UiO : University of Oslo

Web search

Pierre Lison
University of Oslo, Dep. of Informatics

INF3800: Søketeknologi
May 14, 2014

[Note: Some slide diagrams borrowed from C. Manning, P. Nayak and P. Raghavan]

UiO : University of Oslo

Outline of the lecture

- Basics of web search
- Web crawling & indexing
- Link analysis
- Conclusion

Outline of the lecture

- **Basics of web search**
- Web crawling & indexing
- Link analysis
- Conclusion

Web search

- Web search creates a number of challenges to "traditional" IR:
  - *Scale* (billions of web pages)
  - *Heterogeneous* content
  - *Trust* becomes a key factor in ranking
  - Web users different from "traditional" IR users
  - Business aspects (e.g. sponsored search)

Types of web queries

- *Informational*: general info on topic [~50%]

  Italian cuisine

  Britney Spears family life

  Types of nuclear fusion reactions

- *Navigational*: search specific entity [~20%]

  University of Oslo in Norway

  cxense AS

  Research webpage of Pierre Lison

- *Transactional*: want to *do* something [~30%]

  Car rental from Gardemoen

  "House of Cards" online streaming

  Iphone 4S Norway

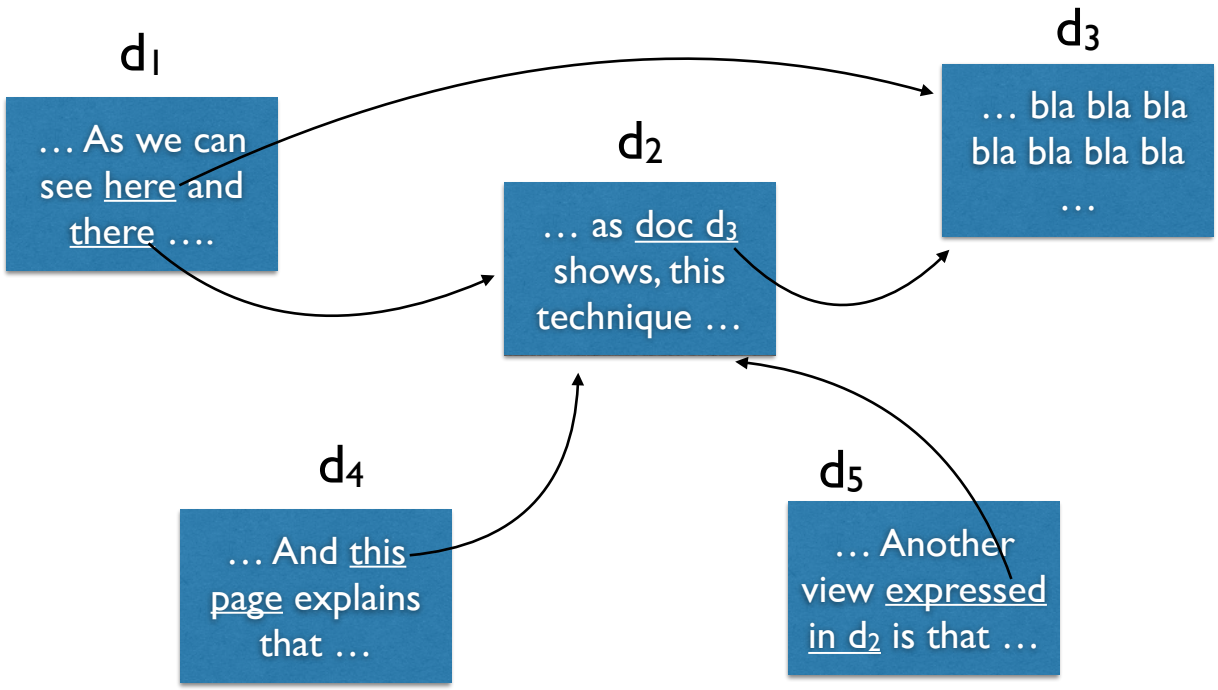
Web queries

- Precision often more important than recall!
  - Especially precision on *top results*
  - Necessary to filter untrusted pages / spam
  - Need to consider other qualities than relevance (trustworthiness, recency of content, etc.)
  - Recall only matters if number of matches very small
- Query language must be lightweight (mostly phrase queries)

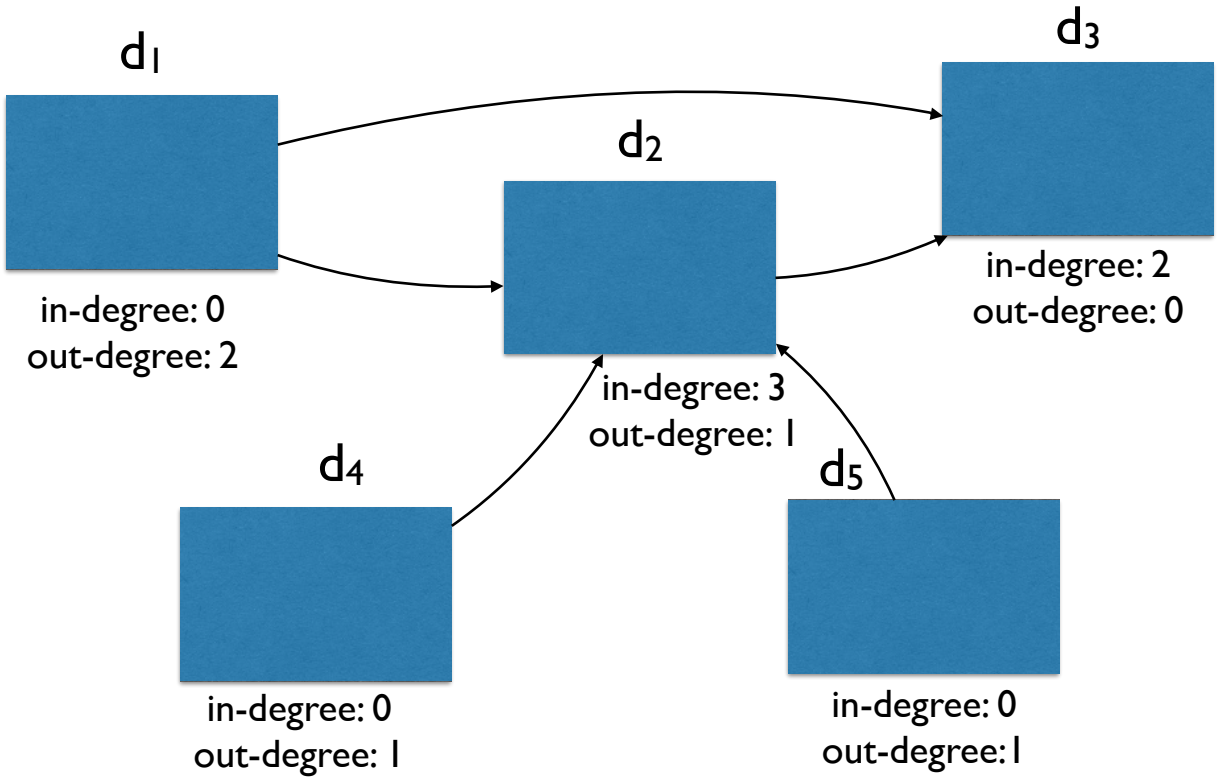
Web content

- Massively *distributed* creation of content
- Need to assess the *trustworthiness* of pages (obsolete information, duplicates, spam, etc.)
- Content may be *unstructured* (text), *semi-structured* (XML), *structured* (databases)
- Mixture of multiple *media* (text, images,video, etc.)
- Dynamically generated webpages (by querying an application server with backend database)

The web graph



The web graph



Spamdexing

- **Spamdexing**: "*manipulation of web content to appear artificially high on search results for particular keywords*"
  - Continuous battle between spammers and search engines (*adversarial* information retrieval)
- Common spamming techniques:
  - Keyword stuffing, invisible text
  - *Cloaking*: server returns fake content to web crawlers
  - *Doorways*: dummy start page carefully crafted for keywords
  - Optimisation of *metadata* on the page (notably URLs)

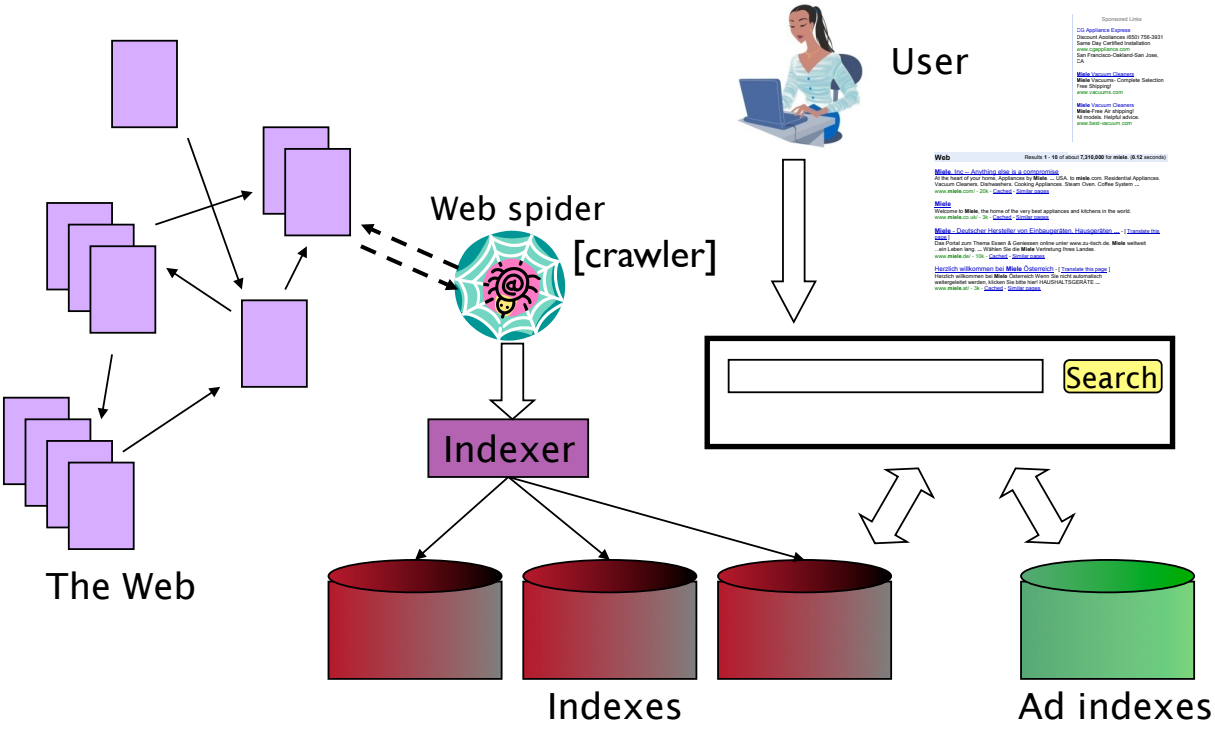
Spamdexing

- Conter-measures:
  - Exploit "quality signals" (from web & from users) to determine whether a webpage is trustworthy
  - Limits on meta-keywords
  - Analysis of web graph to detect suspicious linkages
  - Machine learning to classify spam
  - Editorial intervention (blacklists etc.)

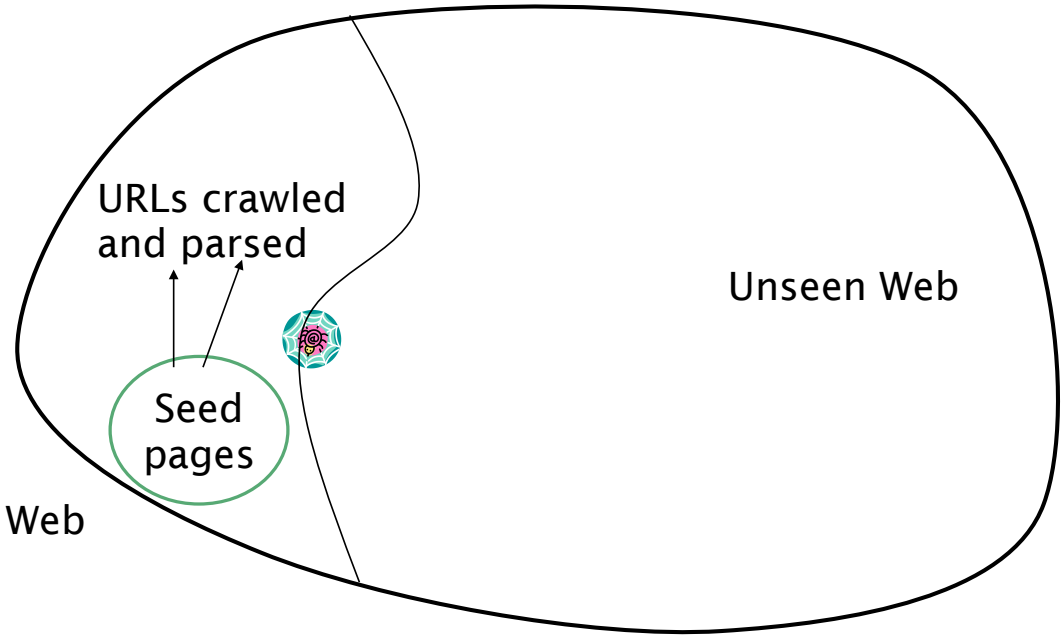
Outline of the lecture

- Basics of web search
- **Web crawling & indexing**
- Link analysis
- Conclusion

Search architecture



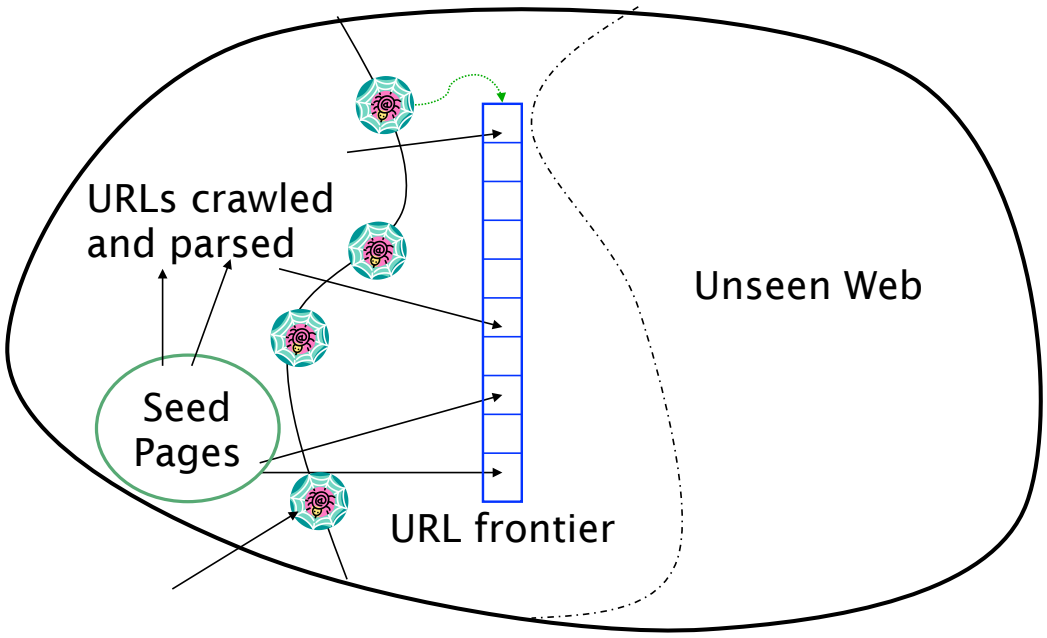
Web crawling



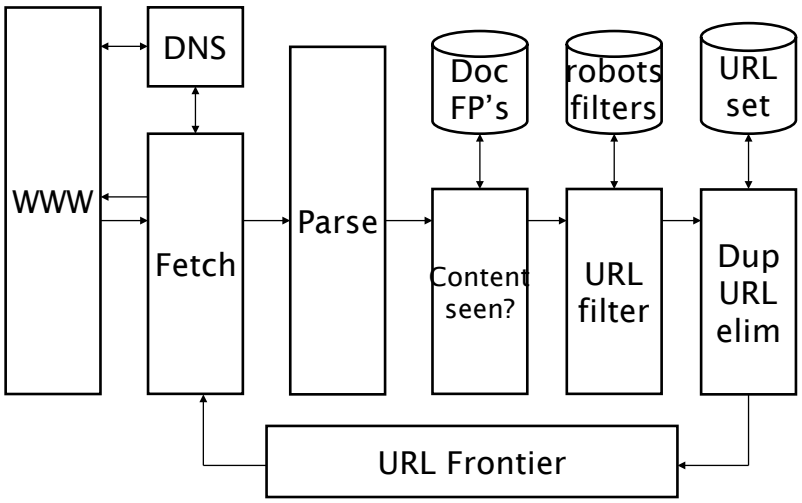
Requirements for web crawlers

- Distributed, scalable, efficient (obviously)
- *Robust* to all types of content
  - Malicious or ill-constructed pages
  - Dynamically generated pages
- *Polite*
  - avoid flooding servers
  - only crawl allowed pages
- Able to *prioritise* content

Web crawling

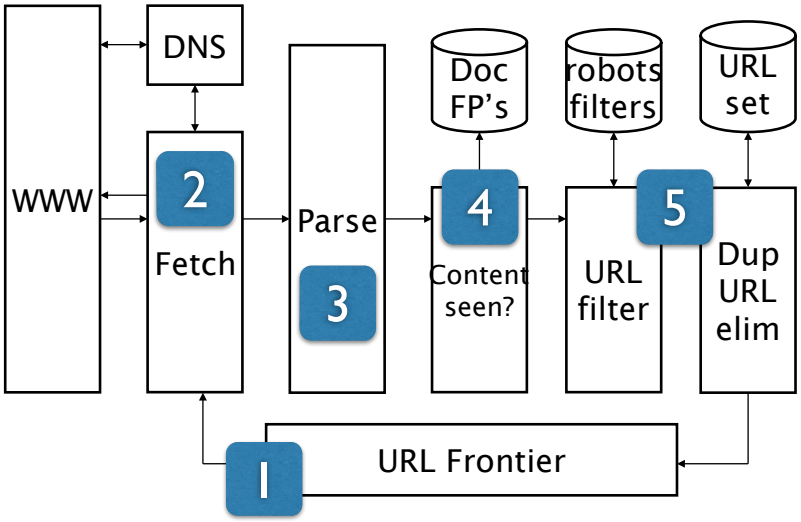


Crawling workflow



URL frontier: data structure containing the set of URLs that have been detected but not yet crawled

Crawling workflow



1 Pick URL from frontier

2 Fetch document

3 Parse its content

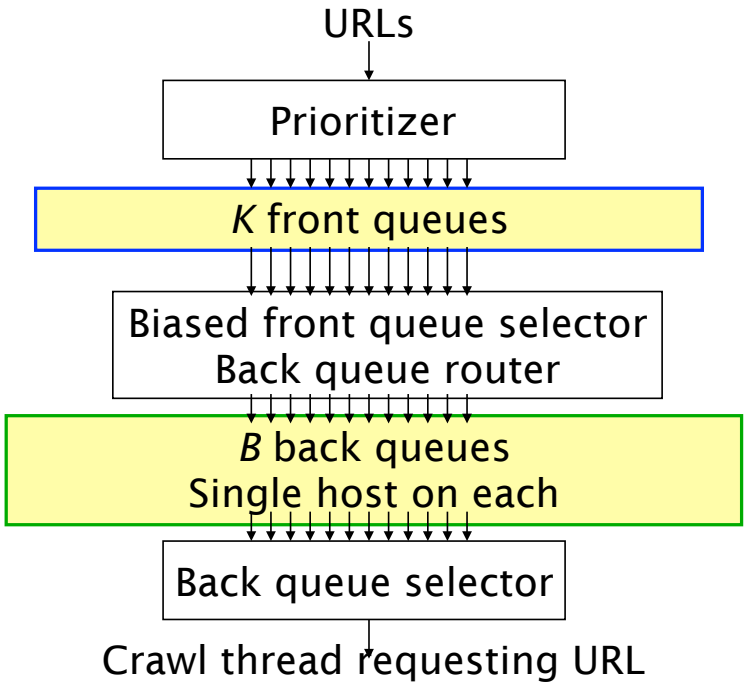
4 Check if content already seen (if not, add it to index)

5 Filter outgoing URLs (enforce politeness, remove duplicates) and add to frontier

URL frontier

- URL frontier must be able to *sort* the next URLs to crawl
- Two criteria:
  - **Politeness**: do not flood web servers with too many requests in short periods of time
  - **Prioritisation**: crawl webpages that are of high-quality and/or are frequently updated more often
  - Conflicts between these two criteria!

URL frontier in Mercator



System of two (FIFO) queues:

- **Front queues** for prioritisation (each queue = a priority level)
- **Back queues** for politeness (each queue = a specific host)

Web indexing

- Two types of index partitioning:
  - **Partitioning by terms**: index terms divided in subsets, and each subset is allocated to a node

Greater concurrency (in theory)

- Must exchange & merge long posting lists across nodes
- Load-balancing

  - **Partitioning by documents**: each node is responsible for a local index for subset of all documents (query sent to each node and the results are merged back)

Often easier to distribute, more efficient I/O on posting lists

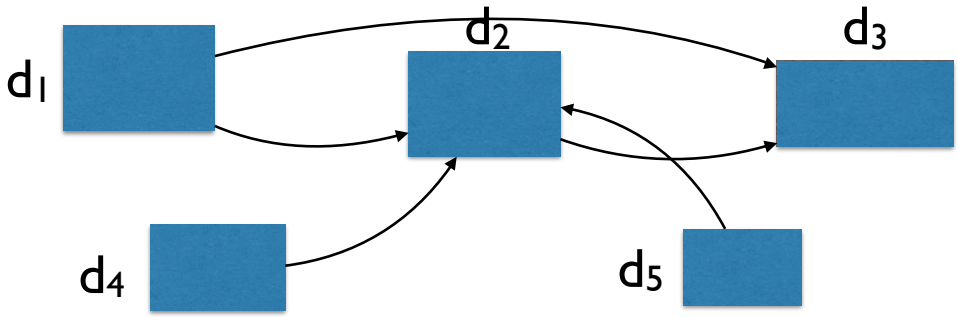
- More disk seeks
- Need to calculate global statistics separately

Outline of the lecture

- Basics of web search
- Web crawling & indexing
- **Link analysis**
- Conclusion

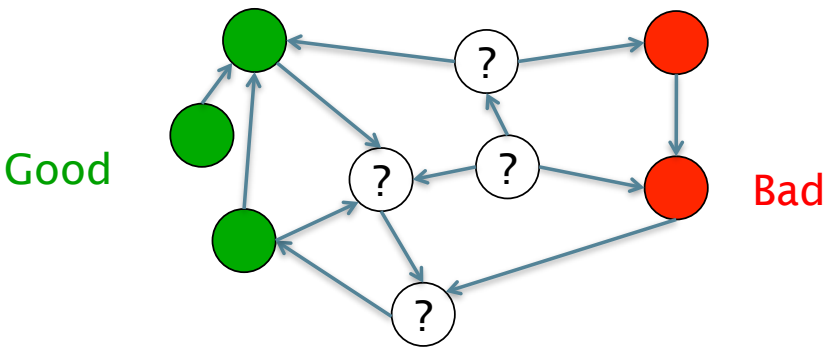
Link analysis

- Document *trustworthiness* at least as important as relevance for web search!
  - How to determine it?
- **Link structure** between documents provides powerful indicators of quality and trust



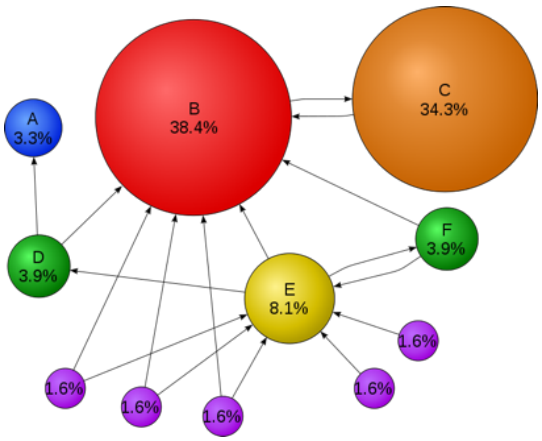
Link analysis

- Key idea: the quality of a webpage can be determined by looking at its neighboring links
  - Link from node A to node B = "conferral of authority" from A to B
  - Can be interpreted as a quality signal from A to B
  - Good nodes will tend to point to good nodes, and bad nodes to bad nodes



PageRank

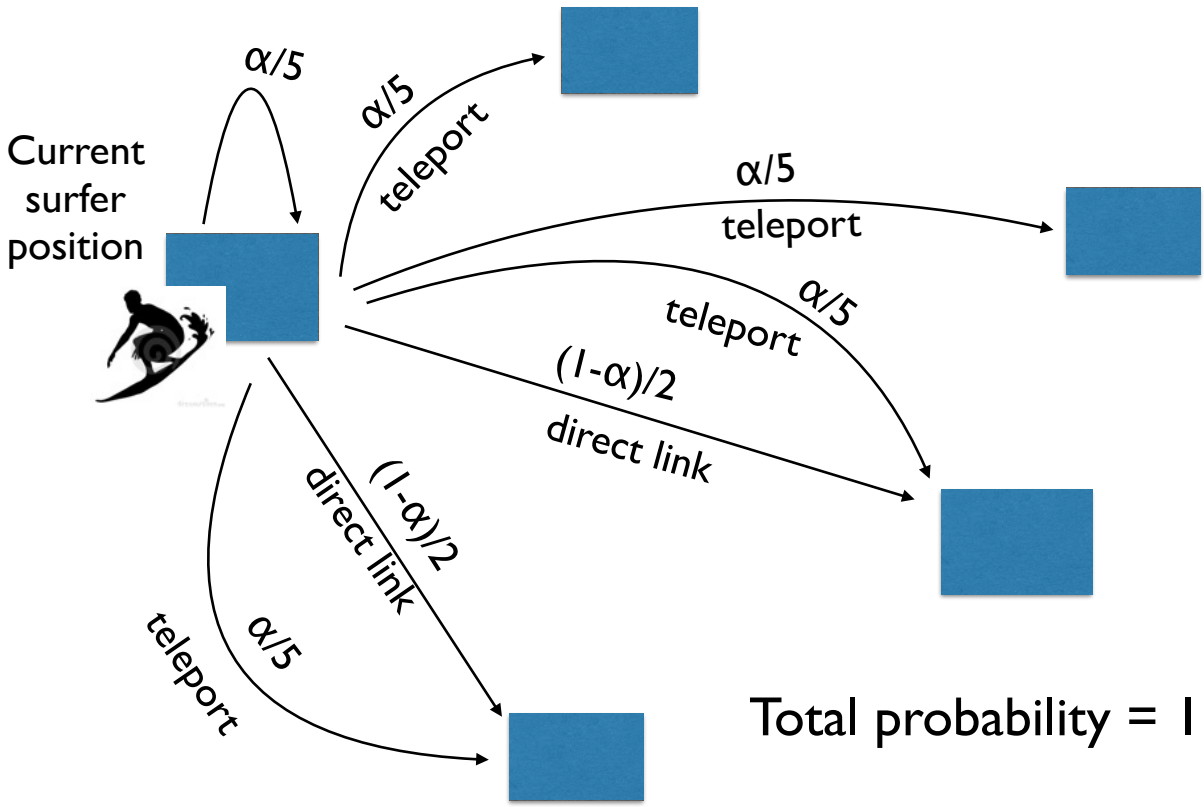
- Most well-known algorithm for ranking the quality of webpages according to their link structure is **PageRank**
  - Used (among many other algorithms) by Google Search
  - Assigns a numerical score (between 0 and 1) to each page



PageRank

- Imagine a web surfer that randomly surfs the web for an infinite amount of time
- Two ways of moving from A to B:
  - **Follow an explicit link** from A to B (all links are equally likely to be followed)
  - **Teleport** from A to B, for example by typing URL in browser (all possible webpages are equally likely)
- Teleportation rate α defines the relative probability of teleport versus link following

PageRank



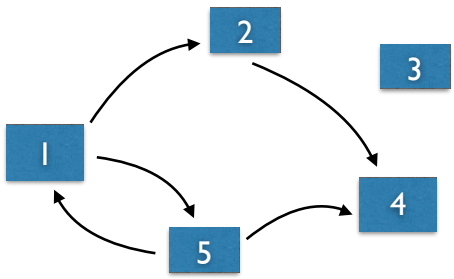
PageRank

- PageRank for *d*: if the surfer was allowed to continue surfing indefinitely, what would be the fraction of the time where he is on page *d*?
- This random walk can be represented as a **Markov Chain**
  - the state is the current position of the surfer
  - the transition matrix P encodes the probability of going from document i to document j for all pairs i,j

PageRank

- Several ways to calculate this PageRank
- One simple technique is the *power iteration method*:
  - Start with some initial distribution $\mathbf{x}_{t0}$ over possible states (documents)
  - Calculate the probability vector for the next state $\mathbf{x}_{t1} = \mathbf{x}_{t0} P$ (matrix multiplication)
  - And continue the iteration until convergence

PageRank example



Assume a rate $\alpha = 0.5$

Transition matrix P =

	1	2	3	4	5
1	0.1	0.35	0.1	0.1	0.35
2	0.1	0.1	0.1	0.6	0.1
3	0.2	0.2	0.2	0.2	0.2
4	0.2	0.2	0.2	0.2	0.2
5	0.35	0.1	0.1	0.35	0.1

Let us start with distribution $\mathbf{x}_{t0} = [1 \quad 0 \quad 0 \quad 0 \quad 0]^T$

→ $\mathbf{x}_{t1} = \mathbf{x}_{t0} P = [0.1 \quad 0.35 \quad 0.1 \quad 0.1 \quad 0.35]^T$

→ $\mathbf{x}_{t2} = \mathbf{x}_{t1} P = [0.2075 \quad 0.145 \quad 0.12 \quad 0.3825 \quad 0.145]^T$

→ $\mathbf{x}_{\infty} \approx [0.19 \quad 0.19 \quad 0.144 \quad 0.286 \quad 0.19]^T$

Outline of the lecture

- Basics of web search
- Web crawling & indexing
- Link analysis
- **Conclusion**

Conclusion

- Challenges for web search
  - *Precision* more important than recall
  - Huge variations in document content and quality
  - *Trustworthiness* of pages must be assessed
  - Need to scale to huge amounts of data (crawling must follow specific priorities)
- Link analysis (for instance *PageRank*) allows us to score the importance of each page according to its link structure