Spam in Blogs and Social Media

Pranam Kolari, Tim Finin
Akshay Java, Anupam Joshi

March 25, 2007
Tutorial Outline

• Spam on the Internet
  – Variants
  – Social Media Spam
• Reason behind Spam in Blogs
• Detecting Spam Blogs
• Trends and Issues
• How can you help?
• Conclusions
Presenters

Pranam Kolari is a UMBC PhD student. His dissertation is on spam blog detection, with tools developed in use both by academia and industry. He has active research interest in internal corporate blogs, the Semantic Web and blog analytics.

Tim Finin is a UMBC Professor with over 30 years of experience in the applying AI to information systems, intelligent interfaces and robotics. Current interests include social media, the Semantic Web and multi-agent systems.

Akshay Java is a UMBC PhD student. His dissertation is on identifying influence and opinions in social media. His research interests include blog analytics, information retrieval, natural language processing and the Semantic Web.

Anupam Joshi is a UMBC Professor with research interests in the broad area of networked computing and intelligent systems. He currently serves on the editorial board of the International Journal of the Semantic Web and Information.
SPAM!
E-mail Spam

• Early form seen around 1992 with MAKE MONEY FAST
• 80-85% of all e-mail traffic is spam
• In numbers

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>(June)</td>
<td>30 billion per day</td>
</tr>
<tr>
<td>2006</td>
<td>(June)</td>
<td>55 billion per day</td>
</tr>
<tr>
<td>2006</td>
<td>(December)</td>
<td>85 billion per day</td>
</tr>
<tr>
<td>2007</td>
<td>(February)</td>
<td>90 billion per day</td>
</tr>
</tbody>
</table>

Sources: IronPort, Wikipedia
(ping!) You have new mail!

Enlarge Your Thesis! Make your Advisor Happier! HOT ANALYSIS TOOLZ!

Click Here!

What are you doing? What? Me? Nothing! I clicked by accident!
What is Spam?

• “Unsolicited usually commercial e-mail sent to a large number of addresses” – Merriam Webster Online

• “Spamming is the abuse of electronic messaging systems to send unsolicited bulk messages, which are almost universally undesired.” – Wikipedia

• As the Internet has supported new applications, many other forms are common, requiring a much broader definition

Capturing user attention unjustifiably on the Internet (E-mail, Web, Social Media etc..)
**Taxonomy of Spam**

**Internet Spam**

- **Direct**
  - E-Mail Spam
  - IM Spam (SPIM)
  - Community Spam

- **Indirect**
  - Tag Spam
  - Comment Spam
  - Spam Blogs (Splogs)
  - General Web Spam

**spamdexing**
Spamdexing

“We use the term spamming (also, spamdexing) to refer to any deliberate human action that is meant to trigger an unjustifiably favorable relevance or importance for a page, considering the page’s true value”
- Gyongyi et al

• Compromises Relevance and Importance Scores used by Search Engines
• Mostly effects the long tail of keywords (more susceptible)
• Spam in Blogs is a form of spamdexing
• 20% of the indexed Web (2005)
2 results stored on your computer - Hide - About

Student Loan Consolidation - loan consolidation: student loan consolidation
Student Loan Consolidation - 2, 2006 Student Loan Consolidation Rates to

Repaying Student Loans Held by the U.S. Department of Education
1 gif Bugal Loan Consolidation 1 gif Bugal Going Back to School "Can I pay my defaulted student loan held by the Department by credit card?" www.ed.gov/offices/OSFPDICSR/repaying.html - 23k - Cached - Similar pages - Note this

Loan Consolidation
US Department of Educations Collection for Defaulted Student Loans Guide - Information about loan consolidation options that are available to borrowers.
www.ed.gov/offices/OSFPDICSR/consolidation.html - 19k - Cached - Similar pages - Note this

Student Loan Solutions: Student College Loan Consolidation ...
StudentLoanSolutions is a premier lender with the Department of Education. Call 1-800-255-5105. StudentLoanSolutions offers some of the best borrower...
www.studentloansolutions.com/ - 9k - Cached - Similar pages - Note this

Student Loan Consolidation at Nelnet: Federal Loan Consolidation ...
Student Loan Consolidation at Nelnet. Specializing in federal loan consolidation, school loan consolidation, student loan debt consolidation and federal...
www.consolidation.nelnet.netDefault.asp - 19k - Cached - Similar pages - Note this

The Student Loan People
For more information on Federal Loan Consolidation, visit our Consolidation pages. Do You Have a Defaulted Loan? If you have a student loan and payment has...
www.khesic.ky.gov/defaults_repayment.html - 20k - Cached - Similar pages - Note this

Defaulted Student Loan Consolidation Services
Our default consolidation services help your credit recover quickly!
www.studentloanconsolidator.com/default/ - 15k - Cached - Similar pages - Note this

Default Student Loan Consolidation Testimonials

Sponsored Links
Lower payments up to 58%
Some of the top benefits available For help call us at (800) 255-5105 www.StudentLoanSolutions.com

4.5% Fixed Consolidation
Lock in rate now as low as 4.5%. Pre-qual in 1 minute (866) 810-4275 www.NextStudent.com

Consolidate with Wachovia
Cut Your Monthly Student Loan Payment In Half. We Make It Easy.
www.Wachovia.com

Defaulted Student Loan
Get out of debt fast by knowing your options. Nonprofit, learn more www.DebtRaleOptions.org

Defaulted Student Loan
Directory Of Student Loan Deals. Find Student Loans Quickly
StudentLoanWealthyGeek.com

Get Out Default
Stop Garnishment and Tax Seizures and reduce your interest by 1.25%
mystudentloanblock.com

Student Loan DEFAULT Help
Stop tax offset, garnishment, etc. Act NOW before it gets worse!
wssn.net

4.5% Fixed Consolidation
Lock in a low 4.5% rate. Reduce your payment by 60%. No Fees.
www.LendFocus.com/Student
Taxonomy of Spam

Internet Spam

Direct

E-Mail Spam
IM Spam (SPIM)
Community Spam

Indirect

Tag Spam
Comment Spam
Spam Blogs (Splogs)
General Web Spam

(Form)

(Mechanisms)

Social Media Spam
DIGG Spam

• Sybil attack
  – “in computer security is an attack wherein a reputation system is subverted by forging identities in peer-to-peer networks” – Wikipedia

• Users that sell
  – http://userssubmitter.com $20, plus $1 per digg
  – Spike the vote, site sold on eBay, bought by digg user, finally shut down by digg
  – Geekforlife, a top 100 digg user, sold his account for $822 on eBay

• In addition to direct spam, DIGG popular page also results in high ranking on search engines (Spamdexing)
A calm carnival crowd

Tuesday, February 27th, 2007, 2:07 pm

[General]

I said before, that a crowd can be violent. But

“Why Are People Fascinated By Photographs of Crowds?”

• Author Submits Story
• 4 hours later only one DIGG on the story
• Seeks out User/Submitter
• 12 hours later a DIGG Popular Story
Tag Spam

• Social Bookmark Tools
  – del.icio.us spam in long tail
  – Furl popular page spam

• Also used for spamdexing
How to Use Simple Tag and Ping Marketing Techniques
How to Use Simple Tag and Ping Marketing Techniques.
How to Use Simple Tag and Ping Marketing Techniques.

Marketing Tips From The Tag And Ping School of Marketing
Marketing Tips From The Tag And Ping School of Marketing.
ezinearticles.com/?Marketing-Tips-From-The-Tag-And-Ping-School-of-Marketing&id=210546 - 37k - Cached - Similar pages

Tag And Ping - The Next Generation of Blog And Ping!
Discover The New Breakthrough Method That Instantly, Surely, & Effortlessly Drives Hordes Of Laser Targeted Visitors To Your Website.
www.tagandping.com/ - 80k - Cached - Similar pages

Tag and Ping Phenomenon - WordPress SEO and Blog Marketing
Tag and ping has been sweeping through the internet marketing community with the release of the Tag and Ping guide by Sean Wu.
www.optiniche.com/blog/174/tag-and-ping/ - 43k - Cached - Similar pages

Tag and Ping Review - WordPress SEO and Blog Marketing
Tag and ping has received a number of reviews online, some favourable and others no so favourable.
www.optiniche.com/blog/179/tag-ping-review/ - 26k - Cached - Similar pages

Blogmarks.net : Public marks with tag "tag and ping"
Tag and Ping fallout, RSS revenues, and Web 2.0 Idea Management. by surftopmind. Coverage of the results of tag and ping spam, and two new web 2.0 sites of... www.blogmarks.net//marks/tag/tag+and+ping - 42k - Cached - Similar pages

Tag and Ping - Software that automates Tag and Ping
MyBlogLog Spam

• Fake Profile Image
• Fake users
• Fake site visits
• Fake co-authors
• Add friends
• Comment Spam

Hi refer,

I would like to add you as a co-author of my MyBlogLog community below:

Blog/Site: Progressive House (http://www.[link removed].com)
MyBlogLog community:
http://www.mybloglog.com/buzz/community/[text removed]

Your MUST click on the link below to accept this request:
http://www.mybloglog.com/buzz/[link removed]

Thanks,
Blogmemes_Belgium
A Smarter Web - Tech Review on the Semantic Web

March 12th, 2007, by Tim Finin, posted in GENERAL, Semantic Web, Web

The current issue of Technology Review has a long article on the Semantic Web, A Smarter Web, with the theme

New technologies will make online search more intelligent—and may even lead to a "Web 3.0."

The article tells the Semantic Web story using Eric Miller's involvement as a thread and mentions many other people and companies along the way.
MySpace Spam

• MySpace is the (5th) 4th highest visited site
• Direct User Targeting
  – Free Products
  – CAMS

<table>
<thead>
<tr>
<th>Date/Time</th>
<th>User</th>
<th>Message Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 8, 2007 3:10 PM</td>
<td>Keep Drea.</td>
<td>FREE COLLEGE SCHOLARSHIPS GIVEAWAY- $10000 SCHOLAR would like to invite you to join the group FREE COLLEGE SCHOLARSHIPS GIVEAWAY- $10000 SCHOLARSHIPS FREE FOR COLLEGE- CLICK HERE</td>
</tr>
<tr>
<td>Mar 6, 2007 9:49 AM</td>
<td>Alexander.</td>
<td>FREE COLLEGE SCHOLARSHIPS EVENT GIVEAWAY- CLICK HE would like to invite you to join the group FREE COLLEGE SCHOLARSHIPS EVENT GIVEAWAY- CLICK HERE</td>
</tr>
</tbody>
</table>
Hey hun, How are you? Everything's going okay on my end. I think we're gonna go out this weekend, maybe check out a movie or something. Wish you could go... that would be soooo much fun! I'm thinking about getting a little kitten from the shelter, maybe. What do ya think? Ohhh, almost forgot to tell ya ... shawn told me to give you this website; said that you might be interested in it. stuff and I have been taking these new diet pills for a couple of weeks now and have been seeing amazing results. We actually have lost 27 lbs between the 2 of us in the past few weeks! you HAVE to try this out... we've been so busy that we haven't been able to hit the gym yet, so i can only imagine how much we'll lose after we start exercising. and we have soooo much energy now. WOW! anyway, the website we got it from is right here, so just click here. Take care hun,

The Hoodia Scam

• Create an “interesting” profile
• Aggressively add friends (+Sybil Attacks)
• Wall Hoodia Comments
Orkut Spam

• Community Oriented Spam extends across almost all tools

Misti Kulkarni

from: Misti Kulkarni
to: TSEC Alumni across the GLOBE
sent: January 31, 2007 12:29 AM 12:29 AM
subject:

If you desperately need an urgent job with good salary, we can help you out with information at:

http://www.jobguarentee.bravehost.com

If you have any queries plz write to me at:
mail2cyber@gmail.com

Happy new year 2007
Other forms of spam

It does not take long for spammers to exploit a new form of social media

- **Wiki(pedia) Spam**
  
  *The wikipedia community deletes spam articles and pages*

- **Guestbook Spam**
  
  *Yes, some still have guestbooks and they attract spam entries*

- **IRC spam**
  
  *Bots on IRC channels send you unwanted messages*

- **Second life spam**
  
  *Notecards pop up with unwanted messages*

- **Twitter spam**
  
  *It appeared about a week after Twitter went big time*

- **Semantic web spam**
  
  *This one has not yet been seen*
Spamdexing and its relation to Blogs
"I had my own blog for a while, but I decided to go back to just pointless, incessant barking."
Anatomy of Spamdexing

- Identify Profitable Contexts
- Create multiple doorway pages (or blogs)
- Spamdex doorways
  - Target Relevance Scores
    - TFIDF
    - Meta-tag
  - Target Importance Scores
    - PageRank
- (Or) Game Sponsored Search
  - Search Engine Arbitrage
- Monetize
  - Affiliate Programs
  - Contextual Advertisements
Identify Profitable Contexts
Doorway Page for “Student Loan Consolidation”
10K in-links!

Oyler Action Photography
Tri City Speedway DSC_6503.JPG
oyleractionphotos.com/wp-gallery2.php?g2_view=
core.ShowItem&g2_ItemId= 36 · Cached page

Oyler Action Photography
Keep up the good work! show full show summary.
Wow that's a great picture! I'm an aspiring photographer myself and just hope some day I'm able to take a quality shot like that.
oyleractionphotos.com/wp-gallery2.php?g2_view=
core.ShowItem&g2_ItemId= 168 · Cached page
+Show more results from oyleractionphotos.com

Temple Stalk.com: January 2005 Archives
templestark.com/archives/2005/01/ index.php · Cached page

2003_Thai_Camp 081
Really great page ! Thanks! <a href=
http://castor.tamu.edu/TCAAAbusiness-search.html
>business search</a> [url=ht show full show
summary. Really great page
inosanto.com/gallery/main.php?g2_ItemId= 6652 .
Cached page

House Improvement | Ohio Alpha Beta Phikaps
Album: Main Stairs Carpet The main set of stairs in the
draw received new carpet in late 2003.
Browser Javascript Redirect for Users
Cryptic JavaScript Redirects

eval(unescape("%64%6f%63%75%6d%65%6e%74%2e%77%72%69%74%65%28%
%22%3c%73%63%72%69%70%74%20%73%72%63%3d%27%68%74%74%22%2b%
%3a%2f%2f%62%69%67%22%70%3a%2f%2f%68%71%2e%69%6e%66%6f%73%74%61%
%73%2e%70%3f%69%22%3e%3c%2f%73%63%72%69%70%74%22%2b%22%3e%22%3b")
)

document.write
("<script src='http://bighq.info/stats2.php?i='+'d'
=12&group=2'>");
LEADS to affiliates highly profitable
Anatomy of Spamdexing

Affiliate Programs
Context Ads

arbitrage

ads/affiliate links

in-links

Spam pages,
Spam Blogs
[DOORWAY]

JavaScript Redirect
[Previously Cloaking]

Affiliate Program
Buyers

Spam pages,
Spam Blogs,
Spam Comments,
Guestbook Spam
Wiki Spam

spamdex

Search Engines

SERP
Why Blogs used in Spamdexing

• Spam Blogs
  – Quickly indexed by search engines (ping servers)
  – Can be hosted on third party services with high authority and trust (e.g. blogspot)
  – Search engines show a higher preference to blogs – crawling and ranking

• Spam Comments
  – In-links from high authoring blog sites indexed frequently
  – “no-follow” not universally used

• In both cases blogs provide quicker return on investment to spammers
Comment Spam in Numbers

want protection from all that orange? Get Akismet.

Totals

Total spam: 857,506,143
Total ham: 48,242,922
56% of all blogs are splogs!

**Splogs in Numbers**

- Sampled weblogs.com during the last week of January
- 8.3 Million Pings
- Used existing splog detection tools
- Around 10% error rate

**Silicon Valley or Splog Valley?**

1. **Mountain View, CA**
2. **Washington DC**
3. **San Francisco, CA**
4. **Orlando, FL**
5. **Lansing, MI**
Splogs – Context

High PPC contexts are primary spam drivers

auto buy california cancer card casino cheap consolidation credit debt diet discount equipment estate finance florida forex free gift golf health hotel insurance jewelry lawyer loan loans medical money mortgage new online phone poker rental sale software texas trading travel used vacation video wedding
The Blogger Episode

I've been wanting to start a blog, but I don't have the technical knowledge.

Dude, just get yourself a free BlogSpot account.

Really? Even though I'm not a Spammer??

© BLAUGH.COM - FITZ & PIRILLO - SAVE 10% AT GODADDY BY USING THE COUPON CODE "BLAUGH"
Why are people complaining?

• Splogs reduce user trust on social web ranking, making them visit web-pages they can well do without
• Splog content is often plagiarized
• Splogs demote value of authentic content
• Splogs steal advertising (referral) revenue from authentic content producers
• Splogs stress the blogosphere infrastructure
• Splogs skew results of market research tools
Source of the Problem
Splog software ?!

“Honestly, Do you think people who make $10k/month from adsense make blogs manually? Come on, they need to make them as fast as possible. Save Time = More Money! It’s Common SENSE! How much money do you think you will save if you can increase your work pace by a hundred times? Think about it…”

“Discover The Amazing Stealth Traffic Secrets Insiders Use To Drive Thousands Of Targeted Visitors To Any Site They Desire!”

“Holy Grail Of Advertising… “

$ 197

“Easily Dominate Any Market, Any Search Engine, Any Keyword.”
Splog bait: young girls need personal injury lawyer to pay for diplomas

« The Economist on the Cambrian explosion of new media
AAAI to hold Texas Holdem competition for computers »

Splog bait: young girls need personal injury lawyer to pay for diplomas

By Tim Finin on Saturday, April 29th, 2006 at 8:11 pm

This splog bait has many terms, such as Royal Caribbean Cruise and Aruba Vacation Package, that make the splog bait post likely to be plagiarized by sploggers. Did you ever wonder what happens when a bus full of young girls get into an accident on their way to an online gambling site? They probably hoped to make millions of dollars playing poker. Texas holdem and blackjack. Now they need a personal injury lawyer to sue the bus company! (Yes, this is splog bait.) The poor girls will have to take brand-name, FDA approved medications for their injuries — drugs like ambien, tramadol, lexapro, phentermine and viagra. Some might even require physical therapy if their limbs are broken or burned. But that’s cheap compared to the millions they lost playing poker.
Our splog bait was picked up and used by dozens of sploggers.
Splog bait: young girls need personal injury lawyer to pay for ... by Tampa Florida real estate

Sat, 29 Apr 2006 20:59:51 -0400

SPLOG BAIT: YOUNG GIRLS NEED PERSONAL INJURY LAWYER TO PAY FOR ... 
by Tampa Florida real estate (tampa-florida-real-estate) @ Sat, 29 Apr 2006 20:59:51 -0400

Now they need a personal injury lawyer to sue the bus company! (Yes, this is splog bait. ... And the real estate market might go soft if the bank rates are not low for mortgage — maybe mortgage insurance will help. ...
How do these tools work?

Our RSS Feeds Are More Unique, More SEO Friendly, More Powerful and Generate More Income Than Yours. Want To Know WHY???
RSS Magician

- Content Source
  - Plagiarized
  - Article Directories
- Content Manipulation
  - Dictionary
  - Word Shuffling
  - Sentence Shuffling
- Multi-lingual Support!

However, Pro-sploggers don’t use these tools
Characteristics of Spam (Blogs)
Where do they appear?
Diet Information

Weight-loss diets restrict the intake of specific foods, or food in general, or reduce body weight. This is what "fad diets" are marketed for. There is a (sometimes confusing) multitude of weight loss techniques, many of which are ineffective. What works for one person will not necessarily work for another, due to metabolic differences and lifestyle factors. Also, it’s important to note that dieting does not actually lead to weight loss in the long term. Reducing the body's food supply causes it to stockpile excess fat as a starvation response.

Copyright 2005-2006 -
November 2006 New No Deposit Rtg Coupon Codes
November 12, 2006 on 2:45 am | In Uncategorized | No Comments

Bone Cancer
proudly hosted by health350

Bob Ryan finds inspiration in his son playing for the Packers

A Watchdog set to block bone cancer drug

Doctor’s Push For Drug Pits Him Against Its Maker

A Doctor’s Push For Drug Pits Him Against Its Maker

Wall Street Journal (subscription), NY - 4 hours ago... 46, a pharmacy specialist in gastrointestinal nutrition, had been trying to help a teenager with cancer who was unable to eat following a bone-marrow transplant ...

IDM Pharma Files For European Approval To Market Hepact For ...

Trading Markets, CA - Nov 7, 2006

Post Stitching
• Excerpts scraped from other sources
Did you know that refurbished power tools could

Did you know that refurbished power tools could continue being the concern your families are deliberating? Surely online tool power tools retailer au should be the advantage you’re trying to get information for. Some bosses say chicago electronic power tools could continue being a good thing. Others say it’s not so good. Believe people in general needs a little craft power tools every now and then? Let us help you as much as we can in finding yours.
If you thought the property market has little to do with football. Think again! A recent study by a magazine shows that about 20% of prime country homes are bought by footballers.

If you have been under the impression that the only property that football is concerned with is its pitch, think again or question an estate agent in places like Cheshire or Surrey.

Country Life, a rural set magazine, annually analyses prime country homes that are sold for more than £2m. But the point to notice is that 40% of such properties are bought by entrepreneurs, 20% by people employed in the service sector and 20% to professional footballers.

The top end of the Cheshire property market is dominated by players from Manchester United, Liverpool, Everton, Blackburn and Manchester City. A Real Estate Consultancy firm working on a confidential document for FA claims that...
Let me know if you need any help paying all those bills, Dad...

My AdSense scraper sites did pretty well last month.
The Splog Detection problem
The Splog Detection Problem

Given a blog identified by its URL X, is it spam?

How is this different from other spam detection problems?
What and When of filtering

E-mail Spam

Web Spam

Spam Blogs

Social Network Spam

- Time
- Posts
- Friends

- Time
- Time
## Constraints on the Filter

<table>
<thead>
<tr>
<th>E-mail Spam</th>
<th>Web Spam</th>
<th>Spam Blogs</th>
<th>Social Network Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>Search Engines</td>
<td>Web Search Engines</td>
<td>Community Sites</td>
</tr>
<tr>
<td>E-mail Service Provider</td>
<td>Page Hosting Services (eg. Tripod)</td>
<td>Blog Search Engines</td>
<td>Blog Hosting Services</td>
</tr>
<tr>
<td>Fast Detection</td>
<td>Batch Detection</td>
<td>Fast Detection</td>
<td>Mostly Offline</td>
</tr>
<tr>
<td>Low Overhead</td>
<td>Mostly Offline</td>
<td>Low Overhead</td>
<td>Batch Detection</td>
</tr>
<tr>
<td>Online</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Who actively uses it?

- **E-mail Spam**: Users, E-mail Service Provider
- **Web Spam**: Search Engines, Page Hosting Services (eg. Tripod)
- **Spam Blogs**: Web Search Engines, Blog Search Engines, Blog Hosting Services, (Ping Servers)
- **Social Network Spam**: Community Sites
### Adversarial Attacks on the Filter

<table>
<thead>
<tr>
<th>E-mail Spam</th>
<th>Web Spam</th>
<th>Spam Blogs</th>
<th>Social Network Spam</th>
</tr>
</thead>
</table>
| - Image Spam  
  - Character Replacement | - Scripts  
  - Doorways | - Scripts  
  - Doorways  
  - Deceptive Behavior | - Deceptive Behavior |

![Diagram](attachment:diagram.png)
Detection Background

• Pre-indexing at Ping Stream
• Post-indexing

• Blacklists/Whitelists
• URL Based Features
• Home-Page Based Features
• Feed Based Features (Temporal)
• Link Based Features
Detection Background

Filters deployed on Ping Streams (typically)...
Detection Approaches that Work
DID YOU GET THAT EMAIL FROM PROF. SMITH?

Uh... no.

Really? He sent it to everyone.

I guess my spam filter must have deleted it.

That's weird. What are your spam settings?

I use a very strict set of keyword heuristics based on subject matter.

Let me guess, anything that doesn't say "free food" gets deleted?

If it's not about food, it's unsolicited.
IP Based Blacklists

• Intuition: Most self-hosted blogs are tied to “spammer-friendly” domain hosting services
• Map Ping URL’s to IP Addresses
• Rank IP Addresses by hosted blogs
• Pick the top 100 (say)
  – Randomly sample for a subset of blogs
  – Verify no false positives
  – Add IP to Blacklist
• Actively used by a2b.cc and many others
Models over name and URL values can be very effective
Regex Filters on (Name) URL

- The Blogspot “?” filter
- The hyphen filter
- .. and many more
- Very effective and fast

- A reactive measure
- Active Supervision
- Highly susceptible to Adversarial Attacks
Ping Stream at Weblogs.com

Text token based features over name and URL can be very effective
URL *Only* Features

- Intuition: Spammers target search engines through context rich URLs
- No page fetches – can be very fast
- Techniques vary
  - How are URLs segmented
    - Hyphens, forward slashes etc.
    - Supervised Segmentation Technique
    - Character N-grams
  - How are detection models constructed
    - SVMs
    - Bayesian Filters
Novel URL Segmentation

- Motivation: Spammers glue words in URL
- Learn Segmentation using spamming contexts as proposed by Salvetti et al

http://dietsthatwork.blogspot.com

Naïve Approach

Enhanced Approach

Salvetti et al
Novel URL Segmentation

- Train a Bayesian Classifier (10K +, 10K -)
- 1k test samples
- F-measure 76% spam, 79% authentic
- Human Baseline 73% spam, 78% authentic
- Classifier beats humans when identifying spam using URLs alone!
URL+Name

- SVM based approach
- Hand Verified Samples of 2K spam URLs and 2K authentic URLs from ping server
- N-character gram tokenization
- Linear Kernel
- Binary Features
- F-measure ~ 90%
Blog Home-Page Based Features

• Intuition: Home-Page can provide a good snapshot of a blog (most recent posts), and also captures blog-rolls and link-rolls
• Only one page fetch per URL – can be very fast
• Techniques vary by selected features
  – Bag-of-words
  – Bag-of-tokenized-outlinks
  – Text Compression
JANUARY 7, 2007

expenses Software
management package, accurately calculates all
claims including VAT and mileage. Ensuring
expenses Sales Agents are expected to
negotiate and secure sale price, implementation
and configuration.

Democrats to Offer Legislation for Medicare to Negotiate Drug Prices
June 27, 2006 Democrats say they will
introduce legislation today to allow Medicare to
negotiate contracts with pharmacies — are too
slow for community pharmacists to cover their
expenses.

Tax Tiger can negotiate an Installment Agreement with the IRS on your
As you can expect, the IRS definition of living
expenses and your credit is different. Why
your approval will vary and have them as you

Scraped content typically lacks the personal nature of blog posts
Bag-of-Words Feature

- SVM based model
  - Binary Features
  - Linear Kernel

- Hand verified training set of blogs and splogs
  - 700 blogs and 700 splogs

- F measure ~ 88%
Bag-of-2-Words Feature

• Features created using two adjacent words
• Less susceptible to adversarial attacks
• SVM based model
  – Binary Features
  – Linear Kernel
• F measure ~ 86%
Bag-of-URL Tokens Feature

• Features constructed by tokenizing outgoing URLs
• Similar to (earlier) URL only models but captures more context
  – URL tokens in the badge
crazy blogosphere
  – Default Wordpress Links
• F-measure ~ 82%
More Home-Page Based Features

- We have been experimenting with multiple approaches
- Dataset available at:
Post 1 from a Spam Blog on Forex.. Trading
Post 2 from the same Spam Blog on Forex.. Trading
Temporal Regularities

• Intuition: Splogs feature highly correlated content across multiple posts (text, out-links etc..)
• Lin et al have proposed and analyzed useful metrics to identify such correlations
• Two fetches per URL – can exploit structured feed entries
• Precision ~ 86%, Recall ~ 47%
Other features that work

- Outgoing Anchors
- Character N-grams and Word N-grams
- Named Entity Ratio
- Text Compression Ratio
- Pronoun Entity Ratio
Detection Approaches that could also Work
Figure 8: Ping Time Series of Italian Blogs on a single day.

Figure 9: Ping Time Series of Italian Blogs over five days.
Sping vs. Ping times

Figure 10: Ping Time Series of Blogs on a single day.

Figure 11: Ping Time Series of Blogs over five days.

Figure 12: Ping Time Series of Splogs on a single day.

Figure 13: Ping Time Series of Splogs over a five day period.
Spings vs. Pings: frequency

blogs vs. their ping frequency follows a power
law, but splogs vs. spings does not
Splogs vs. Blogs – In-degree

Figure 6: In-degree distribution of authentic blogs subscribe to a power-law.
Splogs vs. Blogs – Out-degree

Figure 7: Out-degree distribution of authentic blogs subscribe to a power-law.
Splog Protection – Global

- Test set of 700 authentic blogs and 700 splogs
- In-links from Technorati
- Out-links from blog-home page
- Labeled in-links/out-links to be a blog, splog, news site, other website
- 3 feature-types trained using SVMs
  - In-link distribution
  - Out-link distribution
  - Co-citation distribution
- F-measure ~ 80%
Summary of Existing Techniques

• Local Properties will continue to hold the key
  – Fast
  – Low Cost
  – Can be easily incorporated into blog harvesting (crawling)

• Global Techniques restricted by coverage
  – Less Effective to Blog Search Engines
  – Continue to be effective for web search engines
How Techniques are Used

Ping Stream

Check with known blogs and splogs
Check known regex patterns
Use models over text of URL
Use models over Blog home-page
Use models over post correlation

BLACKLISTS
WHITELISTS
REGEX
URL MODEL
TEXT MODEL
TEMPORAL MODEL

Increasing cost
Elimination Approaches (Most Don’t Work)
“So, what that mean for the current users? It means that we have a few things to iron out. We are working on them as fast as possible. It will insure that we don’t have these issues going forward in the future and have a solid solution.”

SPLOG REPORTER STATUS = RED

“Due to lack of resources, mainly time, we have decided to pursue other other web 2.0 interests. We had a good run and appreciate everyone who supported the splog fighting movement. We have decided to keep the site up so as not to "break the web" and all the links to the site. If anyone is interested in acquiring Splog Reporter please contact us.”
Eliminating Splogs depends directly on blacklisting by search engines, contextual advertisement and affiliates. Sploggers have no more incentive to host such blogs and typically bring it down.

... a rather slow elimination process.
How Splogs Effect Blog Analytics
Nature of Splogs in TREC 2006

- Around 83K identifiable blogs in the collection, with 3.2M permalinks
- We identified 13,542 splogs
- Blacklisted 543K permalinks from these splogs
- This accounts to 16% of the entire collection
- Results tally in % to TREC dataset (Macdonald et al 2006)
Impact of Splogs in TREC Queries

- Distribution of Splogs that appear TREC queries
- Top 100 Search Results ranked using TFIDF Scoring

Keywords: American Idol, Cholesterol, Hybrid Cars
Higher in Spam Prone Contexts

Splog Distribution for 'Spam Terms'

Spam query terms based on analysis by McDonald et al 2006 ..
Input from Groups
Dealing with Splogs
• When did you start seeing spam? **in 2005** – on blog search engine results.

• How much spam? (in percents) **roughly 25%** of top retrieval results

• Is it increasing/decreasing? **increasing**

• How are you tackling it? and How much effort goes into it? **definitely my team considers this an important research effort. In terms of person-hrs, at least 2080 in 2006**

• Do you have any new architectures/frameworks/initiatives? **Yes – work publicly available**

• What would you want researchers to address? **temporal and structural properties**

• What are future trends? **collaborative splog detection**
• When did you start seeing spam? Late 2005. Late summer of 2002 couple of years hand edited black lists, URL pattern filtering
• How much spam? (in percents) big non-blog category, around 70% from weblogs.com,
• Is it increasing/decreasing? No significant change
• How are you tackling it? and How much effort goes into it? Around 4160 person-hrs
• Any message(blurbs) to the ICWSM audience? "At Sphere, we're analyzing the mechanisms used by splogs and finding detectable signatures that can be applied to very large datasets. It's our contention that available tools and protocols must necessarily limit the strategies available to successful sploggers. While that set of strategies continues to grow, we're continually innovating to find new and effective ways of detecting those strategies, suppressing spam, and keeping the blogosphere a rich, easily navigable universe of information"
When did you start seeing spam? *In early 2005, we started to see the #spam blogs and pings from non-weblogs sources ramp out rapidly.*

How much spam? (in percents)
*There's a big range, depends on the day. There can be huge spam attacks on some days. For today, for example: 15% spam and 30% non-blog (but today's 15% spam number is on the low side). This doesn't include other kinds of bad pings: duplicates, out-dated, urls that are forbidden or have no data.*

Is it increasing/decreasing? *It's increasing in absolute counts, just like the total number of weblogs ever created is increasing. However, in terms of the number of active spam weblogs, it's fairly constant, just like the number of active English-language weblogs is fairly flat now.*

Do you have any new architectures/frameworks/initiatives? *Our architecture is in place and working well.*
What would you want researchers to address?
- There is a minority of sophisticated spammers that create sleeper spam blogs that look just like regular blogs and then months later start linking to affiliate sites. How to catch these early?
- How to detect repurposed blog content (downright plagiarism either for use in spam blog or simply for the age-old reason of building one's own reputation on false premises).
- Spam in foreign languages: structural similarities/differences/extent of the problem?

What are future trends?
- PayForPost type services: in-post advertisements that indistinguishable from blogger content and may or may not be disclosed => need for spam detection at the post level instead of simply at the blog level; need to filter out/discount posts whether than entire blogs.
What Next?
Adversarial Classification
"Oh, like you know something the Internet doesn't know."
Adversarial Classification

• All forms of spam detection fit into a general class of classification problem
  – Classification is considered a “game” between classifier and adversary
  – Adversaries adapt to evade filters

• It’s clear that a classifier has to learn to classify new forms of spam
  – Update training sets
  – Identify new detection techniques
“New Niche”
Update Training Sets, Use Language Independent Techniques
Blogspot "1" technique

- A form of spam blog that seems to be quite common on blogger
  - Create 1 post
  - Use blog as doorway

- Filters find it difficult to detect spam without avoiding false positives
- Sandboxing technique used by blog search engines

- Technique will go out of fashion once web search engines use similar strategies
Blogspot “1” technique

SATURDAY, MARCH 10, 2007

Student Loan. All about of Student Loan

Hey, a Student Loan is far less soft than a immoral Student Loan. Ah, the Student Loan is more inconsiderate than some blank Student Loan.

Uh, some Student Loan is much more poignant than one busy Student Loan.

POSTED BY WUKUHOKOQU AT 9:38 AM - 0 COMMENTS

ABOUT ME

WUKUHOKOQU

VIEW MY COMPLETE PROFILE

LINKS

PREVIOUS POSTS

Student Loan. All about of Student Loan.

ARCHIVES

March 2007

Blogger
Non-English Splogs

• Are splogs appearing in Non-English?
  – Not yet?
  – Seen many cases of spam blogs in Japanese

• Creating training examples for multiple languages not feasible

• Identifying language independent techniques that capture stereotypes in splog create tools will be important
Other Requirements

• Evaluation of effectiveness of filters
  – Generic Precision and Recall
  – Precision and Recall with a when?
• Continued QoS guarantees in an adversarial setting
  – Methodology
  – Semi-automatic techniques
• Methodology for overall maintenance of the filter
Continuing Work

Ping Stream

- Check with known blogs and splogs
- Check known regex patterns
- Use models over text of URL
- Use models over Blog home-page
- Use models over post correlation

PROCESS LOGS

COMMUNITY INPUT

CONTINUING WORK
Conclusion
Conclusion

• Trends suggest that the problem of splogs will continue
• Research needs to continue identifying new detection techniques
• The community has to incorporate collaborative techniques for detection
• Research needs to understand filter evolution in a more principled way
Resources

• http://icwsm-blog-spam.blogspot.com/
  – For (too) late breaking resources for this tutorial
• http://planet.socialMediaResearch.org/
  – A feed aggregator for blogs relevant to social media
  – Suggest new blogs to planet-smr@cs.umbc.edu
• http://sus.picious.info/
  – Prototype spam reporting service
• http://ebiquity.umbc.edu/memeta/resources/
  – Blog and splog datasets, bibliography, etc.
“I still don’t have all the answers, but I’m beginning to ask the right questions.”