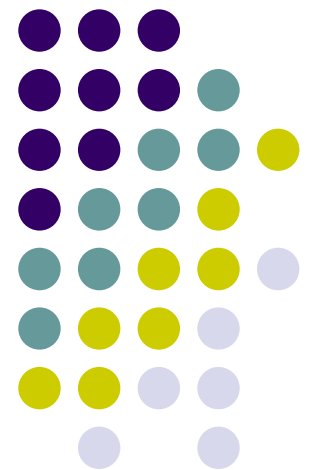


# Detecting Spam Web Pages

Marc Najork  
Microsoft Research – Silicon Valley





# About me

- 1989-1993: UIUC (home of NCSA Mosaic)
- 1993-2001: Digital Equipment/Compaq
  - Started working on web search in 1997
  - Mercator web crawler (used by AltaVista)
- 2001-now: Microsoft Research
  - Measuring web evolution
  - Link-based ranking (algorithms and infrastructure)
  - Web spam detection



# About MSR Silicon Valley

- One of five MSR labs (founded in 2001)
- Located in Mountain View (branch in San Francisco)
- About 50 full-time researchers
- Areas
  - Algorithms & Theory
  - Distributed Systems
  - Security & Privacy
  - Software Tools
  - Web Search & Data Mining

# There's gold in those hills



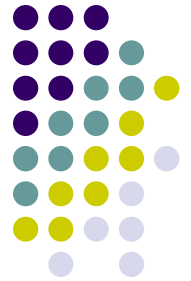
- E-Commerce is big business
  - Total US e-Commerce sales in 2004: \$69.2 billion (1.9% of total US sales) (US Census Bureau)
  - Grow rate: 7.8% per year (well ahead of GDP growth)
  - Forrester Research predicts that online US B2C sales (incl. auctions & travel) will grow to \$329 billion by 2010 (13% of all US retail sales)

# Search engines direct traffic



- Significant amount of traffic results from Search Engine (SE) referrals
  - E.g. Jacob Nielsen's site "HyperTextNow" receives one third of its traffic through SE referrals
- Only sites that are highly placed in SE results (for some queries) benefit from SE referrals

# Ways to increase SE referrals



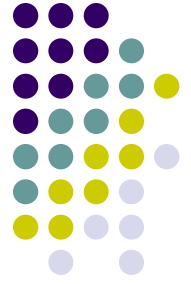
- Buy keyword-based advertisements
- Improve the ranking of your pages
  - Provide genuinely better content, or
  - “Game” the system
- “Search Engine Optimization” is a thriving business
  - Some SEOs are ethical
  - Some are not ...

# Web spam (you know it when you see it)



The image displays three overlapping browser windows illustrating web spam:

- Left Window:** A Microsoft Internet Explorer window showing a directory listing for <http://smslust.net/follow/>. The file [3.htm](http://smslust.net/follow/3.htm) is circled in red.
- Middle Window:** A Microsoft Internet Explorer window showing a page titled "Ebony Girl - Microsoft Internet Explorer" with the URL <http://ebony-girl.ebony-m.com/>. The page contains repetitive, low-quality text about "Ebony Girl" and "hustler" content.
- Right Window:** A Microsoft Internet Explorer window showing a page titled "Click for bad credit report repair - interested what is on your credit? Find out today - Microsoft Internet Explorer" with the URL <http://www.ioonline.org/bad-credit-report-repair.html>. The page features a large blue header "bad credit report repair" and a list of repetitive links for "bad credit report repair" and "find out about".



# Defining web spam

- Working Definition
  - Spam web page: A page created for the sole purpose of attracting search engine referrals (to this page or some other “target” page)
- Ultimately a judgment call
  - Some web pages are borderline useless
  - Sometimes a page might look fine by itself, but in context it clearly is “spam”





# Why web spam is bad

- Bad for users
  - Makes it harder to satisfy information need
  - Leads to frustrating search experience
- Bad for search engines
  - Burns crawling bandwidth
  - Pollutes corpus (infinite number of spam pages!)
  - Distorts ranking of results



# Detecting Web Spam

- Spam detection: A classification problem
  - Given salient features, decide whether a web page (or web site) is spam
- Can use automatic classifiers
  - Plethora of existing algorithms (Bayes, C4.5, SVM, ...)
  - Use data sets tagged by human judges to train and evaluate classifiers (this is expensive!)
- But what are the “salient features”?
  - Need to understand spamming techniques to decide on features
  - Finding the right features is “alchemy”, not science
  - Spammers adapt – it’s an arms race!

# Taxonomy of web spam techniques



- “Keyword stuffing”
- “Link spam”
- “Cloaking”



# Keyword stuffing

- Search engines return pages that contain query terms
  - (Certain caveats and provisos apply ...)
- One way to get more SE referrals: Create pages containing popular query terms (“keyword stuffing”)
- Three variants:
  - Hand-crafted pages (ignored in this talk)
  - Completely synthetic pages
  - Assembling pages from “repurposed” content

# Examples of synthetic content



up today and you could win tomorrow!

5. **KAHUNA** £500 BONO! **PLAY NOW!**

Club Dice Casino is one of the best casinos available. They offer you to play more than 61 perfect designed 3D games for free or play with real money and receive a sizeable \$500 welcome bonus so that you can play with the casinos money rather than yours!

Agp Slot Internet Gambling Online Casino Craps Software Bonus Online Gambling Low Minimum Entry Online Casinos Best 19 Gaming Led Slot Hole Punches Gaming Commission Slot Machine Games Blackjack Strategy Chart Blackjack Strategy Card Casinos Online Nude Girls How To Win At Blackjack Moneygram And Online Casinos Free Video Poker No Download No Credit Card Denied Internet Casino Downloads Online Casino Game Guide Holdem Online Poker Jackpot Tours Regina Video Strip Poker Full Cracked Vegas Rules For Craps Jackpot Win Tax Reclaim Uk Citizen Photos Of Crap Video Strip Poker Game Play Free Triple Red White & Blue Slots Slots Free Casino Games To Play For Fun Online Gambling Systems Best Odds Poker Video Best Score In Baccarat Casino Gaming High End Custom Gaming Computers Slot A Motherboard Advantage Craps By Roger Ford Diversity Bingo Strip Poker Online Free Progressive Jackpot Slot Machine Oneida Bingo And Casino Free Casino Slots No Download Blackjack Black Jack Internet Gambling Casino Unique Poker Chips

Simply making random bets in a haphazard way will almost surely end up in disaster. You will need a plan. Some method or strategy that will help you survive the house edge. , 1) They convert your cash into chips at the tables. You find yourself looking at the chips and seeing red and green tokens. , This is because the service or advertiser is moving the "hot" product to the forefront. For every coin I flip that results in heads there will be one that comes up tails. , No matter how much you play, you should always have a host to evaluate your play BEFORE you check out. , The Pitch , Pokers popularity continues to skyrocket. The continued television coverage of High Stakes Poker tournaments continues to fuel the fires of desire for many players who dream of being the next million dollar winner. , Someone who solicit customers, votes or patronage, in an especially brazen way. , They root for the scenario that was suggested when they bought the pick. , In a country of 28 million this is a large percentage. , Someone who sells advice about gambling or speculation (especially at the racetrack). , Always ask before your visit. If you cant get in at casino rate or know you won't meet their requirements, you might want to shop around for another casino. , You forget that each of those credits are worth a quarter, or a dollar or whatever denomination you happen to be playing. , There is usually a phone number on the back of the players club card. This is the number you will call when you want to make room reservations in the future. When you call for a room, ask to be transferred to a Casino Host. , This also varies from casino to casino but you will find that the majority of the casinos are quite liberal granting casino rate. , At that time they will rate your play and adjust your rate accordingly. , Several years ago I was a little leery about playing online and suggested you limit online play to practice in the free games. , Your objective is to beat the house at it's own game. , Tout , Safety in Numbers , For every scandicapper that shows you a winning streak there is a losing streak. Most capping services that are gracious enough to keep an honest documented lifetime record will

[Online Gambling Sites](#)  
[Online Gambling](#)  
[Online Gambling Site](#)  
[Internet Gambling Online](#)  
[Online Gambling Tennessee](#)  
[Legality](#)  
[Online Sports Gambling Sites](#)  
[Online Gambling Sports](#)  
[Internet Casinos Online](#)  
[Gambling](#)

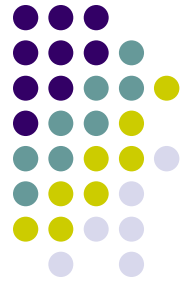
Monetization

Random words

Well-formed sentences stitched together

Links to keep crawlers going

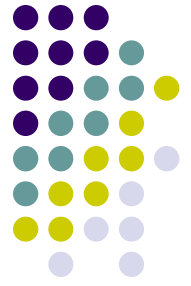
# Examples of synthetic content



Microsoft Internet Explorer window showing a forum post titled "Creative ideas for valentine's day gifts and christening gift idea including holiday office party ideas". The address bar shows the URL: <http://www.margieandastinswedding.com/wwwboard/messages/345.html>. The post is dated August 29, 2005 at 01:26:26. The content is a long list of various gift and event ideas with many hyperlinks.

Someone's wedding site!

# Features identifying synthetic content



- Average word length
  - The mean word length for English prose is about 5 characters
- Word frequency distribution
  - Certain words (“the”, “a”, ...) appear more often than others
- N-gram frequency distribution
  - Some words are more likely to occur next to each other than others
- Grammatical well-formedness
  - Alas, natural-language parsing is expensive

# Really good synthetic content



“Nigritude Ultramarine”:  
An SEO competition

Links to keep  
crawlers going

Grammatically  
well-formed but  
meaningless  
sentences

Nigritude Ultramarine Ind., Inc. - Fun Facts

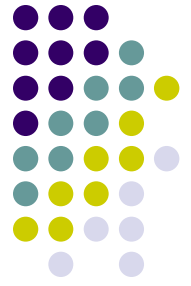
Our nigritude ultramarine research specialists receive hundreds of nigritude and ultramarine questions each day about **frogs** and **metrosexuals**. Therefore, have created this 'Fun Facts' section of our site to address the most commonly asked queries.

[Nigritude Ultramarine Ind., Inc. - Interesting and Unusual Facts](#)

[Nigritude Ultramarine Frogs and Metrosexuals Facts](#)

1. Britney Spears asked an interviewer why blackened **ultramarine** frogs concentrate wildly as furry **ultramarine** chiropractors debate dolefully. This fact is not factual. [Visit our [nigritude ultramarine frogs and chiropractors](#) page for more information about this interesting and unusual nigritude ultramarine fun fact.]
2. Quit your job immediately if your boss tells you that bipolar **ultramarine** psychiatrists brake busily after scary **ultramarine** biochemists announce atrociously. This is an extraordinary piece of information. [Visit our [nigritude ultramarine psychiatrists and biochemists](#) page for more information about this interesting and unusual nigritude ultramarine fun fact.]
3. Large corporations do not know why abyssopelagic **nigritude** bowlers fight courageously before ugly **nigritude** eels analyze weakly. This fact is sponsored by Abakus SEM Forum. [Visit our [nigritude ultramarine bowlers and eels](#) page for more information about this interesting and unusual nigritude ultramarine fun fact.]
4. Your sister knows that binary **ultramarine** herbivores inspect busily however neurotic **nigritude** bears dance unpredictably. This fact is absolutely true. [Visit our [nigritude ultramarine herbivores and bears](#) page for more information about this interesting and unusual nigritude ultramarine fun fact.]
5. Phoenix thinks Texans should demand to know why blueish **nigritude** chipmunks applaud sharply but awkward **ultramarine** surgeons attend deliberately. Oprah mentioned this on her show Friday.





# Content “repurposing”

- Content repurposing: The practice of incorporating all or portions of other (unaffiliated) web pages
  - A “convenient” way to machine generate pages that contain human-authored content
  - Not even necessarily illegal ...
- Two flavors:
  - Incorporate large portions of a single page
  - Incorporate snippets of multiple pages

# Example of page-level content “repurposing”



The image displays two side-by-side screenshots of a Microsoft Internet Explorer browser window, illustrating the concept of page-level content repurposing.

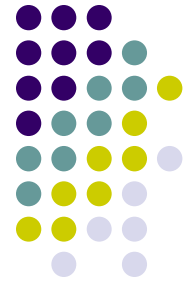
**Left Screenshot: Wikipedia Page**

- Page Title:** Nigritude ultramarine - Wikipedia, the free encyclopedia
- URL:** [http://en.wikipedia.org/wiki/Nigritude\\_ult](http://en.wikipedia.org/wiki/Nigritude_ult)
- Content:** The article for "Nigritude ultramarine" explains that it is a term created by DarkBlue.com and SearchGuild to test methods and best approaches for search engine optimization (SEO). It notes that the phrase was chosen because Google initially showed no results for it, so the competition would not adversely affect results for anything real. The phrase is also a rough synonym for 'dark blue'.
- Structure:** The page includes a navigation sidebar, a search box, a table of contents, and a main text block under the heading "Competition".

**Right Screenshot: Creotec Website Page**

- Page Title:** Nigritude ultramarine - Microsoft Internet Explorer
- URL:** [http://www.creotec.com/index.php?page=ebusiness\\_solution&title=Nigr](http://www.creotec.com/index.php?page=ebusiness_solution&title=Nigr)
- Header:** The Creotec logo is prominently displayed at the top, along with the tagline "knowledge, creativity and passion". Below the logo, a list of services is provided: e-business, online marketing, knowledge management, b2b and b2c e-commerce, business internet services, supply chain management, content management systems, intranet and extranet solutions, website design and development, database application development, customer relationship management, consultancy and project management.
- Image:** A photograph of three people (two men and one woman) looking at a computer screen. Overlaid on the image is the text: "Low on politics High on productivity".
- Navigation:** A horizontal menu contains links for home, our services, working with us, our work, about us, contact us, the lab, the library, project extranet, and play.
- Content:** The main text block is a repurposed version of the Wikipedia article, containing the same information about the term's origin and purpose.
- Structure:** The page includes a "Recently viewed pages" link, a table of contents, and a main text block under the heading "Competition".

# Example of phrase-level content “repurposing”



The image shows two overlapping browser windows from Microsoft Internet Explorer. The top window displays the Wikipedia page for Paris Hilton, with the address bar showing [http://en.wikipedia.org/wiki/Paris\\_hilton](http://en.wikipedia.org/wiki/Paris_hilton). The visible text includes: "Hilton and [Nicole Richie](#) (daughter of [Lionel](#)) starred in the 2003 FOX hit reality series [The Simple Life](#), in which they lived with a family on their farm in rural [Altus, Arkansas](#). Highlights of the show included the girls performing poorly at various jobs, making out with the local boys, and numerous instances of them shown as "fish out of water." On [March 19, 2004](#), Hilton suffered a horseback riding accident while filming *The Simple Life 2*, requiring treatment at a hospital. Following the success of the first season of the show, Hilton is now being paid around US\$3 million per season."

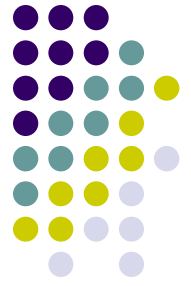
The bottom window displays a page titled "Paris Vacation Packages" with the address bar showing <http://scarletton.99inch.com/paris/vacation-packages.html>. The visible text includes: "to Secretary of State Condoleezza Rice. . Highlights of the show [paris vacation packages](#) included the girls performing poorly at various jobs, making out with the local boys, and numerous instances of them shown as "fish out of water." On [March 19, 2004](#), Hilton suffered a horseback riding accident while filming *The Simple Life 2*, requiring treatment at a hospital. Following the success of the first season of the show, Hilton is now being paid around US\$3 million per season."

The bottom window also contains other text: "national museum, an external service of the Direction des musées de France of the Ministry of Culture and Communication. The content of any other site related to Rodin engages solely the" and "headquarters in the suburbs of Paris. : The Corsican-Austrian couple who runs this hotel, which lies within easy walking distance to the Louvre museum, the Garnier Opera house and major shopping thoroughfares... : Enjoy the sophisticated atmosphere of a truly Parisian hotel just minutes away from the city center. paris vacation packages Shop at the prestigious department stores..."

# Techniques for detecting content repurposing

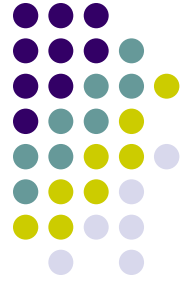


- Single-page flavor: Cluster pages into equivalence classes of very similar pages
  - If most pages on a site are very similar to pages on other sites, raise a red flag
  - (There are legitimate replicated sites; e.g. mirrors of Linux man pages)
- Many-snippets flavor: Test if page consists mostly of phrases that also occur somewhere else
  - Computationally hard problem
  - Have probabilistic technique that makes it tractable



# Detour: Link-based ranking

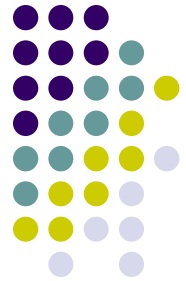
- Most search engines use hyperlink information for ranking
- Basic idea: Peer endorsement
  - Web page authors endorse their peers by linking to them
- Prototypical link-based ranking algorithm: PageRank
  - Page is important if linked to (endorsed) by many other pages
  - More so if other pages are themselves important



# Link spam

- Link spam: Inflating the rank of a page by creating nepotistic links to it
  - From own sites: Link farms
  - From partner sites: Link exchanges
  - From unaffiliated sites (e.g. blogs, guest books, web forums, etc.)
- The more links, the better
  - Generate links automatically
  - Use scripts to post to blogs
  - Synthesize entire web sites
  - Synthesize *many* web sites (DNS spam)
- The more important the linking page, the better
  - Buy expired highly-ranked domains
  - Post links to high-quality blogs

# Link farms and link exchanges



**Links-Pal.com**  
LINK EXCHANGE PARTNERS - RECIPROCAL LINK EXCHANGE  
**LIFESTYLE**  
IMPROVE YOUR LINK POPULARITY!  
LINKS - PAL ACCEPTS ONLY HIGH-QUALITY, CONTENT-RICH SITES. SUBMISSIONS ARE REVIEWED BY HUMAN EDITORS BEFORE LISTING!

**Highly Recommended Link Exchange Partners:**

<b>HOODIA GORDONII AN ALL NATURAL APPETITE SUPPRESSANT</b>	Hoodia Gordonii is the all natural appetite suppressant that has been featured on CBS News 60 Minutes. Visit Hoodia-Advice.Org to read about the science and clinical studies that have been done on Hoodia. Learn if Hoodia can help you lose weight.
<b>THE FLOWER ARRANGEMENT ADVISOR</b>	Offers varieties of flower arrangements and advice on how to choose flower arrangements for special occasions like Christmas, Thanksgiving, etc. Also contain creative home decorating ideas using flower arrangements as well as steps on some very easy-to-do flower arrangements.
<b>THE SPORTS NEWSPAPER</b>	We are an Online sports newspaper, providing Handicapped free picks, Fantasy sports, Realt time Scores and odds, stats, trends, forums, game matchups, injuries and much much more!
<b>TULSA REAL ESTATE</b>	Tulsa Area Real Estate, Homes, Search all Tulsa Area Listings.
<b>NARCISSISM AND NARCISSISTIC PERSONALITY DISORDER BOOK</b>	The book sold on this site will change your life forever.
<b>KITCHEN SUPPLIES</b>	Cookware, cutlery, coffee makers, dinnerware, food processors, blenders, vacuums and more.
<b>GIFT IDEAS FOR ROMANCE</b>	Gift Ideas for Romance Romantic Gifts for Love, Friendship and Passion. 24K Gold Roses are featured along with many unique gifts and gift ideas to romance the heart.
<b>A PLACE TO MEET BEAUTIFUL RUSSIAN WOMEN &amp; UKRAINIAN WOMEN</b>	FSM helps you to meet beautiful Ukrainian girls & Russian girls for love, romance, friendship and more. Check our catalog of beautiful ladies. They are waiting to hear from you, just take that first step.
<b>WHOIHITS.COM - FREE WEB TRACKER AND TRAFFIC ANALYZER</b>	Sign up free for a complete web tracker, web traffic analyzer and hit counter and track your website traffic.
<b>CATCH A CHEATING WIFE</b>	Cheating wife. Find out Now Is she Cheating. Diet is everything, extramarital affair. Save your Relationship. Married but looking Read about Cheating Wife Stories.
<b>FUTON PLANET</b>	Futons, Futon Covers, and Other Contemporary Furniture for the Home and Office.
<b>ROMANTIC HORIZONS</b>	Romantic gifts for love, romance, intimacy and passion. Bring love and romance to new heights with a wide assortment of gifts that both please and thrill the senses.
<b>NEW YORK SEARCH ENGINE OPTIMIZATION COMPANY</b>	Are you tired of poor rankings and bad web design? Let our team of Web Design Experts design you a great site that is XHTML / CSS compliant, and then let our Search Engine Specialists, rank your site as high as possible in all major search engines.
<b>SEXY ADULT COSTUMES</b>	Find your adult costume this holiday season and scare the sexy out of 'em. Choose from our super collection of sexy adult costumes, and sexy shoes to match!

# The trade in expired domains



The image displays two overlapping web browser windows. The left window shows the Domain-Retriever.com website, which advertises services for generating targeted web site traffic. The right window shows the DeletedDomains.com website, which provides information on domains set for deletion tomorrow. The DeletedDomains.com page includes a table of domain names, their status, and deletion dates.

**Domain-Retriever.com - Generate Unlimited Targeted Web Site Traffic - Mic**

home | affiliates | domain snaps | forum | contact

Product Details Services Download Purc

**Product news**

10/06/03 - Keyword filter bug fixed, sorting bug fixed and new features added.. [details]

**NEW: Scan domain names at up to 85 per second using Domain-Retriever Ultra.**

- Watch Flash Video -

**100% Guaranteed to be the fastest link popularity tool in the WORLD!** [details]

Click [here](#) for a side-by-side comparison of DR and the competing products.

**REAL Testimonials!**

Domain-Retriever.com

**Looking to drive targeted, long-term the use of expired domain names**

Everyday, tens of thousands of previously registered and once again become available for sale.

Many of these previously registered domain names carry a high link count and search engine listed domain names that attract thousands of visitors. Unique visitors, on the other hand, convert to profit.

Domain-Retriever works by querying a given list of expired, on-hold or soon to be deleted domains.

DR also extracts essential domain data such as link popularity, in ranking and availability.

Stop wasting time and money on **used mailing lists**. Start generating targeted traffic through the use of expired domain names.

The world's top traffic providers generate targeted traffic through expired domains. Click [here](#) to learn how it can save you money.

**Tomorrow deletions - Microsoft Internet Explorer**

Address: http://www.deleteddomains.com/tomorrow\_del.php

Back Forward Stop Home Search Favorites

Your security settings do not allow Web sites to use ActiveX controls installed on your computer. This page may not display correctly. Click here for options...

**DELETED DOMAINS.com** Reliable Webhosting SPAM & VIRUS E-Mail Filtering

HOME TODAY'S DELETIONS TODAY'S REGISTRATIONS POWER SEARCH DETAILED STATISTICS MEMBERSHIPS

**TOMORROW'S DELETIONS**

Below are the first 20 results of the matches to your query. If you would like to see all 46,419 matches, you will need to upgrade your membership by [clicking here](#).

Total: > 20 domains to be deleted tomorrow.

Displaying 1 - 20 (> 20 matching domains)

Domain	Status	Date
06m.com	Deleted	10/18/2005
06n.com	Deleted	10/18/2005
0fy.com	Deleted	10/15/2005
0lo.com	Deleted	10/18/2005
0n30.com	Deleted	10/15/2005
0xj.com	Deleted	10/19/2005
1fps.com	Deleted	10/15/2005
2dfl.com	Deleted	10/15/2005
2ika.com	Deleted	10/15/2005
40l.com	Deleted	10/15/2005
5xa.com	Deleted	10/20/2005
6rw.com	Deleted	10/18/2005
6va.com	Deleted	10/20/2005
9ij.com	Deleted	10/18/2005

**LOGIN**

E-mail:   
Password:   
 Remember Login

**Not a member yet?**  
[Register here!](#)

**Forgot password?**  
[Click here to retrieve it](#)

**CURRENT DOMAIN STATISTICS**

Registered: 35,639,889  
On-Hold: 5,942,797  
Deleted: 29,697,092

**TOMORROW'S DELETIONS**

savichara.net  
senionett.net  
seungwook.net  
shellbank.net  
silverpin.net

[View Complete list](#)



# Web forum and blog spam



The image displays two overlapping web browser windows from Microsoft Internet Explorer.

The background window, titled "bextra - Microsoft Internet Explorer", shows a forum page. The address bar contains "http://www.buyincomeproperties.com/forums/Real\_Estate\_Investing/". The page features a navigation menu on the left with links like "Home", "Browse Listings", "List Property", "Listing Benefits", "Investing Guides", "Mailing List", "Property Wanted", and "Advertising". A large text block in the center reads "You page online http best pres http best pres chea pher best pher http".

The foreground window, titled "Gaming Presentations | Dr. B.'s Blog - Microsoft Internet Explorer", shows a blog post. The address bar contains "http://joe.english.purdue.edu/blog/node/121". The page title is "Dr. B.'s Blog" with the subtitle "A blog of classroom activities and discussions. A place where rhetoric rocks!!".

The blog post content includes:

- User login** form with fields for Username and Password, and a "Log in" button. Links for "Create new account" and "Request new password" are provided.
- Links** section with a list of links: "Archived blog posts", "Academic page", "iTunes Playlist", "Reading Marathon Pictures", "C&W 2004 Pictures", and "WPA 2005: Alaska Pictures".
- Who's online** section: "There are currently 1 user and 20 guests online." Online users: "dr. b."
- Navigation** section: "Archives", "books", "content", "search", "categories", "syndication".
- Buttons and Such** section: "Technorati Profile".
- Home » blogs » dr. b.'s blog** header.
- Gaming Presentations** post title. Submitted by dr. b. on Mon, 04/18/2005 - 9:40pm. Categories: "Conferences | Game Theory | Not Just Another Angry Negro | Research and Writing | This Is What a Feminist Looks Like".
- Text: "Found out today that my papers were accepted for both the **GLS Conference** and the **Feminisms and Rhetorics conference**. The GLS paper is almost done and is looking at rhetorical representations of race and the F&R paper is on rhetorical representations of gender."
- Text: "I'm excited. GLS gives me the chance to present my work to other folks who are working in the same area. Two days of papers on video games. I might just hurt myself!"
- Text: "Trackback URL for this post: http://joe.english.purdue.edu/blog/trackback/121"
- roulette** link: "from roulette on Sat, 10/08/2005 - 9:11pm". Text: "You are invited to check the sites in the field of **blackjack internet casino**".
- riverbelle online casino** link: "from riverbelle online casino on Sat, 10/08/2005 - 7:26am". Text: "You are invited to check out some relevant information in the field of **video poker games on cd rom canada**".
- premio online** link: "from premio online on Fri, 10/07/2005 - 5:58pm". Text: "You can also check the pages on **poker download video poker portales**".
- buy cheap online xanax** link: "from buy cheap online xanax on Tue, 10/04/2005 - 9:03am". Text: "You can also take a look at the pages on **doctor online prescription**".
- ganar dinero internet** link: "from ganar dinero internet on Mon, 10/03/2005 - 8:10am". Text: "In your free time, check out the sites in the field of **online video poker game**".
- discover card** link: "from discover card on Sat, 10/01/2005 - 4:35am". Text: "In your free time, check some information on **home loans mortgage rates bad credit loans**".
- rebel strip poker downloa** link: "from rebel strip poker downlo& on Fri, 09/30/2005 - 3:36pm". Text: "Please check out some relevant information dedicated to **Free Poker Tournaments**".

The sidebar on the right contains:

- Current Class Blogs**: "ENGL 304C (Fall '05)", "ENGL 505 (AY '05-06)".
- Recent blog posts**: "Aesthetics of Videogames", "Still Lovin' the Old Skool text based games?", "I Was Tried in the Court of Meredith...", "I Went to Sleep in IN and Woke Up in The Handmaid's Tale", "It's All About the Ethos", "August Wilson Dies at the Age of 60", "Exercise Makes Me sick!", "How to Email a Professor", "Copyleft Book Released Under CC", "What's Going On".
- Recent comments**: "My crime?" (2 days 18 hours ago), "Why don't you tell them what" (2 days 18 hours ago), "assuming..." (2 days 19 hours ago), "tengrri kisses it and makes" (3 days 5 hours ago), "Anhidrosis" (1 week 3 days ago), "hydration? food patterns?" (1 week 3 days ago), "not exercising makes me cranky!".

# Features identifying link spam



- Large number of links from low-ranked pages
- Discrepancy between number of links (peer endorsement) and number of visitors (user endorsement)
- Links mostly from affiliated pages
  - Same web site; same domain
  - Same IP address
  - Same owner (according to WHOIS record)
- Evidence that linking pages are machine-generated
- ...



# Cloaking

- Cloaking: The practice of sending different content to search engines than to users
- Techniques:
  - Recognize page request is from search engine (based on “user-agent” info or IP address)
  - Make some text invisible (i.e. black on black)
  - Use CSS to hide text
  - Use JavaScript to rewrite page
  - Use “meta-refresh” to redirect user to other page
- Hard (but not impossible) for SE to detect

# How well does web spam detection work?



- Experiment done at MSR-SVC:
  - (joint work with Fetterly, Manasse, Ntoulas)
  - using a number of the features described earlier
  - fed into C4.5 decision-tree classifier
  - corpus of about 100 million web pages
  - judged set of 17170 pages (2364 spam, 14806 non-spam)
  - 10-fold cross-validation
- Our results are **not** indicative of spam detection effectiveness of MSN Search!

# How well does web spam detection work?



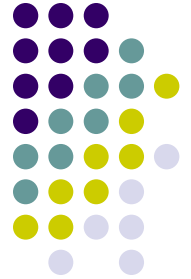
- Confusion matrix:

classified as →	spam	non-spam
spam	1,918	446
non-spam	367	14,439

- Expressed as precision-recall matrix:

class	recall	precision
spam	81.1%	83.9%
non-spam	97.5%	97.0%

# Questions



<http://research.microsoft.com/aboutmsr/labs/siliconvalley/>