What is WEB SPAM

Web Spammers target the last step

Web spam (you know it when you see it)

Defining Web Spam

- Spam web page is...
  - 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
  - Some pages look fine in isolation, but in context are clearly "spam"
Spamming Techniques

- **Boosting Rank:**
  - Term Spamming
  - Link Spamming

- **Hiding Spam:**
  - Content Hiding
  - Cloaking
  - Redirecting

Boosting Rank by Term Spamming

- Editing the textual content
  - The Search engine looks for relevant terms in various fields
  - Different fields are weighed different

Term Spam: 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
  - Completely synthetic pages
  - Assembling pages from "repurposed" content

Synthetic content for keyword stuffing

- Monetization
- Random words
- Well-formed sentences stitched together
- Links to keep crawlers going
More examples of synthetic content

Someone’s wedding site!

Really good synthetic content

“Nigritude Ultramarine”: An SEO competition

Links to keep crawlers going

Grammatically well-formed but meaningless sentences

Boosting Rank by Link Spamming

- Link structure ➔ importance
- Outgoing links
- Incoming links
- Use Directories
- Link exchange and spam farms

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
Inflate rank: Link farms, link exchanges

Inflate rank: Expired domains

Inflate rank: Web forum and blog spam

Hiding Spam

- Invisible content
- Cloaking:
  serve different page to a crawler than to a browser
- Techniques:
  - Recognize page request is from search engine (based on “user-agent” info or on IP address)
  - Make some text invisible (i.e., black on black)
  - Use CSS to hide text
  - Use JavaScript to rewrite page (dynamically created)
  - Use “meta-refresh” to redirect user to other page
Why should we care

* We depend on search engines and trust them
* Web Spam undermines the reputation of a trusted information source