



# Topics from CS 315 Web Search and Data Mining: Web Spamming

Lisa Ventura

Takis Metaxas, Adviser



## What is Web Spam?

Web spam is the attempt to modify the web, its structure and contents, in order to influence search engine results in ways beneficial to web spammers. Web spammers deliberately manipulate search engine indexes using unethical methods to raise a website's rank in search engine results. These spammers create web spam pages for the sole purpose of attracting search engine referrals.

## Spamming Techniques

Spammers seek to inflate their page's rank by typically altering either the content or the links on the page.

### Links

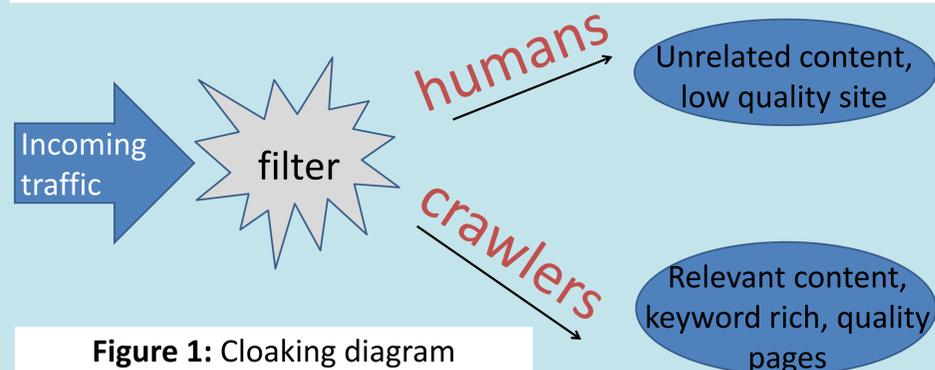
- Link farms (Figure 2)
- Hidden links

### Content

- Keyword stuffing (Figure 3)
- Invisible text
- Article spinning

### Other

- Cloaking (Figure 1)
- Page hijacking



## Why Should We Care About Spam?

We depend on search engines to provide us with credible information. Web spam undermines the reputation of a trusted information source. According to a study by Chitika, the first result is clicked in 32.5% of Google searches. If spammers are successful and elevate their rank in search results, users receive untrustworthy and potentially harmful information.

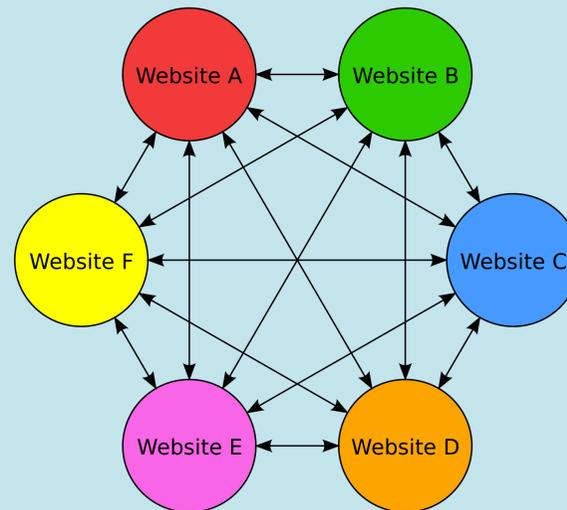


Figure 2: Link farm example diagram. Photo from MithrandirMage.

no hands seo review, download No Hands SEO, no hands seo, auto appro download, no hands seo software download, linxbot alternative, no ha LinxBot, TweeterNaire , auto seo backlink software, No hands SEO rev autobacklink bomb do, [GET] no hands seo, no hands seo backlinksforu backlinks softtware, No Hands SEO, Auto Backlink Bomb dl, nohandsseo forum, software for SEO link building, backlinks software download, No Hands SEO software, get no hands seo, auto backlink bomb index fa backlink bomb tutorials, no hand seo, no hand seo in mediafire, Down linkbot#sclient=psy, nohandseo price, LinxBot â, LinxBot â, LinxBot hand seo, LinxBot megaupload , auto backlink bomb download, top back scrapebox, no hands seo free, banned no hands seo, no hand seo softw Hands SEO software, Tweeternaire review, No Hands SEO megaupload, fr software, "no hands seo" use approve , no hands seo blogspot, no han

Figure 3: Keyword stuffing "no hands SEO" on a spam page. Photo from Matt Cutts, Google Webmaster Central Blog.

## Detecting Spam

Search engines can detect spam web pages with statistical analysis by employing the power law.

Power law: Fraction  $f(k)$  of items with popularity  $k$  is proportion to  $k^{-c}$ .

$$f(k) \approx k^{-c}$$

$$\log f(k) \approx \log(k^{-c})$$

$$\log f(k) \approx -c \log k$$

$$y \approx -c x$$

A power law is a straight line on a log-log plot.

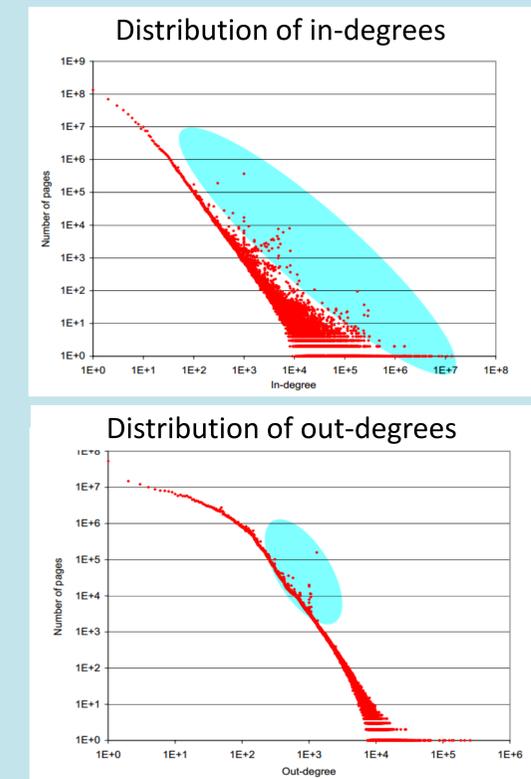


Figure 4: Both graphs illustrate the power law. The blue areas, outliers from the expected linear distribution, indicate probable link farms.

## Acknowledgements

Takis Metaxas, "What is WEB SPAM"; Marc Najork, "Detecting Spam Web Pages"; Gyongyi, Molina "Spam: It's Not Just for Inboxes Anymore"; Fetterly et al "Spam, Damn Spam, and Statistics."