15 – WEB INFORMATION RETRIEVAL

CS 1656

Introduction to Data Science

Alexandros Labrinidis – http://labrinidis.cs.pitt.edu
University of Pittsburgh

WEB CRAWLING

Web Information Retrieval

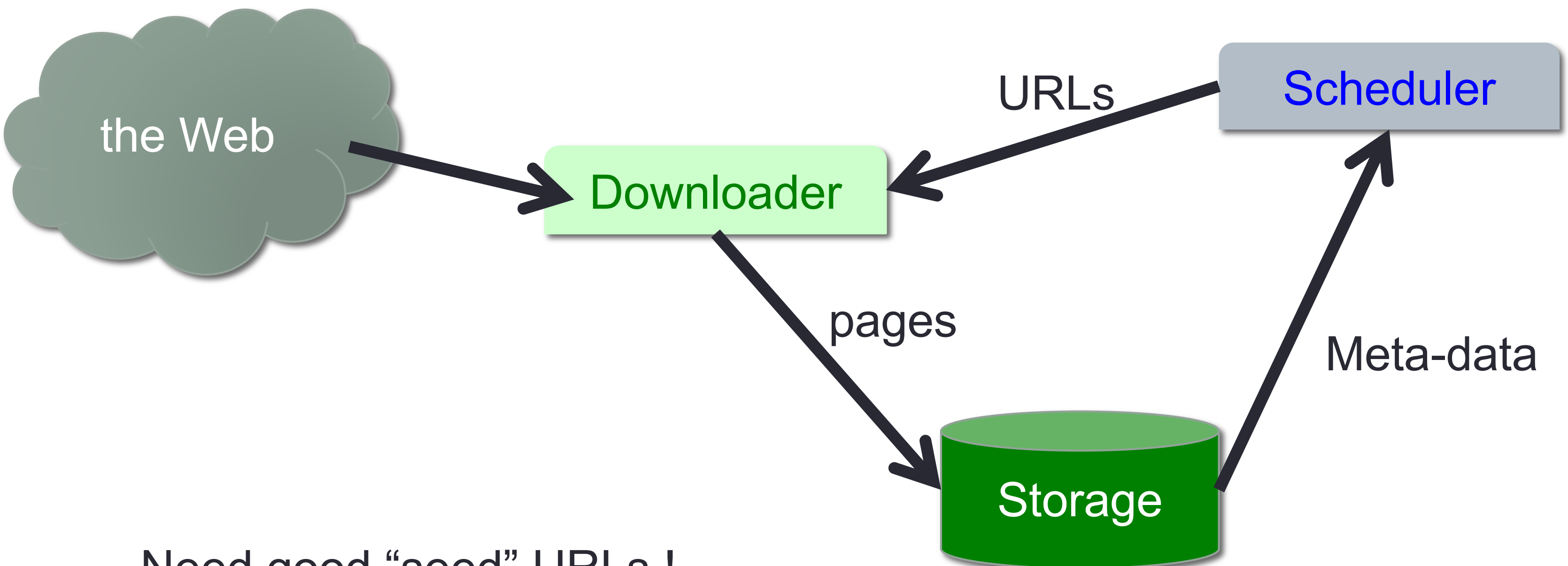
- Q) How to locate relevant web pages?

- A) Maintain a **directory**
- A1) Human-generated web directories
  - Rely on human users (experts or not experts) to submit web pages
  - Examples: Yahoo Web Directory https://dir.yahoo.com/
  - Open Directory Project http://www.dmoz.org

    http://www.dmoz.org/Society/People/Personal_Homepages/L/

- A2) Machine-generated web directories or index
  - Use web crawler to collect web pages and store locally
  - Examples: Google, Bing, Yahoo Search

Anatomy of a Web Crawler



Need good "seed" URLs !

Can we store the entire Web?

- No, but why?
- Reason #1: Storage
- Estimate for size of Web is ASTRONOMICAL
  - 45 billion pages
  - [Source: http://www.worldwidewebsize.com/]
- Solution:
  - Store metadata, instead of storing entire web page

Can we store the entire Web?

- Reason #2: Cannot access **entire** Web!

- Because not sure where all pages are
  - E.g., if a web page has no other page linking to it
  - Could be on purpose (dark net)

- Because data is access-controlled
  - E.g., password protected slides

- Because data is made available only after filling out forms
  - Also known as the Deep Web / Hidden Web
  - E.g., Pitt user directory (http://find.pitt.edu)

Mechanics of Web Crawlers

- HTTP Protocol – Request

```
telnet db.cs.pitt.edu 80
```

> Trying 136.142.50.166...
> Connected to db9.cs.pitt.edu.
> Escape character is '^]'.

```
GET /courses/cs1656/fall2016/test12.html HTTP/1.0
```

(need to hit return twice)

- HTTP Protocol – Response
  - (continued on next page)

Mechanics of Web Crawlers – II

- HTTP/1.1 200 OK ← STATUS CODE
- Date: Wed, 14 Sep 2016 03:16:12 GMT
- Server: Apache/2.2.3 (CentOS)
- Last-Modified: Wed, 14 Sep 2016 03:15:35 GMT
- ETag: "1f773c70-af-53c6f2480a3c0"
- Accept-Ranges: bytes
- Content-Length: 175
- Connection: close
- Content-Type: text/html; charset=UTF-8

- <html>
- <head> <title>This is a test file</title> </head>
- <body>
- <h1> This is a test file </h1>
- Created exclusively for the CS1656 class on September 14, 2016.
- </body> </html>
- Connection closed by foreign host.

DNS Lookup

- DNS = Domain Name Service
  - holds mapping of domain names (e.g., db.cs.pitt.edu) to IP addresses (e.g., 136.142.50.166) and vice versa

```
> nslookup 10.228.27.74
Server:         136.142.57.10
Address:        136.142.57.10#53

74.27.228.10.in-addr.arpa     name = ipsec-10-228-27-
74.vpn.pitt.edu.

> nslookup db.cs.pitt.edu
Server:         136.142.57.10
Address:        136.142.57.10#53

db.cs.pitt.edu canonical name = db9.cs.pitt.edu.
Name:   db9.cs.pitt.edu
Address: 136.142.50.166
```

Web Server Access Logs

```
> ssh elements.cs.pitt.edu
> wget
http://db.cs.pitt.edu/courses/cs1656/fall2016/test12.html
```

```
> grep test12 db.access_log
```

136.142.227.11 - - [13/Sep/2016:23:23:45 -0400] "GET /courses/cs1656/fall2016/test12.html HTTP/1.0" 200 175

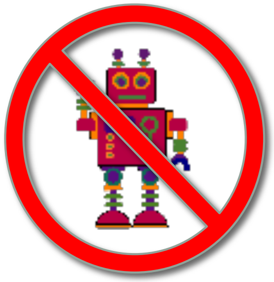
```
> nslookup 136.142.227.11
```

Server: 136.142.57.10

Address: 136.142.57.10#53

11.227.142.136.in-addr.arpa name = hydrogen.cs.pitt.edu.

The Robots Exclusion Protocol



- Web crawlers are not always wanted by web site owners
  - Q: can you think of examples?
  - A: craigslist, eBay, other shopping web sites

- As a result, /robots.txt indicates what is allowed and what is not allowed to be fetched by web crawlers
  - Example:

```
User-agent: *
Disallow: /
```

- **Note**: web crawlers can misbehave and ignore robots.txt
- **Note**: robots.txt file is publicly available

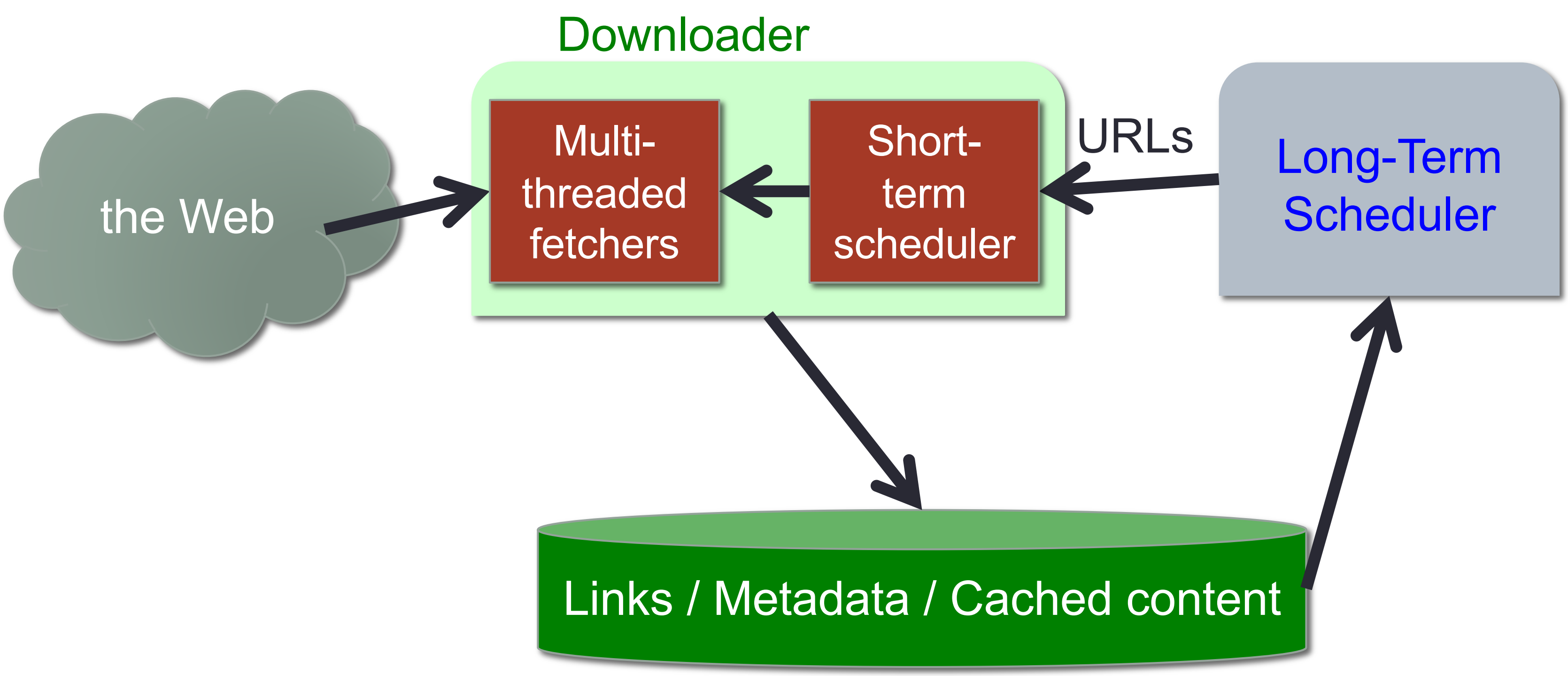
TYPES OF WEB CRAWLERS

Not all crawlers are created equal

- Create a full index of the Web
  - E.g., googlebot, bingbot
- Vertical crawling
  - E.g., news bots, spambots
  - E.g., based on file types
- Focused crawling
  - E.g., based on driving query
- Mirror a specific page / check for updates
  - E.g., https://www.changedetection.com/

SCHEDULING

Anatomy of a Web Crawler – II



How to schedule web page crawls?

- **Q1**: How often to crawl a page?
  - Too often ➔ wasted bandwidth and crawler resources
  - Too sparse ➔ missed potentially important updates
- Long-term scheduling:
  - Determine based on page quality/freshness estimations
- Short-term scheduling:
  - Determine based on politeness policy
    - In order not to overwhelm web site
    - Spread requests over time
    - Spread requests across multiple web sites

Crawling the Deep Web

- Also called the *Hidden Web*
- Need to provide form input for page to be generated
  - E.g., weather for `15213` zip code

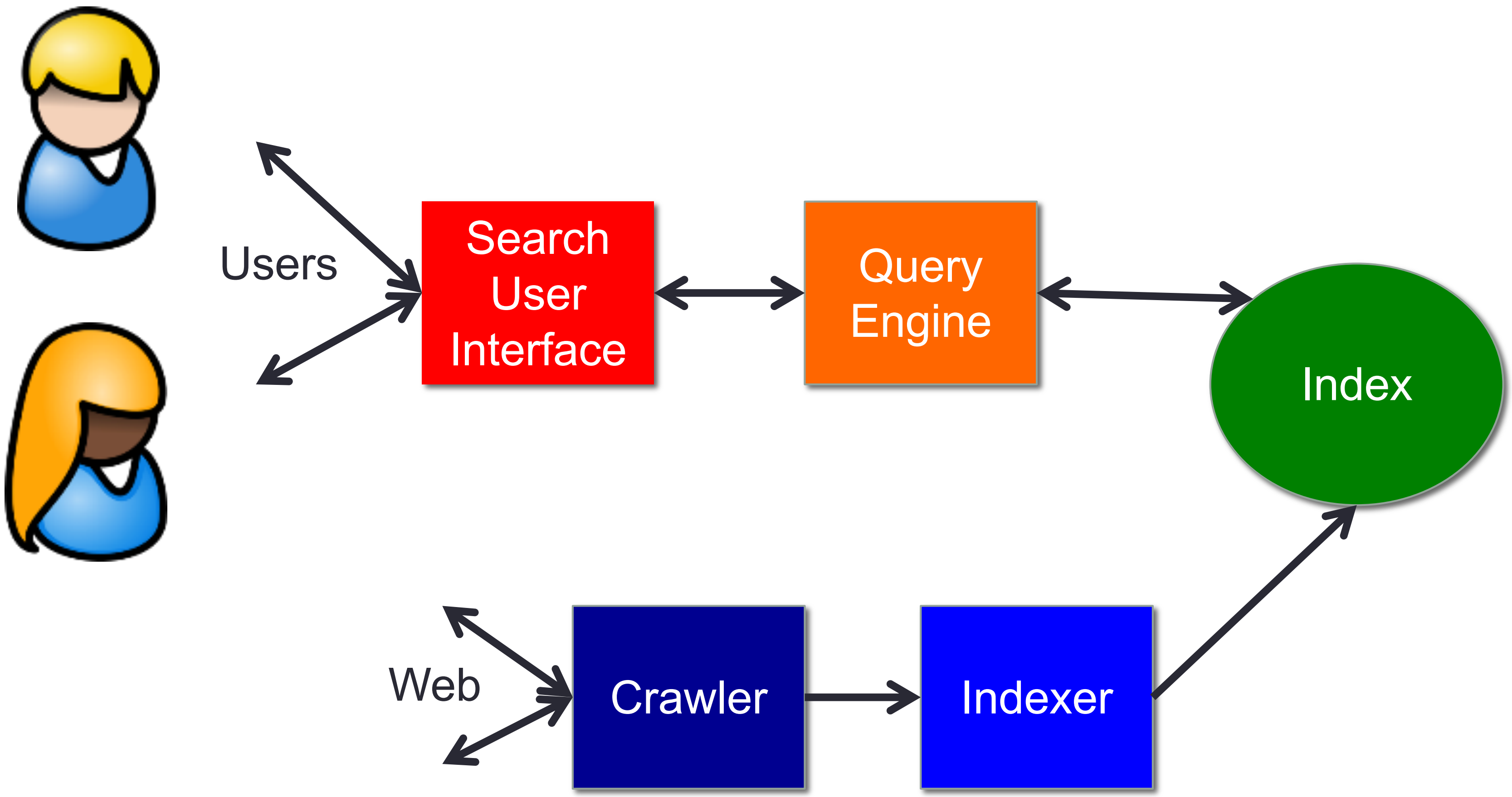
- General web crawling of Deep Web is very difficult

- Special cases for known forms/input:
  - E.g., can request same form and see if results changed
  - Called *screen scraping*
  - Challenging for pages with Javascript, especially AJAX

[More info: http://en.wikipedia.org/wiki/Web_crawler]

WEB INFORMATION RETRIEVAL

Web Information Retrieval Architecture



Search Engine Ranking

- Simple approach: use relevance ranking (TF-IDF)

- Question: What are the challenges / shortcomings?

- Answers:
  - Web site owners can "pad" pages with more keywords
  - Web site owners can present different pages to crawler vs to humans
  - Are not considering web page importance
  - Are not taking advantage of hypertext link structure

How about using link structure?

- Main idea:
  - View incoming links as "**votes**" of confidence for web page
  - Similar to citations in academic publishing

- Important question remains:
  - How to count incoming links – are all equally important?

- Must have different levels of importance:
  - Link from authoritative web site (e.g., nytimes.com)

    vs
  - Link from one's own web site (e.g., iloveapples.com)

In-classroom activity

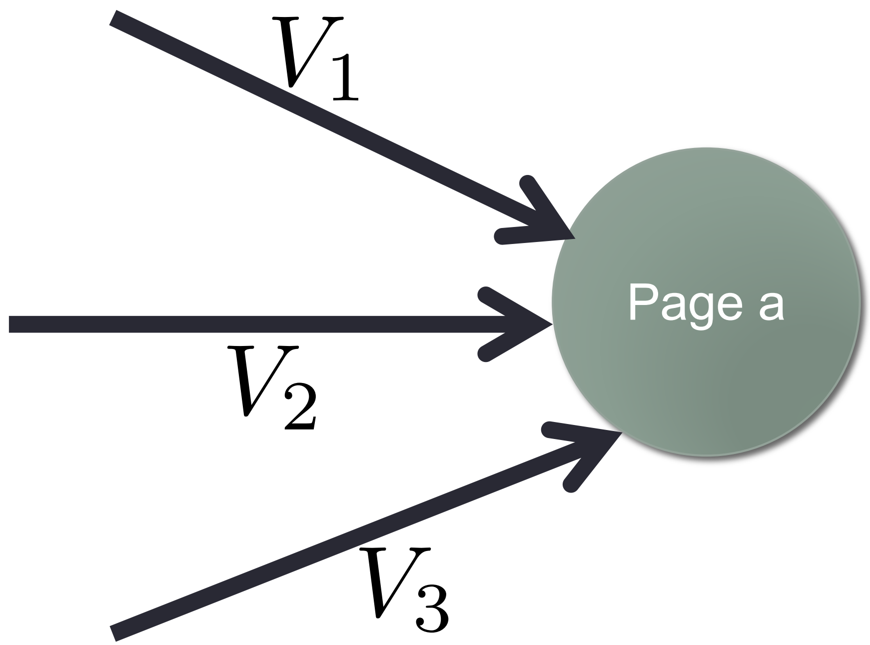
- Students organize into teams
  - The most popular team wins
- Each team gets **unlimited “votes”**
- Each team decides **how many votes** to “cast”
  - One or multiple votes per team
- Each team decides **where to give** votes to
  - Vote can go to any team (including own / duplicates)
- What voting strategy do you think is better?

In-classroom activity

- Vote only for your team once
- Give multiple votes to your team
- Give multiple votes to another team
- Give one vote to your team and another to another team
- Give multiple votes to multiple teams (not including your team)
- Give multiple votes to multiple teams (also voting for your team)

Google's PageRank – Receivers

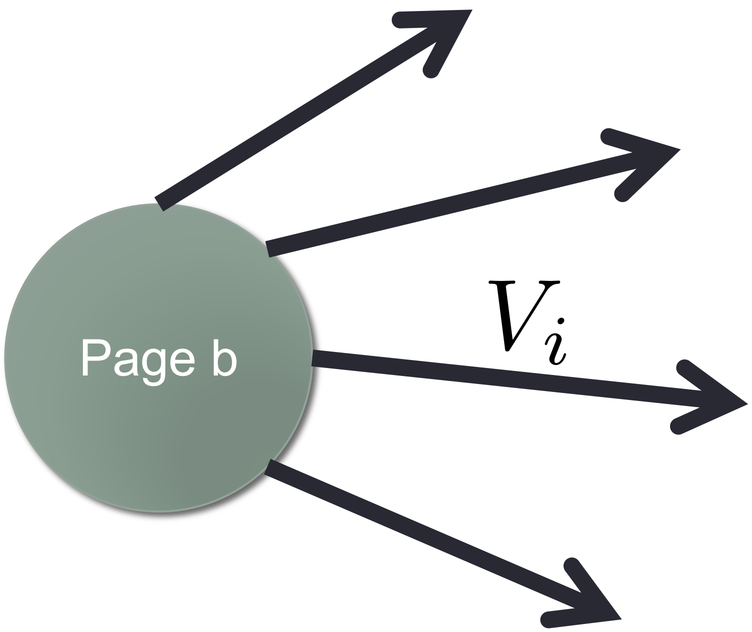
Consider only incoming links!



$$PR(a) = (1 - d) + d(V_1 + V_2 + \ldots + V_n)$$

PageRank of page a is the sum of values of incoming links

Google’s PageRank – Senders



How much “value” should each page give?

PageRank
of page b

$$V_i = \frac{PR(b)}{C(b)}$$

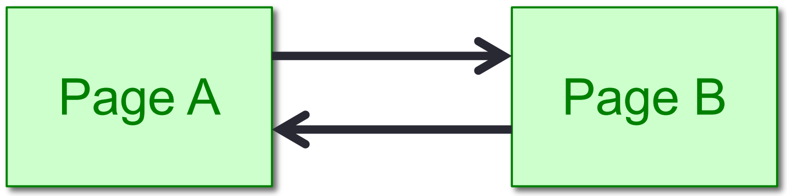
Number of outgoing
links from page b

Putting it all together

- PageRank of page **a** that has **n** incoming links from pages $\mathbf{T_1}$ to $\mathbf{T_n}$ is given by the following formula:

$$PR(a) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \ldots \frac{PR(T_n)}{C(T_n)})$$

A simple example

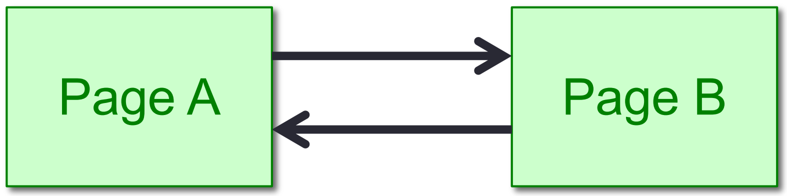


Notice the directional edges!

- Guess initial values of 1.0, i.e., PR(A) = PR(B) = 1.0

- d=0.85
- PR(A) = (1-d) + d(PR(B) / 1) = 0.15 + 0.85*1 = 1
- PR(B) = (1-d) + d(PR(A) / 1) = 0.15 + 0.85*1 = 1

A simple example – Take 2



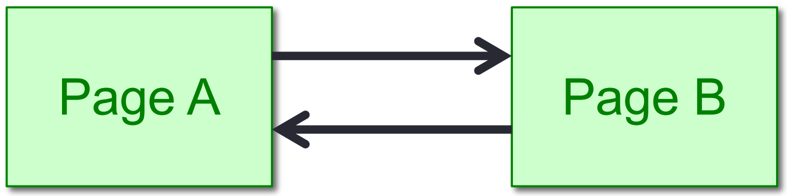
Notice the directional edges!

- Guess initial values of 0.0, i.e., PR(A) = PR(B) = 0.0
- **d=0.85**
- PR(A) = (1-d) + d(PR(B) / 1) = 0.15 + 0 = 0.15
- PR(B) = (1-d) + d(PR(A) / 1) = 0.15 + 0.85*0.15 = 0.2775

- PR(A) = 0.15 + 0.85 * 0.2775 = 0.385875
- PR(B) = 0.15 + 0.85 * 0.385875 = 0.47799375

- PR(A) = 0.15 + 0.85 * 0.47799375 = 0.5562946875
- PR(B) = 0.15 + 0.85 * 0.5562946875 = 0.622850484375

A simple example – Take 3



Notice the directional edges!

- Guess initial values of 40, i.e., PR(A) = PR(B) = 40.0
- **d=0.85**

- PR(A) = (1-d) + d(PR(B) / 1) = 0.15 + 0.85*40 = 34.15
- PR(B) = (1-d) + d(PR(A) / 1) = 0.15 + 0.85*34.15 = 29.1775

- PR(A) = 0.15 + 0.85 * 29.1775 = 24.950875
- PR(B) = 0.15 + 0.85 * 24.950875 = 21.35824375

- Seems numbers will get to 1.0 and stop

More examples

- https://webworkshop.net/seo-tools/pagerank_calculator

Modern search engine ranking

- Modern search engines utilize multiple signals:
  - PageRank
  - Name of query term in domain name
  - Location
  - https vs http

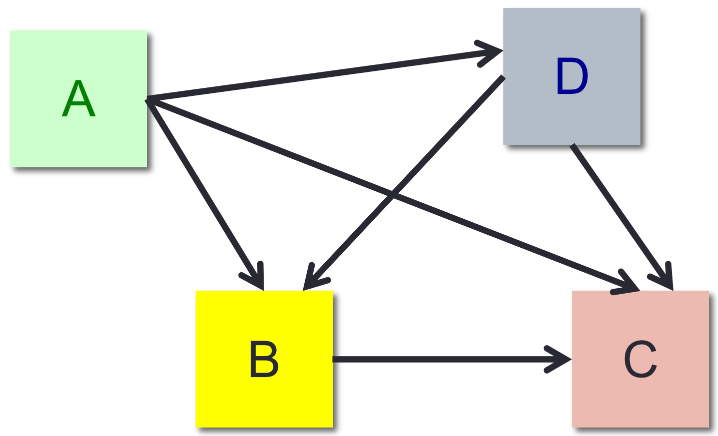
- This is a result of many years of improvements
- And reacting to Search Engine Optimization techniques!

[More info: http://en.wikipedia.org/wiki/Search_engine_optimization]

  - E.g., link farms

[More info: http://en.wikipedia.org/wiki/Link_farm]

Understanding Question



- **Question**:
  - Which of these pages will have a higher PageRank value?

- **Possible Answers**:
  - A
  - B
  - C
  - D