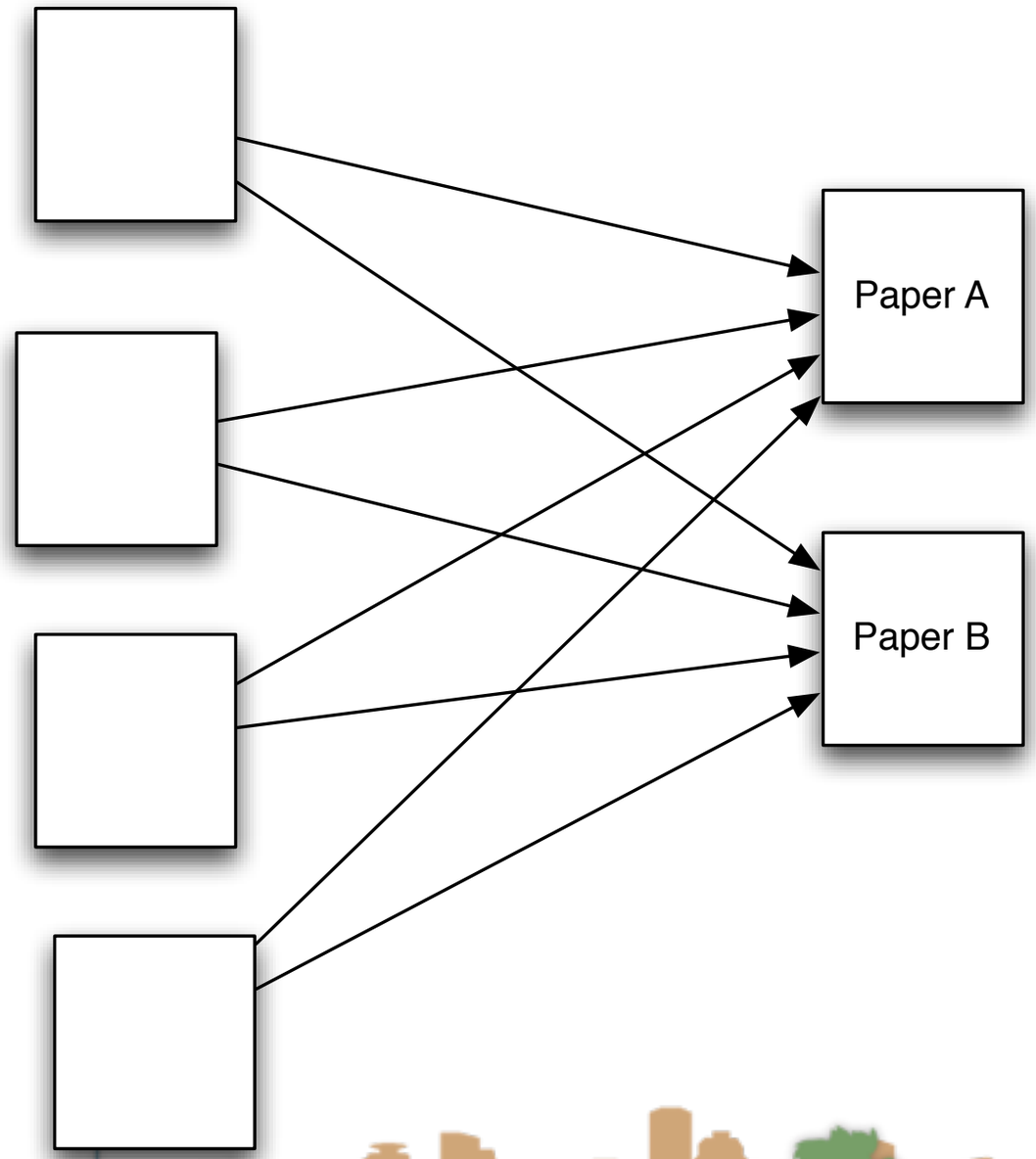
The web as a directed graph

- Pages are nodes
- Links are directed edges
  - ICS
  - ...this links my page to the ICS home page
- LinkAnalysis/PageRank has its origins in bibliometrics
  - "Measurement of influence among publications based on citations"
- Just as citing a paper confers authority upon it, linking to a page confers authority to it.

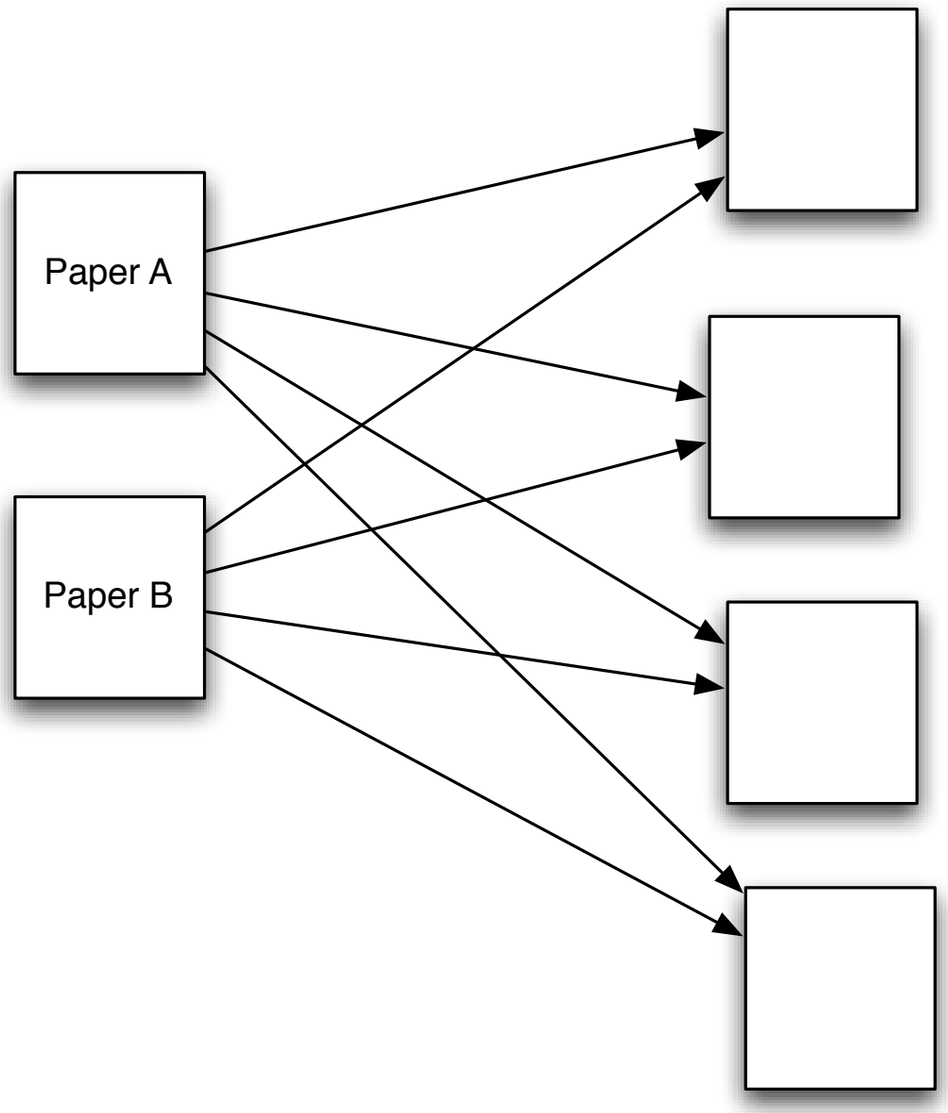
Bibliometrics

- Two ways of measuring similarity of scientific articles:
  - Cocitation similarity: The two articles are cited by the same articles
  - Bibliographic coupling similarity: The two articles site the same articles

Co-citation similarity



Bibliographic coupling similarity

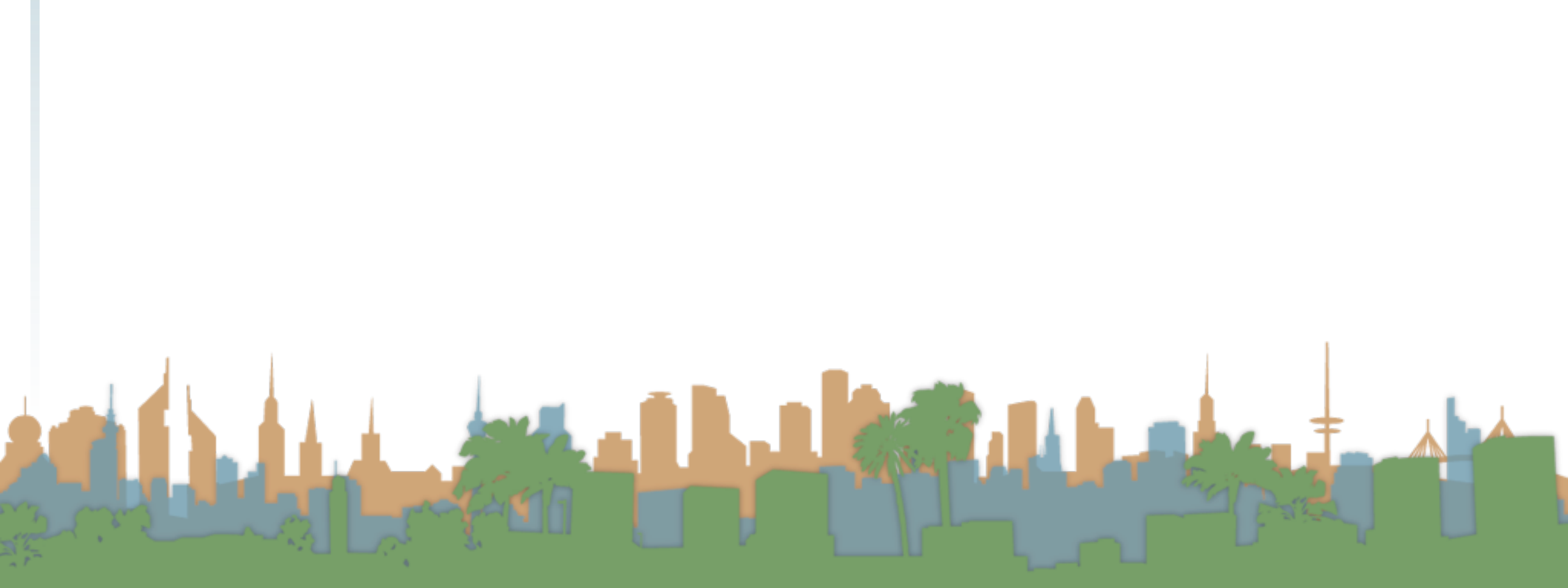


Bibliometrics

- Citation frequency can be used to measure impact
  - Each article gets one vote
  - Not a very accurate measure
- Better measure: weighted citation frequency/ citation rank
  - An article's vote is weighted according to its citation impact.
  - Sounds circular, but can be formalized in a well-defined way
  - This is basically PageRank
  - Invited for citation analysis in the 1960's by Pinsker and Narin

Key Observation

- A citation in scientific literature is like a link on the web



Link based query processing

- First retrieve all pages meeting the query
- First generation:
  - Then order them by their link popularity
    - citation frequency
  - Easy to spam. Why?
- Second generation:
  - Order them by their weighted link popularity
    - PageRank

Link based query processing

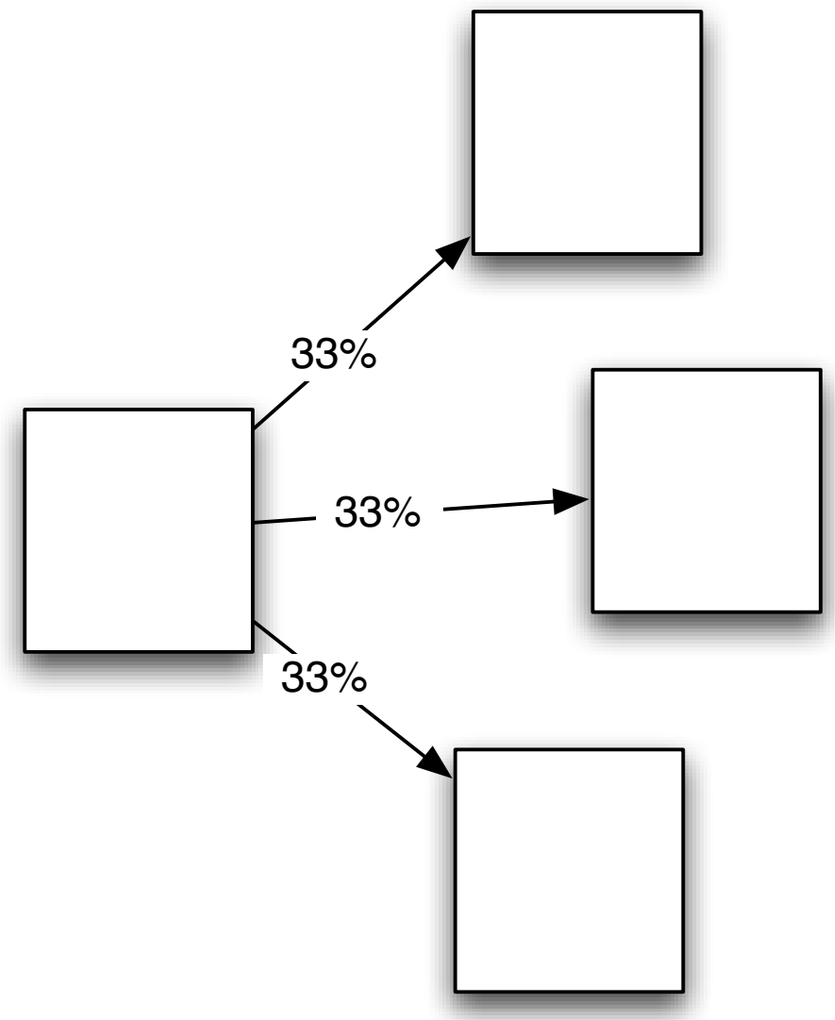
- First retrieve all pages meeting the query
- First generation:
  - Then order them by their link popularity
    - citation frequency
  - Easy to spam. Why?
- Second generation:
  - Order them by their weighted link popularity
    - PageRank

Link based query processing

- A full search engine balances many different scores
  - cosine similarity
  - term proximity
  - zone scoring
  - PageRank
  - contextual relevance (implicit queries)

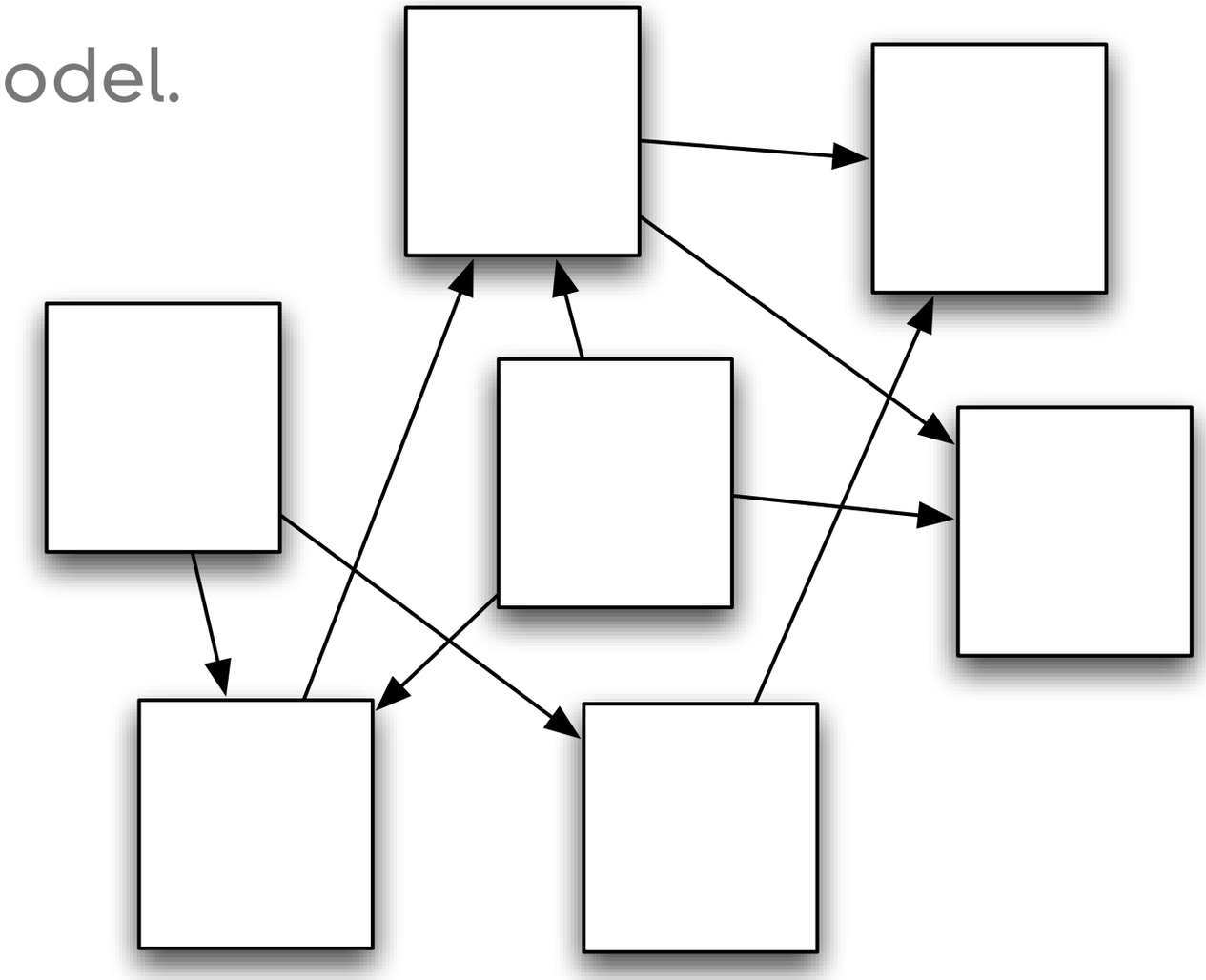
PageRank

- Every webpage gets a score
  - between 0 and 1
  - it's PageRank
- The random walk
  - Start at a random page
  - Follow an out edge with equal probability
- In the long run each page has a long-term visit rate.



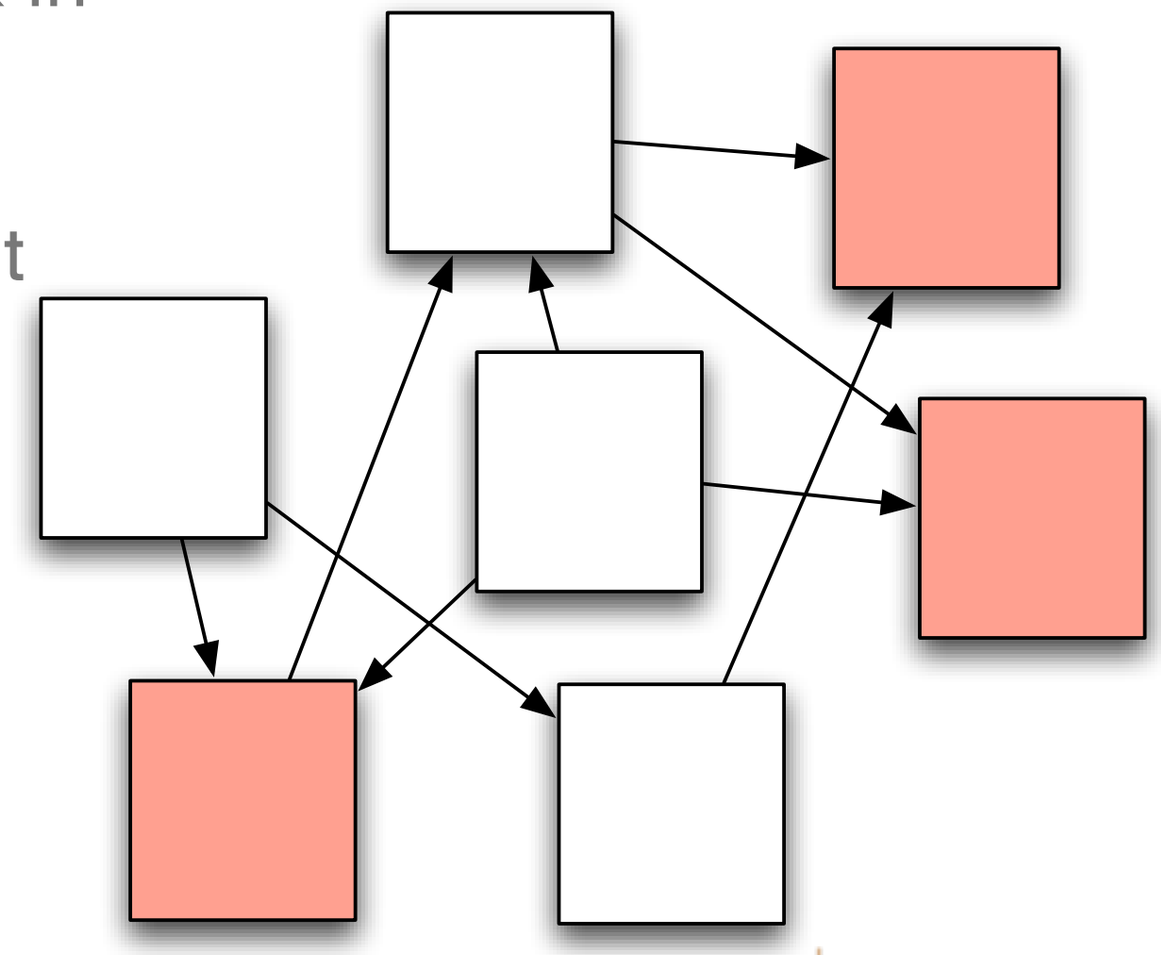
PageRank

- PageRank is a page's long-term steady state visit rate based on a random walk model.



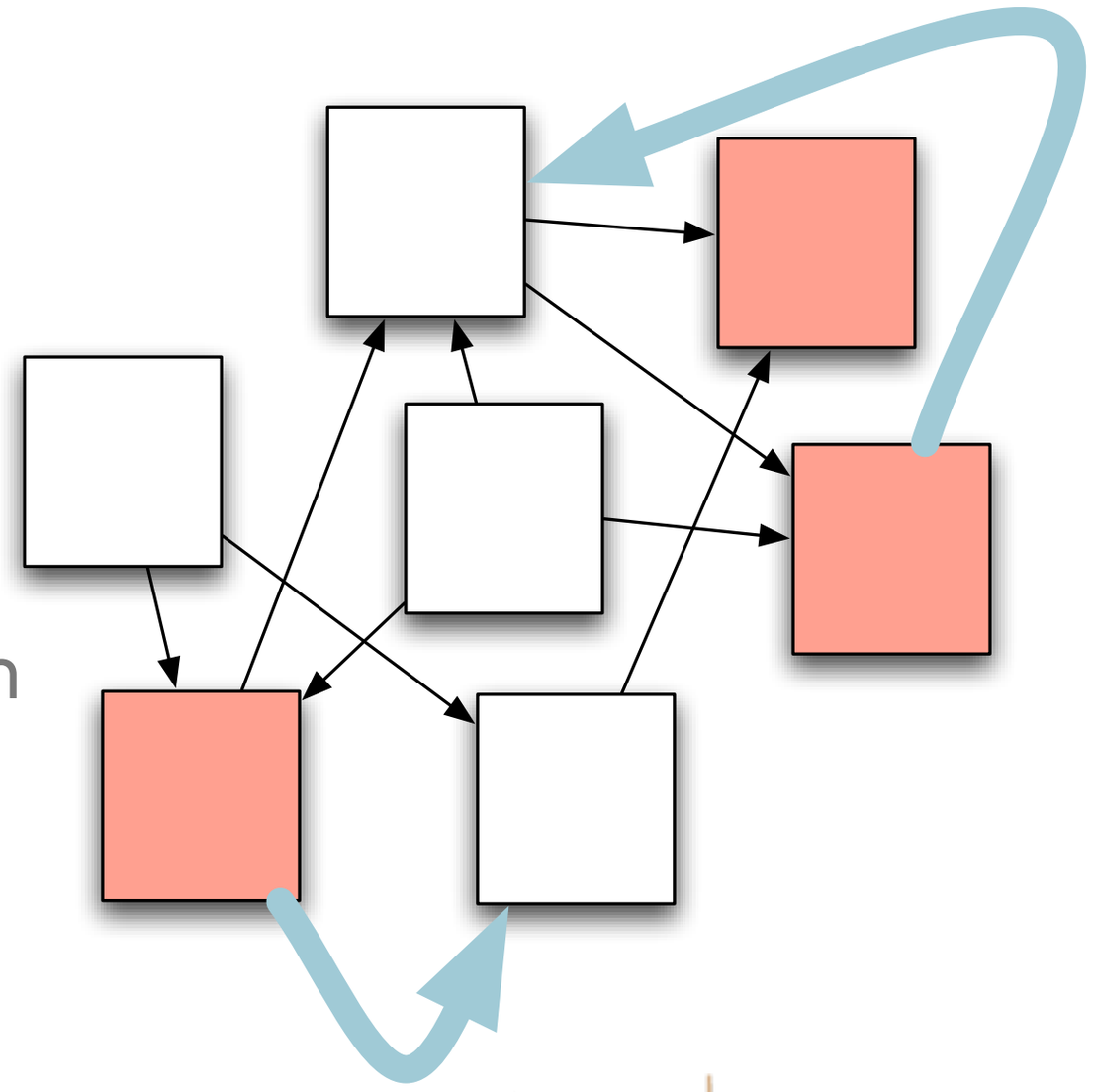
Visit Rate not quite enough

- The web is full of dead-ends
  - A random walk can get stuck in dead-ends
  - Makes no sense to talk about long-term visit rates



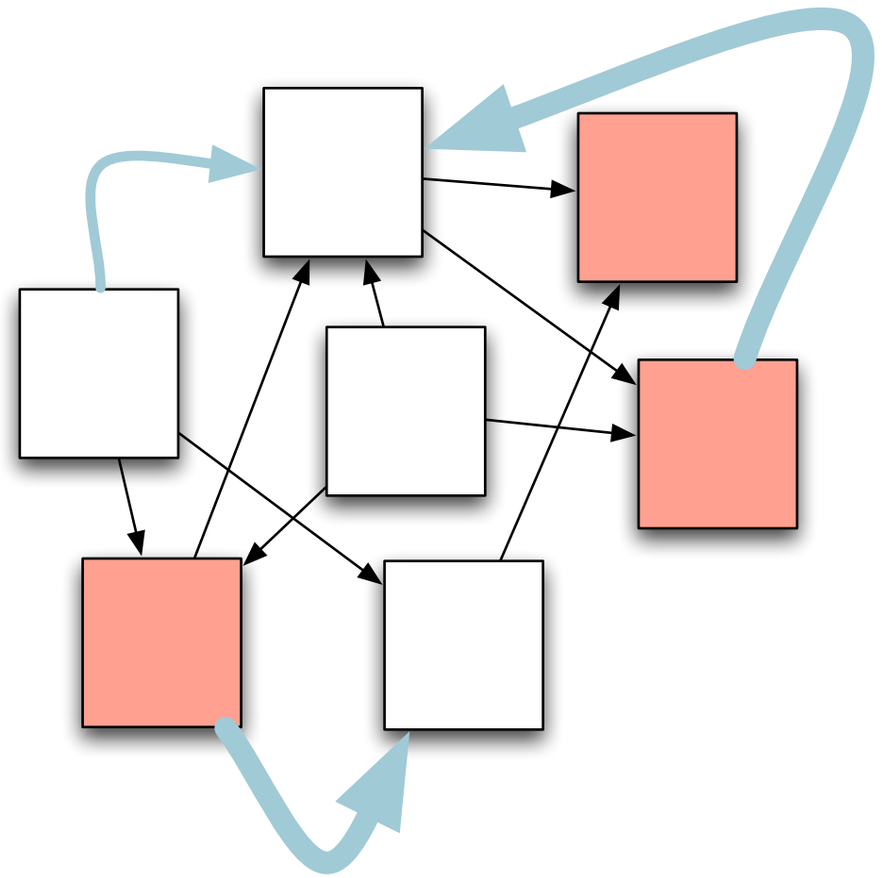
Teleporting

- At a dead end, jump to a random web page
  - at any non-dead end, with probability 10% jump to a random web page anyway
  - the other 90% choose a random out link
  - “10%” is a tunable parameter



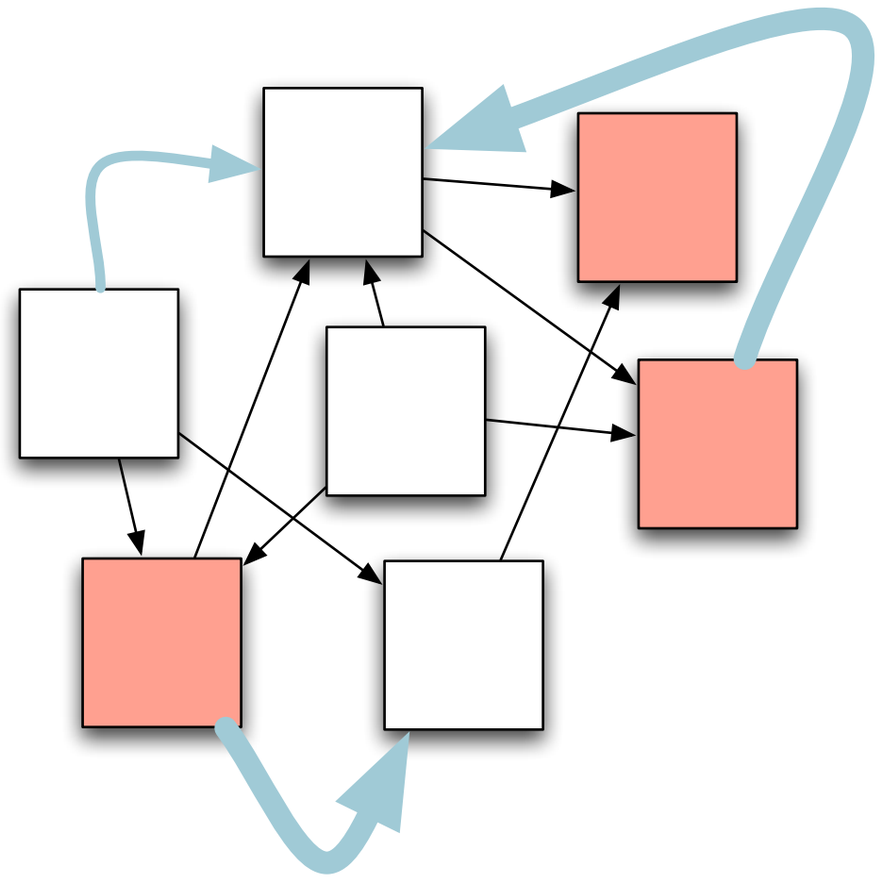
Teleporting

- Now we cannot get stuck locally
- There is a long-term visit rate at which any page is visited.
- How do we computer the visit rate?
  - aka How do we compute PageRank?
- (By the way this is a Markov Model)



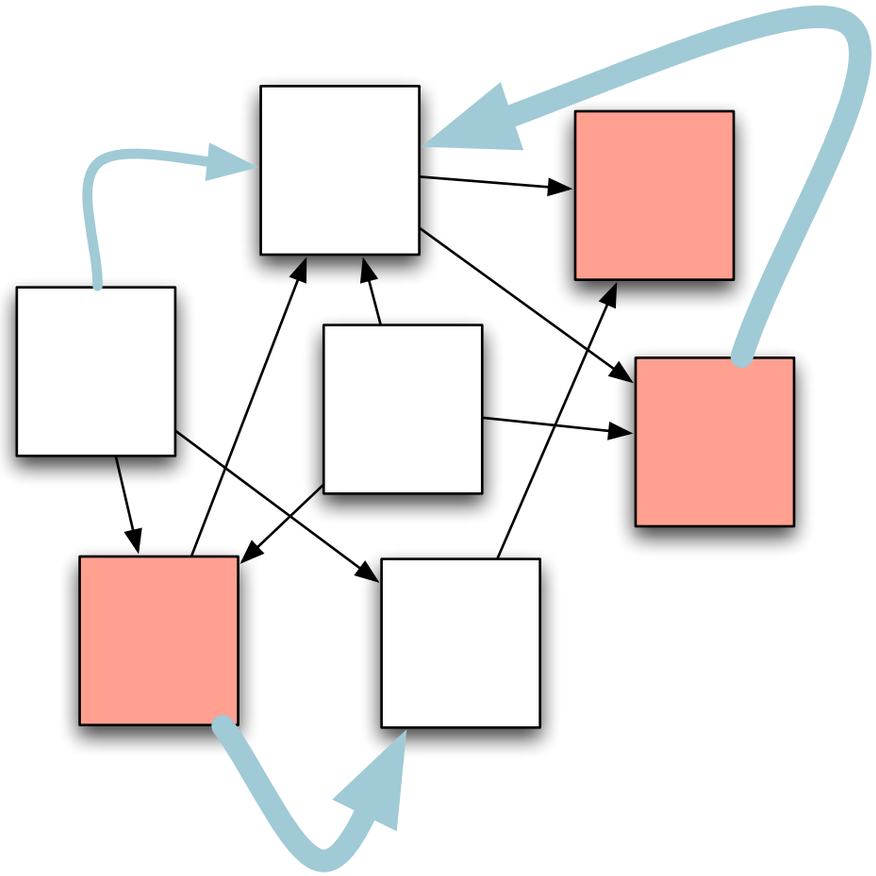
Teleporting

- Now we cannot get stuck locally
- There is a long-term visit rate at which any page is visited.
- How do we computer the visit rate?
  - aka How do we compute PageRank?
- (By the way this is a Markov Model)



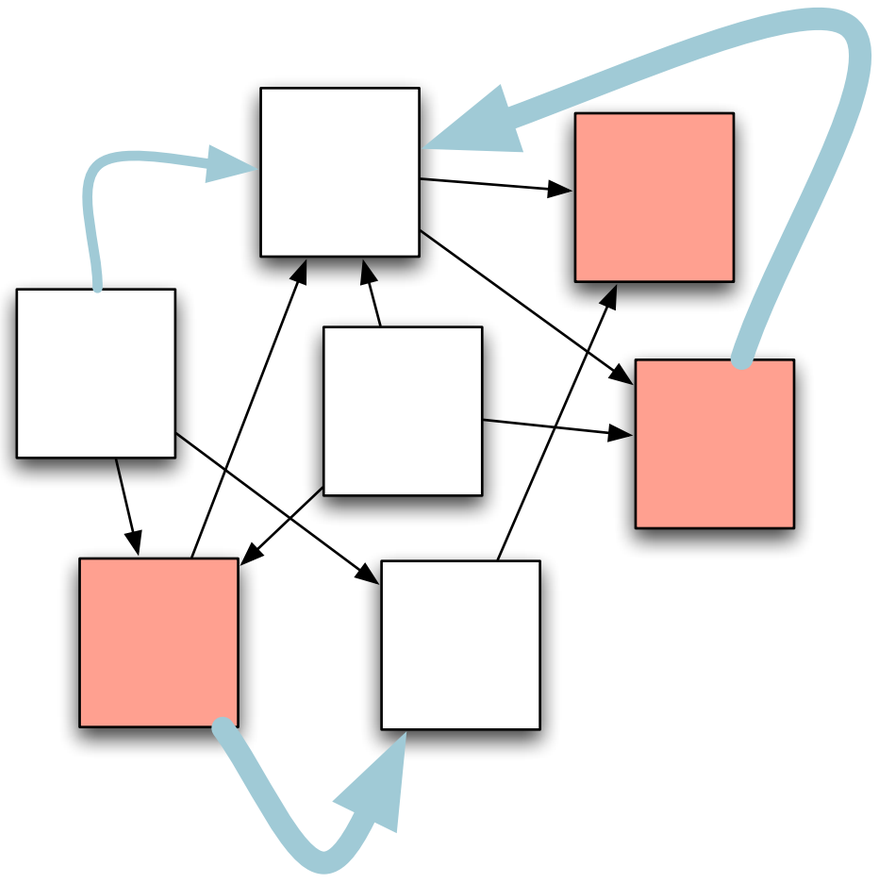
Markov Chains

- A Markov Chain is a mathematical "game"
- It consists of n states
  - corresponds to web pages
- And a transition probability matrix
  - corresponds to links
  - it is like an adjacency matrix



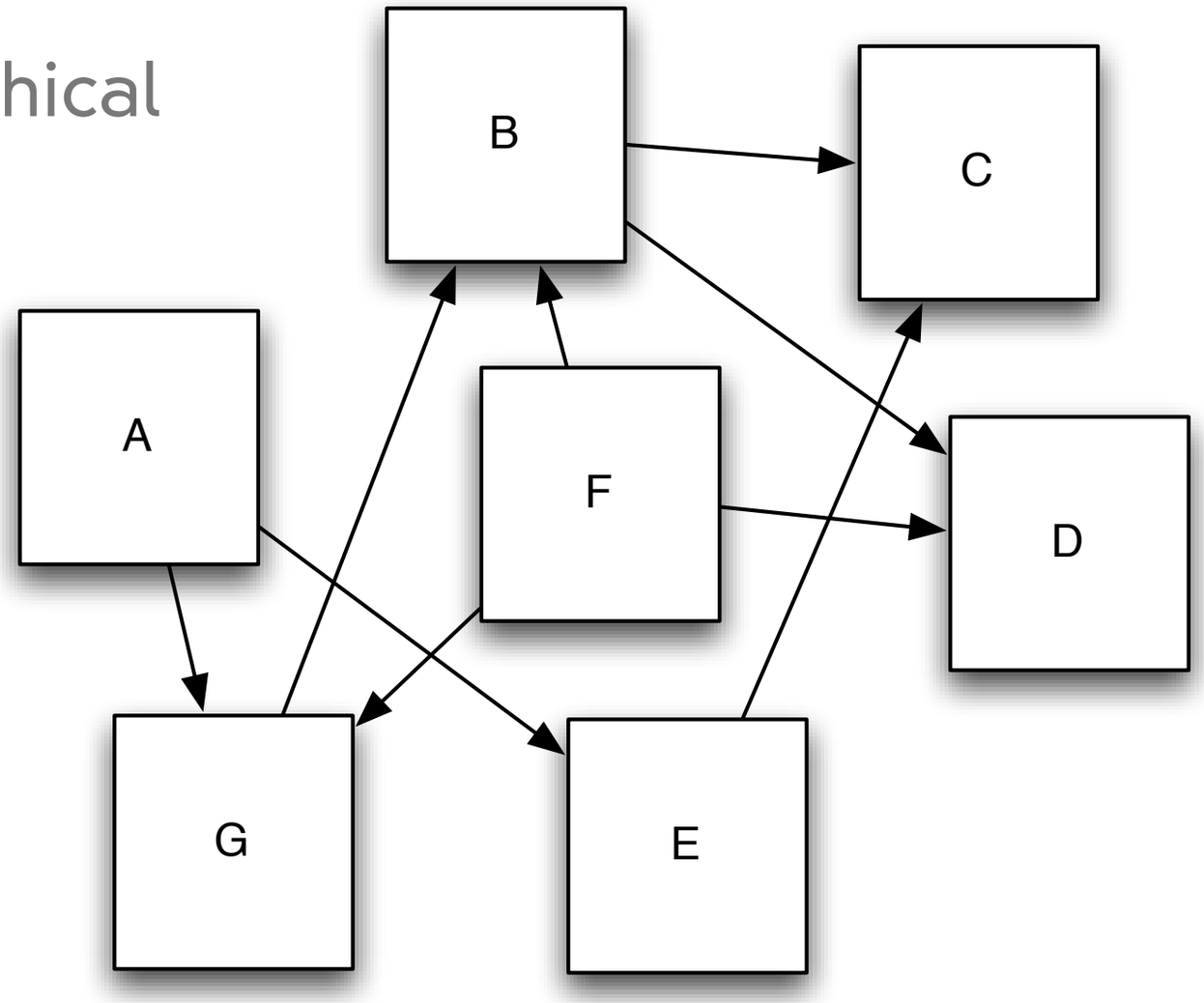
Markov Chains

- At any moment in the game we are in one of the states
- In the next step we move to a new state
- We use the transition matrix to decide which state to move into.
- If you are in state “i” then the probability of moving into state “j” is P(i -> j)

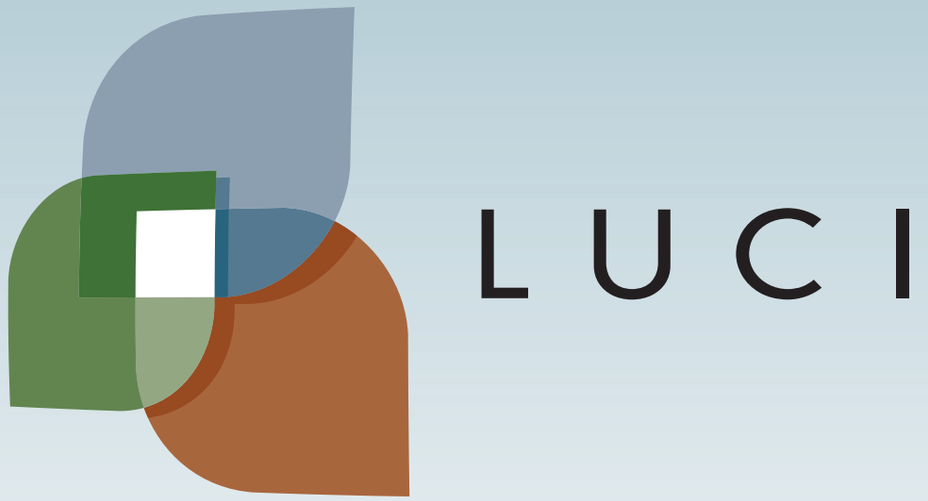


Exercise

- Compute the parameters of the Markov Chain for this graphical model



next...



LUCI