DATA MINING
LINK ANALYSIS RANKING

PageRank -- Random walks

The HITS algorithm

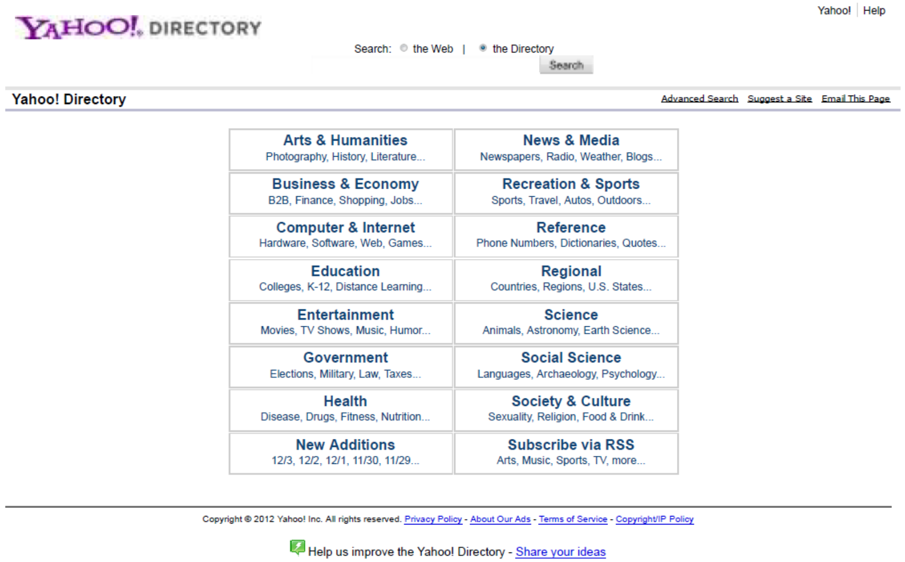
- Network Earth
  - https://www.bilibili.com/video/BV1ss41117Tg
  - https://www.youtube.com/watch?v=xZ3OmlbtaMU

Network Science

- A number of complex systems can be modeled as networks (graphs).
  - The Web
  - Social Networks
  - Biological systems
  - Communication networks (internet, email)
  - The Economy
  - Citation network
- We cannot truly understand such complex systems unless we understand the underlying network.
  - Everything is connected, studying individual entities gives only a partial view of a system
- Data mining for networks is a very popular area
  - Applications to the Web is one of the success stories for network data mining.

How to organize the web

- First try: Manually curated Web Directories



How to organize the web

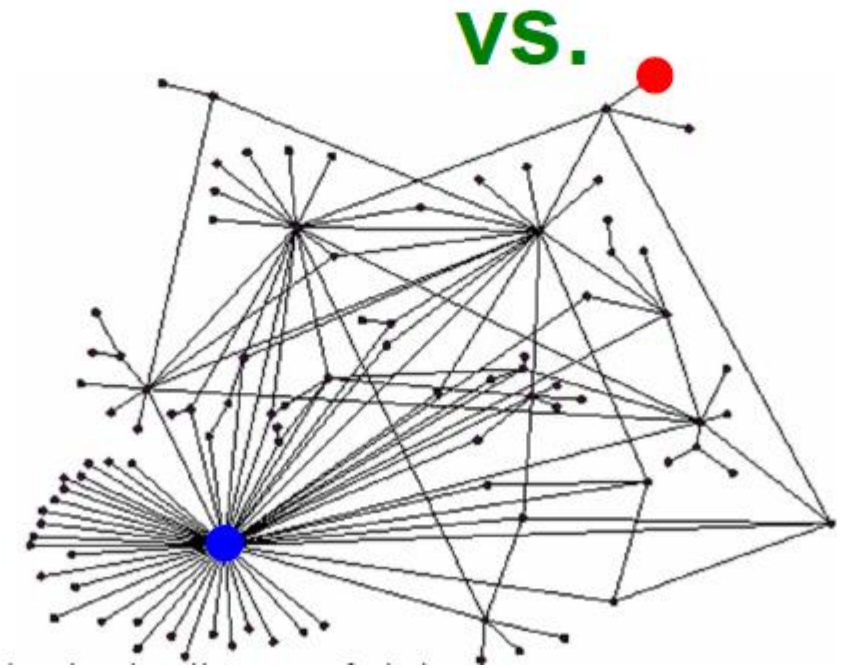
- Second try: Web Search
  - Information Retrieval investigates:
    - Find relevant docs in a small and trusted set e.g., Newspaper articles, Patents, etc. (“needle-in-a-haystack”)
    - Based on textual and semantic similarities
  - But: Web is huge, full of untrusted documents, random things, web spam, etc.
    - Everyone can create a web page of high production value
    - Rich diversity of people issuing queries
    - Dynamic and constantly-changing nature of web content

How to organize the web

- Third try (the Google era): using the web graph
  - Shift from relevance to authoritativeness
  - It is not only important that a page is relevant, but that it is also important on the web

Link Analysis Ranking

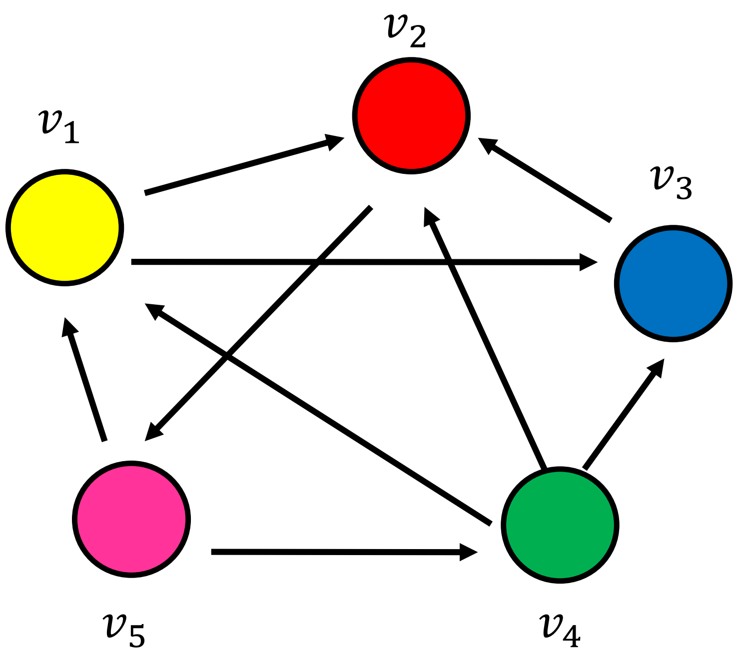
- Use graph structure to determine relative importance of nodes
  - Applications: Ranking on graphs (Web, social media, etc)
- Intuition: An edge from node p to node q denotes endorsement
  - Node p endorses/recommends/confirms the authority/centrality/importance of node q
  - Use the graph of recommendations to assign an authority value to every node



What is the simplest way to measure importance of a page on the web?

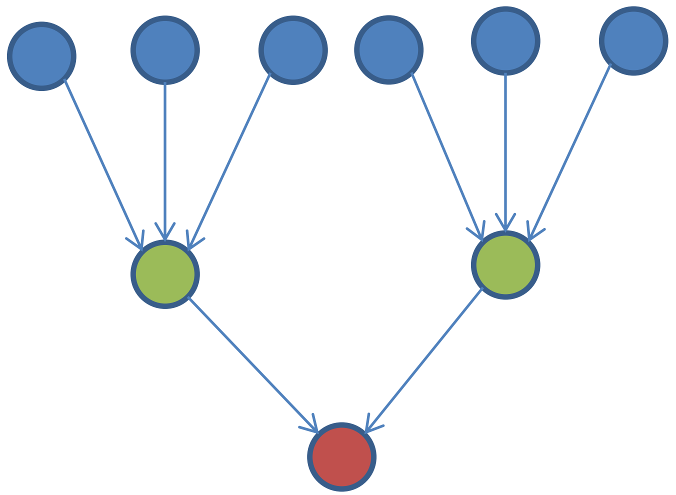
Rank by Popularity

- Rank pages according to the number of incoming edges (in-degree, degree centrality)



1. Red Page
2. Yellow Page
3. Blue Page
4. Purple Page
5. Green Page

Popularity



- It is not important only how many link to you, but how important are the people that link to you.
- Good authorities are pointed by good authorities
  - Recursive definition of importance

PAGERANK

The PageRank Citation Ranking: Bringing Order to the Web

by Larry Page and Sergey Brin and R. Motwani and T. Winograd, 1999

PageRank

- Good authorities should be pointed by good authorities
  - The value of a node is the value of the nodes that point to it.
- How do we implement that?
  - Assume that we have a unit of authority to distribute to all nodes.
  - Node i gets a fraction w_i of that authority weight
  - Each node distributes the authority value it has to its neighbors
  - The authority value of each node is the sum of the authority fractions it collects from its neighbors.

$$w_i = \sum_{j \to i} \frac{1}{|N_{out}(\mathrm{j})|} w_j$$

Recursive definition

Example

$$w_i = \sum_{j \to i} \frac{1}{|N_{out}(j)|} w_j$$

$w_1 = 1/3\ w_4 + 1/2\ w_5$

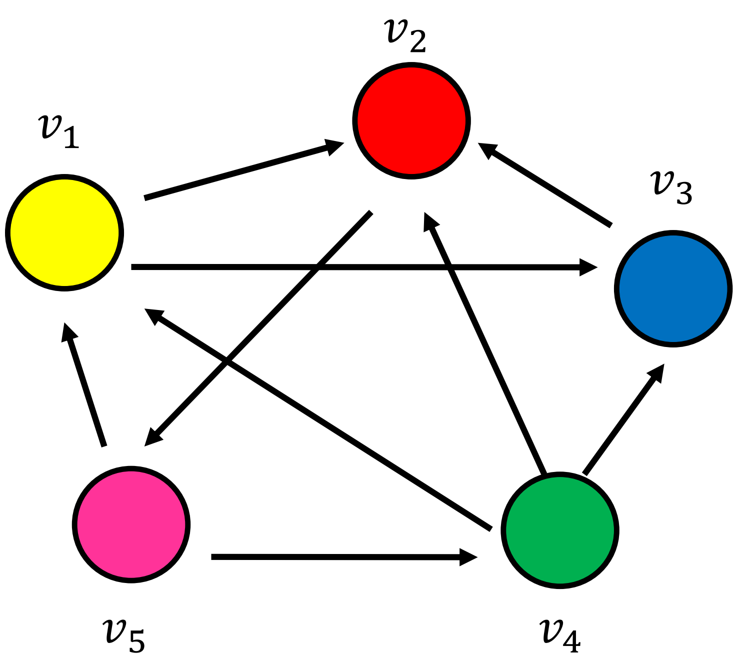
$w_2 = 1/2\ w_1 + w_3 + 1/3\ w_4$

$w_3 = 1/2\ w_1 + 1/3\ w_4$

$w_4 = 1/2\ w_5$

$w_5 = w_2$

$w_1 + w_2 + w_3 + w_4 + w_5 = 1$



We can obtain the weights by solving this system of equations

Computing PageRank weights

- A simpler way to compute the weights is by iteratively updating the weights using the equations
- PageRank Algorithm

Initialize all PageRank weights to $\mathrm{w}_{\mathrm{i}}^{0} = \frac{1}{n}$

Repeat:

$$w_i^t = \sum_{j \to i} \frac{1}{|N_{out}(j)|} w_j^{t-1}$$

Until the weights do not change

- This process converges

Example

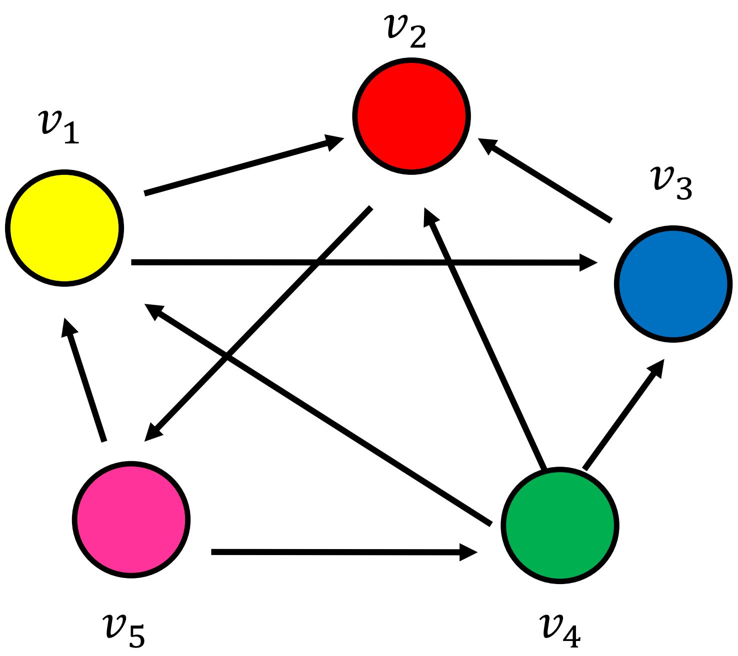
$w_1 = 1/3\ w_4 + 1/2\ w_5$

$w_2 = 1/2\ w_1 + w_3 + 1/3\ w_4$

$w_3 = 1/2\ w_1 + 1/3\ w_4$

$w_4 = 1/2\ w_5$

$w_5 = w_2$



	w_1	w_2	w_3	w_4	w_5
t=0	0.2	0.2	0.2	0.2	0.2
t=1	0.16	0.36	0.16	0.1	0.2
t=2	0.13	0.28	0.11	0.1	0.36
t=3	0.22	0.22	0.1	0.18	0.28
t=4	0.2	0.27	0.17	0.14	0.22

Example

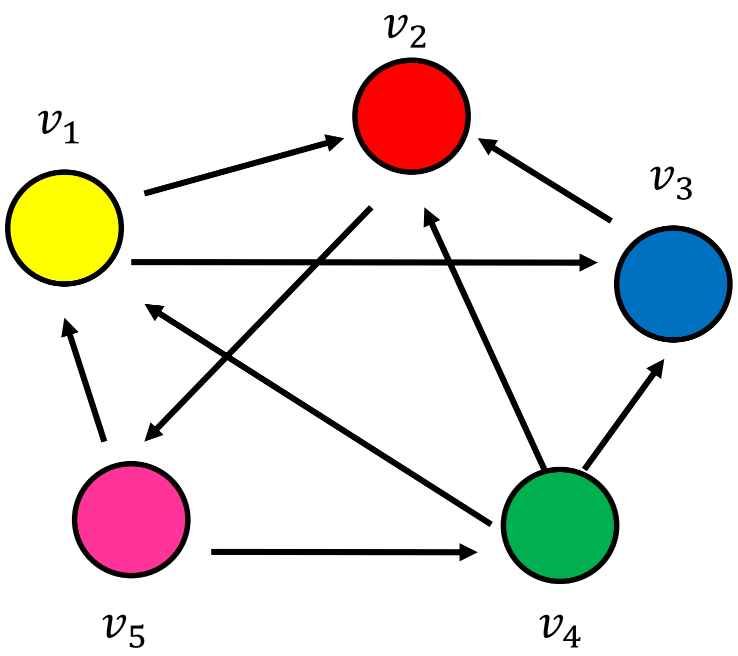
$w_1 = 1/3\ w_4 + 1/2\ w_5$

$w_2 = 1/2\ w_1 + w_3 + 1/3\ w_4$

$w_3 = 1/2\ w_1 + 1/3\ w_4$

$w_4 = 1/2\ w_5$

$w_5 = w_2$

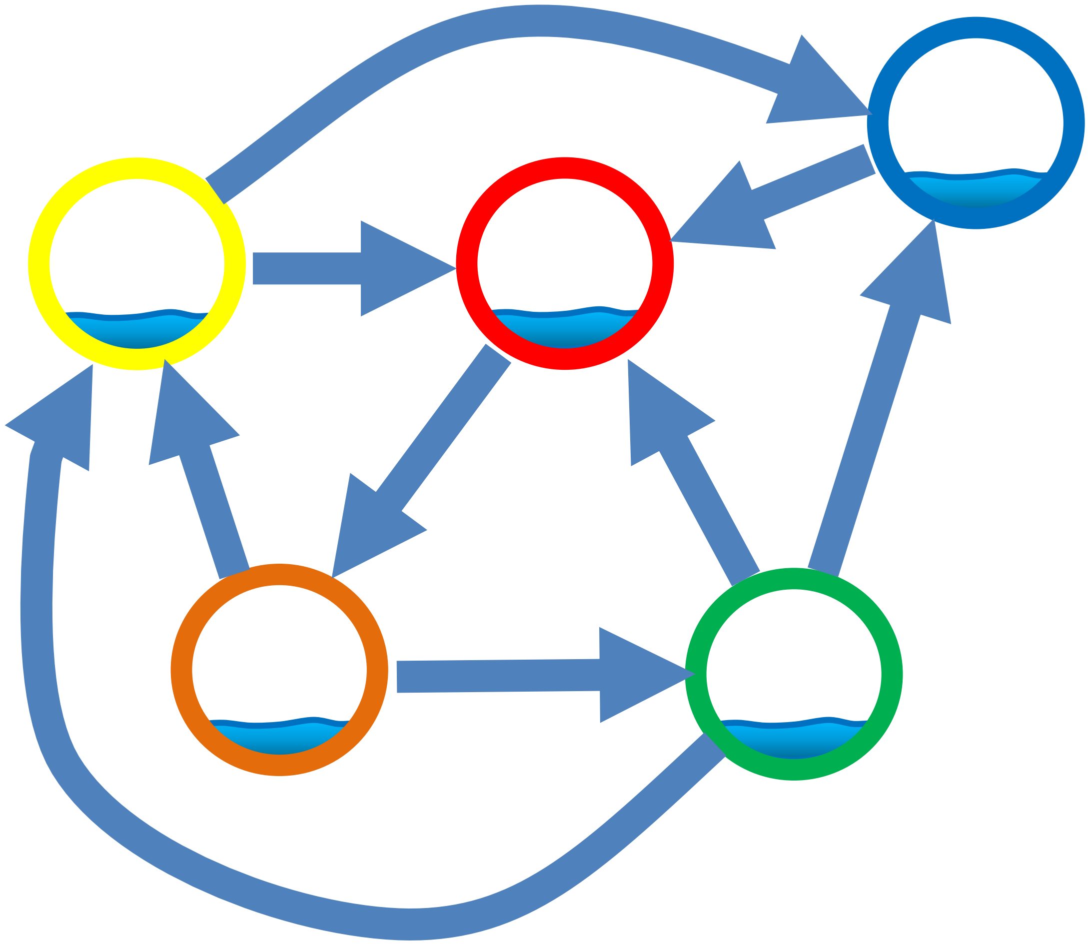


	w_1	w_2	w_3	w_4	w_5
t=25	0.18	0.27	0.13	0.13	0.27

The PageRank algorithm

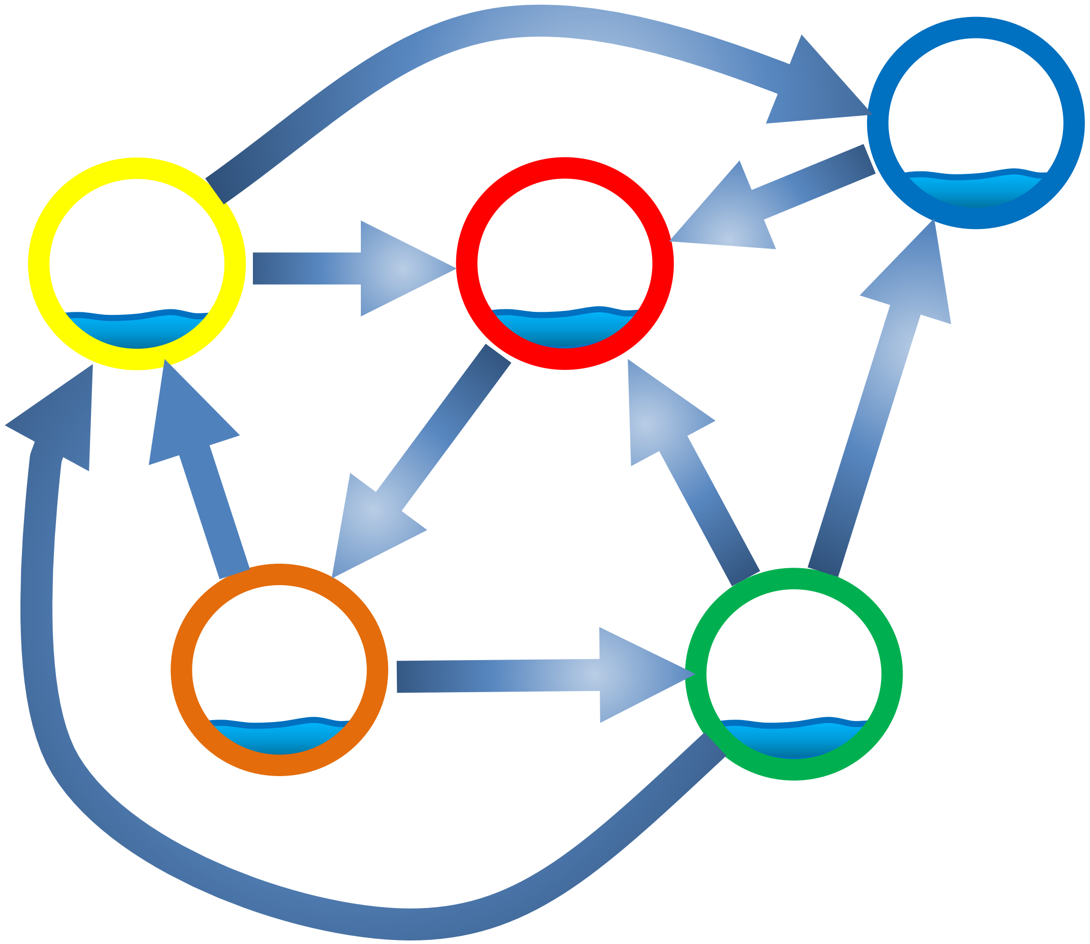
Think of the nodes in the graph as containers of capacity of 1 liter.

We distribute a liter of liquid equally to all containers



The PageRank algorithm

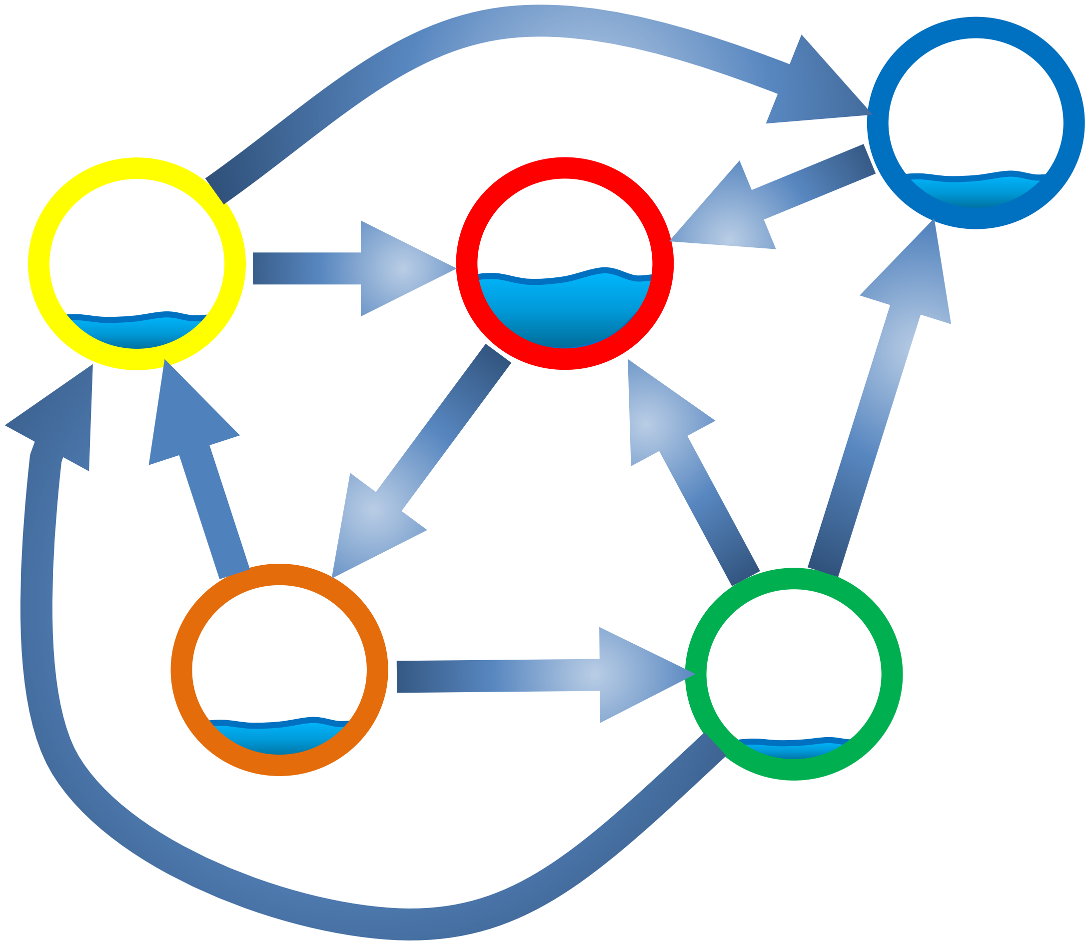
The edges act like pipes that transfer liquid between nodes.



The PageRank algorithm

The edges act like pipes that transfer liquid between nodes.

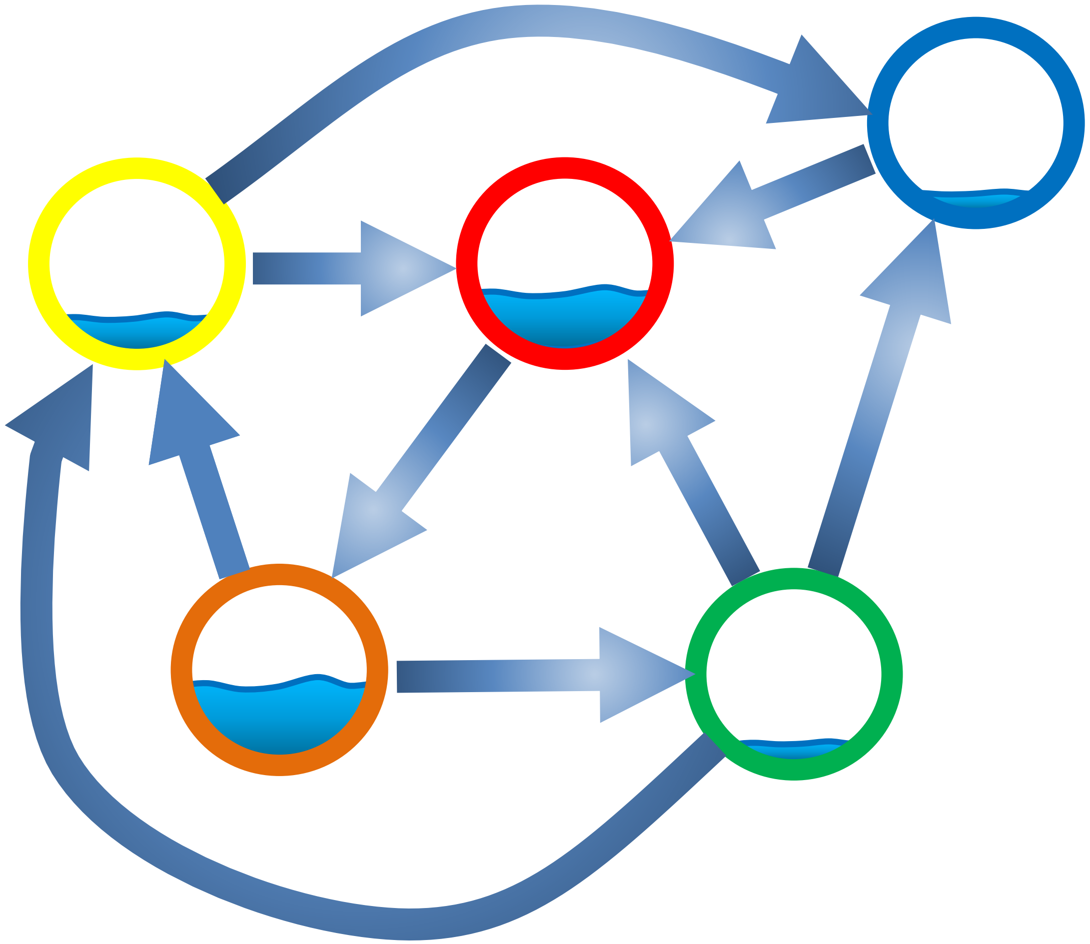
The contents of each node are distributed to its neighbors.



The PageRank algorithm

The edges act like pipes that transfer liquid between nodes.

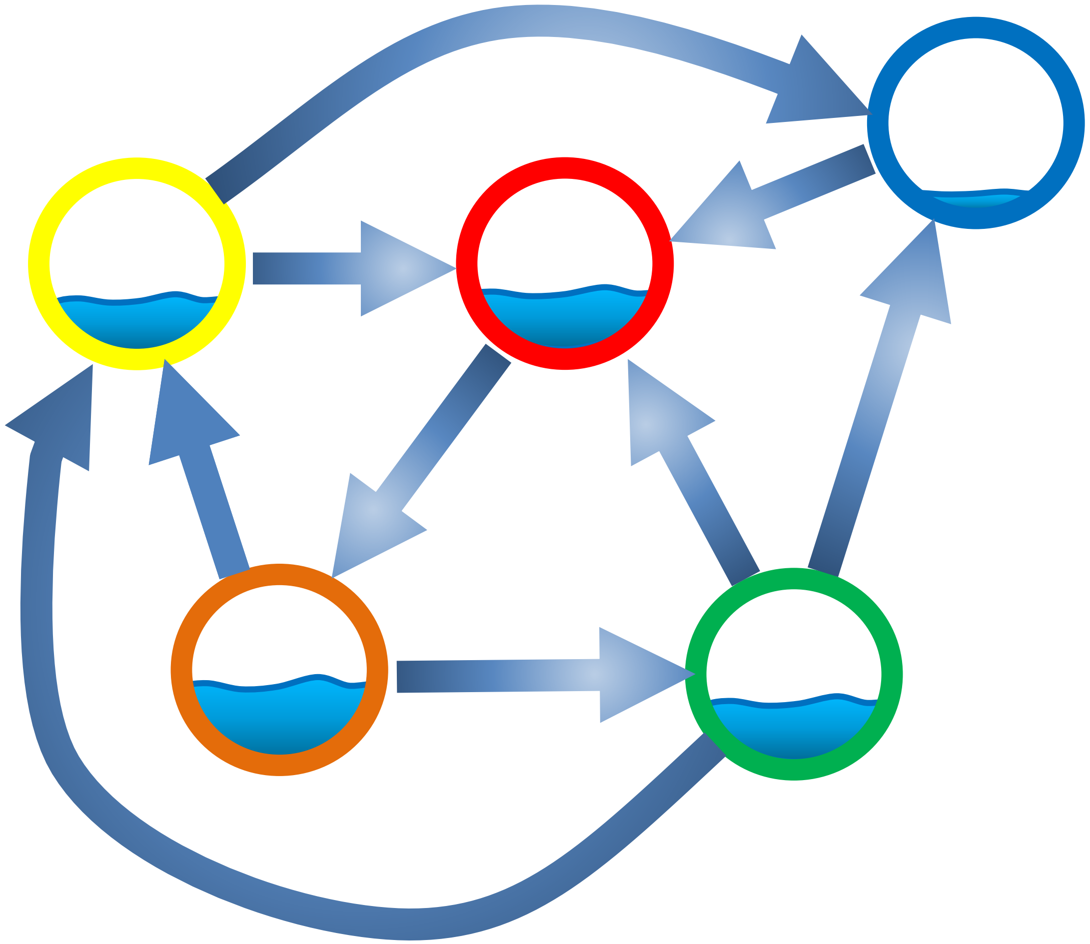
The contents of each node are distributed to its neighbors.



The PageRank algorithm

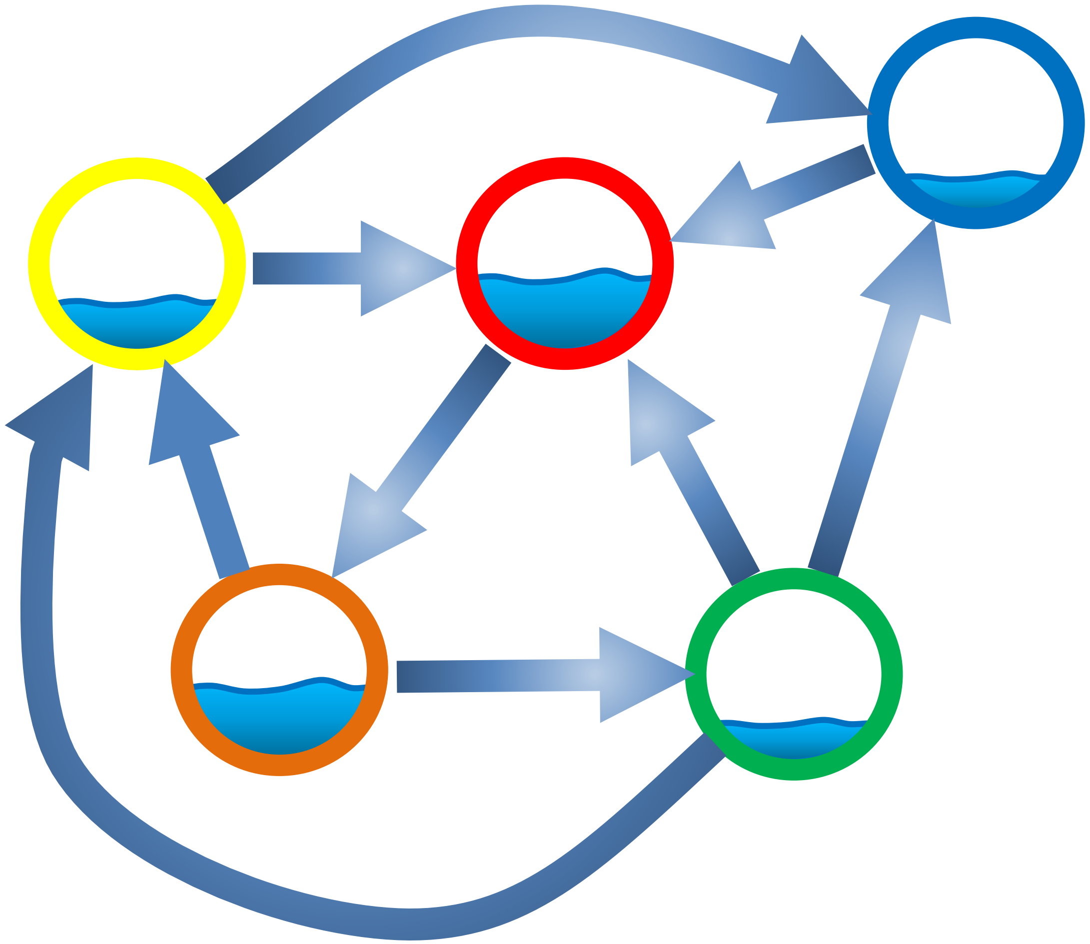
The edges act like pipes that transfer liquid between nodes.

The contents of each node are distributed to its neighbors.



The PageRank algorithm

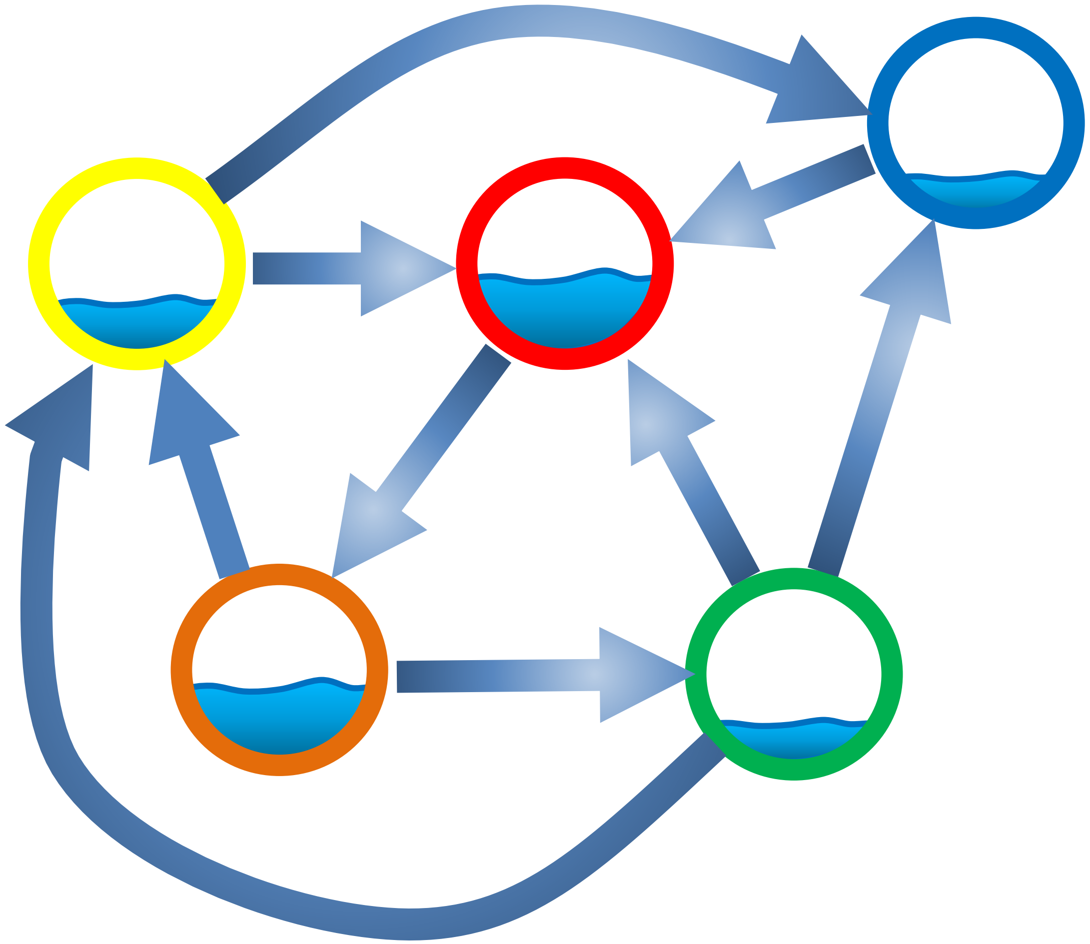
The system will reach an equilibrium state where the amount of liquid in each node remains constant.



The PageRank algorithm

The amount of liquid in each node determines the importance of the node.

Large quantity means large incoming flow from nodes with large quantity of liquid.



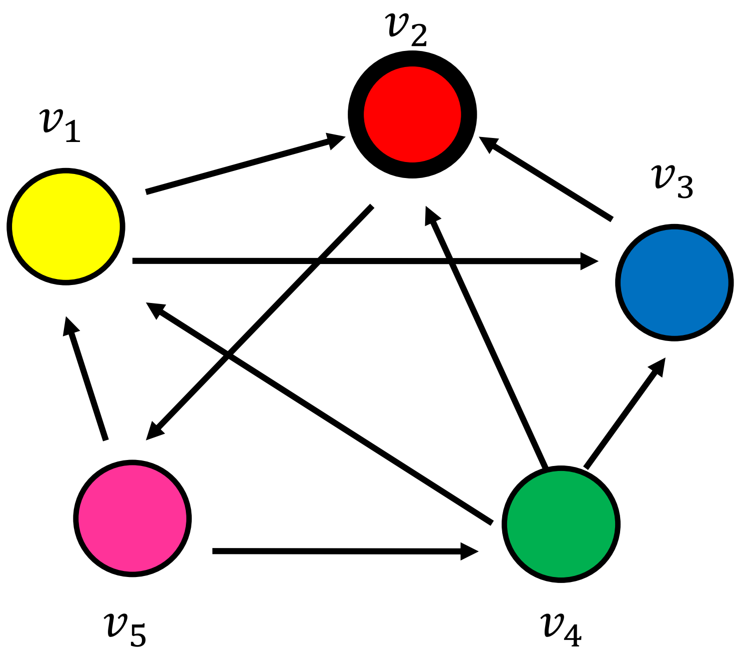
Random Walks on Graphs

- The algorithm defines a random walk on the graph

- Random walk:
  - Start from a node chosen uniformly at random with probability $\frac{1}{n}$.
    - Pick one of the outgoing edges uniformly at random
    - Move to the destination of the edge
    - Repeat.

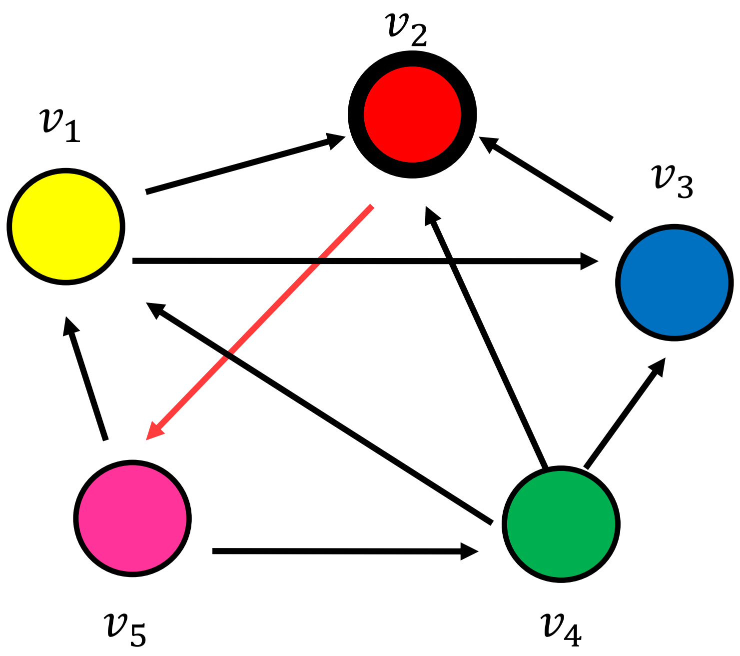
Example

- Step 0



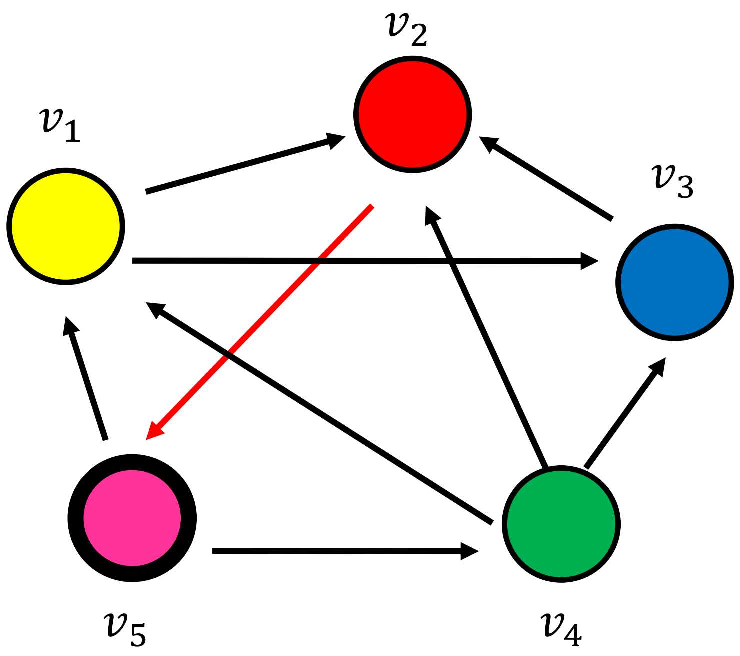
Example

- Step 0



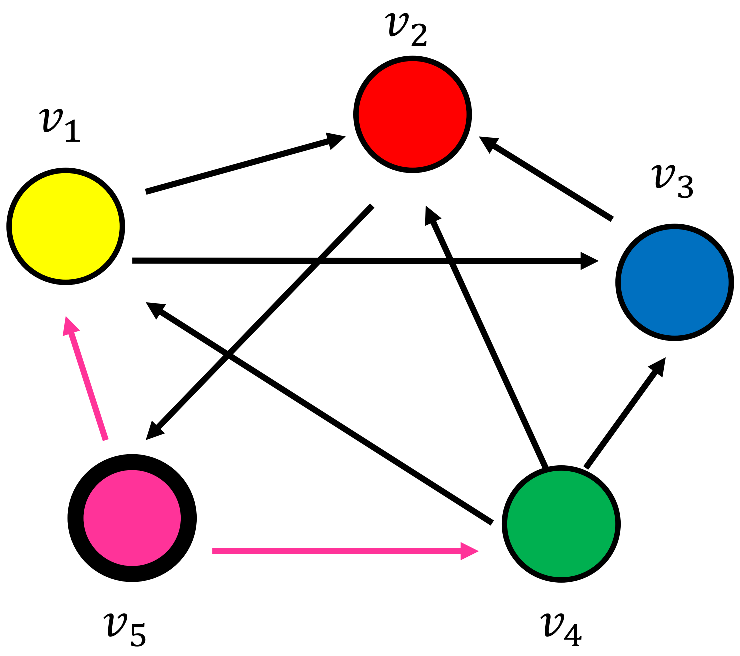
Example

- Step 1



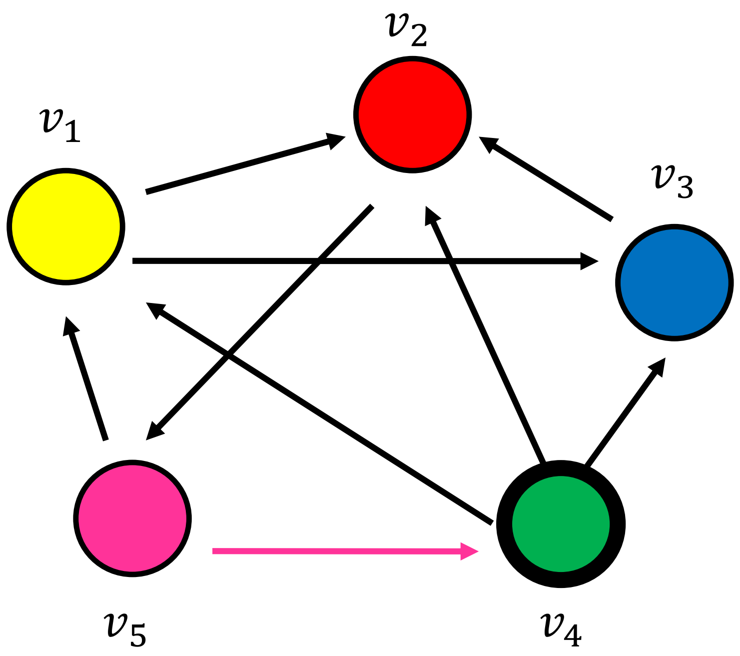
Example

- Step 1



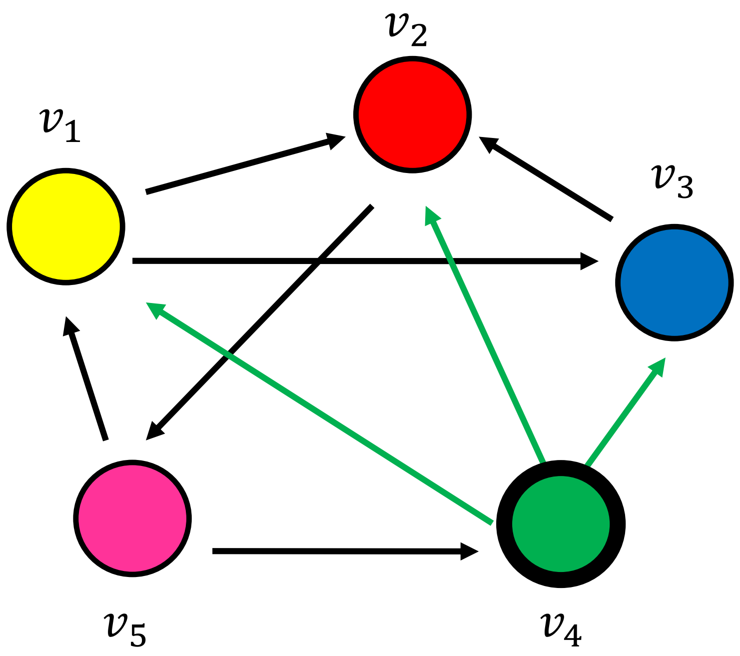
Example

- Step 2



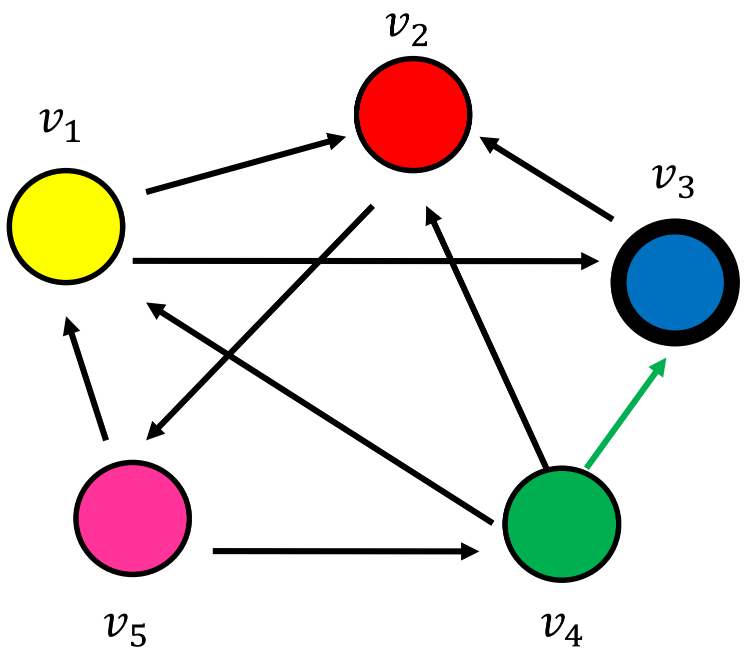
Example

- Step 2



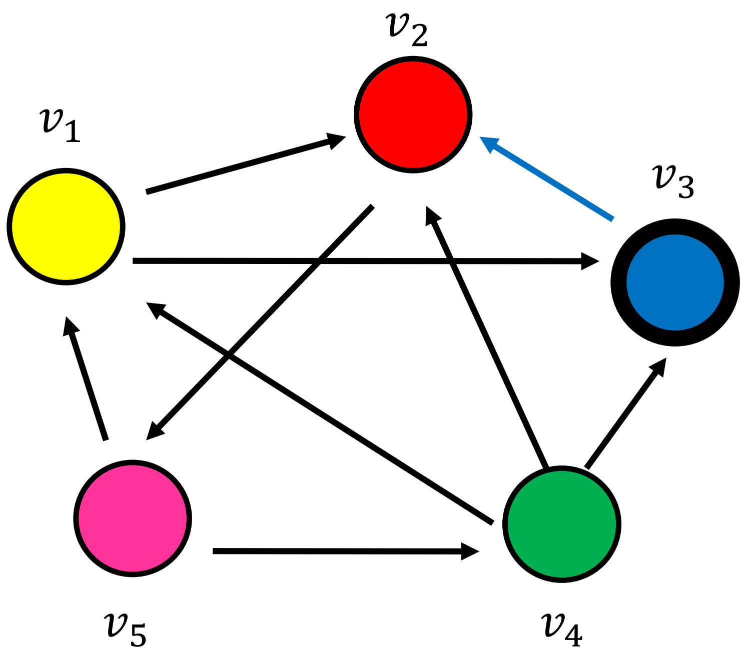
Example

- Step 3



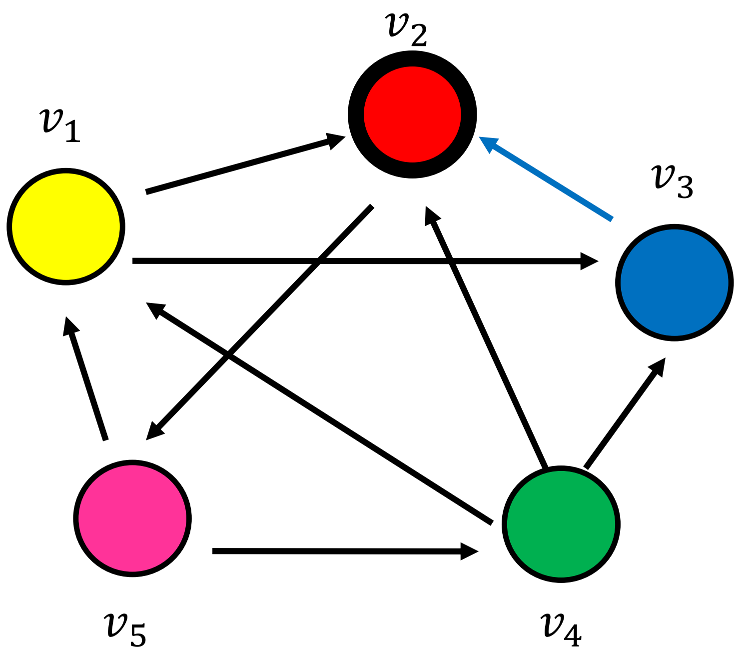
Example

- Step 3



Example

- Step 4…



Random walk

- Question: what is the probability p_i^t of being at node i after t steps?

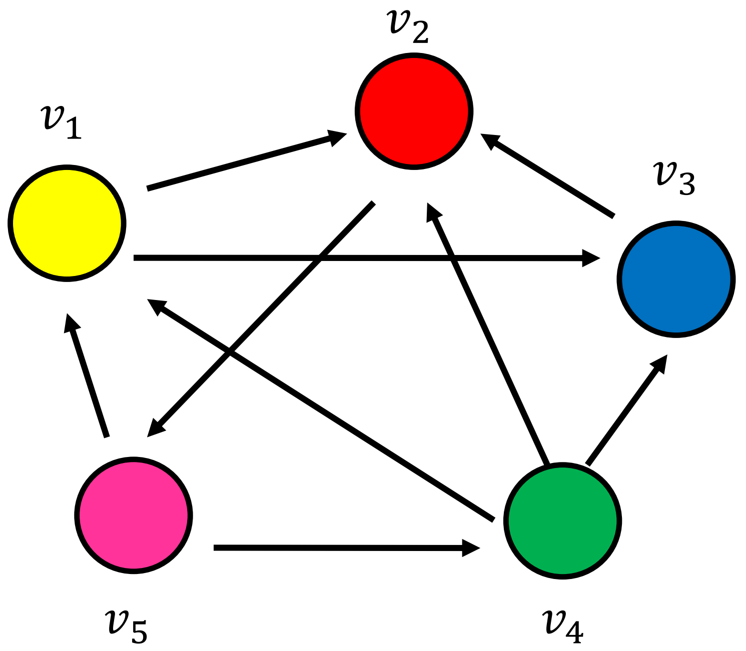
$$p_1^0 = \frac{1}{5} \qquad p_1^t = \frac{1}{3}p_4^{t-1} + \frac{1}{2}p_5^{t-1}$$

$$p_2^0 = \frac{1}{5} \qquad p_2^t = \frac{1}{2}p_1^{t-1} + p_3^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_3^0 = \frac{1}{5} \qquad p_3^t = \frac{1}{2}p_1^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_4^0 = \frac{1}{5} \qquad p_4^t = \frac{1}{2}p_5^{t-1}$$

$$p_5^0 = \frac{1}{5} \qquad p_5^t = p_2^{t-1}$$



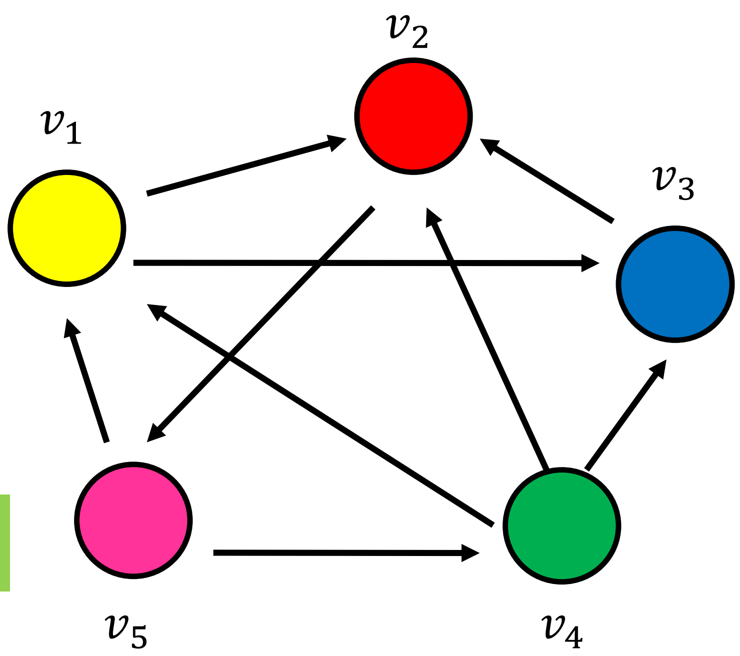
$$w_i^t = \sum_{j \to i} \frac{1}{|N_{out}(\mathrm{j})|} w_j^{t-1}$$

The equations are the same as those for the PageRank iterative computation

Random walk

- At convergence:

$$p_i = \sum_{j \to i} \frac{1}{|N_{out}(\mathrm{j})|} p_j^{t-1}$$



We get the same equation as for PageRank

The PageRank of node i is the probability that the random walk is at node i after a very large number of steps

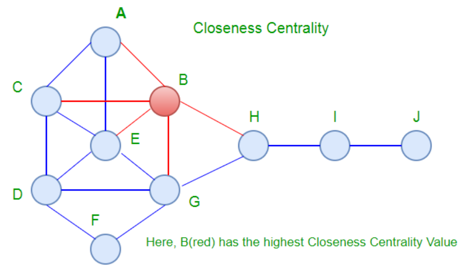
PageRank history

- Huge advantage for Google in the early days
  - It gave a way to get an idea for the value of a page, which was useful in many different ways
    - Put an order to the web.
  - After a while it became clear that the anchor text was probably more important for ranking
  - Also, link spam became a new (dark) art

OTHER ALGORITHMS

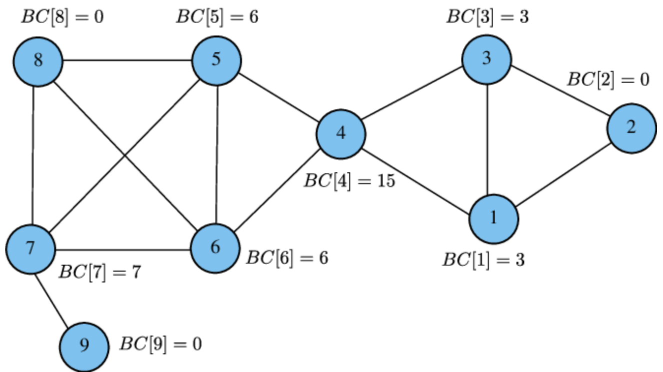
Social network analysis

- Evaluate the centrality of individuals in social networks
  - degree centrality
    - The (weighted) degree of a node
  - Distance centrality (closeness centrality)
    - The reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. The more central a node is, the closer it is to all other nodes.



$$D_c(v) = \frac{1}{\sum_{u \neq v} d(v,u)}$$

  - betweenness centrality
    - Represents the degree to which nodes stand between each other.



$$B_c(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

其中σ_{st}是节点s到节点t的最短路径之数量，而$\sigma_{st}(v)$这些路径经过v的次数。

THE HITS ALGORITHM

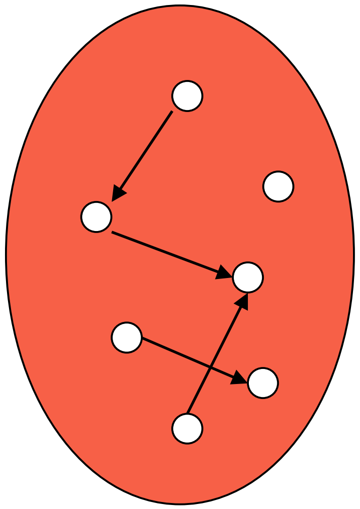
Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, *46*(5), pp.604-632.

The HITS algorithm

- Hyperlink-Induced Topic Search (HITS), another algorithm proposed around the same time as PageRank for using the hyperlinks to rank pages
  - Jon Kleinberg: then an intern at IBM Almaden
    - Member of the National Academy of Sciences, the National Academy of Engineering, and the American Academy of Arts and Sciences
  - IBM never made anything out of it

Query dependent input

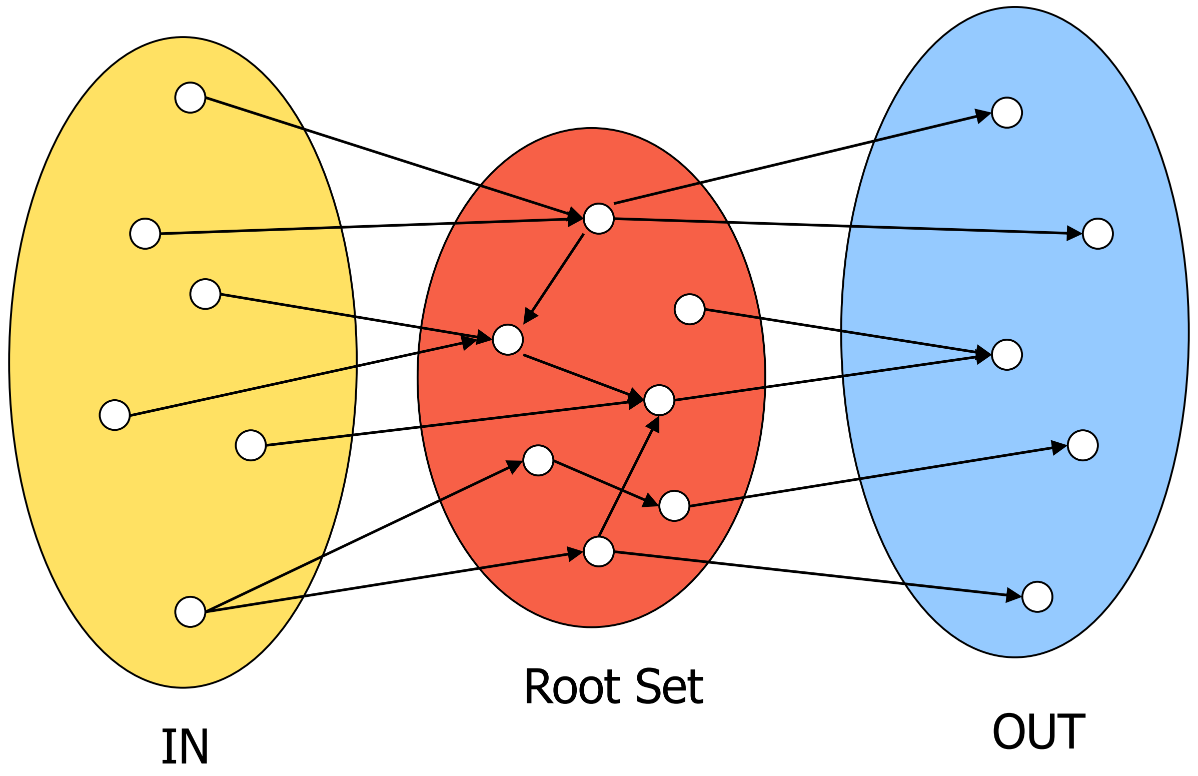
Root set obtained from a text-only search engine



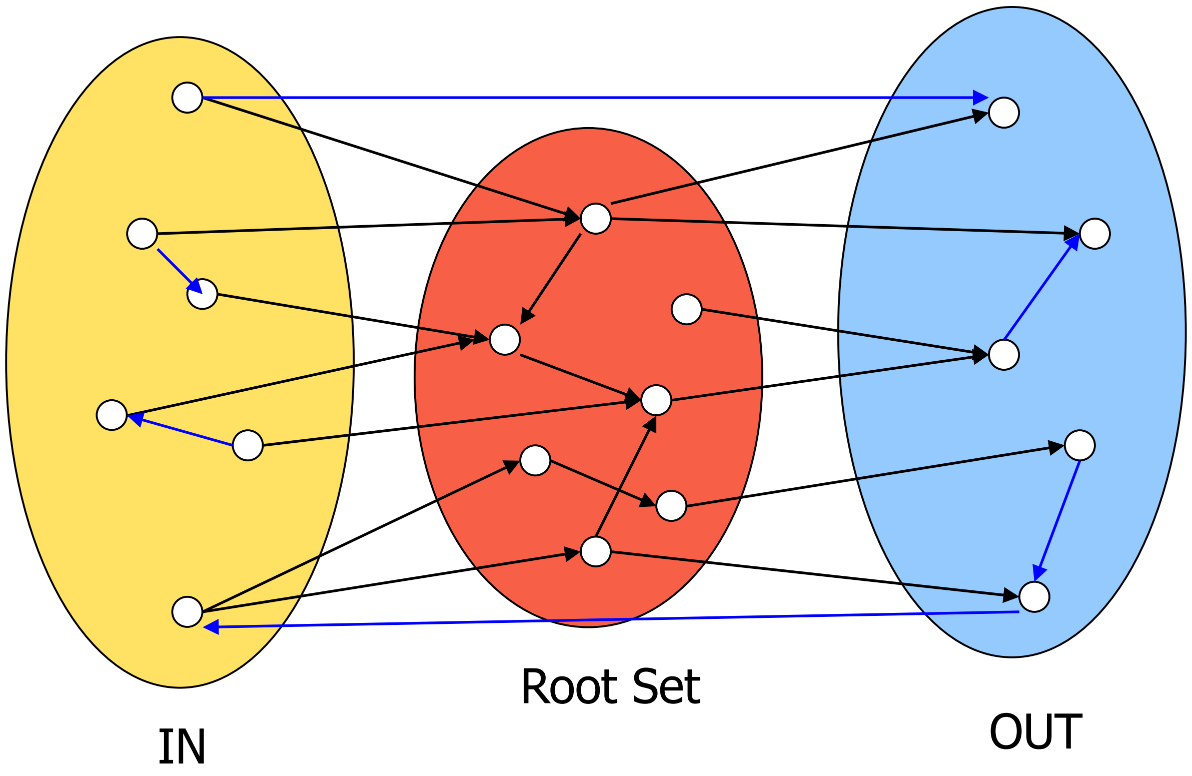
Root Set

(the search results of a given query, e.g. ‘shanghaitech’. Some resulting pages are linked.)

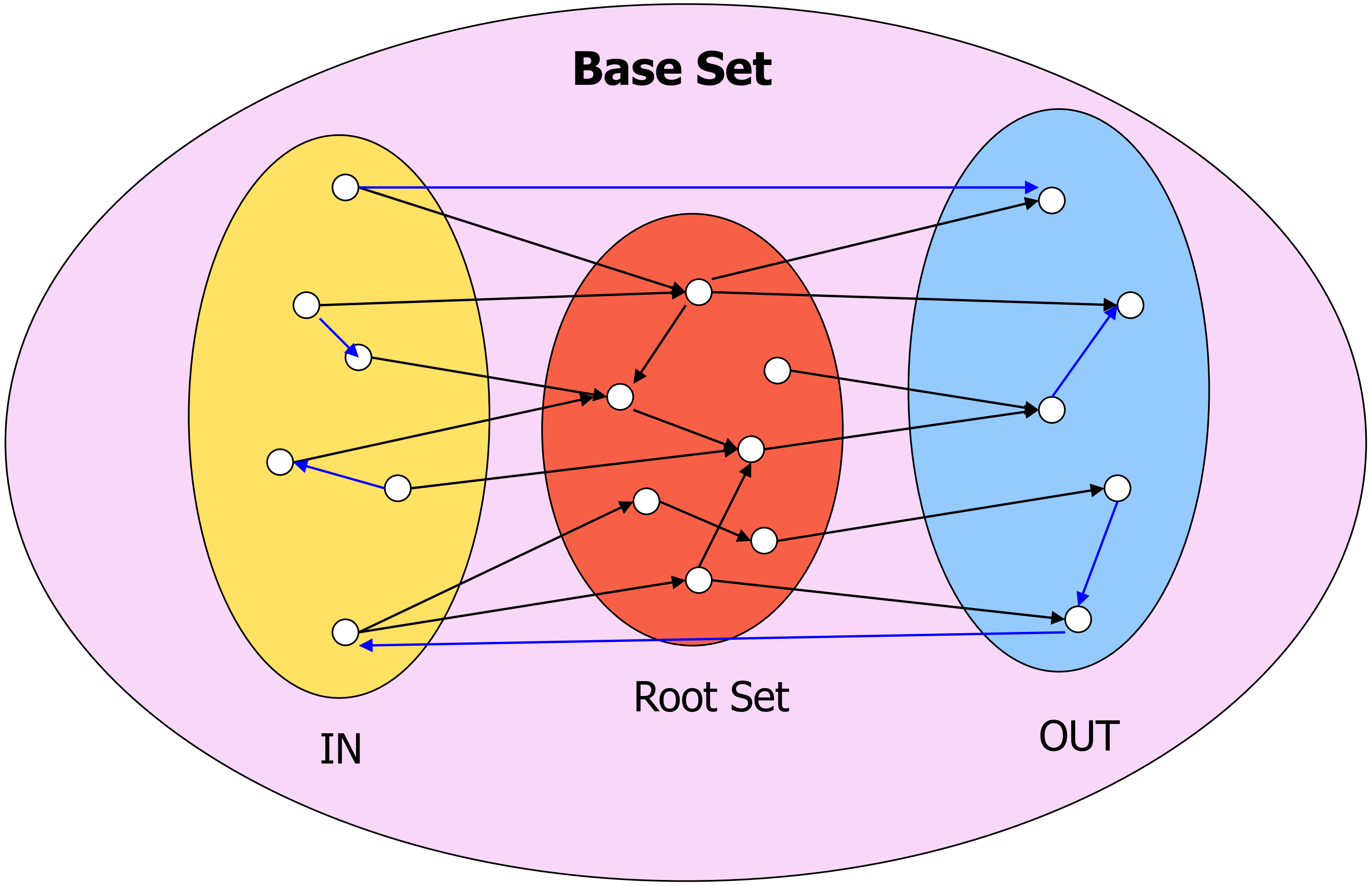
Query dependent input



Query dependent input

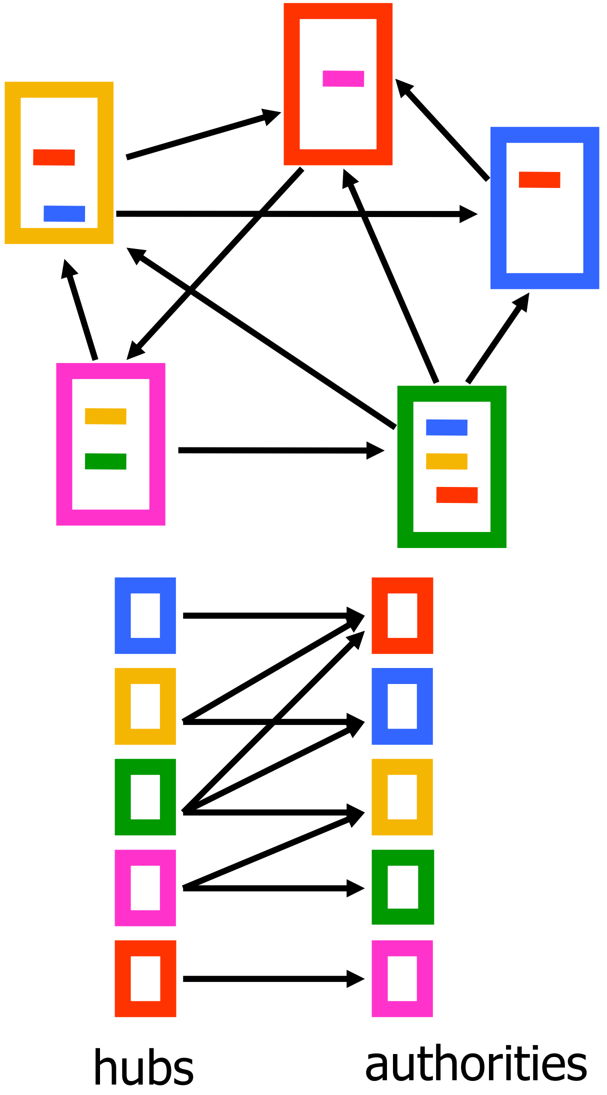


Query dependent input



Hubs and Authorities

- Authority is not necessarily transferred directly between authorities
- Pages have double identity
  - hub identity
  - authority identity
- Good hubs point to good authorities
- Good authorities are pointed by good hubs



Hubs and Authorities

- Two kind of weights:
  - Hub weight
  - Authority weight

- The hub weight is the sum of the authority weights of the authorities pointed to by the hub

- The authority weight is the sum of the hub weights that point to this authority.

HITS Algorithm

- Initialize all weights to 1.
- Repeat until convergence
  - O operation : hubs collect the weight of the authorities

$$h_i^t = \sum_{j:i\to j} a_j^{t-1}$$

  - I operation: authorities collect the weight of the hubs

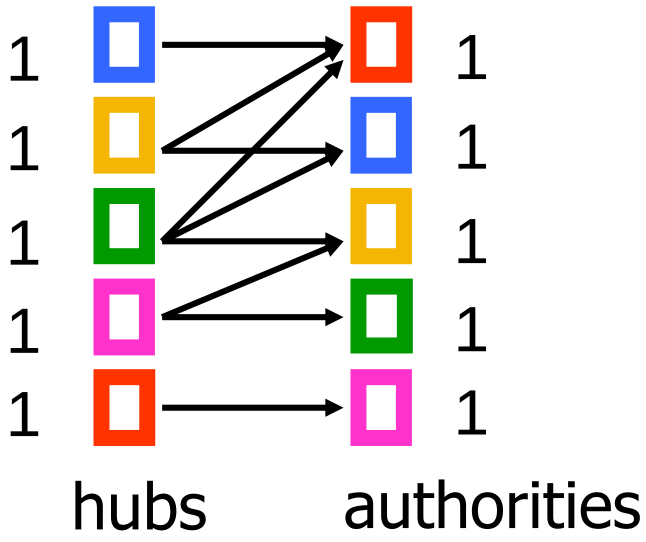
$$a_i^t = \sum_{j:j\to i} h_j^{t-1}$$

  - Normalize weights under some norm

The order of updates does not matter after many iterations.

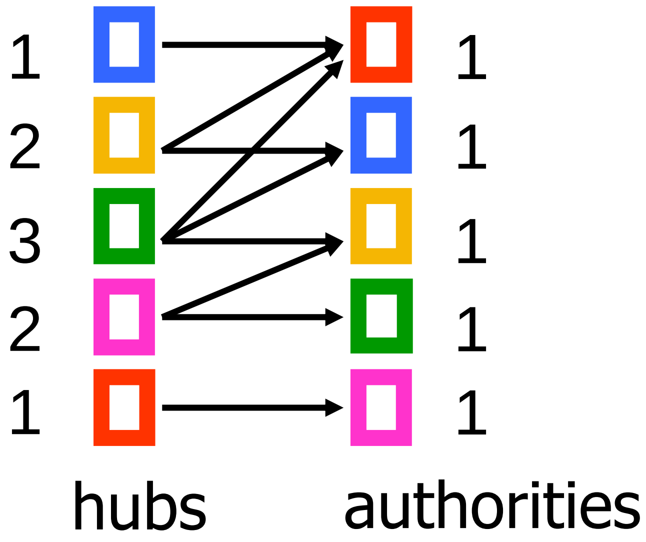
Example

Initialize



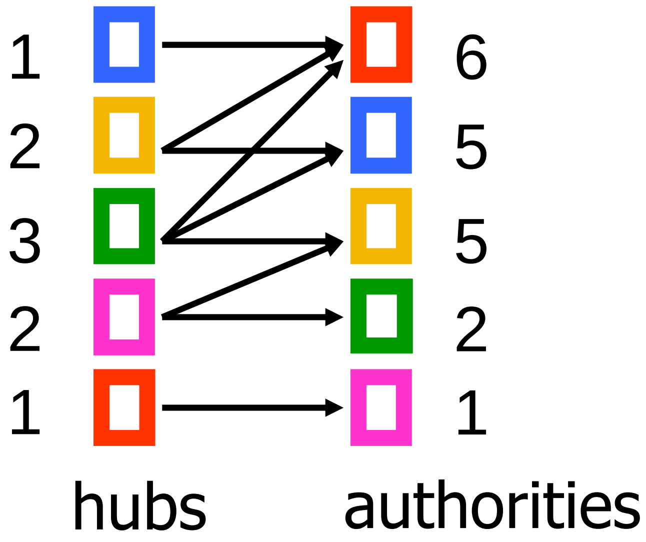
Example

Step 1: O operation



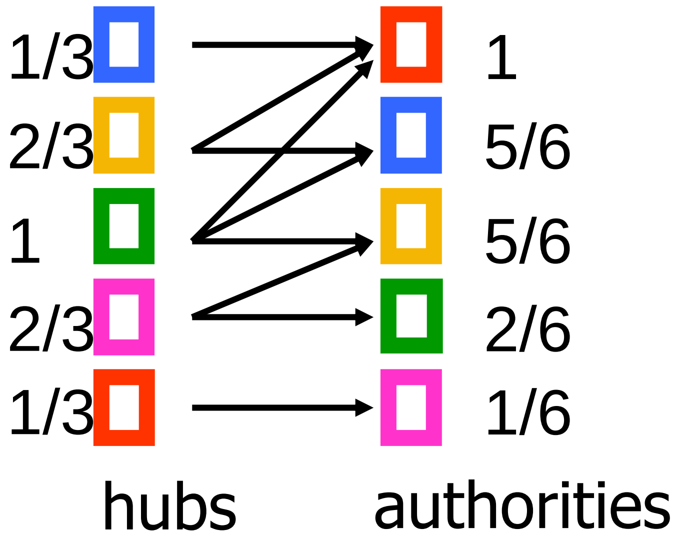
Example

Step 1: I operation



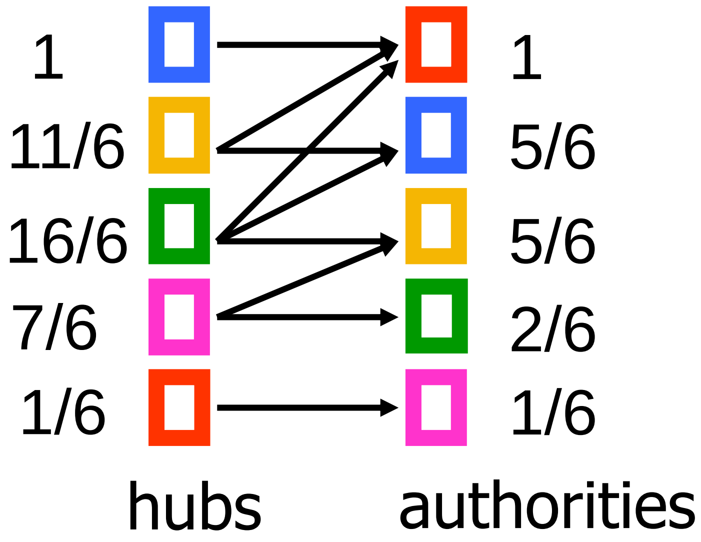
Example

Step 1: Normalization (Max norm)



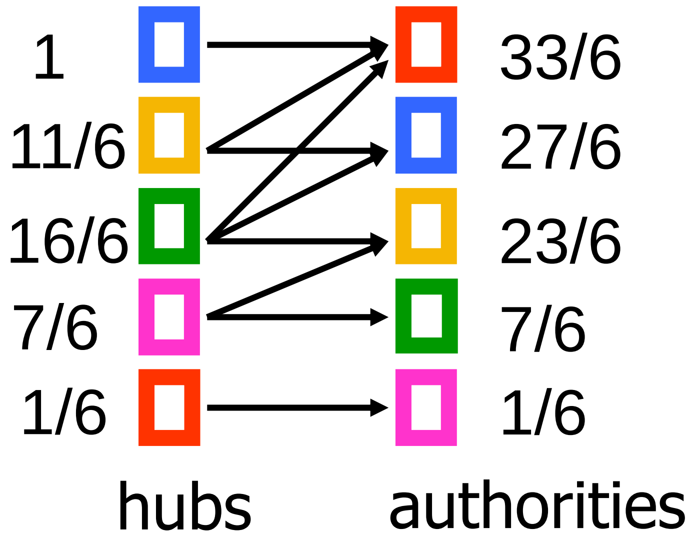
Example

Step 2: O operation



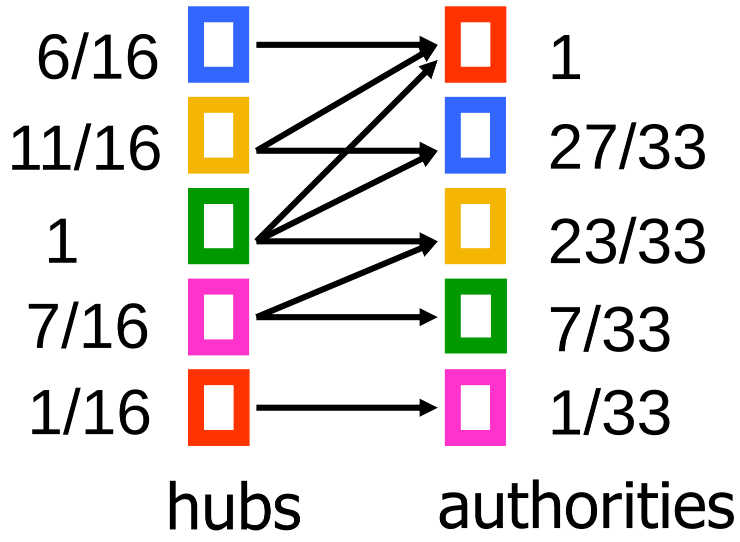
Example

Step 2: I operation



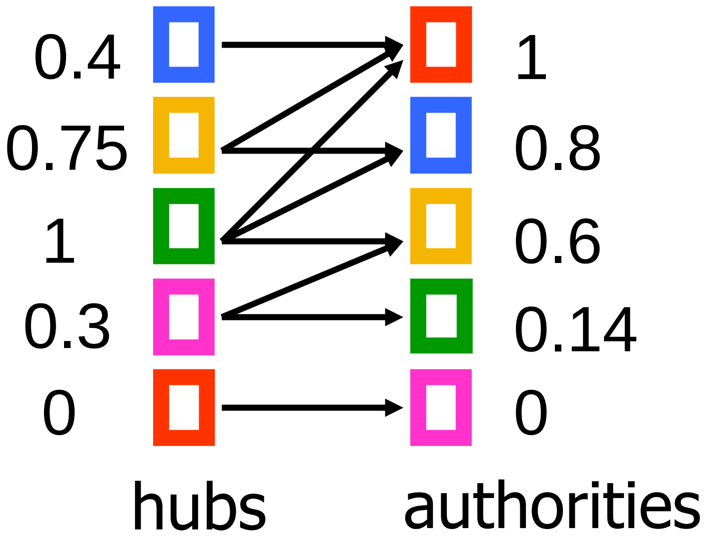
Example

Step 2: Normalization (Max norm)



Example

Convergence



HITS vs PageRank

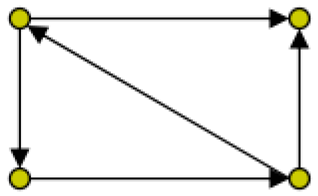
- HITS
  - Authority is not necessarily transferred directly between authorities
  - Typically not web-scale (unlike PR), **based on search results of a query**, a subset of Web, usually not precomputed
  - ‘Topic drift’ problem: when expanding the root set to base set, may include authoritative pages of other topics that affect the result

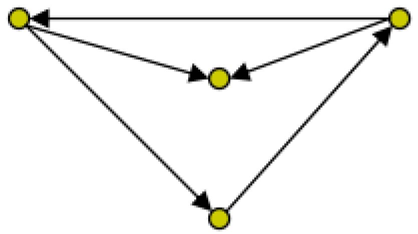
Graph Similarity

- Comparing biological networks
  - Deriving phylogenetic trees from metabolic pathway data [Heymans, Singh, 2003].
- Social network mapping
  - Small world phenomena [Milgram, 1967; Watts, 1999].
- Chemical structure matching
  - Finding similar structures in a chemical database [Hattori et al., 2003].

Graph Similarity Metrics

- Isomorphism (同构) – identifying a bijection (双射、一一映射) between the nodes of two graphs which preserves (directed) adjacency.

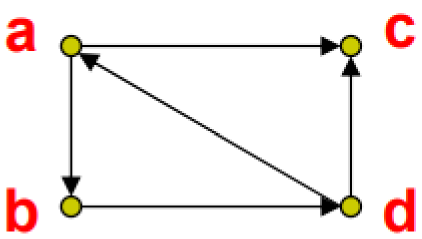


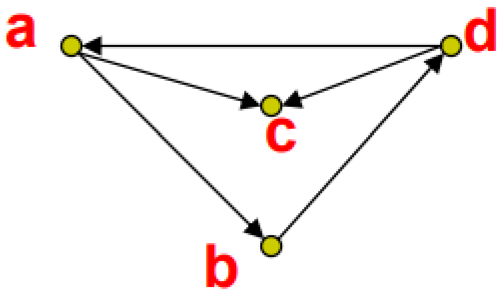


- Corneil & Gotlieb, *Journal of the ACM*, 1970.
- Pelillo, *Neural Computation*, 1999.
- Ullman, *Journal of the Assoc. of Computing Machinery*, 1976.

Graph Similarity Metrics

- Isomorphism (同构) – identifying a bijection (双射、一一映射) between the nodes of two graphs which preserves (directed) adjacency.

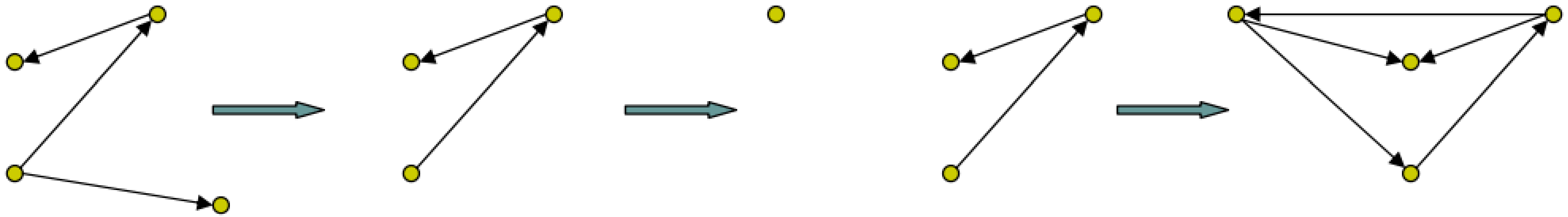




- Corneil & Gotlieb, *Journal of the ACM*, 1970.
- Pelillo, *Neural Computation*, 1999.
- Ullman, *Journal of the Assoc. of Computing Machinery*, 1976.

Graph Similarity Metrics

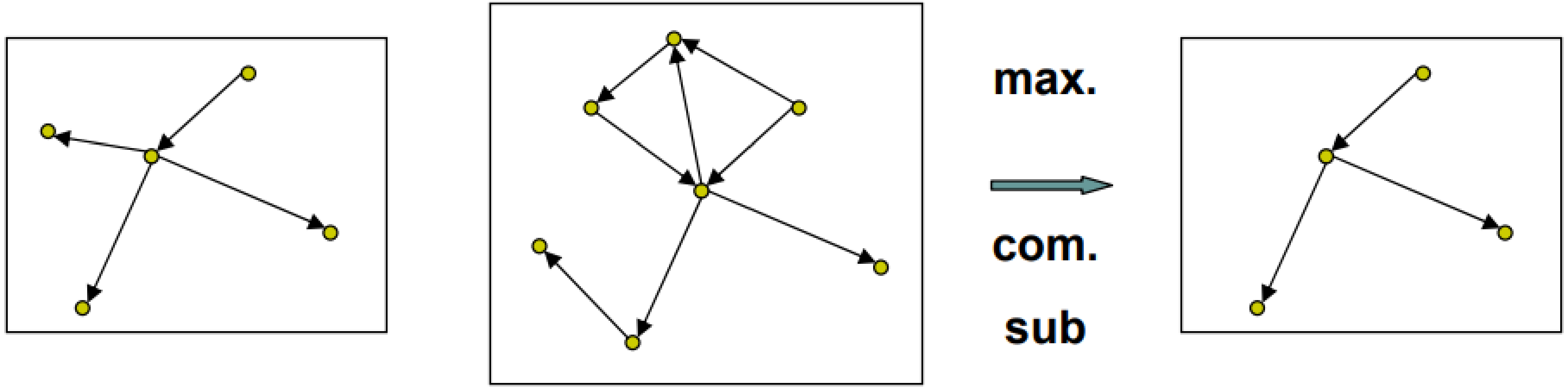
- Edit distance – given a cost function on *edit operations* (e.g. addition/deletion of nodes and edges), determine the <u>minimum cost transformation</u> from one graph to another.



- Bunke, *IEEE Trans. Pattern Analysis and Machine Int.,* 1999.
- Messmer & Bunke, *IEEE Trans. Pattern Analysis and Machine Int.*, 1998.

Graph Similarity Metrics

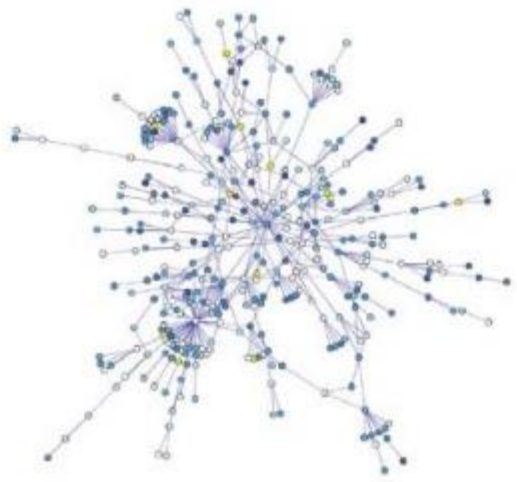
- Maximum common subgraph – identifying the `largest' isomorphic subgraphs of two graphs.
- Minimum common supergraph – identifying the `smallest' graph that contains both graphs.

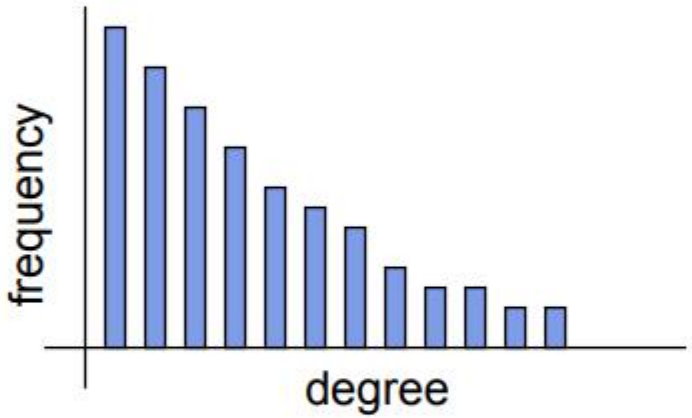


- Fernandez & Valiente, *Pattern Recognition Letters*, 2001.
- Bunke, Jiang & Candel, *Computing*, 2000.

Graph Similarity Metrics

- Statistical methods – assessing *aggregate measures* of graph structure (e.g. degree distribution, diameter, betweenness measures).





- Albert, Barabasi, *Reviews of Modern Physics*, 2002
- Dill, Kumar, et al., *ACM Transactions on Internet Technology*, 2002.
- Watts, Small Worlds, 1999.