Topics from CS 315 Web Search and Data Mining:
Web Spamming
Lisa Ventura

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.

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.

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

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 \( k^{-c} \).

\[
f(k) \approx k^{-c} \approx \log (k^{-c}) \approx -c \log k \approx y \approx -c x
\]

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

Figure 1: Cloaking diagram

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

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

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

Acknowledgements