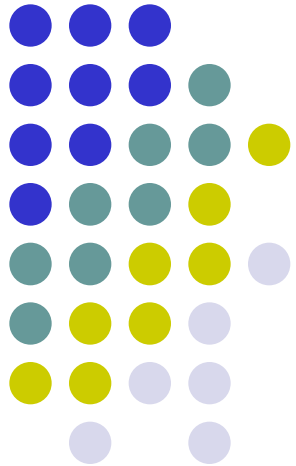
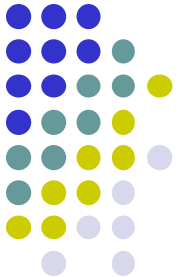
Web Algorithms – Web Search

Part 5: Spamming

Eng. Fabio Persia, PhD

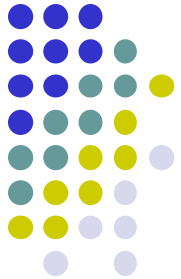


What is Web Spam?



- **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

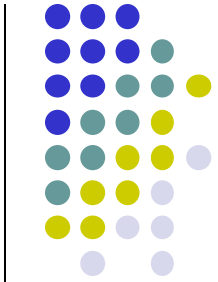
Web Search



- **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

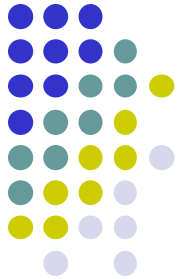


- As people began to use search engines to find things on the Web, those with commercial interests tried to **exploit search engines** to bring people to their own site – whether they wanted to be there or not

- **Example:**
  - Shirt-seller might pretend to be about “movies”

- **Techniques for achieving high relevance/importance for a web page**

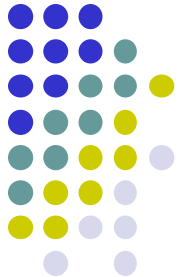
First Spammers: Term Spam



- **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**

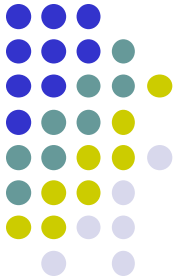
Google's Solution to Term Spam



- **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

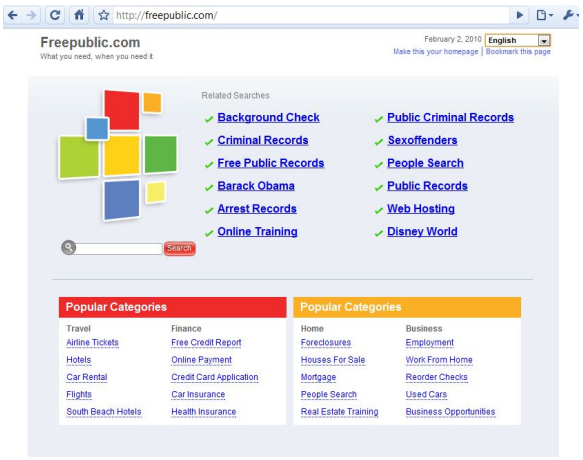
Why It Works?



- **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

Google vs. Spammers: Round 2!

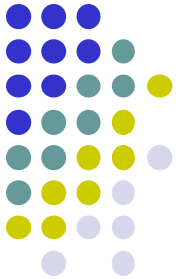
- Once Google became the dominant search engine, spammers began to work out ways to fool Google
- **Spam farms** were developed to concentrate PageRank on a single page
- **Link spam:**
  - Creating link structures that boost PageRank of a particular page



Link Spamming

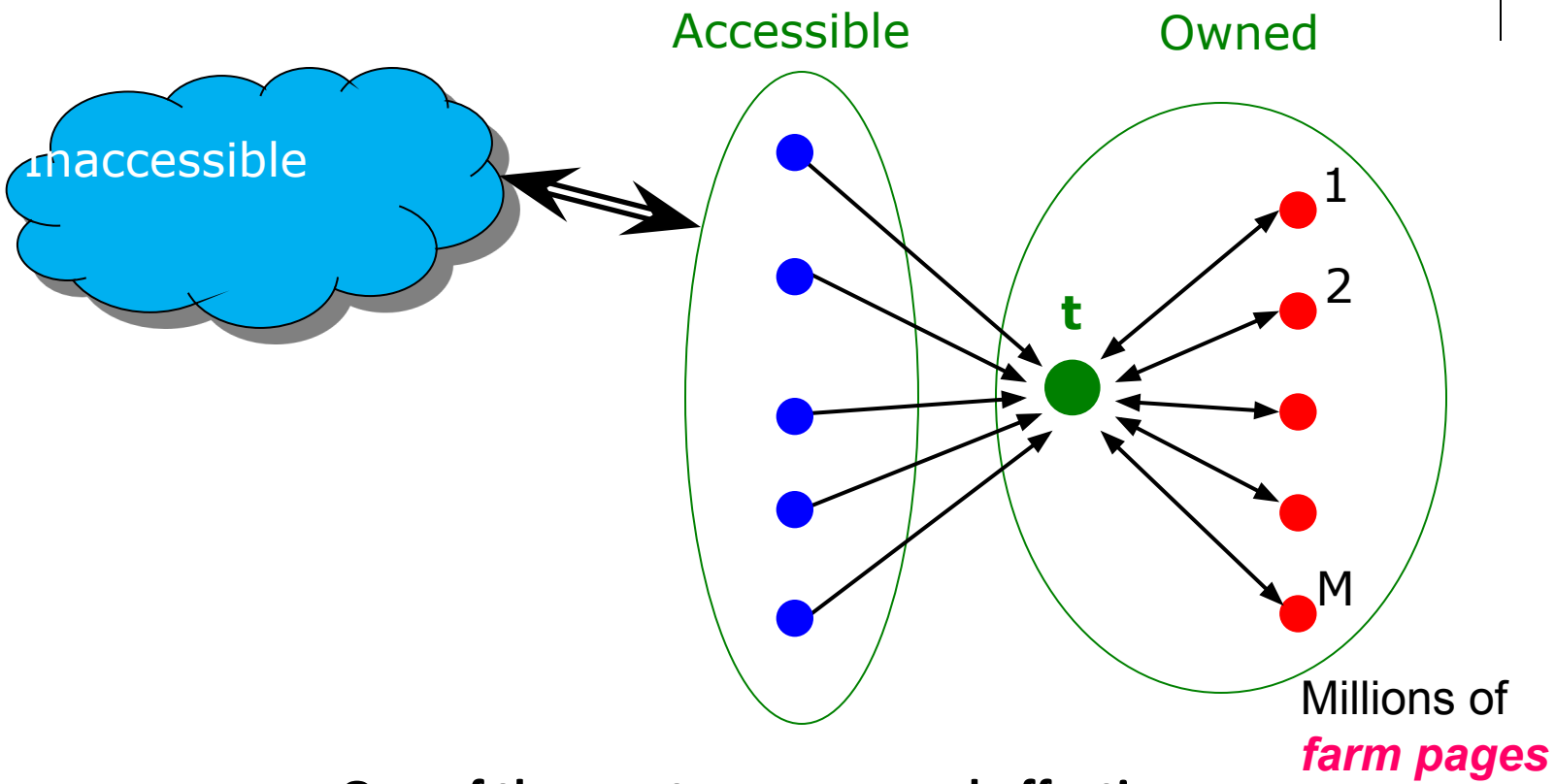
- **Three kinds of web pages from a spammer's point of view**
  - **Inaccessible pages**
  - **Accessible pages**
    - e.g., blog comments pages
    - spammer can post links to his pages
  - **Owned pages**
    - Completely controlled by spammer
    - May span multiple domain names

Link Farms



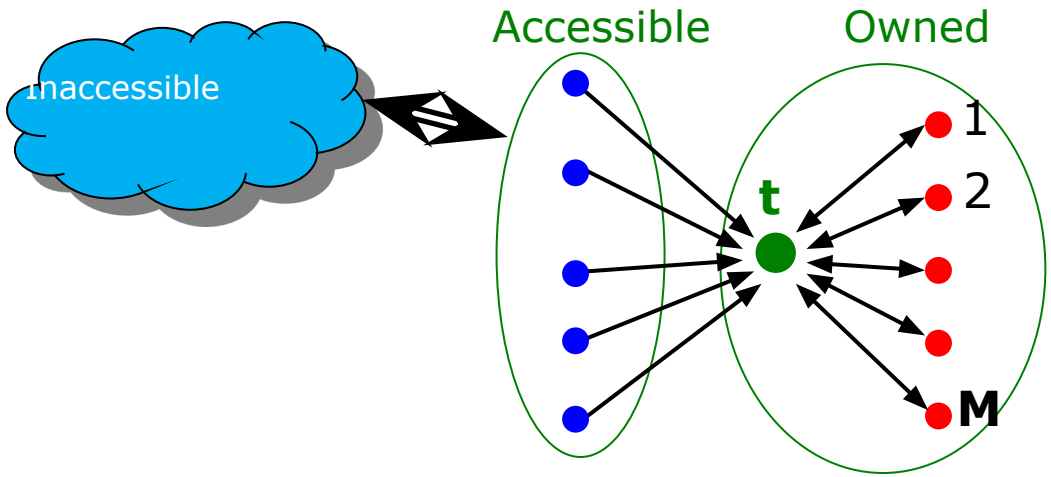
- **Spammer's goal:**
  - Maximize the PageRank of target page ***t***
- **Technique:**
  - Get as many links from accessible pages as possible to target page ***t***
  - Construct "link farm" to get PageRank multiplier effect

Link Farms



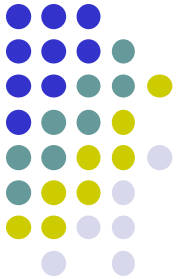
One of the most common and effective organizations for a link farm

Analysis



N…# pages on the web
M…# of pages spammer owns

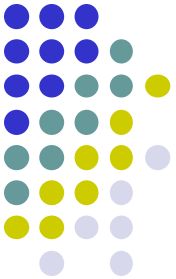
TrustRank: Combating the Web Spam



- **Combating term spam**
  - Analyze text using statistical methods
  - Similar to email spam filtering
  - Also useful: Detecting approximate duplicate pages

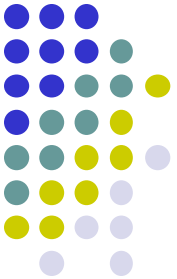
- **Combating link spam**
  - **Detection and blacklisting of structures that look like spam farms**
    - Leads to another war – hiding and detecting spam farms
  - **TrustRank** = topic-specific PageRank with a teleport set of **trusted pages**
    - Example: .edu domains, similar domains for non-US schools

TrustRank: Idea



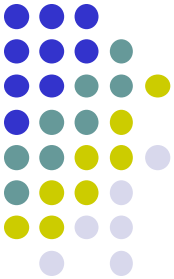
- **Basic principle: Approximate isolation**
  - It is rare for a "good" page to point to a "bad" (spam) page
- Sample a set of **seed pages** from the web
- Have an **oracle** (**human**) to identify the good pages and the spam pages in the seed set
  - **Expensive task**, so we must make seed set as small as possible

Trust Propagation



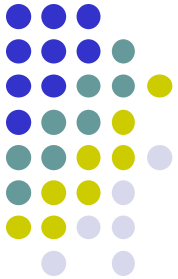
- Call the subset of seed pages that are identified as **good** the **trusted pages**
- Perform a topic-sensitive PageRank with **teleport set = trusted pages**
  - **Propagate trust through links:**
    - Each page gets a trust value between **0** and **1**
- **<u>Solution 1:</u> Use a threshold value and mark all pages below the trust threshold as spam**

Simple Model: Trust Propagation



- **Set trust of each trusted page to 1**
- Suppose trust of page ***p*** is t_p
  - Page ***p*** has a set of out-links o_p
- For each $q \in o_p$, ***p* confers the trust** to ***q***
  - $\beta\, t_p / |o_p|$ for $0 < \beta < 1$
- **Trust is additive**
  - Trust of ***p*** is the sum of the trust conferred on ***p*** by all its in-linked pages
- **Note similarity to Topic-Specific PageRank**
  - Within a scaling factor, **TrustRank = PageRank** with trusted pages as teleport set

Why is it a good idea?

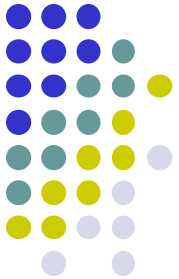


- **Trust attenuation:**
  - The degree of trust conferred by a trusted page decreases with the distance in the graph
- **Trust splitting:**
  - The larger the number of out-links from a page, the less scrutiny the page author gives each out-link
  - Trust is **split** across out-links

Picking the Seed Set

- **Two conflicting considerations:**
  - Human has to inspect each seed page, so seed set must be as small as possible
  - Must ensure every **good page** gets adequate trust rank, so need make all good pages reachable from seed set by short paths

Approaches to Picking Seed Set



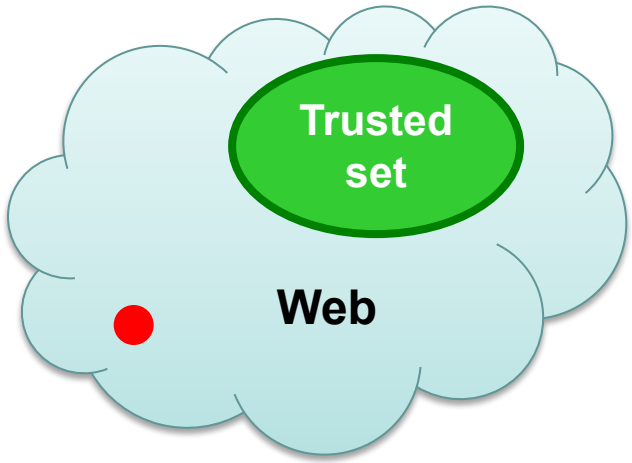
- Suppose we want to pick a seed set of ***k*** pages

- **How to do that?**
  1. **PageRank:**
     - Pick the top ***k*** pages by PageRank
     - Theory is that you can't get a bad page's rank really high
  2. **Use trusted domains** whose membership is controlled, like .edu, .mil, .gov

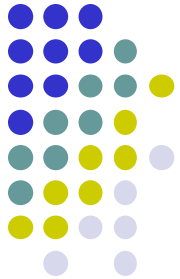
Spam Mass

- In the **TrustRank** model, we start with good pages and propagate trust
- **Complementary view:**

  What fraction of a page's PageRank comes from **spam** pages?
- In practice, we don't know all the spam pages, so we need to estimate

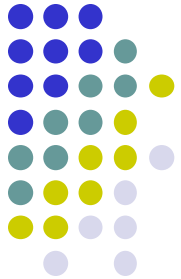


The Web graph



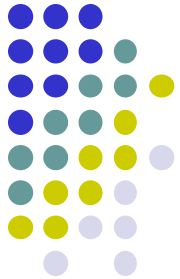
- We are interested in the graph of the web for several reasons
  - link analysis for
    - data mining (ex. PageRank)
    - determination of communities
    - sociological aspects of the creation of content
  - determination of models for
    - formal proof of ownership of the algorithms
    - determine peculiar regions of the graph
    - predict the evolution of new phenomena

The Structure of the Web



- Several studies have tried to extrapolate the size of the Web and, while still providing data for different periods of observation, all agree that the number of published papers is now several billion
- As regards the rate of growth of the Web, it was observed that its size doubles every 15 months
- more detailed and recent studies have tried to identify properties of the web graph not only quantitative, with particular reference to the connectivity features
- The interest for the connectivity is essentially linked to the possibility of using software tools to automatically explore large Web portions

The Structure of the Web

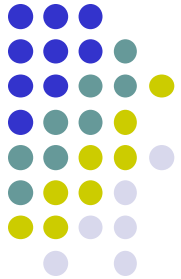


- A ***weakly connected component*** is a set of related pages so that each is accessible from all others if hyperlinks can be followed forward or backward

- A ***strongly connected component*** is a set of related pages so that for any pair of pages ***(u, v)*** there exists a directed path from ***u*** to ***v***

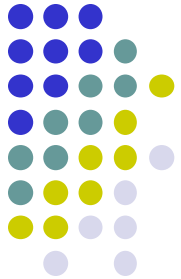
- From experiments, about 90% of the Web has proven to be a weakly-connected component

The Structure of the Web

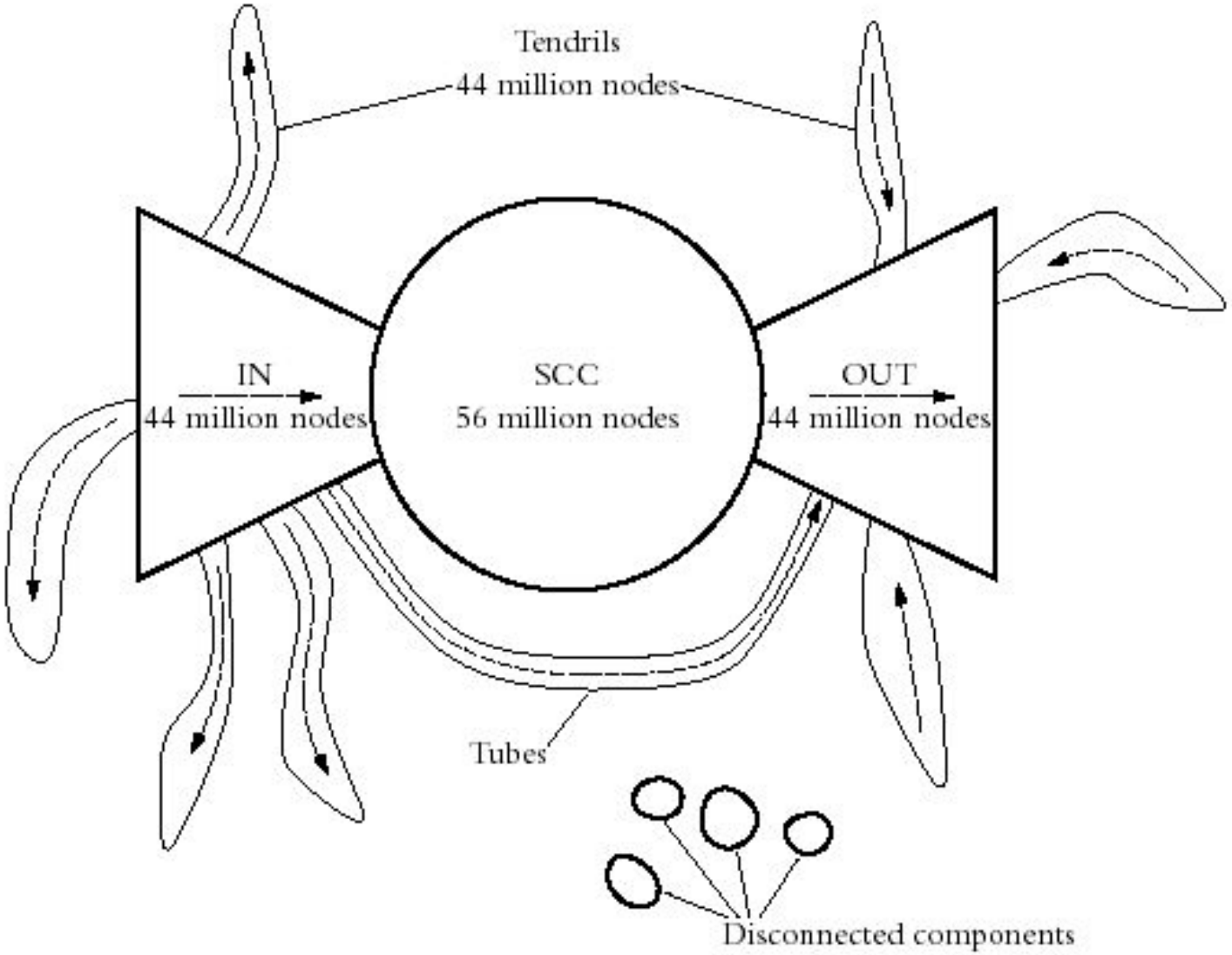


- A it is now known, by crawling experiments carried out by known academic institutions, that the structure of the Web includes:
  - a ***strongly connected central core (SCC)***
  - a ***subgraph (IN)*** with direct routes that lead SCC
  - a ***component (OUT)*** that leads to the outside of SCC
  - ***tendrils*** relatively isolated attached to one of the big three subgraphs

The Structure of the Web



- The core represents about 30% of the total and is made up portals, the search engines, from information and big companies sites
- The subgraph IN represents about 24% and consists of pages that allow to reach the nucleus but that are not reachable from it, then personal pages or smaller sites
- The subgraph OUT represents about 24% and consists of pages that can be accessed from the nucleus but which are not connected to it in the opposite direction and is formed mainly of pages for universities, companies, research centers etc.
- the remaining 22% is formed by those pages that are linked to each other but completely disconnected from the rest of the Web



Region:	SCC	IN	OUT	Tendrils	Disconnected	Total
Size:	56,463,993	43,343,168	43,166,185	43,797,944	16,777,756	203,549,046

Due to the nature and cardinality of the of the four identified regions, the model shown in the drawing has been named "Bow-tie model"