

Web Spam

Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University



What is Web Spam?

- **Spamming:**
 - Any deliberate action to boost a web page's position in search engine results, incommensurate with page's real value
- **Spam:**
 - Web pages that are the result of spamming
- This is a very broad definition
 - **SEO** industry might disagree!
 - SEO = search engine optimization
- Approximately **10-15%** of web pages are spam

Web Search

- **Early search engines:**

- Crawl the Web
- Index pages by the words they contained
- Respond to search queries (lists of words) with the pages containing those words

- **Early page ranking:**

- Attempt to order pages matching a search query by “importance”
- **First search engines considered:**
 - (1) Number of times query words appeared
 - (2) Prominence of word position, e.g. title, header

First Spammers

- As people began to use search engines to find things on the Web, those with commercial interests tried to **exploit search engines** to bring people to their own site – whether they wanted to be there or not
- **Example:**
 - Shirt-seller might pretend to be about “movies”
- **Techniques for achieving high relevance/importance for a web page**

First Spammers: Term Spam

- **How do you make your page appear to be about movies?**
 - **(1)** Add the word “movie” 1,000 times to your page
 - Set text color to the background color, so only search engines would see it
 - **(2)** Or, run the query “movie” on your target search engine
 - See what page came first in the listings
 - Copy it into your page, make it “invisible”
- **These and similar techniques are term spam**

Google's Solution to Term Spam

- Believe what people say about you, rather than what you say about yourself
 - Use words in the anchor text (words that appear underlined to represent the link) and its surrounding text
- PageRank as a tool to measure the “importance” of Web pages

Why It Works?

- **Our hypothetical shirt-seller loses**
 - Saying he is about movies doesn't help, because others don't say he is about movies
 - His page isn't very important, so it won't be ranked high for shirts or movies
- **Example:**
 - Shirt-seller creates 1,000 pages, each links to his with "movie" in the anchor text
 - These pages have no links in, so they get little PageRank
 - So the shirt-seller can't beat truly important movie pages, like IMDB

Why it does not work?



Web

Results 1 - 10 of about 969,000 for [miserable failure](#). (0.06 seconds)

[Biography of President George W. Bush](#)

Biography of the president from the official White House web site.

www.whitehouse.gov/president/gwbbio.html - 29k - [Cached](#) - [Similar pages](#)

[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)

[More results from www.whitehouse.gov »](#)

[Welcome to MichaelMoore.com!](#)

Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...

www.michaelmoore.com/ - 35k - Sep 1, 2005 - [Cached](#) - [Similar pages](#)

[BBC NEWS | Americas | 'Miserable failure' links to Bush](#)

Web users manipulate a popular search engine so an unflattering description leads to the president's page.

news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - [Cached](#) - [Similar pages](#)

[Google's \(and Inktomi's\) Miserable Failure](#)

A search for **miserable failure** on Google brings up the official George W.

Bush biography from the US White House web site. Dismissed by Google as not a ...

searchenginewatch.com/sereport/article.php/3296101 - 45k - Sep 1, 2005 - [Cached](#) - [Similar pages](#)