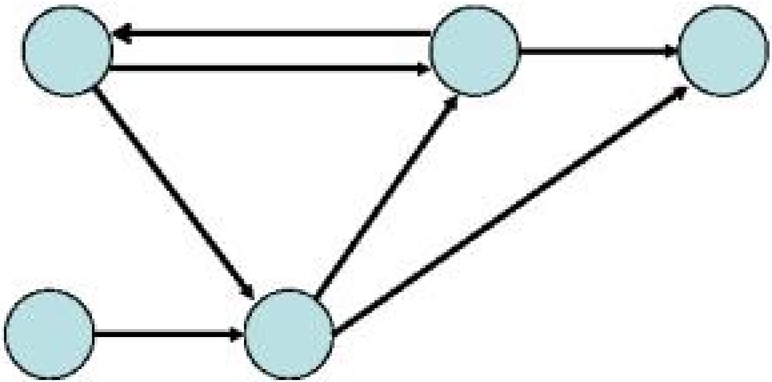
The Work of Jon Kleinberg

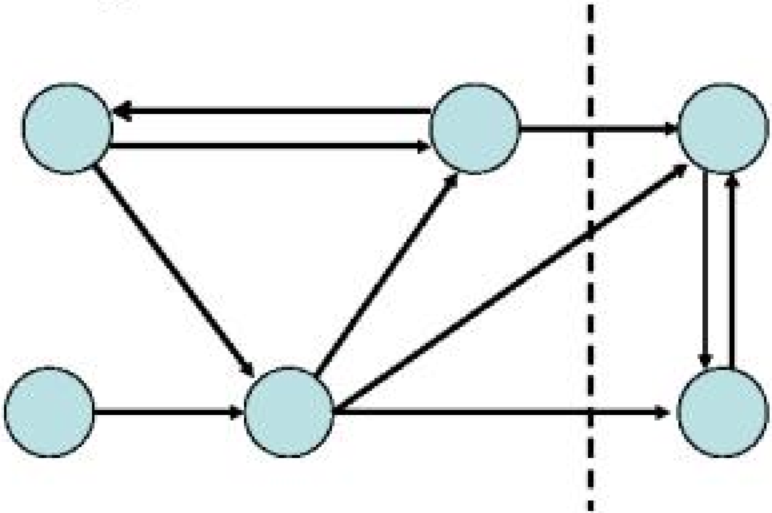
- Hubs and Authorities
- Small Worlds
- Bursts
- Nearest Neighbor
- Collaborative Filtering

Brin and Page



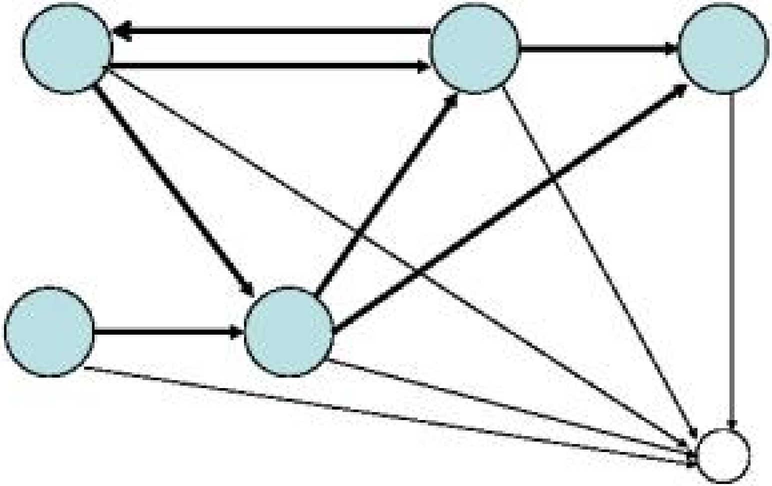
Random walk on a directed graph

Brin and Page

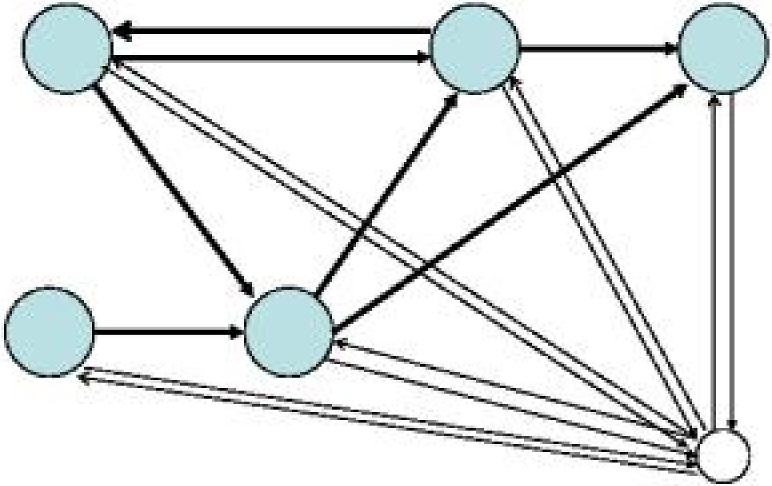


Random walk on a directed graph

Brin and Page

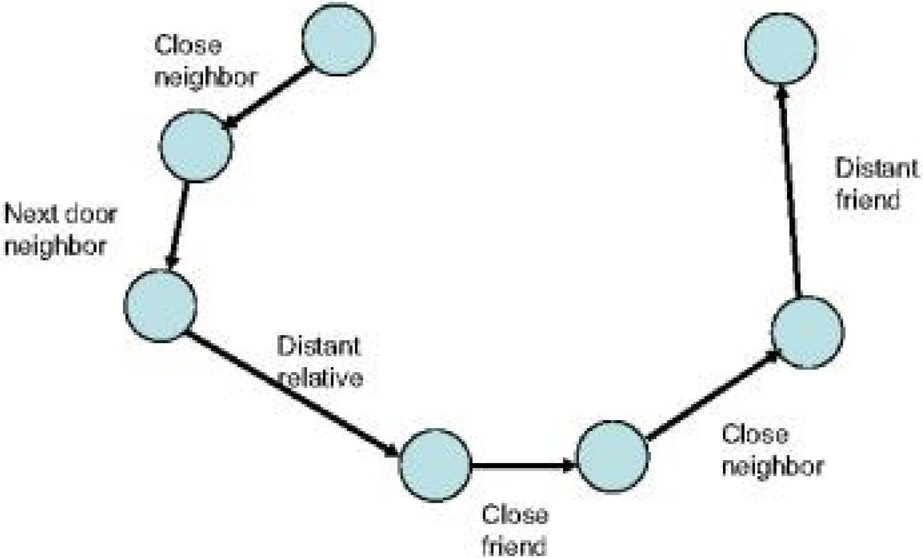


Brin and Page



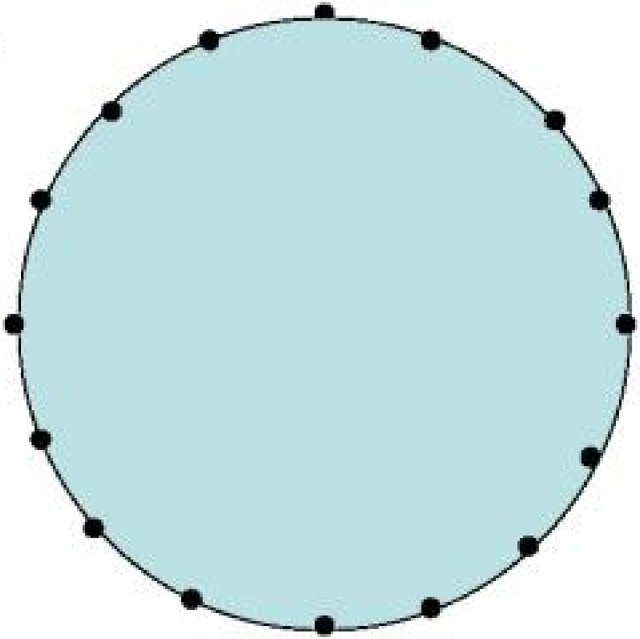
Small Worlds

Stanley Milgram (1967)

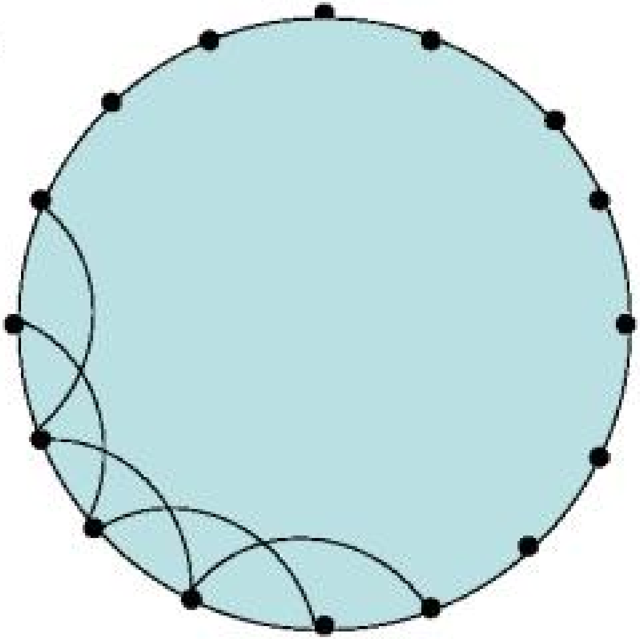


Any two people connected in six steps

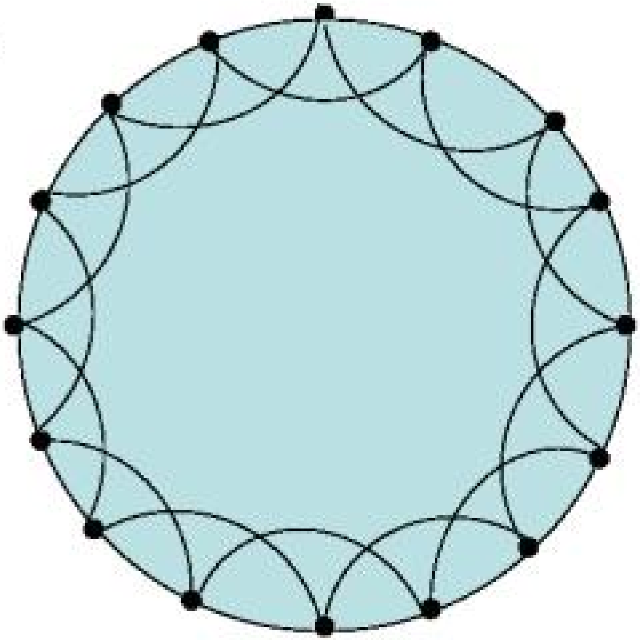
Watts and
Strogatz



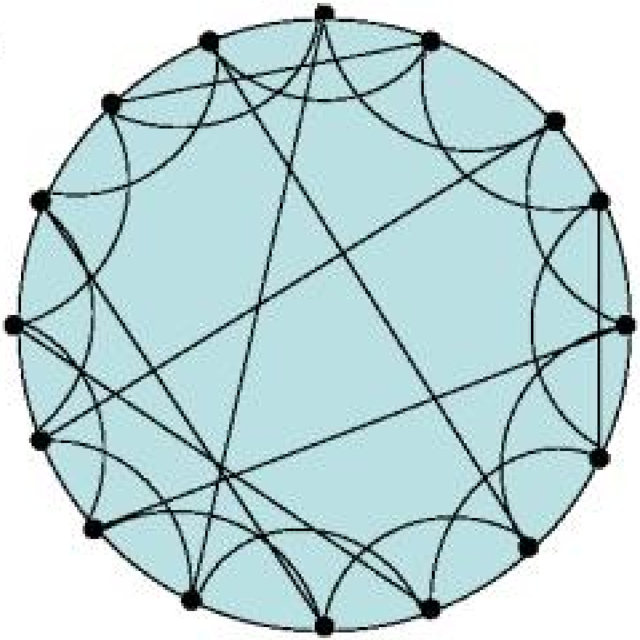
Watts and Strogatz



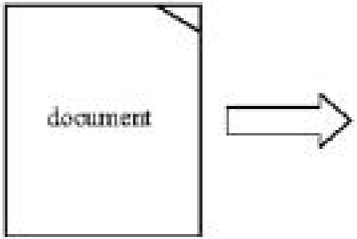
Watts and Strogatz



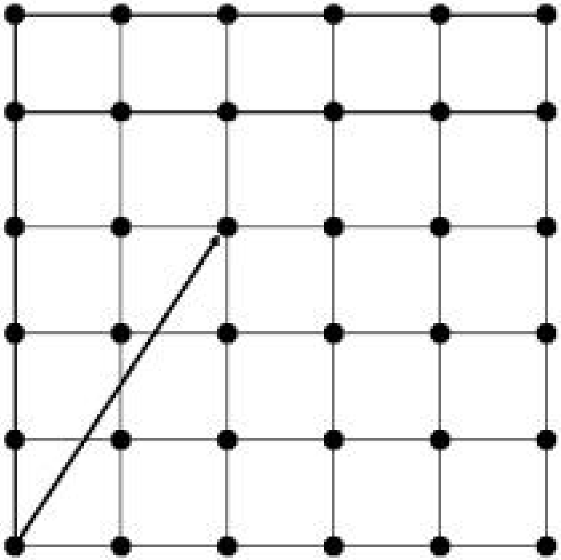
Watts and
Strogatz



Access to Information
SMART Technology

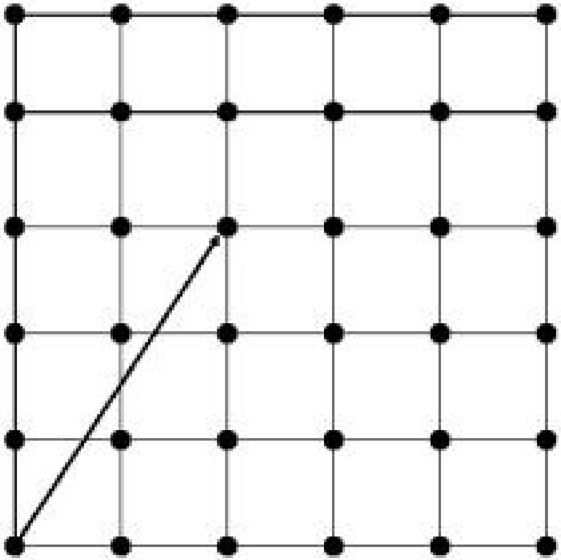


aardvark	0
abacus	0
⋮	
antitrust	42
⋮	
CEO	17
⋮	
microsoft	61
⋮	
windows	14
wine	0
wing	0
winner	3
winter	0
⋮	
zoo	0
zoology	0
Zurich	0



- Early work following that of Stanley Milgram focused on the existence of short paths

- Kleinberg asked the question “How does one find short paths between two neighbors using only local information?”



- Short edges to all neighbors in the plane
- Long range edges random with probability of the edge from u going to v given by

$$\mathrm{Prob}(u,v) = \frac{c}{\mathrm{dist}^r(u,v)}$$

- For r=2 Kleinberg showed that there is a polynomial time algorithm using only local information for finding short paths.
- For r<2 there are many long random edges but no efficient algorithm using only local information to find short paths.
- For r>2 there are fewer long range edges but still no efficient algorithm to find short paths even when they exist.

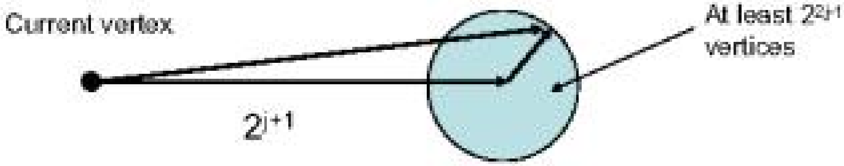
Proof

- Define phases
- Phase j - lattice distance from current vertex to destination is in interval $(2^j, 2^{j+1}]$
- At most log n phases
- If time at most log n in a given phase, then $O(\log^2 n)$ algorithm

- $\text{Prob}(u,v) = \dfrac{c}{\text{dist}^2(u,v)}$

- For large n there exists constant c_1 such that $\sum d^{-2}(u,v) \le c_1 \ln n$

- It follows that there exists constant c_2 such that each vertex within distance 2^j of u has probability at least $c_2 \dfrac{2^{-2j}}{\ln n}$ of being the long distance contact of u



- In phase j the current vertex is within distance 2^{j+1} of the destination.
- There are at least 2^{2j-1} vertices within distance 2^j of the destination.
- Therefore at least 2^{2j-1} vertices within distance $2^{j+1}+2^j < 2^{j+2}$ of current location that are within 2^j of destination.

- Thus there are 2^{j+1} vertices close to the destination and close to the current vertex

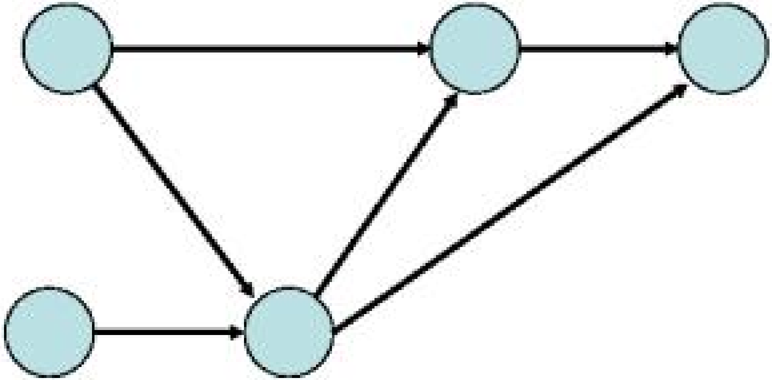
- Each of these has probability of at least $\frac{c_2}{2^{2j+4}\ln n}$ of being the long distance contact

- Phase ends with probability at least $\frac{c_2}{8\ln n}$

- Expected number of steps is O(ln n) per phase

- Total is $O(\ln^2 n)$

- This research has many applications outside the social sciences in areas such as peer to peer file sharing systems.
- Many real sets of data have the needed quadratic decrease in probability distributions, e.g., distance between individuals as measured by distance between their zip codes.



World Wide Web

Detecting bursts in data streams

- Information in a data stream can be organized by topic, time, frequency, or some other parameter

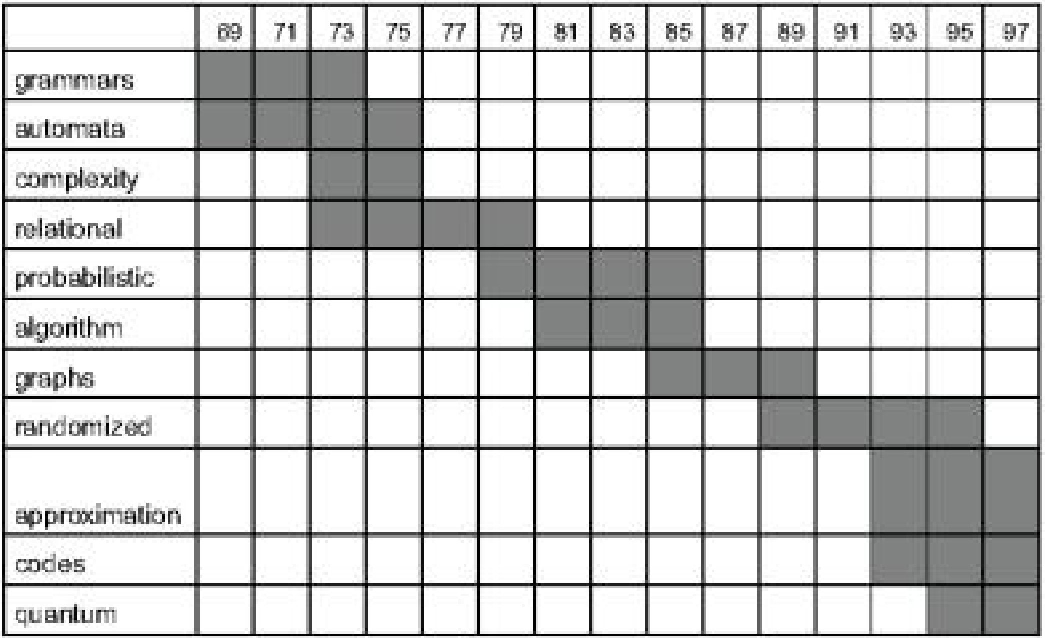
- Bursts of activity provide structure for identifying or organizing data

Kleinberg's model

- Gaps between events satisfy distribution

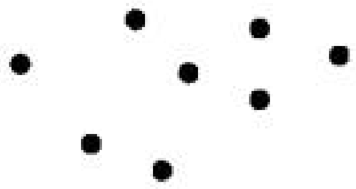
$$P(x)=ae^{-alpha\ x}$$

- Expected value of gap is 1/alpha
- Kleinberg's model
  - Infinite number of states q_0 ,q_1,q_2, . . . each q_i having an arrival rate
  - q_0 is the base state with rate 1/g
  - Each q_i has rate $(1/g)s^i$ where s is a scale parameter
  - Each state has two transitions
  - Cost associated with each transition to a higher state



STOC and FOCS Conferences 1969-1997

Nearest neighbor



Given n points in d dimensions the obvious brute force algorithm uses time n^2d. If n and d are say 10^6 each the time would be 10^{18}

Project points onto line

- If two points are close together they will remain close together.
- However, many points far apart may become very close together or even coincident.

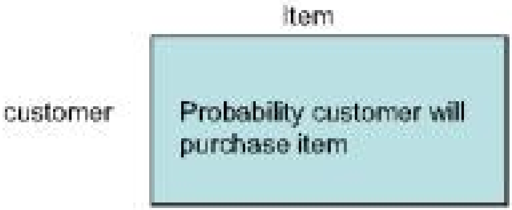
- Queries in high dimensions
  - If point x is closer to a query q than a point y, then with probability greater than ½ the projection of x will be closer to the projection of q than the projection of y.
  - If x is closer to q than y then with a sufficient number of projections the probability that x is closer to q than y in a majority of the projections will be true with high probability.
  - Using VC-dim argument Kleinberg proved that when

$$(1+\varepsilon)\text{dist}(x,q) \leq \text{dist}(y,q)$$

    the test will fail for a majority of projections with high probability.

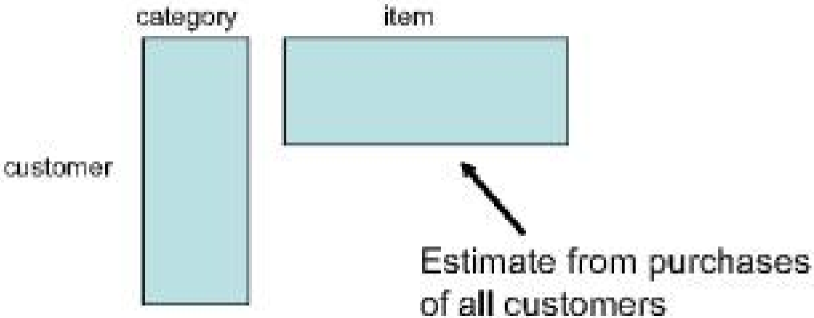
Collaborative filtering

- The process of targeting an ad or recommending a purchase based on a small amount of information.



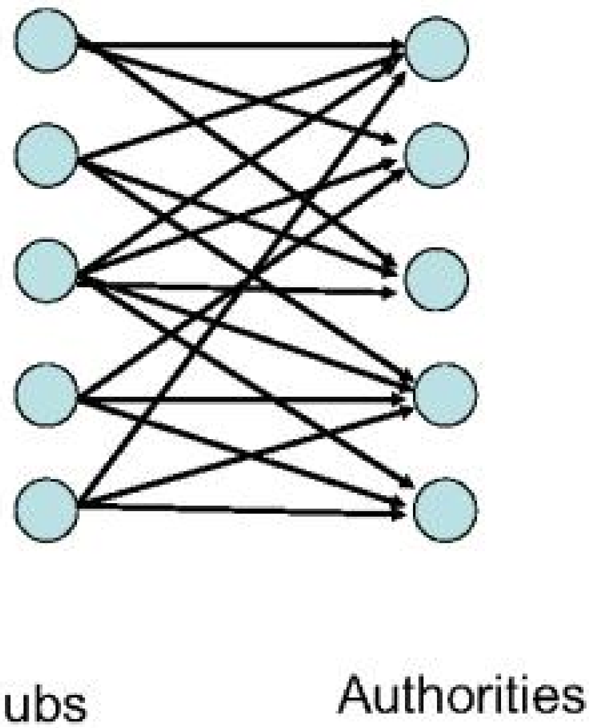
- For a given customer would like to recommend the highest probable item.

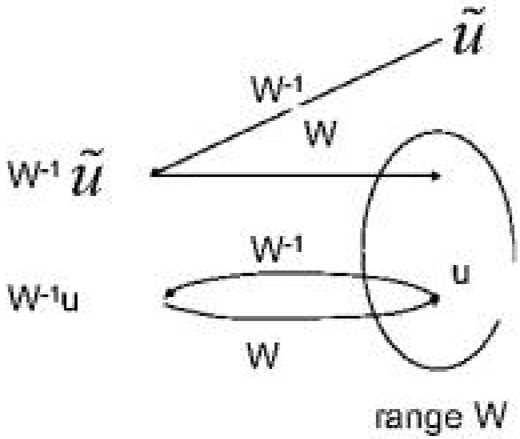
Probability matrix may have structure



- Let A=PW.
- Let u be row of A and let $\tilde{u}$ be the estimate of u obtained from s samples.
- How close is $\tilde{u}$ to u?

- Suppose we know W
- u is in the range of W but $\tilde{u}$ most likely is not.
- Projecting $\tilde{u}$ onto the range of W might improve the estimate of u


Hubs
Authorities



$\tilde{u}$
W-1
W
W-1 $\tilde{u}$
W-1
u
W-1u
W
range W

- Applying WW' to $\tilde{u} - u$ we get

$$WW'(\tilde{u} - u) = WW'\tilde{u} - u$$

- We need to bound how far $WW'\tilde{u}$ can be from u.

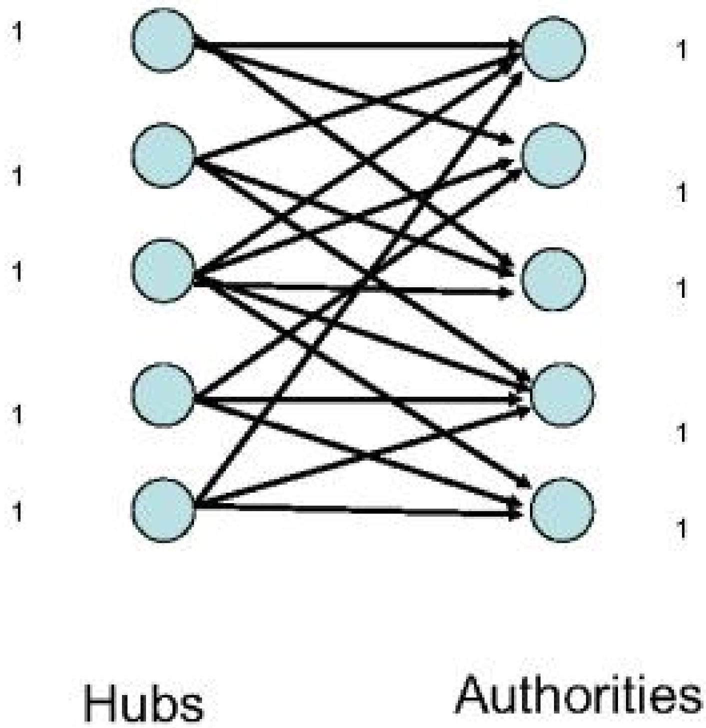
- We would like each component of $WW'\tilde{u}$ to be within ε of the corresponding component of u with high probability.

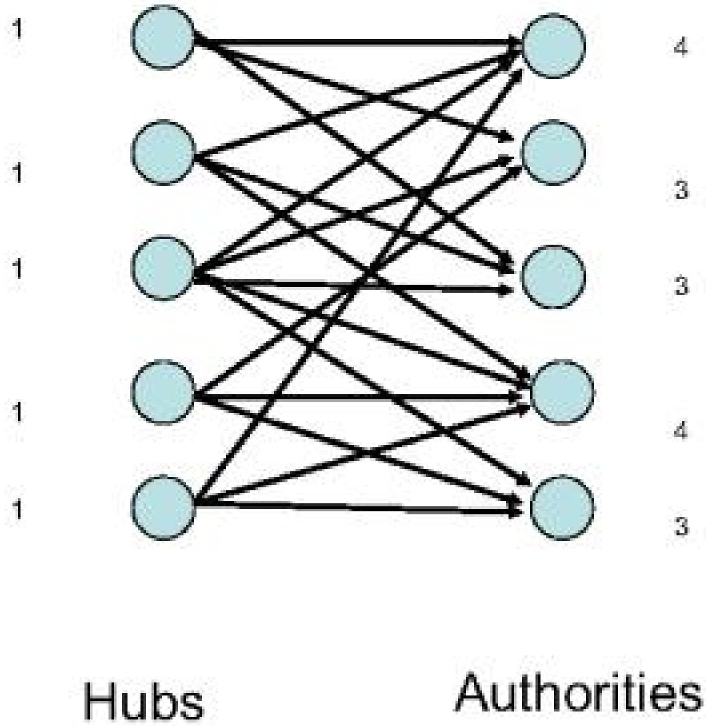
- Item corresponding to largest component is optimal recommendation.

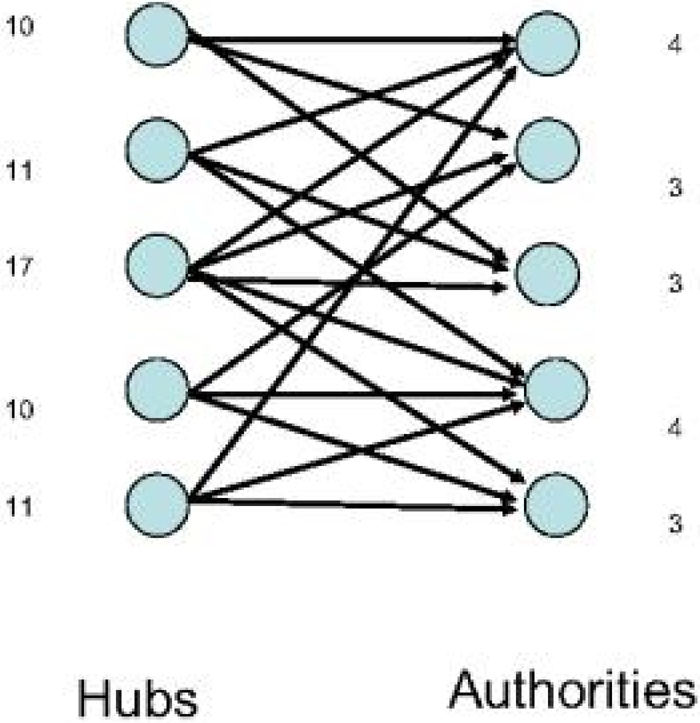
- What projection should we use?
- Obvious projection is to project orthogonally. The original projection minimizes the total error but not the individual errors.
- Kleinberg and Sandler showed how to keep individual error small at the expense of increasing total error.

Conclusions

- Kleinberg's research has impacted our understanding of social networks, information, and the www.
- Its impact has influenced every modern search engine and thus the daily lives of researchers through the world.
- It has laid the theoretical foundation for the future of computer science.

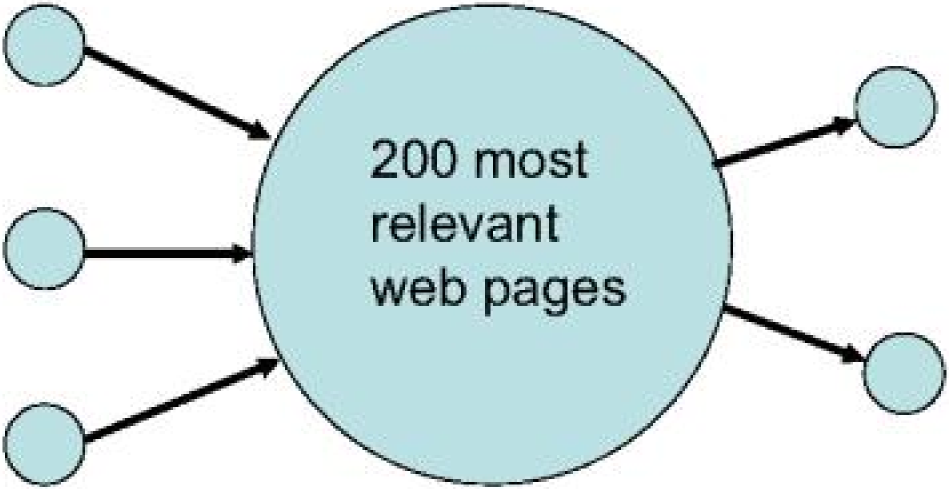

1
1
1
1
1
1
1
1
1
1
Hubs
Authorities


1
1
1
1
1
4
3
3
4
3
Hubs
Authorities


10
11
17
10
11
4
3
3
4
3
Hubs
Authorities

The iterative procedure converges so that

- Hub weights are coordinates of the major eigenvector of A^TA.
- Authority weights are coordinates of the major eigenvector of AA^T.



Use in and out links to extend the set of pages