishing
  $\longrightarrow$ already a single successful attempt can do enormous damage

The Co-evolution Problem (2)

- They improve methods to hide spam from trainable filters $\longrightarrow$ image spam:
  - Percentage of image-based spam increased from 30% in early 2006 to 65% in late 2006
  - CAPTCHA-style techniques are used to hinder OCR
  - Images are slightly varied to prevent hash- or fingerprint-based detection

Spam Leaks Out

Spammers move beyond e-mail and newsgroups into more and more domains:

- The Web:
  - Search results esp. for commercially attractive terms are distorted by worthless spam sites
  - Open or semi-open sites (blogs, wikis, forums, guestbooks, etc.) all become victims of spammers
- Chat and Instant Messaging (SPIM)
- Phones and cell phones: VoIP and SMS spam

Web Spam

- Web Spam: “Web pages that are created to manipulate search engines and deceive Web users” [Web06]
- How much?
  estimated $\approx 8\%$ of Web pages
  and $\approx 18\%$ of Web sites
- Large part of web spam is targeted as search engines (get a higher ranking for spammer’s pages), not at users
  $\longrightarrow$ **Spamdexing**

Detecting Web Spam

[Web06] use spam mails to detect Web spam:

- Most (85%–95%) spam mails contain URLs
- Assumption: "URLs found in email spam messages are reliable indicators of Web spam pages"

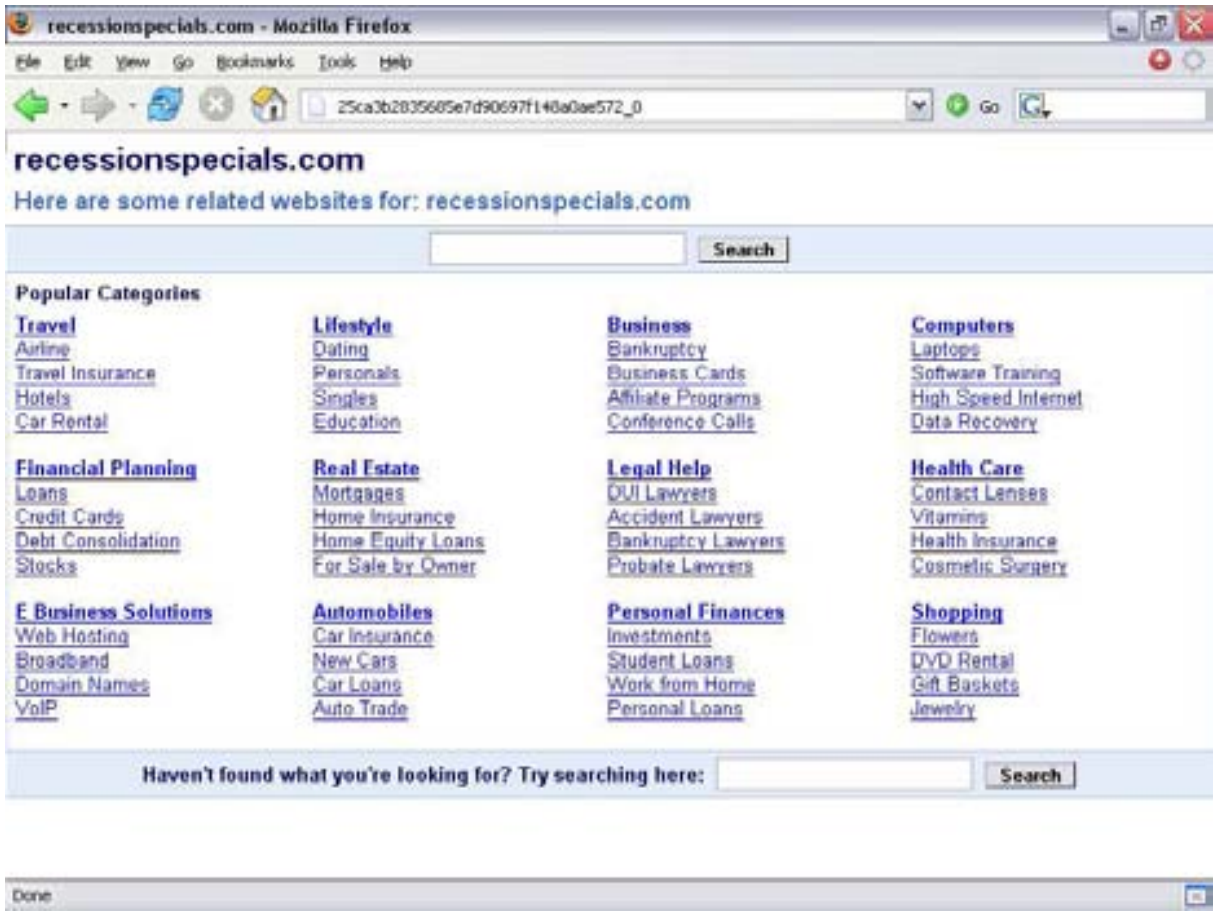
```
You've Won!

Click to see what it is:
http://click.recessionspecials.com/sp/t.pl?id=92408:57561182
------------------------------------------------

Remove yourself from this recurring list by sending a blank email to
mailto:unsub-53821024-5237@recessionspecials.com
```

Detecting Web Spam (2)

- Assumption: “URLs found in email spam messages are reliable indicators of Web spam pages”



Detecting Web Spam (3)

Method:

1. Extract URLs from e-mail spam.
   Tried with 1.4 million spam messages from the SpamArchive (collected from 11/2002–01/2006)
2. Crawl URLs (1.2 mio.), resolving any redirects.
   Some URLs triggered up to 13 redirects; often various different URLs point to the same final redirect.
3. Download files from final URLs ($\approx 408,000$), except for non-textual files
   (app, audio, image, video, etc.: $\approx 25,000$)
4. Identify duplicates
   ($\approx 101,000$ of the remaining $\approx 383,000$ files)

Detecting Web Spam (4)

5. Filter against lists of known-to-be-good URLs (Alexa, SiteAdvisor) to remove false positive.
   (2 of the 3 most frequent final URLs are *fp*: www.yahoo.com and www.msn.com ← results of redirects when spammer's user pages no longer exist.)
   $\approx 349,000$ probable spam pages remain after removing $34,000$ false positives

→ Resulting "Webb Spam Corpus" ($>$ 1 GB, including redirect pages + metadata) available at http://www.webbspamcorpus.org/.

Using Web Spam

- Identified Web spam can be used to enhance e-mail spam filtering: classify “spamicity” of URLs found in messages
- Search engines could improve their ranking algorithms (PageRank) and their result lists by excluding Web spam and degrading pages linking to or linked from Web spam $\longrightarrow$ *ParentPenalty* and *BadRank* algorithms [Wu05]
- Analysis might increase understanding of spammers’ techniques and lead to improved methods of fighting spam

Web Spam: Co-evolution Issues

How might spammers adapt to the analysis of Web spam from e-mail spam?

- Better hide URLs in spam mails (e.g., in images)
  $\longrightarrow$ would make them less accessible and thus be a good thing
- Add URLs to good pages to contaminate collection

Researchers will need to improve methods for detecting URLs and for filtering false positives when such techniques appear.

The Past

The Development of Spam

The Development of Spam Filters

The Present

Challenges

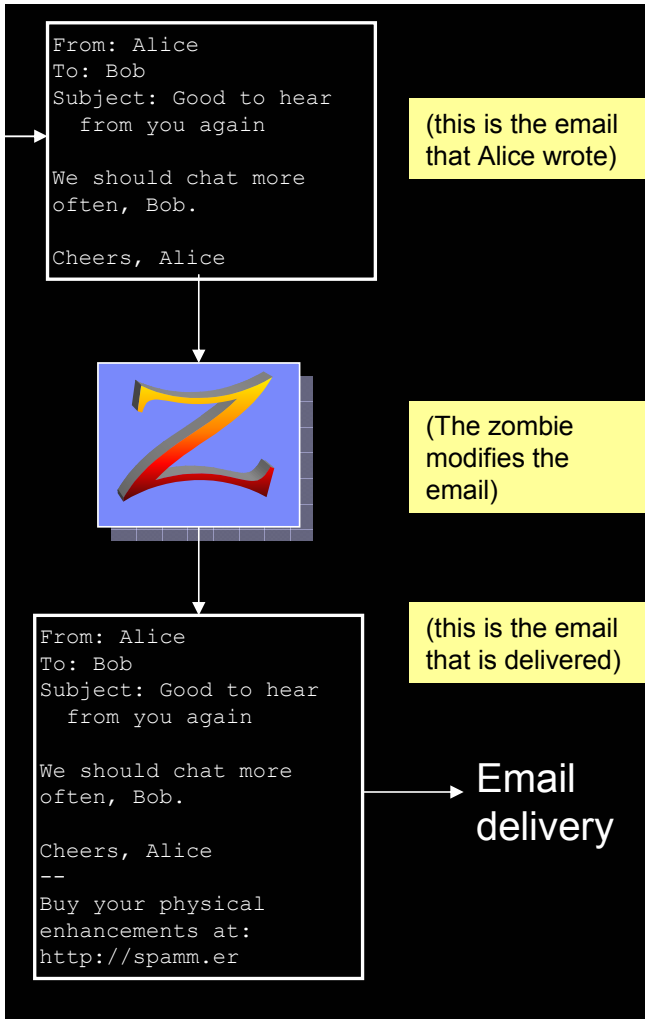
Web Spam

The Future?

The Future?

Parasitic Spam

- Most recent spam is sent from bot nets (networks of infected Windows PCs), a trend that emerged in 2003.
  Vint Cerf: “one in four Internet PCs is part of a bot net.”
- Anticipating co-evolution: How might spammers react if the quota of spam that makes it though spam filters gets “too low” for their taste? [Swi06]
- They could use the zombies (bots) to insert spam content into legitimate (ham) messages (outgoing or incoming)



Parasitic Spam: Methods

1. Spammers might add signatures
   → inconspicuous and hard to detect
2. Or even modify or insert regular content:

```
From: Alice
To: Bob
Subject: Good to hear from you again

We should chat more often, Bob.
Check out: http://spamm.er for your physical
enhancements.

Cheers, Alice
```

   → more effective, but also more likely to raise suspicions
3. Or add new `multipart/*` sections
   → might even be invisible to readers, but would ruin training models of spam filters

Dealing With Parasitic Spam

- Such parasitic spam would be very difficult for filters to handle:
  - mixes ham and spam
    $\longrightarrow$ classifying as either won't do
  - both sender and recipient are legitimate
    $\longrightarrow$ sender authentication won't do
  - PC is infected
    $\longrightarrow$ Message checksums/digital signatures won't do since they could be tampered with
- Spam filters could classify mails *by region* and discard spam regions
- More effective solution: Get rid of zombies!
  $\longrightarrow$ increase PC security, not spam filtering

The Future?

Will we ever get rid of spam?

- Spam is a business model, and a working one
- The only way to get rid of spam would be to destroy the business model
- But so far spammers have managed to adapt to all challenges (legal or technical) to their business model
- We can expect them to do so in the future, by developing new methods (such as parasitic spam)
- We may even be able to anticipate such future moves (which gives us a headstart in fighting them), but this doesn't mean we can prevent them

The Future? (2)

- Co-evolution is a game where (usually) neither side wins, since each is able to counter the developments of the other side

→ Most likely, neither will e-mail collapse nor will the spammers business model cease to work
(though, if anything, the former outcome appears more likely than the latter)

- In any case, *due to the high and increasing effectiveness of spam filters,* mail servers will continue to suffer under even higher loads, and dealing with spam will continue to cost lots of time and money (regarding both e-mail and other domains)

The Future? (3)

The bad news:

- Spam is a business model that works—it will be with us until the end of capitalism

The good news:

- Spam filtering can make spam almost invisible
- . . . poses lots of challenging research problems
- . . . offers various viable business opportunities, too :-)

References

[Fri06] Laura Frieder and Jonathan Zittrain.
Spam Works: Evidence from Stock Touts and Corresponding Market Activity.
Research Paper 43/2006, Oxford Legal Studies, 2006.
URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=920553.

[Sie04] Christian Siefkes, Fidelis Assis, Shalendra Chhabra, and William S. Yerazunis.
Combining Winnow and orthogonal sparse bigrams for incremental spam filtering.
In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, eds., *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2004)*, vol. 3202 of *Lecture Notes in Artificial Intelligence*, pp. 410–421. Springer, 2004.
URL http://www.siefkes.net/papers/winnow-spam.pdf.

[Swi06] Morton Swimmer, Ian Whalley, Barry Leiba, and Nathaniel Borenstein.
Breaking anti-spam systems with parasitic spam.
In *Third Conference on Email and Anti-Spam (CEAS)*. 2006.
URL http://www.ceas.cc/2006/9.pdf.

[Web06] Steve Webb, James Caverlee, and Calton Pu.
Introducing the Webb spam corpus: Using email spam to identify web spam automatically.
In *Third Conference on Email and Anti-Spam (CEAS)*. 2006.
URL http://www.ceas.cc/2006/6.pdf.

[Wu05] Baoning Wu and Brian D. Davison.
Identifying link farm spam pages.
In *14th International World Wide Web Conference*, pp. 820–829. 2005.
URL http://www.cse.lehigh.edu/~brian/pubs/2005/www/.

Sources

- http://www.templetons.com/brad/spamreact.html
- http://www.templetons.com/brad/spamterm.html
- http://en.wikipedia.org/wiki/SpamAssassin
- http://www.search-lab.de/blog/google/2007/01/26/bush-im-luftschutzkeller/
- http://redtape.msnbc.com/2007/01/spam_is_back_an.html
- http://spamlinks.net/about.htm
- http://en.wikipedia.org/wiki/E-mail_spam
- http://www.avertlabs.com/research/blog/?p=170
- http://portal.spidynamics.com/blogs/msutton/archive/2007/01/04/A-Tour-of-the-Google-Blacklist.aspx