



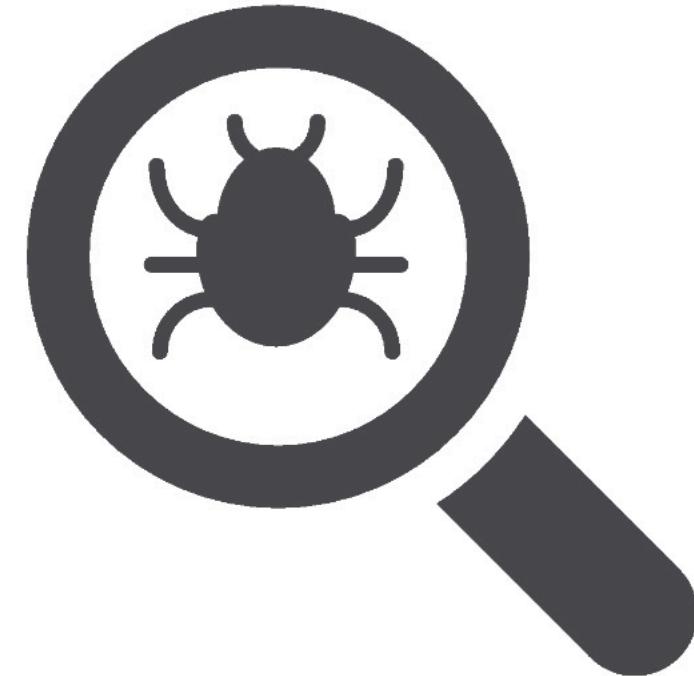
Information Retrieval

Web Information Retrieval

Danilo Montesi
Stefano Giovanni Rizzo

Web Crawling

- Web crawling is the process by which **we gather pages from the Web**
- Goal: **quickly** and efficiently gather as many useful Web pages as possible, together with the **link structure** that interconnects them

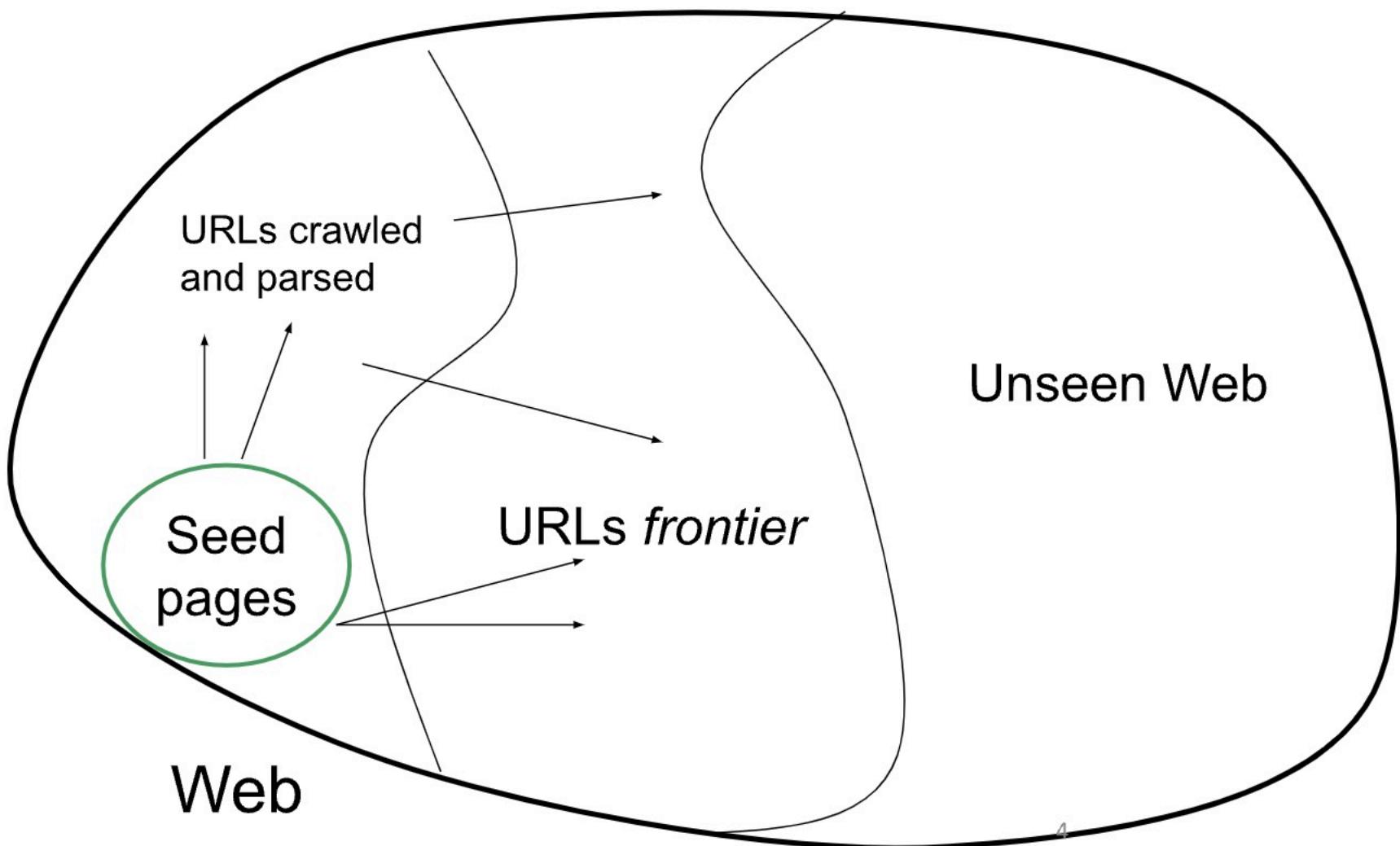




Basic crawler operation

- A crawler (a.k.a. spider)
 - Begin with known “seed” URLs
 - Fetch and parse them
 - Extract URLs they point to
 - Place the extracted URLs on a queue (the URLs *frontier*)
 - Fetch each URL on the frontier and repeat

Crawling picture





Crawler requirements

- A crawler must respect some requirements:
 - **Robustness:** MUST avoid spider traps (fetching an infinite number of pages in a particular domain)
 - **Politeness:** MUST respect Web servers policies, regulating the rate at which crawlers can visit them



Robustness

- Web crawling isn't feasible with one machine
 - All of the above steps are distributed
- Malicious pages
 - Spam pages
 - Spider traps – including dynamically generated
- Even non-malicious pages pose challenges
 - Latency/bandwidth to remote servers vary
 - Webmasters' stipulations
 - How "deep" should you crawl a site's URL hierarchy?
 - Site mirrors and duplicate pages



Politeness

- **Explicit politeness:** specifications from webmasters on what portions of site can be crawled
 - robots.txt
- **Implicit politeness:** even with no specification, avoid hitting any site too often



Robots.txt

- Protocol for giving spiders (“robots”) limited access to a website, originally from 1994
 - www.robotstxt.org
- Website announces its request on what can (or cannot) be crawled
 - For a server, create a file named robots.txt
 - This file specifies access restrictions
- Robots.txt contains set of rules that **should** be followed by clients



Robots.txt: Example

- **Example:**

- No robot should visit any URL starting with "/yoursite/temp/", except the robot called "searchengine":

User-agent: *

Disallow: /yoursite/temp/

User-agent: searchengine

Disallow:

For **all** user-agents
(client names) forbid
access to directory
/yoursite/temp

For user-agents named
“searchengine” do not
forbid nothing,
everything is thus
accessible

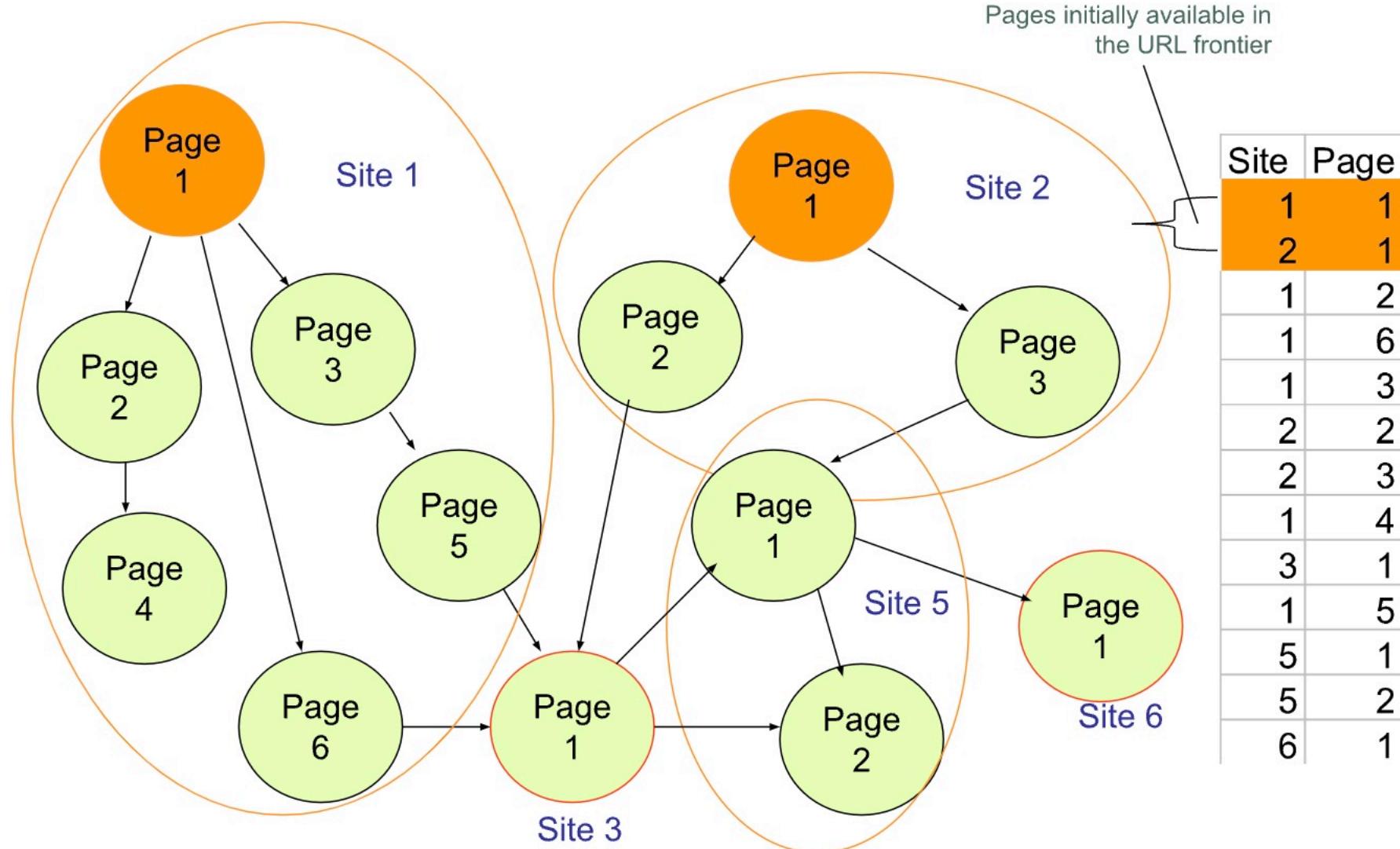


URL frontier

- Pages are **added** to the URL frontier according to the following strategies:
- **Breadth first** strategy: given a Web page in the URL frontier, add **all pages linked by the current page**. Coverage is wide but superficial.
- **Depth first** strategy: given a Web page in the URL frontier, **follow the first link in the current page** until the first page without links.

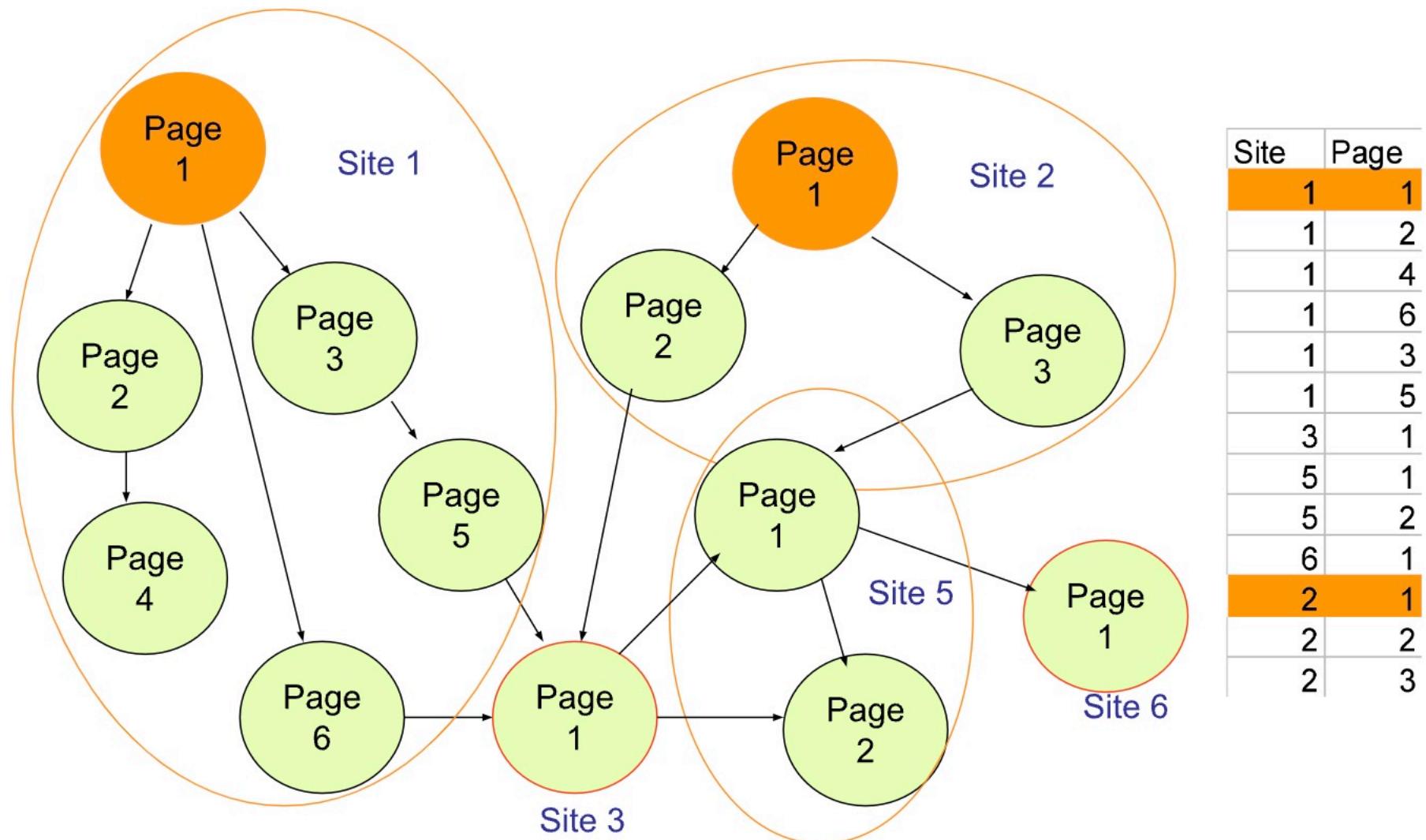


URL frontier: Breadth first strategy





URL frontier: Depth first strategy





Searching the web

- There are thousands of billions of pages on the web, but most of them are not very interesting.
- Suppose you have to visit the site for eBay and you don't know that www.ebay.com is the URL
 - There are millions of web pages that contain the term "eBay"
 - There can be websites with more frequency on the term "eBay" than eBay itself.
- We need a notion of **popularity**, together with a notion of relevance.

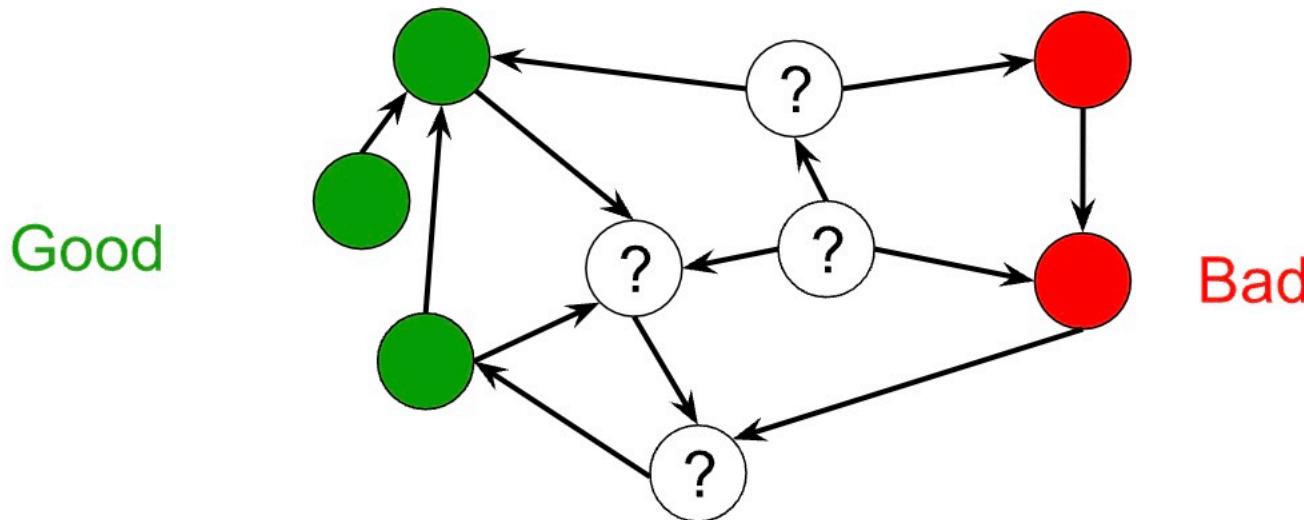


Web information retrieval

- With respect to traditional textual search engines, Web information retrieval systems build ranking by combining at least two evidences of relevance:
 - the degree of matching of a page: the **content score**
 - the degree of importance of a page: the **popularity score**
- While the content score can be calculated using one of the information retrieval models described so far, the popularity score can be calculated from an analysis of the indexed pages' **hyperlink structure** using one or more ***link analysis*** models.

Simple link analysis

- Links are powerful sources of authenticity and authority.
- The **Good**, The **Bad** and The Unknown
 - Good nodes won't point to Bad nodes
 - If you point to a Bad node, you're Bad
 - If a Good node points to you, you're Good





Citation Analysis

- Citation frequency is an estimation of a researcher popularity.
- Bibliographic coupling frequency
 - Articles that co-cite the same articles are related
- Citation indexing
 - Who is this author cited by? (Garfield 1972)
- Pagerank preview: Pinski and Narin '60s
 - Asked: which journals are authoritative?



PageRank

- The PageRank technique for link analysis assigns a **numerical score between 0 and 1** to every node in the web graph.
- The PageRank of a node depends on the **link structure** of the web graph.
- Given a query, a web search engine computes a **composite score** for each web page that combines hundreds of features such as cosine similarity, together with the PageRank score.
- This composite score is used to provide a **ranked list of results** for the query.



The random surfer (1/3)

- Consider a *random surfer* Alice who begins a random walk on the web, starting from a page.
 - Alice is extremely bored, she wanders aimlessly between web pages.
 - Her browser has a special “**surprise me**” button at the top that will jump to a random web page when clicked.
 - Each time a web page loads she chooses whether to
 - Click on a **random** link on the page
 - Click the surprise me button
 - Alice is sufficiently bored that she intends to keep browsing the Web like this forever.



The random surfer (2/3)

- To put this in a more structured form, Alice browses the Web using this algorithm:
 1. Choose a random number r between 0 and 1.
 2. If $r > \lambda$:
 - Click the “surprise me” button
 3. If $r \leq \lambda$:
 - Click a link at random on the current page.
 4. Start again.
- Because of Alice’s special “surprise me” button, we can be guaranteed that eventually she will reach every page on the Internet.

The random surfer (3/3)

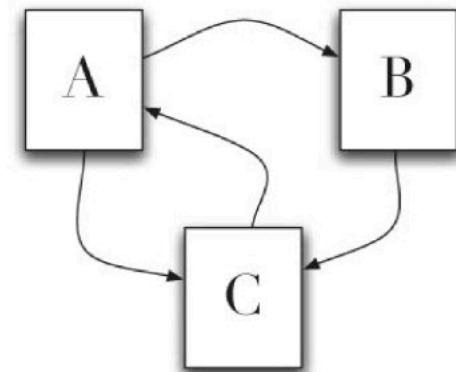
- Now suppose that while Alice is browsing, you walk in and glance at the web page on her screen. **What is the probability** that she will be looking at the eBay website?



- That probability is eBay's **PageRank**.

PageRank calculation

- The PageRank calculation corresponds to finding the stationary probability distribution of a *random walk* on the graph of the Web. A random walk is a special case of a *Markov chain* in which the next state depends solely on the current state.
- If the web consists of the 3 pages in figure (A,B,C), the PageRank of C depends on the PageRank of A and B:



$$PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1}$$

- We start by assuming that the PageRank values for all pages are the same, then we iterate the calculation. After few iterations, the PageRank values converge to the final values of
 - $PR(C)=0.4$
 - $PR(A)=0.4$
 - $PR(B)=0.2$



Use of PageRank in Google

- PageRank is now **only one of the many** factors that determine the final score of a Web page in Google.
- It is now a part of a much larger ranking system that it is believed to account for more than **200 different “signals”** (ranking variables):
 - *language* features (phrases, synonyms, spelling mistakes, etc.)
 - *query* features that relate to language features, trending terms/phrases
 - *time-related* features (e.g., “news” related queries might be best answered by recently indexed documents, while factual queries are better answered by more “resilient” pages)
 - *personalization* features, which relate to one’s search history, behavior, and social surrounding.



Hyperlink-Induced Topic Search (HITS)

- In response to a query, instead of an ordered list of pages each meeting the query, find two sets of inter-related pages:
 - *Hub pages* are good lists of links on a subject.
 - e.g., “Bob’s list of cancer-related links.”
 - *Authority pages* occur recurrently on good hubs for the subject.
- Best suited for “broad topic” queries rather than for page-finding queries.
- Gets at a broader slice of common *opinion*.



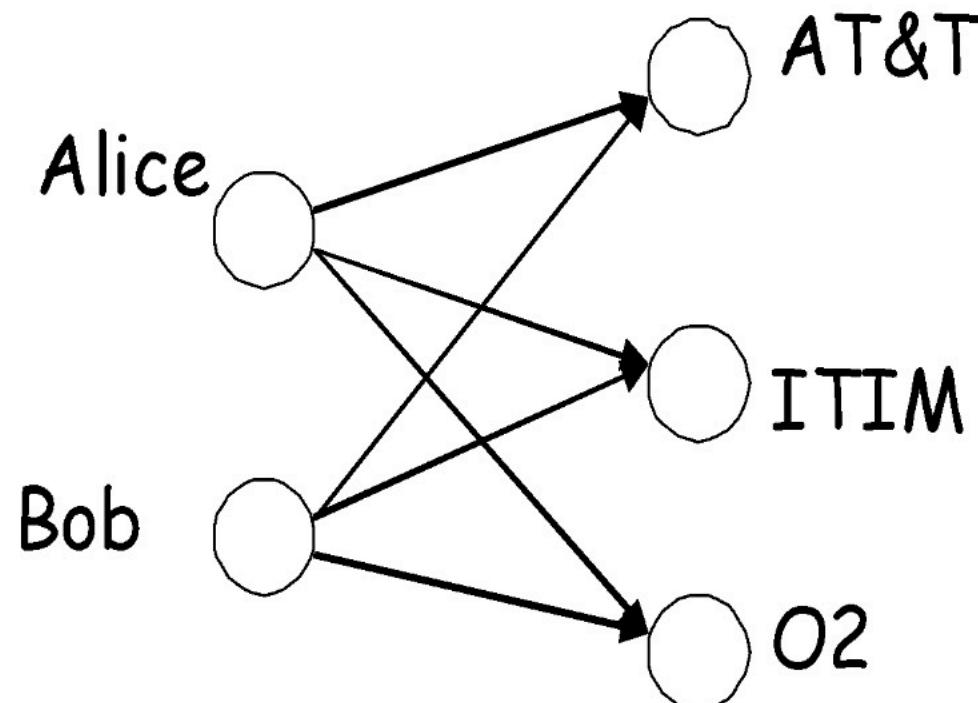
Hubs and Authorities

- Thus, a good hub page for a topic *points* to many authoritative pages for that topic.
- A good authority page for a topic is *pointed to* by many good hubs for that topic.
- Circular definition - will turn this into an iterative computation.



HITS Example

Query: “***Mobile telecom companies***”





Semantic Search

- The name “information retrieval” is standard, but as traditionally practiced, it’s not really right
- All you get is **document retrieval**, and beyond that the job is up to you
- **Semantic Search:** Doing graph search over structured knowledge rather than traditional text search:
 - Google Knowledge Graph
 - Facebook Graph Search
 - Bing’s Satori
 - Things like Wolfram Alpha





References

Bruce Croft, Donald Metzler, Trevor Strohman

Search Engines: Information Retrieval in Practice

Pearson (2010)

Christopher D. Manning, Prabhakar Raghavan and
Hinrich Schütze

Introduction to Information Retrieval

Cambridge University Press. (2008)

Stefano Ceri, Alessandro Bozzon

Web Information Retrieval

Springer. (2013)