SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)
Re-accredited by NAAC with A+ grade, Accredited by NBA(CSE, IT, ECE, EEE & Mechanical)
Approvedy by AICTE, New Delhi, Recognized by UGC, Affiliated to Anna University, Chennai

Department of Computer Applications

Random Walk on Graphs

Course: NoSQL Database system

Class / Semester: II MCA / III Semester

Ranking Webpages

- The problem statement:
  - Given a query word,
  - Given a large number of webpages consisting of the query word
  - Based on the hyperlink structure, find out which of the webpages are most relevant to the query
- Similar problems:
  - Citation networks, Recommender systems

Mixing rate

- How fast the random walk converges to its limiting distribution
- Very important for analysis/usability of algorithms
- Mixing rates for some graphs can be very small: O(log n)

Mixing Rate and Spectral Gap

- Spectral gap: $1 - \lambda_2$
- It can be shown that

For a random walk starting at node i,

$$|P_t(j) - \pi(j)| \leq \sqrt{\frac{d(j)}{d(i)}}\lambda^t.$$

- Smaller the value of λ_2 larger is the spectral gap, faster is the mixing rate

Recap: Pagerank

- Simulate a random surfer by the power iteration method
- Problems
  - Not unique if the graph is disconnected
  - 0 pagerank if there are no incoming links or if there are sinks
  - Computationally intensive?
  - Stability & Cost of recomputation (web is dynamic)
  - Does not take into account the specific query
  - Easy to fool

PageRank

- The surfer jumps to an arbitrary page with non-zero probability (escape probability)

$$M' = (1-w)M + wE$$

- This solves:
  - Sink problem
  - Disconnectedness
  - Converges fast if w is chosen appropriately
  - Stability and need for recomputation
- But still ignores the query word

HITS

- Hypertext Induced Topic Selection
  - By Jon Kleinberg, 1998
- For each vertex v Є V in a subgraph of interest:
  - a(v) - the authority of v
  - h(v) - the hubness of v
- A site is very authoritative if it receives many citations. Citation from important sites weight more than citations from less-important sites
- Hubness shows the importance of a site. A good hub is a site that links to many authoritative sites

HITS: Constructing the Query graph

Subgraph(σ,$\mathcal{E}$,t,d)

σ: a query string.

$\mathcal{E}$: a text-based search engine.

t, d: natural numbers.

Let R_σ denote the top t results of $\mathcal{E}$ on σ.

Set $S_\sigma := R_\sigma$

For each page $p \in R_\sigma$

Let $\Gamma^+(p)$ denote the set of all pages p points to.

Let $\Gamma^-(p)$ denote the set of all pages pointing to p.

Add all pages in $\Gamma^+(p)$ to S_σ.

If $|\Gamma^-(p)| \leq d$ then

Add all pages in $\Gamma^-(p)$ to S_σ.

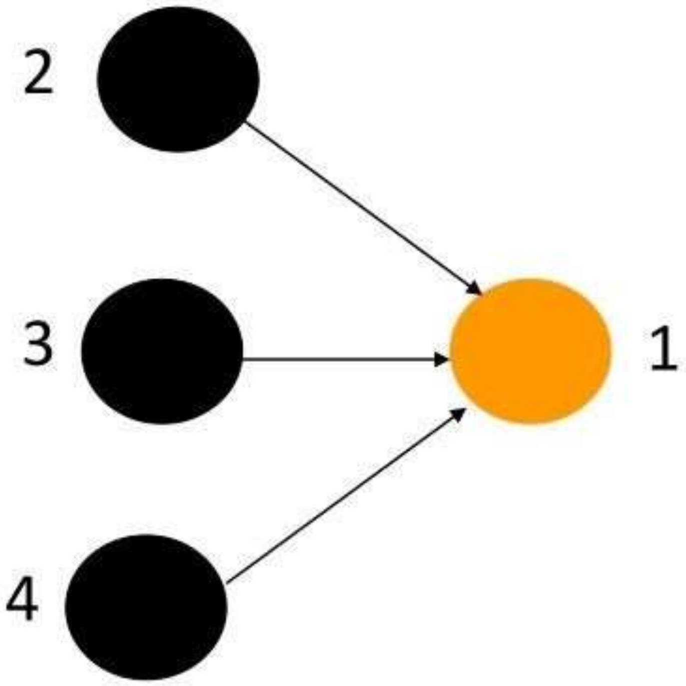
Else

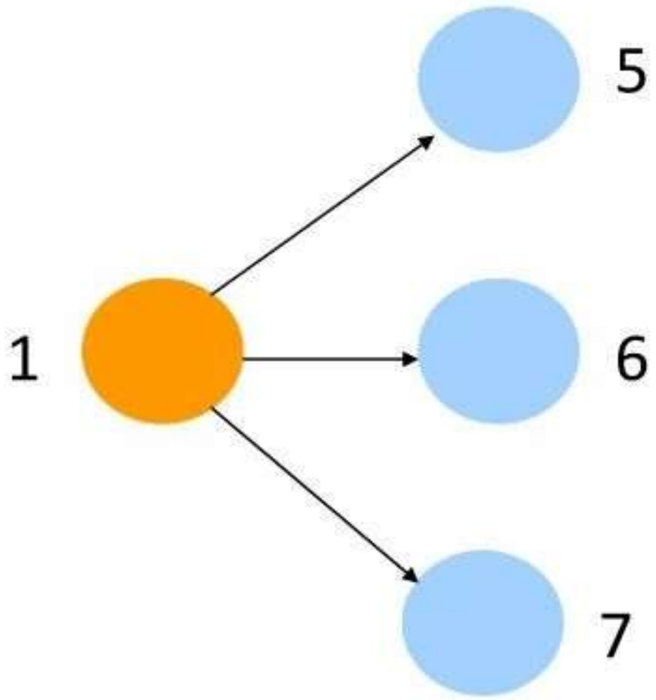
Add an arbitrary set of d pages from $\Gamma^-(p)$ to S_σ.

End

Return S_σ

Authorities and Hubs





$a(1) = h(2) + h(3) + h(4)$

$h(1) = a(5) + a(6) + a(7)$

The Markov Chain

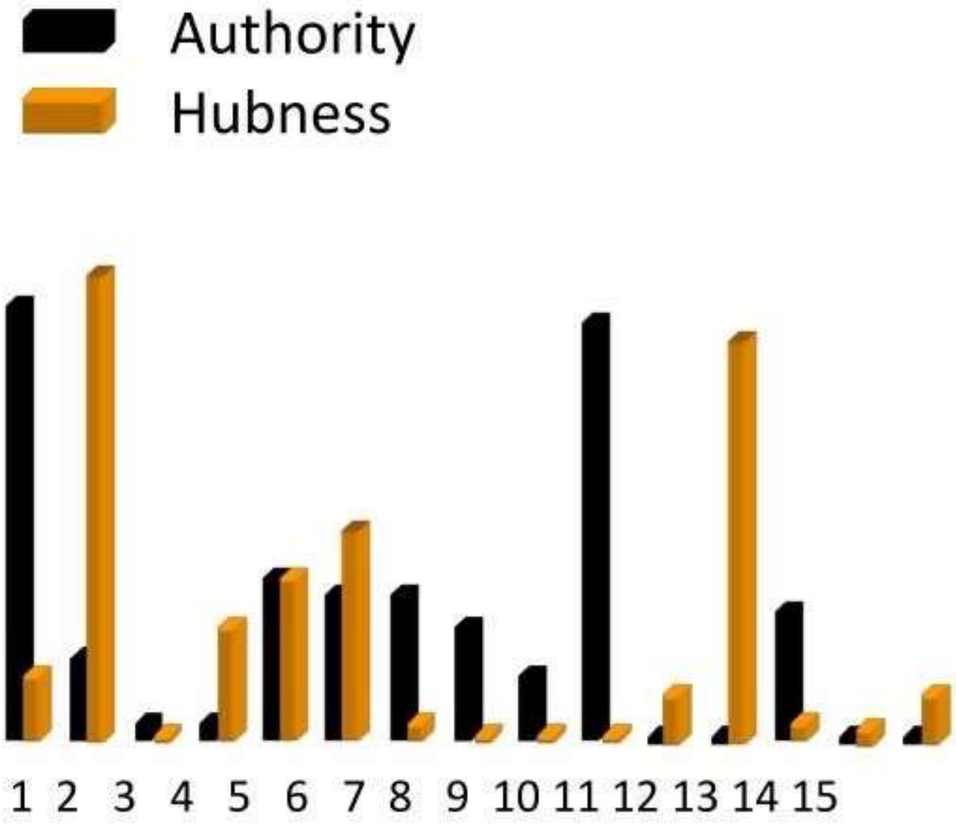
- Recursive dependency:

$$a(v) \leftarrow \sum_{w \in pa[v]} h(w)$$

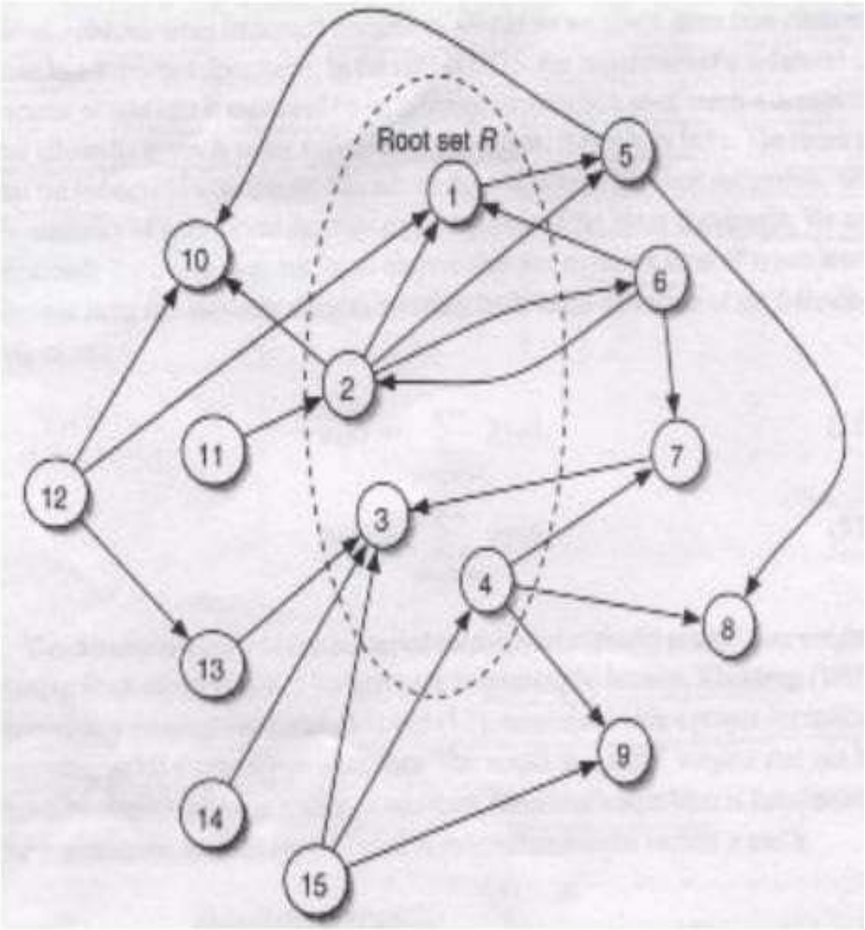
$$h(v) \leftarrow \sum_{w \in ch[v]} a(w)$$

Can you prove that it will converge?

HITS: Example



Authority and hubness weights



Limitations of HITS

- Sink problem: Solved
- Disconnectedness: an issue
- Convergence: Not a problem
- Stability: Quite robust

- You can still fool HITS easily!
  - Tightly Knit Community (TKC) Effect

Acknowledgements

- Some slides of these lectures are from:
  - *Random Walks on Graphs: An Overview*
    Purnamitra Sarkar
  - "Link Analysis Slides" from the book
    Modeling the Internet and the Web
    Pierre Baldi, Paolo Frasconi, Padhraic Smyth

References

- Basics of Random Walk:
  - L. Lovasz (1993) Random Walks on Graphs: A Survey
- PageRank:
  - http://en.wikipedia.org/wiki/PageRank
  - K. Bryan and T. Leise, The $25,000,000 Eigenvector: The Linear Algebra Behind Google (www.rose-hulman.edu/~bryan)
- HITS
  - J. M. Kleinberg (1999) Authorative Sources in a Hyperlinked Environment. *Journal of the ACM* **46** (5): 604–632.

HITS on Citation Network

- *$A = W^TW$ is the co-citation matrix*
  - What is *A*[i][j]?
- *$H = WW^T$ is the bibliographic coupling matrix*
  - What is H[i][j]?

- H. Small, Co-citation in the scientific literature: a new measure of the relationship between two documents, *Journal of the American Society for Information Science* **24** (1973) 265–269.
- M.M. Kessler, Bibliographic coupling between scientific papers, *American Documentation* **14** (1963) 10–25.

SALSA: The Stochastic Approach for Link-Structure Analysis

- Probabilistic extension of the HITS algorithm
- Random walk is carried out by following hyperlinks both in the forward and in the backward direction
- Two separate random walks
  - Hub walk
  - Authority walk
- R. Lempel and S. Moran (2000) The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks* 33 387-401

The basic idea

- Hub walk
  - Follow a Web link from a page u_h to a page w_a (a forward link) and then
  - Immediately traverse a backlink going from w_a to v_h, where $(u,w) \in E$ and $(v,w) \in E$
- Authority Walk
  - Follow a Web link from a page w(a) to a page u(h) (a backward link) and then
  - Immediately traverse a forward link going back from v_h to w_a where $(u,w) \in E$ and $(v,w) \in E$

Analyzing SALSA

(1) *The hub matrix* $\tilde{H}$, defined as follows:

$$\tilde{h}_{i,j} = \sum_{k|(i_h,k_a),(j_h,k_a)\in\tilde{G}} \frac{1}{\deg(i_h)} \times \frac{1}{\deg(k_a)}.$$

(2) *The authority matrix* $\tilde{A}$, defined as follows:

$$\tilde{a}_{i,j} = \sum_{k|(k_h,i_a),(k_h,j_a)\in\tilde{G}} \frac{1}{\deg(i_a)} \times \frac{1}{\deg(k_h)}.$$

Analyzing SALSA

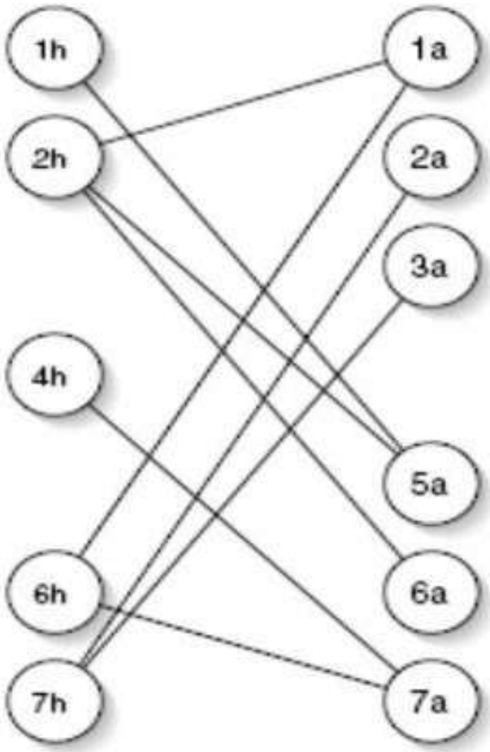
Hub Matrix: $\tilde{H} = W_r W_c^{\mathrm{T}}$

Authority Matrix: $\tilde{A} = W_c^{\mathrm{T}} W_r$

$$d_{\text{in}}(i) \triangleq \sum_{k \in H | k \to i} w(k \to i).$$

$$d_{\text{out}}(k) \triangleq \sum_{i \in A | k \to i} w(k \to i).$$

$$\mathcal{W} = \sum_{i \in A} d_{\text{in}}(i) = \sum_{k \in H} d_{\text{out}}(k).$$



SALSA ranks are degrees!

Proposition 1. *Whenever M_A is an irreducible chain (has a single irreducible component), it has a unique stationary distribution $\pi = (\pi_1, \ldots, \pi_{|A|})$ satisfying:*

$$\pi_i = \frac{d_{\text{in}}(i)}{\mathcal{W}} \text{ for all } i \in A.$$

Similarly, whenever M_H is an irreducible chain, its unique stationary distribution $\pi = (\pi_1, \ldots, \pi_{|H|})$ satisfies:

$$\pi_k = \frac{d_{\text{out}}(k)}{\mathcal{W}} \text{ for all } k \in H.$$

Is it good?

- It can be shown theoretically that SALSA does a better job than HITS in the presence of TKC effect
- However, it also has its own limitations

- Link Analysis: Which links (directed edges) in a network should be given more weight during the random walk?
  - An active area of research

Limits of Link Analysis (in IR)

- META tags/ invisible text
  - Search engines relying on meta tags in documents are often misled (intentionally) by web developers
- Pay-for-place
  - Search engine bias : organizations pay search engines and page rank
  - Advertisements: organizations pay high ranking pages for advertising space
    - With a primary effect of increased visibility to end users and a secondary effect of increased respectability due to relevance to high ranking page

Limits of Link Analysis (in IR)

- Stability
  - Adding even a small number of nodes/edges to the graph has a significant impact
- Topic drift – similar to TKC
  - A top authority may be a hub of pages on a different topic resulting in increased rank of the authority page
- Content evolution
  - Adding/removing links/content can affect the intuitive authority rank of a page requiring recalculation of page ranks

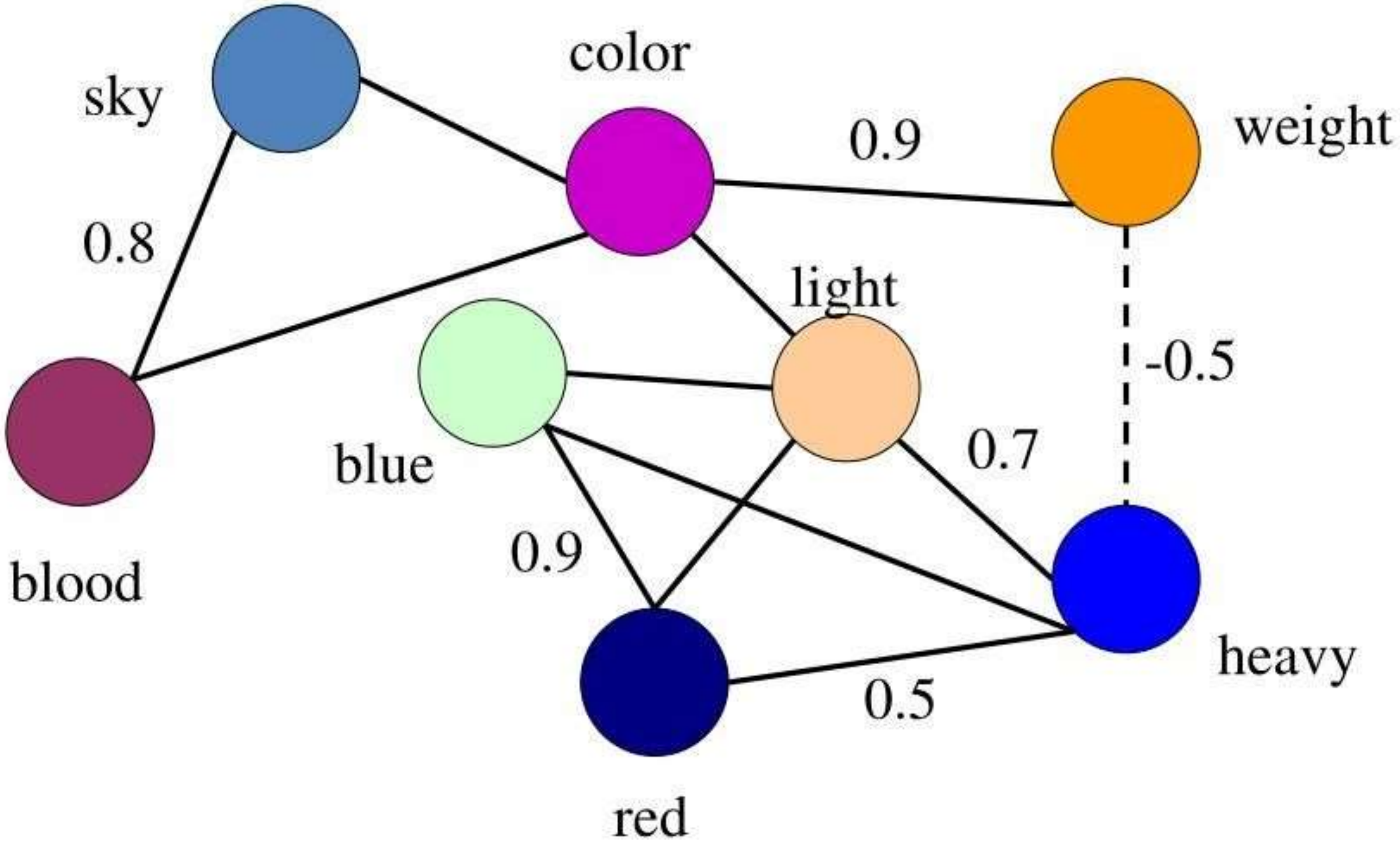
Clustering Using Random Walk

Chinese Whispers

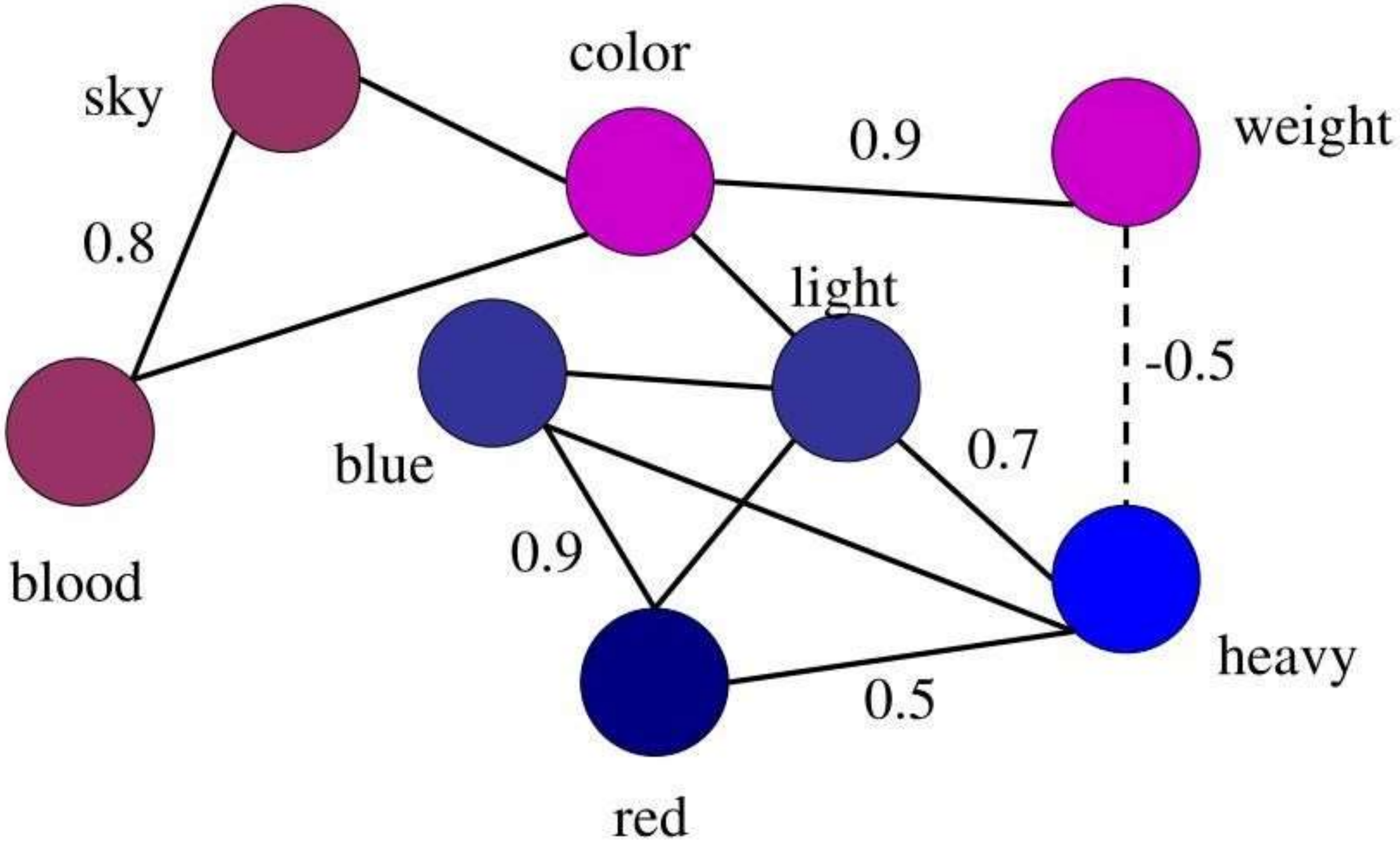
- C. Biemann (2006) Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proc of HLT-NAACL'06 workshop on TextGraphs*, pages 73–80

- Based on the game of "Chinese Whispers"

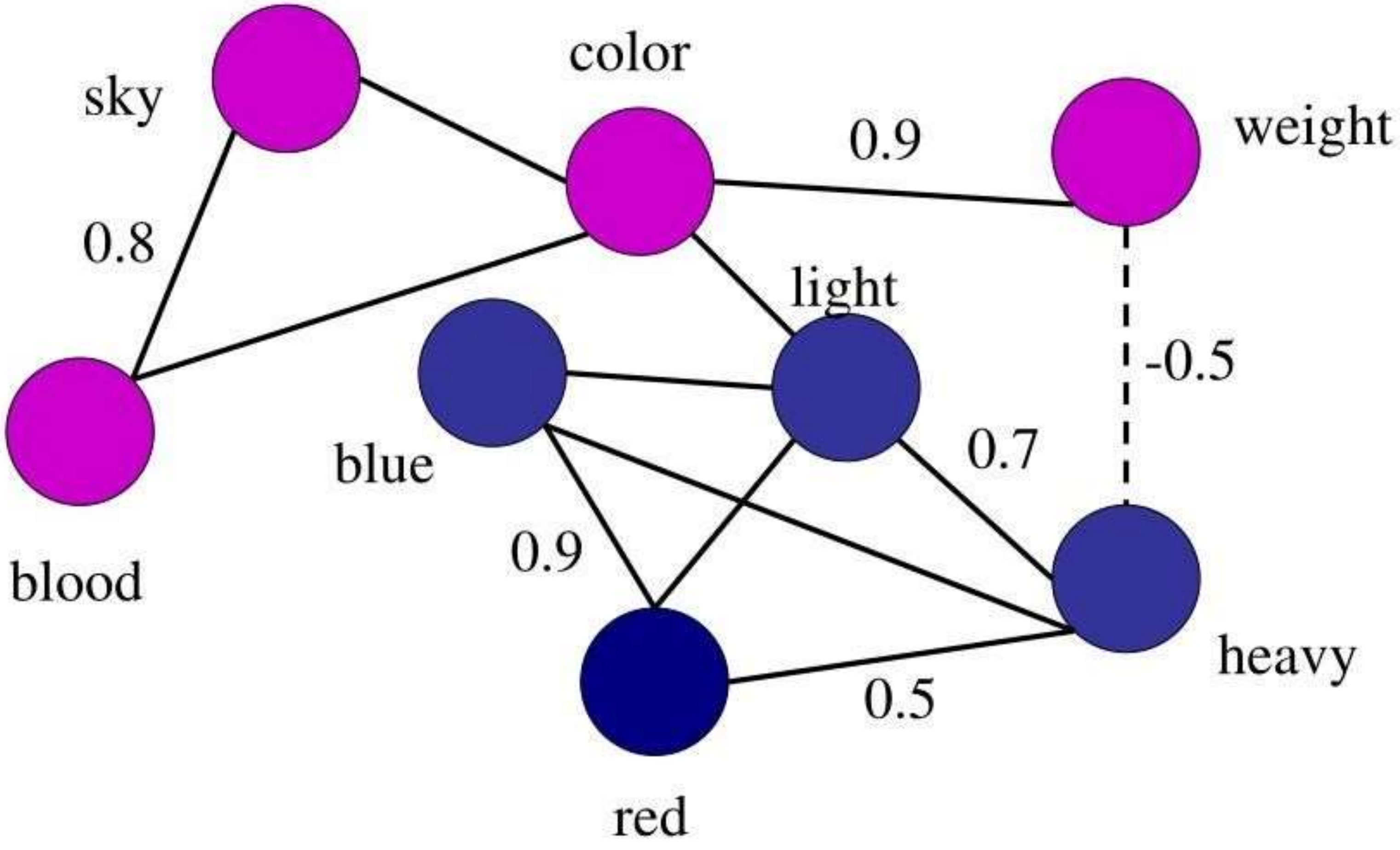
The Chinese Whispers Algorithm



The Chinese Whispers Algorithm



The Chinese Whispers Algorithm



Properties

- No parameters!
- Number of clusters?
- Does it converge for all graphs?
- How fast does it converge?
- What is the basis of clustering?

Affinity Propagation

- B.J. Frey and D. Dueck (2007) Clustering by Passing Messages Between Data Points. *Science* **315,** 972
- Choosing exemplars through real-valued message passing:
  - Responsibilities
  - Availabilities

Input

- n points (nodes)
- Similarity between them: s(*i*,*k*)
  - How suitable an exemplar *k* is for *i*.
- s(*k*,*k*) = how likely it is for k to be an exemplar

Thank You