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>Traffic per Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>(June)</td>
<td>30 billion</td>
</tr>
<tr>
<td>2006</td>
<td>(June)</td>
<td>55 billion</td>
</tr>
<tr>
<td>2006</td>
<td>(December)</td>
<td>85 billion</td>
</tr>
<tr>
<td>2007</td>
<td>(February)</td>
<td>90 billion</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

(Form)
(Mechanisms)

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 girl Bullet Loan Consolidation 1 girl Bullet Going Back to School ... "Can I pay my defaulted student loan held by the Department by credit card? ...
www.ed.gov/offices/OSFAP/DCS/repaying.html - 25k - Cached - Similar pages - Note this

Loan Consolidation
US Department of Educations Collections for Defaulted Student loans Guide - Information about loan consolidation options that are available to borrowers.
www.ed.gov/offices/OSFAP/DCS/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 1-800-256-5105. StudentLoanSolutions offers some of the best borrower...
www.studentloansolutions.com - 9k - Cached - Similar pages - Note this

student_loan_consolidation_program's bookmarks tagged with ... - 2:20pm
Are you looking for defaulted student loan consolidation? Please click here for more information about defaulted student loan consolidation/consolidation ...
del.icio.us/student_loan_consolidation_program/defaulted-student-loan-consolidation - 20k - 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.net/Default.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/adults_repayment.html - 20k - Cached - Similar pages - Note this

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

Default Student Loan Consolidation Testimonials
Facing Loan Default
Department of Education Collections for Defaulted Student loans Guide - Information about what to do if you are facing loan default ...
Loan Consolidation
Going Back to School
www.ed.gov/offices/OSFAP/DCS/ default.html

Repaying Student Loans Held by the U.S.
Department of Education
Addressing Your Defaulted Student Loan. If you default on your student loan, the maturity date of each ...
Federal Family Education Loan (FFEL)
loan consolidation program or the ...
www.ed.gov/offices/OSFAP/DCS/ repaying.html

consolidation defaulted loan private student
consolidation defaulted loan private student.
consolidation loan private student, loan many private ...
consolidation defaulted loan private student: to be marketed by E-LOAN. "E-LOAN is a highly ...
hometown.aol.com/gampxasf/consolidation-defaulted-loan-private-student.html

Private student loan consolidation
student loan consolidation rule consolidation
defaulted student loan. What's Iowa student loan consolidation, consolidation legislation loan student.
www.hometown.aol.com/studentloanghj/private-student-loan-consolidation.htm
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)
“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.

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 Vu.
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 surfthemind. Coverage of the results of tag and ping spam, and two new web 2.0 sites of... www.blogmarks.net/tags/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 [rafer],

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</th>
<th>Time</th>
<th>Profile</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 8, 2007 3:10 PM</td>
<td></td>
<td>Keep Drea..</td>
<td><strong>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</strong></td>
</tr>
<tr>
<td>Mar 6, 2007 9:49 AM</td>
<td></td>
<td>Alexander..</td>
<td><strong>FREE COLLEGE SCHOLARSHIPS EVENT GIVEAWAY- CLICK HE would like to invite you to join the group FREE COLLEGE SCHOLARSHIPS EVENT GIVEAWAY- CLICK HERE</strong></td>
</tr>
</tbody>
</table>
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

---

From: Misti Kulkarni
To: TSEC Alumni across the GLOBE
Sent: January 31, 2007 12:29 AM
Subject: [ignore this person]

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
High Paying Keywords on Adsense to Make Money
Google Adsense is definitely a high paying money making advertisement program on the web. But the money that you make...
www.quickonlinetips.com/archives/2005/02/high-paying-keywords-on-adsense-to-make-money/ - 38k - Cached - Similar pages - Note this

How to make money on your news content website
Can independent online news sites make money... and a difference?... curious about a specific keyword, who are more likely to click on a targeted ad... www.ojr.org/ojr/wiki/make_money/ - 27k - Cached - Similar pages - Note this

Google AdSense - an excellent way to make money free online
It's a new ebook that shows step-by-step how to build websites focused to make money online from adverts - gather keywords, build pages around them...
www.buildwebsite4u.com/income/adsense.shtml - 9k - Cached - Similar pages - Note this

How To Make Money From Micro Stock Photography | DIYPhotography.net
You can edit the keywords in the micro stock agency site, or you can embed the... finally a way to make money so I can buy "half life 2:episode 1"...
www.diyphotography.net/how-to-make-money-from-micro-stock-photography - 21k - Cached - Similar pages - Note this

Identify Profitable Contexts
Top or highly targeted
www.bizvest.com
Super Ads
Take those ads and put them not to make money...
www.aqon.com/enough-bullshitting-lets-make-money-in-24hrs-or-less/ - 28k - Cached - Similar pages - Note this
Federal Student Loan Consolidation

StudentFirst is an industry leading Federal Student Loan Consolidation Company. StudentFirst provides cutting edge technology with quality sales and customer service to make your cost ... Student Loan Consolidation - Guaranteed lowest interest rates allowed by Government. ... Consolidating student loans Federal student loan consolidation, Sallie Mae federal student loan ... Do you have $10,000 or more in student loan debt? Yes No Are any of your student loans in default? Yes No Are you currently out of school or will you graduate within six months? Yes ... Student loan consolidation through the Academic Lending Center. Interest rates on this federal ... NOW is the time to take advantage of consolidating student loans through the Federal ... Consolidating is worth your time! You can lower your monthly payments and save thousands - I am! " Jen B. Borrowers can apply online for consolidation of their federally insured student loan debt. The Higher Education Act (HEA) provides for a loan consolidation program under both the Federal ... U. Consolidate offers the Federal Student Consolidation from the Massachusetts Educational ... Massachusetts Educational Financing Authority ... Consolidating time! You can lower your monthly payments at thousands - I am! " Jen B. Here you will find the information you need to know about Direct Consolidation Loans. Find it here, Ask Us ... Refinance your educational loans with us. Apply today online for federal student consolidation ... Refinance your student and parent loans with industry leading borrower benefits. Rates ranging ...
Browser Javascript Redirect for Users
“LEADS” to affiliates highly profitable
Anatomy of Spamdexing

- Affiliate Programs
- Context Ads
  - arbitrage
  - ads/affiliate links
- in-links
- Spam pages,
  Spam Blogs,
  Spam Comments,
  Guestbook Spam
  Wiki Spam

- spamdex

- JavaScript Redirect
  [Previously Cloaking]
- Affiliate Program Buyers

- SERP
  Search Engines
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
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

56% of all blogs are splogs!

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 **Blogger** 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
“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

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, pethadene and viagra. Some might even require therapy.
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

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. ...

Original post: Splog bait: young girls need personal injury lawyer to pay for ... by at

Google Blog Search: real estate lawyer
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?
Burger King Calories

Add by Classpop.org
Advertise on this site

Weight Loss
Feel Better. Look Better. Lose weight safely and naturally.
dietfame.com

Quick Weight Loss Program
Weight loss made available online. Programs, snacks, herbs and more.
quickweightlosstrainingprogram.com

Glucostar for Weight Loss
glucostar.com

The Cambridge Diet Works
I lost 50lbs 4 years ago and kept it off. Grant Hyde Distributor.
www.camediated.co.uk

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 -

Non-Blog pings at weblogs.com

- No RSS Feeds
- No Dated Entry, no comments
November 2006 New No Deposit Rtg Coupon Codes

November 12, 2006 on 2:45 am | In Uncategorized | No Comments


New rtg online casino no deposit bonus bonus code new online. New casinos online casino coupon codes. Rtg no free online casino poker. Deposit free bonus code no deposit. New online casinos online. Casino coupon no deposit online casino in. Codes free online casino the 2006 new. Poker no deposit no deposit. Casino bonus feel to. No deposit online new rtg online casino.

Casino in the no deposit bonus if. 2006 new no deposit on the no deposit casino. Casino bonus feel to new bonuses free.
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 Imapet For ...
Trading Markets, CA - Nov 7, 2006Osteosarcoma

Post Stitching
• Excerpts scraped from other sources

responses to this entry through the RSS 2.0 feed. You can leave a response, or trackback from your own site.
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 footballers' homes. The Allos group, headed by an Iraqi developer Mayod Allos, has constructed houses for Jamie Redknapp, Graeme Le Saux and the Netherlands' Ronald de Boer.

**Links**

- herfirstlesbiansex.com
- her first lesbian sex
- Lesbian sex
- her first lesbian sex com
- her first lesbian sex video
- lesbian sex
- free lesbian sex video
- lesbian sex picture
- lesbian bisexual
- lesbian sex chat
- lesbian sex tip
- lesbian sex video clip
- clip lesbian sex video
- first time lesbian sex story
- ebony lesbian sex
- lesbian cheerleader having sex
- blonde lesbian sex
- free lesbian sex picture
- first her lesbian sex
- lesbian sex video
- lesbian sex story
- Lesbian sex movie
- dollarforporn
- Teen movies
- porn teen
- lesbian having sex
- hardcore lesbian sex

**Link-roll spam**

- With fully plagiarized text
Splog Experiment

- Captcha Breaker
- Computer
- Content
- Affiliate Accounts
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

<table>
<thead>
<tr>
<th>Time</th>
<th>Posts</th>
<th>Friends</th>
<th>time</th>
<th>time</th>
<th>time</th>
<th>time</th>
</tr>
</thead>
</table>

- E-mail Spam
- Web Spam
- Spam Blogs
- Social Network Spam
Constraints on the Filter

Who actively uses it?

<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></td>
</tr>
<tr>
<td>Batch Detection</td>
<td>Mostly Offline</td>
<td>Blog Hosting Services</td>
<td>(Ping Servers)</td>
</tr>
<tr>
<td>Fast Detection</td>
<td>Low Overhead</td>
<td>Fast Detection</td>
<td>Low Overhead</td>
</tr>
<tr>
<td>Batch Detection</td>
<td>Mostly Offline</td>
<td>Fast Detection</td>
<td>Low Overhead</td>
</tr>
<tr>
<td>Fast Detection</td>
<td>Batch Detection</td>
<td>Fast Detection</td>
<td>Batch Detection</td>
</tr>
</tbody>
</table>
# 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>
<tbody>
<tr>
<td>• Image Spam</td>
<td>• Scripts</td>
<td>• Scripts</td>
<td>• Deceptive Behavior</td>
</tr>
<tr>
<td>• Character Replacement</td>
<td>• Doorways</td>
<td>• Doorways</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Deceptive Behavior</td>
<td></td>
</tr>
</tbody>
</table>

![Diagram](image)
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
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.
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
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%

we  weblog
what motion
was  me
my  thank
org  go
flickr january
paper trackback
words archives

find perfect
info products
news uncategorized
your hot
another resources
website inc
best three
articles copyright
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
Ping times – Italian Blogs

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 does 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 insures 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.”

Listing Spam blogs, LinkFarms, "just for ads!" blogs, "Click my adsense" blogs.

Email autoresponder | Furniture Removals and Backloading in Australia | Image upload | Polls Generator

FIGHTING SPLOG
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
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
NEC – Blog Analytics

- 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 WUKUHKOQOQI AT 9:38 AM - 0 COMMENTS
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

Feedback

COMMUNITY INPUT

LEARNER

BLACKLISTS
WHITELISTS
REGEX
URL MODEL
TEXT MODEL
RSS MODEL
suspicious? bookmarks watchlist add a bookmark

my bookmarks

Search this user's bookmarks for

lillolita - MyBlogLog
16-03-2007 to mybloglog, spam - Edit - Delete

www.myspace.com/lillolita
16-03-2007 to myspace, spam - Edit - Delete

www.myspace.com/168237013
15-03-2007 to myspace, spam - Edit - Delete

orkut - Alyssé's Profile
16-03-2007 to orkut, spam - Edit - Delete

www.myspace.com/148642660
15-03-2007 to myspace, spam - Edit - Delete

www.myspace.com/163650214
04-03-2007 to myspace, spam - Edit - Delete

Religious Articles » What makes a Wikipedia article good?
01-03-2007 to plagiarized, splog - Edit - Delete

Web Messenger » Solvent helps extract data on web pages and materialize it as RDF
01-03-2007 to plagiarized, splog - Edit - Delete

pranam kolari

Profile →
Tags →
Watchlist →

popular tags

dating   delicious   mybloglog
myspace
plagiarized
spam splog   tag technorati
user weblogs.com

rss
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.”