The Structure of Information Networks: An Introduction

Yannet Interian

Fellow at IPAM Fall 2007

Overview Information Networks

- Web
  - Link structure powerful source of information about the underlying content in the network.
  - Ranking web documents (PageRank other Link analysis algorithms)
  - Web communities

Overview Information Networks (Cont)

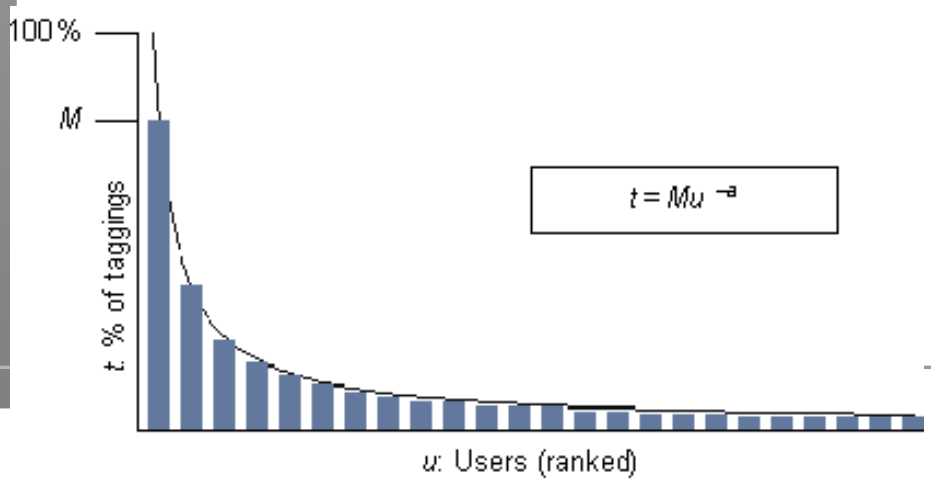
- Small-World Properties in Networks, decentralized Search.
  - Small Diameter.
  - Variety of domains exhibit common structure at a qualitative level
  - Random Models.
  - Application in the design of decentralized peer-to-peer systems.

Overview Information Networks (Cont)

- Cascading Behavior in Networks
  - Diffusion of information can happen rapidly or slowly.
  - Random Models.
  - Investigated when a network is more or less susceptible to information spread.

Overview Information Networks (Cont)

- Power-Law Distributions
  - The degree of a node in a network is the number of neighbors it has.
  - The fraction of nodes with degree d decays like d to some fixed power.
  - What processes are capable of generating such power laws, and why should they be ubiquitous in large networks?



Overview of the Talk

- Part 1. Small World Phenomena and Decentralized Search.
- Part 2. The Web: Link Analysis.

Part 1

Small-World Phenomena and Decentralized Search

Pictures:

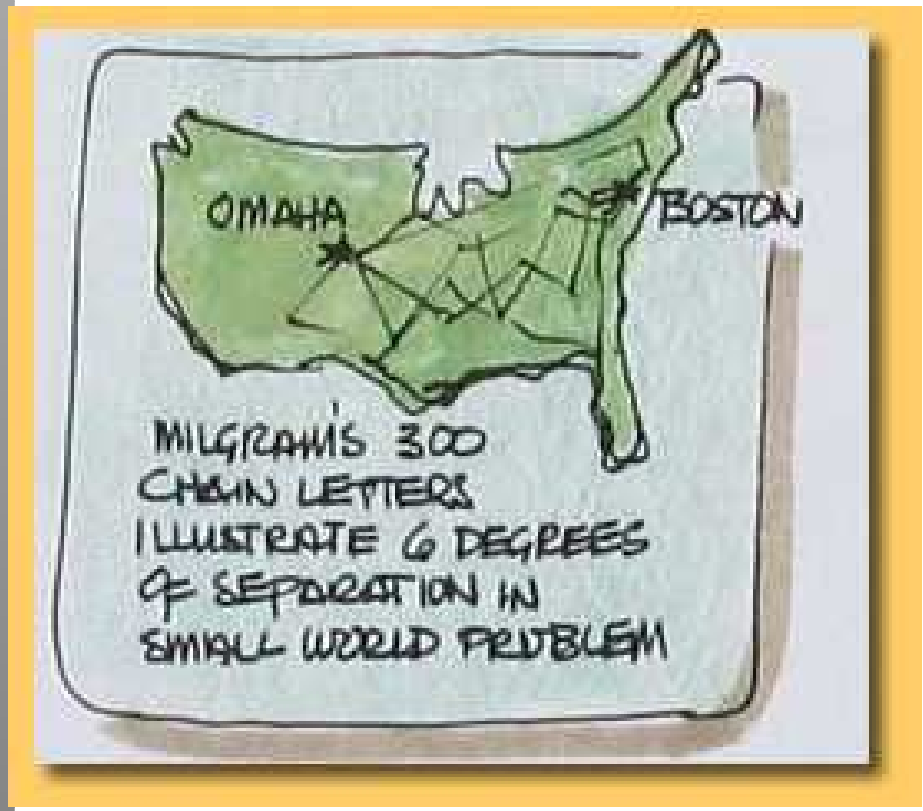
http://www.leggmason.com/thoughtleaderforum/2004/conference/transcripts/

Small-World Phenomena

- The principle that we are all linked by short chains of acquaintances.
- “Six degrees of separation”
- Stanley Milgram (1933-1984)
  - Obedience to authority (1963)
  - Small world experiment (1967)

Milgram Experiment

Goal: find short chains of acquaintances to link people who did not know each other.



Milgram Experiment

- *Target person*: a stockbroker living in a suburb of Boston.
- Randomly chosen “starter” individuals forward a letter to the target.
- Provided the target’s name, address, occupation, and some personal information.
- People were instructed to pass on the letters to someone they knew on first-name basis.
- The letters that reached the destination followed paths of length around 6.

Milgram Experiment (Cont)

- Network
  - Node: each person.
  - Edge: know each other on a first-name basis.
- Some Conclusions
  - (a) Social networks tend to exhibit very short paths between arbitrary pair of nodes.
  - (b) People using their own acquaintances, were able to collectively construct paths to the target.

Milgram Experiment (Cont)

- Network
  - Node: each person.
  - Edge: know each other on a first-name basis.
- Some Conclusions
  - (a) Social networks tend to exhibit very short paths between arbitrary pair of nodes.
  - (b) People using their own acquaintances, were able to collectively construct paths to the target.
- Question: Which kind of networks exhibit properties (a) and (b).
- Watts and Strogatz Model (a)
- Kleinberg what about (b)?

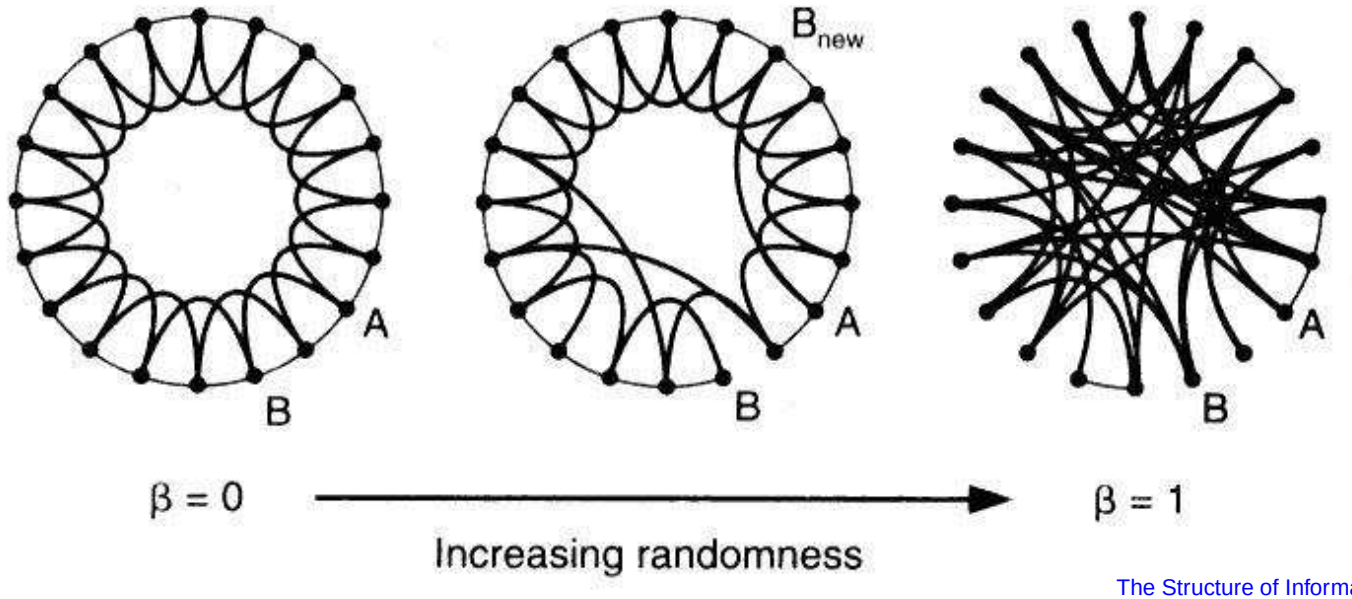
Modelling Milgram Experiment

Why Random Networks?

- Probabilistic model of a Real Network.
- Stylized network model produced by a random mechanism.
- Show the model reproduces properties observed in the real network.
- A finding based on a random model —> observed properties may have a simple underlying basis.
- Goal: extracting the model some qualitative properties that distinguish networks which has type of properties.

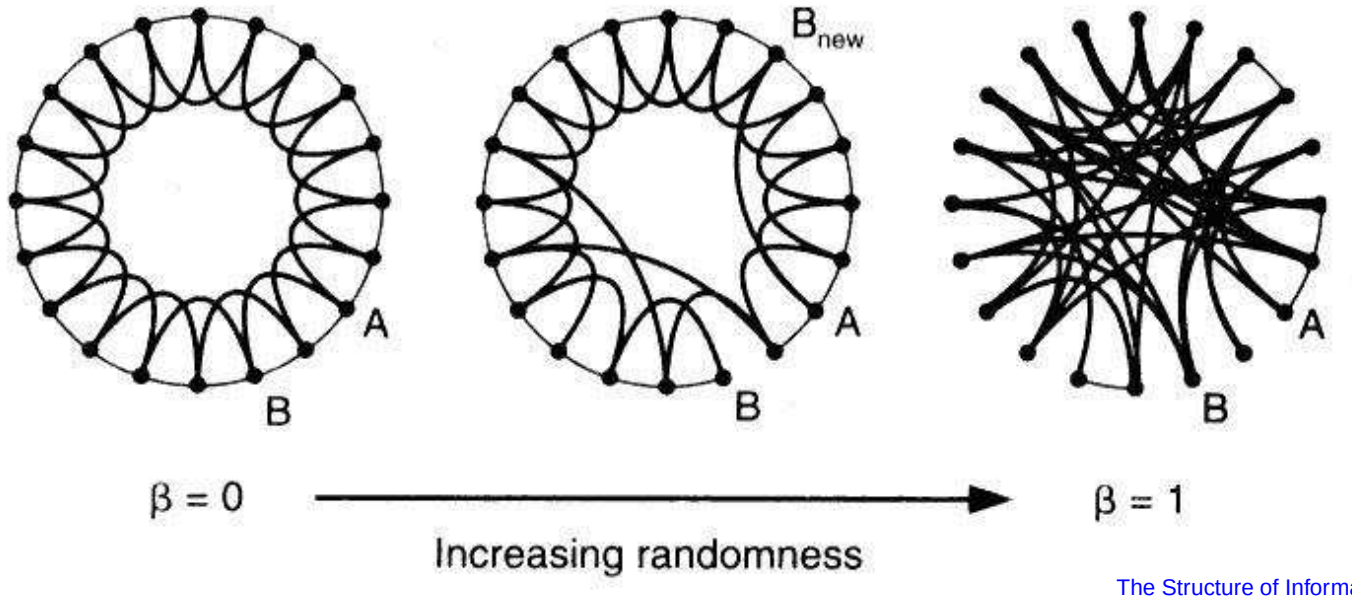
A Random Model Social Networks

- Watts and Strogatz Model.
- n points on a circle, joints the k nearest neighbors (local contacts).



A Random Model Social Networks

- Watts and Strogatz Model.
- n points on a circle, joints the k nearest neighbors (local contacts).
- For every edge, with probability p, rewire the edge to a uniformly chosen destination (long-range contacts).



Watts-Strogatz's Model

- A better model for Social Network.
- Standard Random Model:
  - None of the neighbors of a given node v are neighbors of one another
- Real Network: Many of my friends know each other.
- Has property (a): Diameter is log of the number of vertices.
- Does not have property (b).

What about property (b)?

Recall

- Property (a) Short paths exist.
- Property (b) People, using knowledge only of their own acquaintances, were able to collectively construct short paths to the target (Decentralized Search).

Decentralized Search

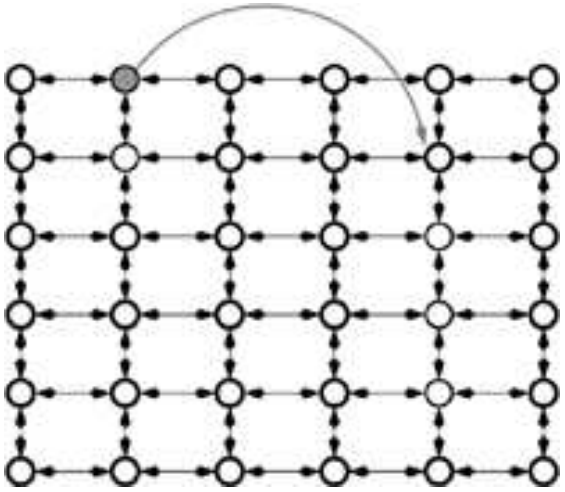
- Why should arbitrary pair of strangers be able to find short chains of acquaintances that link them together?
- No knowledge of global network.
- “Local Cues”.
- Typical strategy (Milgram experiment)
  - Choose someone geographically closer or someone professionally closer.

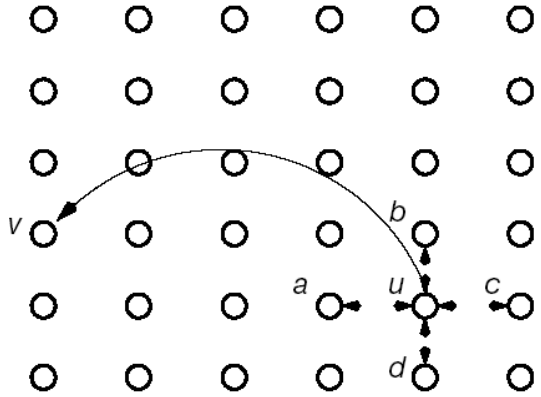
Modelling Decentralized Search

- Networks that support efficient search
  - Contain short paths among all pairs of nodes.
  - Structure that is partially known and partially unknown.

Decentralized Search: Kleinberg Model

- A grid (example 2-dimensional $n \times n$), known to every node.
- Long range connections:
  - $Pr(u, v) \sim d(u, v)^{-\alpha}$.





Decentralized Algorithm: Definition

- Context of previous model.
- Message from a node s to a node t.
- Message holder pass the message to one of its neighbors.
- Every node "knows" its neighbors and the "grid" local structure.
- Knows the location of the node t on the grid.
- Does not know the long-range contacts of any other nodes.

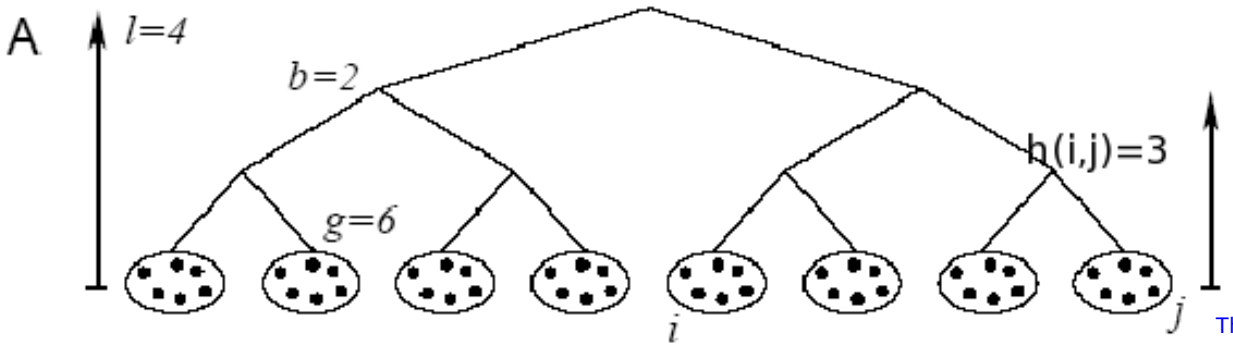
Delivery Time: The expected number of steps required to reach the target.
Good Time: Polynomial in $logn$.

Kleinberg Results

- How to choose among hundreds of acquaintances?
- For $\alpha = 2$ Simple greedy algorithm works. Each node sends message to the neighbor that is closest to target in grid distance.
- For $\alpha = 2$ greedy strategy takes $(logn)^2$ steps.
- For $\alpha < 2$ any decentralized algorithm takes at least $n^{(2-\alpha)/3}$ steps.
- For $\alpha > 2$ any decentralized algorithm takes at least $n^{(\alpha-2)/(\alpha-1)}$ steps.

Hierarchical Model [Kleinberg NIPS 01]

- Lattice captures geographic distance. How do we capture social distance (e.g. occupation)?
- Hierarchical organization of groups.
  - A b-ary tree T, distance $h(i,j) =$ height of least common ancestor.
  - G by creating $k = c\,log^2 n$ links out of each node.
  - $Pr(u,v) \sim b^{-\alpha h(u,v)}$.
  - Note: T is just use for generating G.



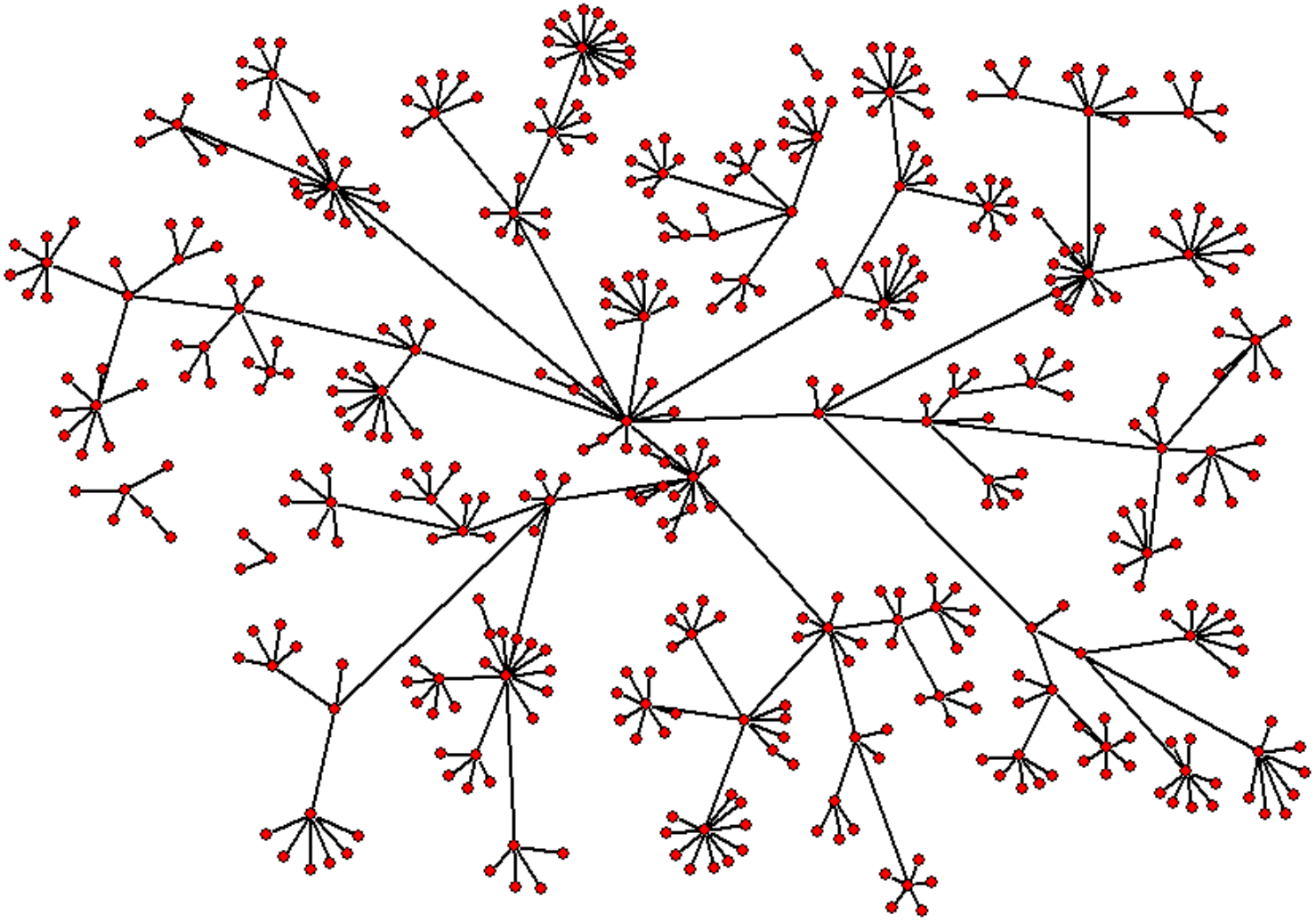
Hierarchical Model: Results

- For $\alpha = 1$ there is a polylogarithimic search algorithm.
- For $\alpha \neq 1$ there is no decentralized algorithm with poly-log time.

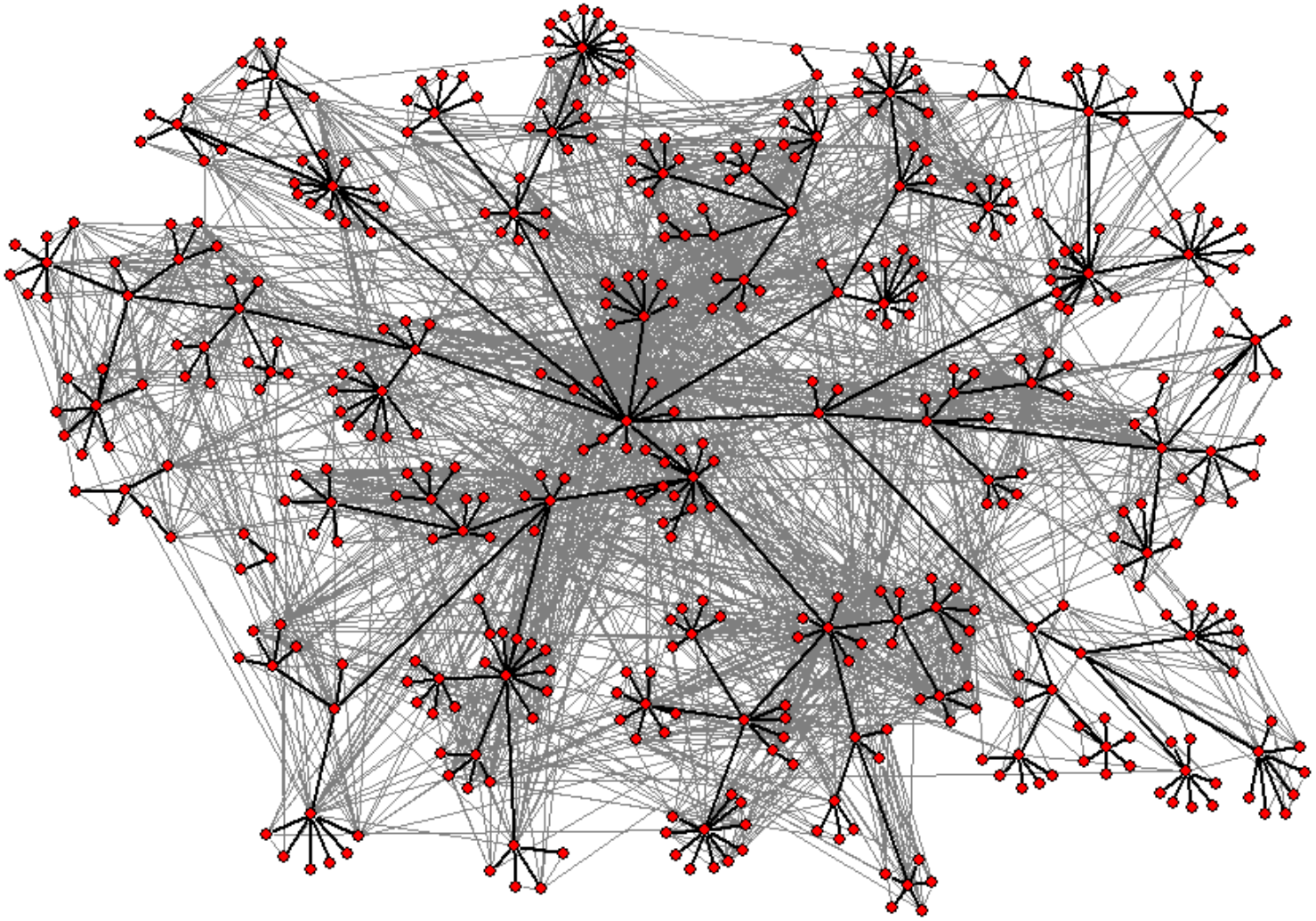
Testing search models on social net-works

- Adamic et al.
- HP Labs email correspondence over 3.5 months
- Edges are between individuals who sent at least 6 email messages each way
- 450 users,median degree = 10, mean degree = 13 average shortest path = 3
- Node properties specified:
  - Degree
  - Geographical location
  - Position in organizational hierarchy
- Can greedy strategies work?

HP Organizational Hierarchy



Email Network Superimposed on the Organizational Hierarchy

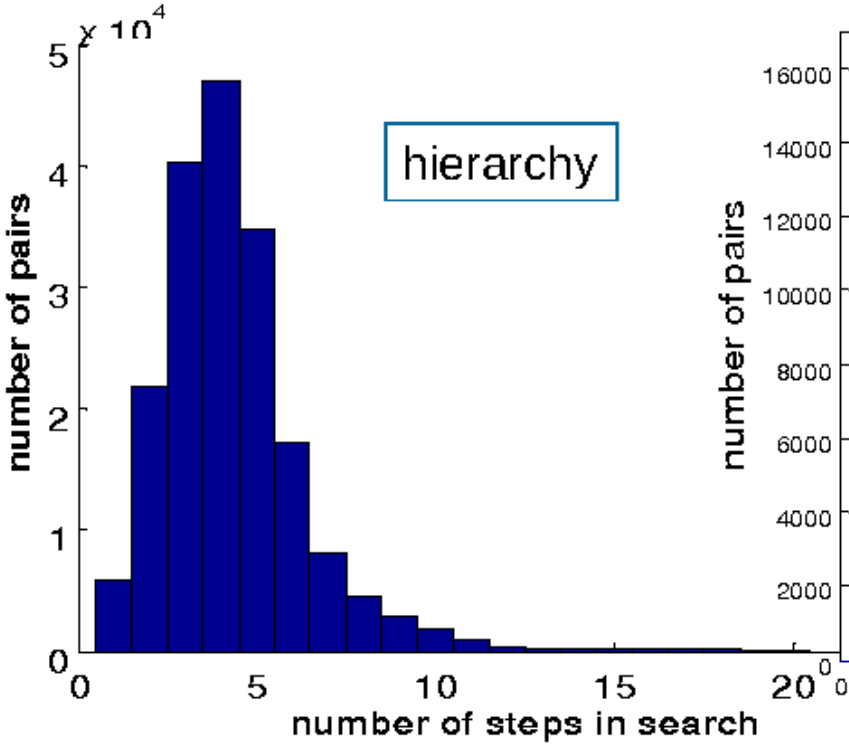


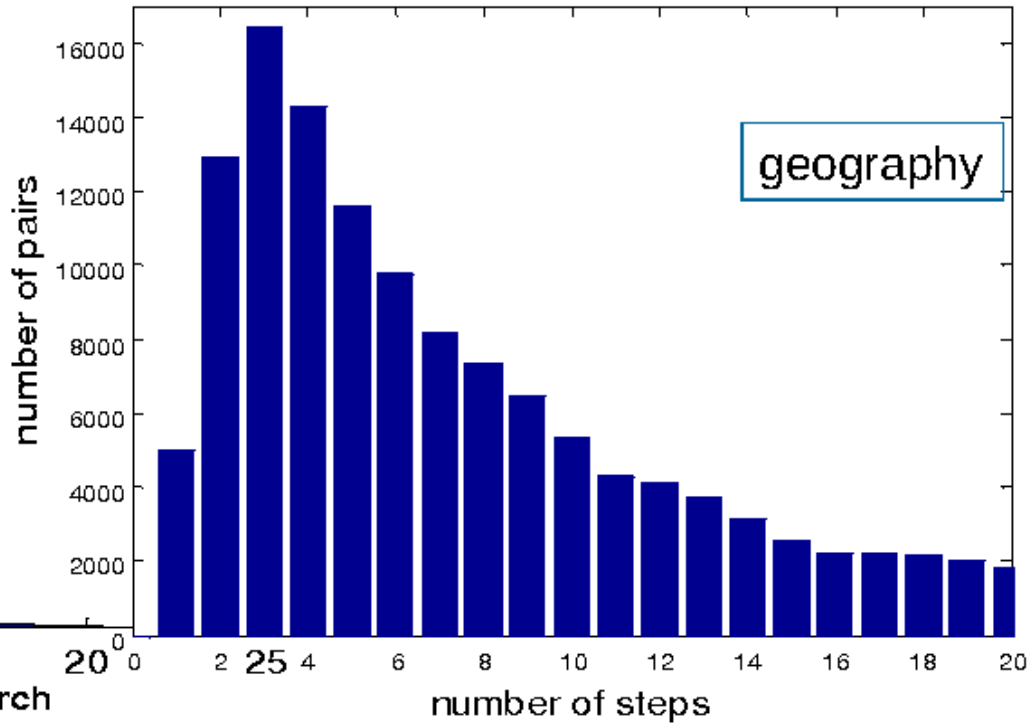
Can greedy strategies work?

- People with the same manager are at distance 1.
- Pick the “email” neighbor that is closer in the Hierarchical distance to the target.

Results

distance	hierarchy	geography	high degree
median	4	6	16
mean	5	11	43.2





Decentralized Routing: Applications

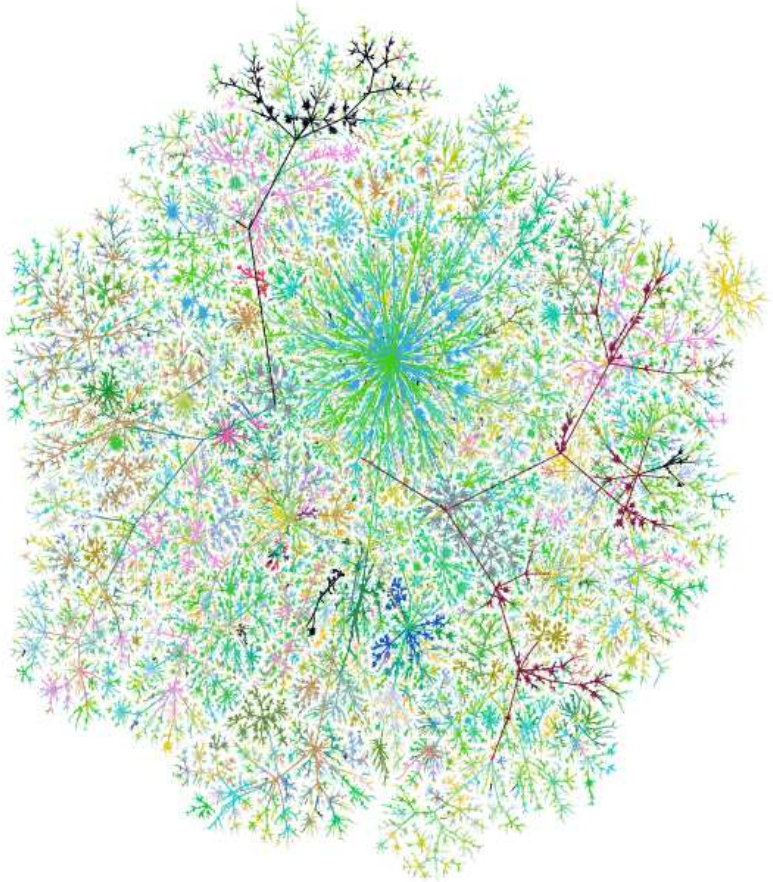
- Browsing behavior on the World Wide Web.
- Design of peer-to-peer file-sharing systems on the Internet.
  - Lookup Problem
  - `www.napster.com` centralized index.
  - `gnutella.wego.com` sent message to all your neighbours.
  - How to desing something in between?

References

- J. Kleinberg. The small-world phenomenon: An algorithmic perspective. Proc. 32nd ACM Symposium on Theory of Computing, 2000
- J. Kleinberg. Small-World Phenomena and the Dynamics of Information. Advances in Neural Information Processing Systems (NIPS) 14, 2001.
- Lada A. Adamic and Eytan Adar. How to search a social network. Social Networks 2005.
- Watts, D. J., Dodds, P. S., Newman, M. E. J.. Identity and search in social networks. Science, 2002

Part 2

Link Analysis



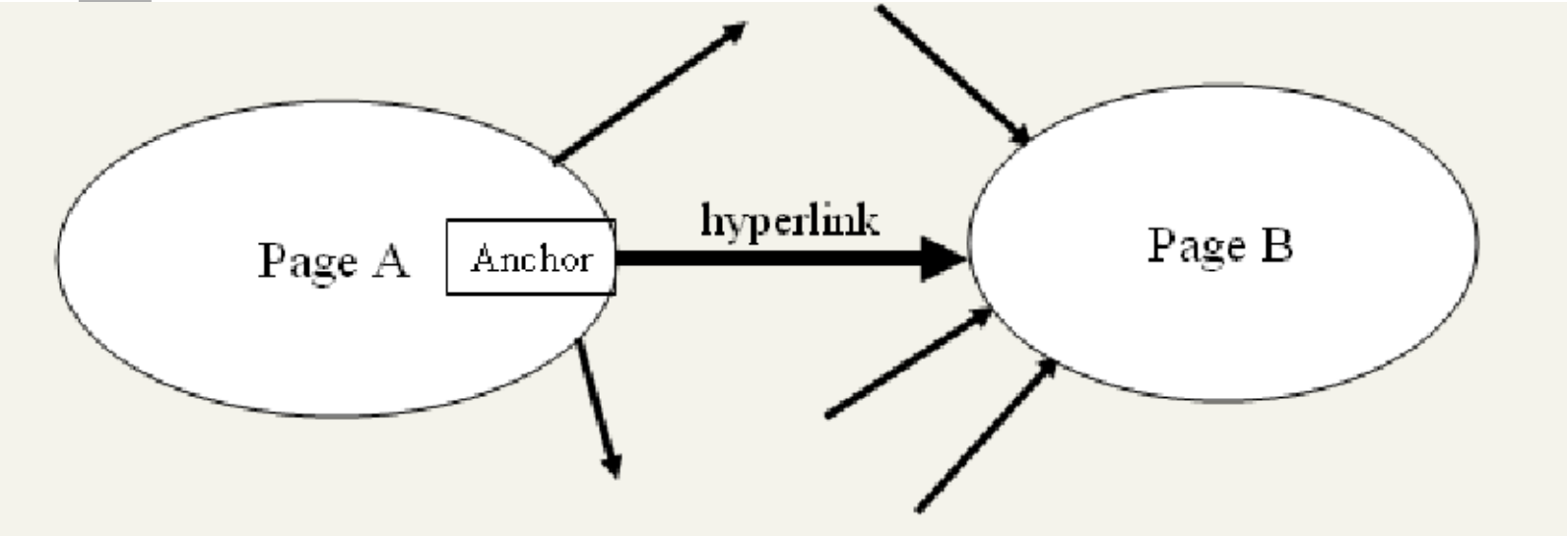
Problem Formulation

- Suppose we are given a collection of documents on some *broad topic*
  - e.g., abortion, evolution, movies, Iraq
  - perhaps obtained through a text search
- Can we organize these documents in some manner?
  - PageRank offers one solution
  - HITS (Hypertext-Induced Topic Selection) is another
  - proposed at approx the same time

Why the Problem is Difficult

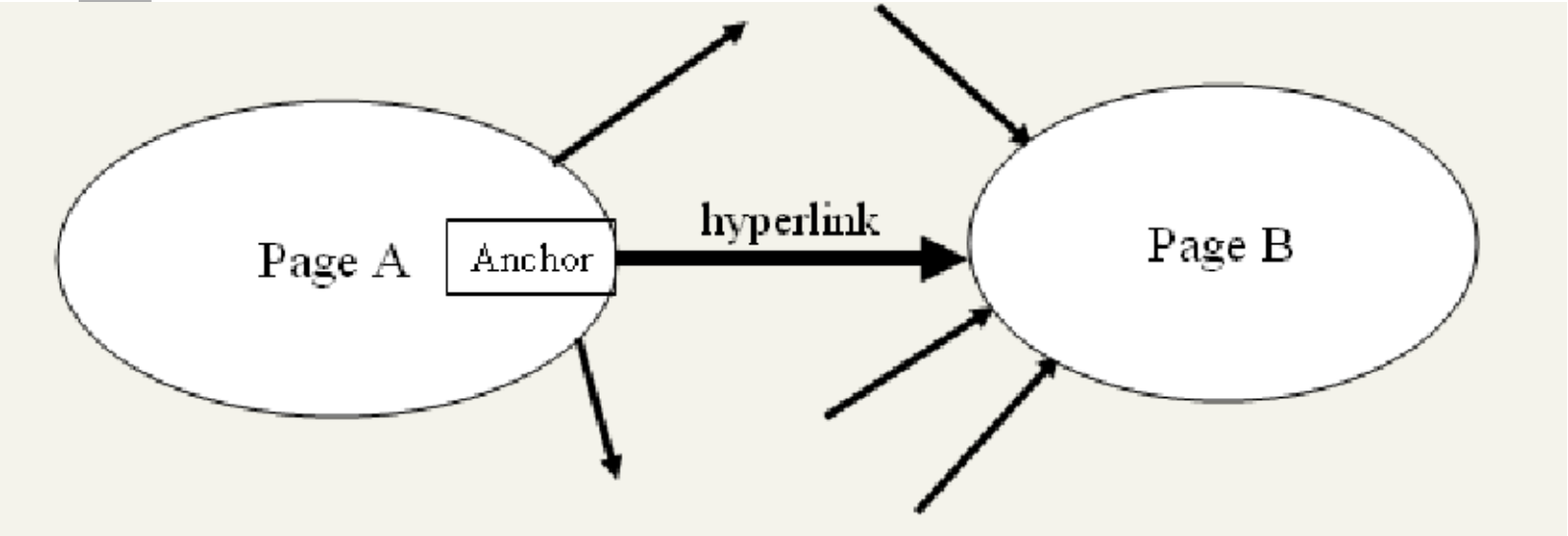
- For *broad-topic* queries find many relevant pages.
- Too large for a human user to digest.
- Authoritative pages for a query often do not use the term in the query.
  - e.g.query "search engines". Many of the natural authorities do use the term on their pages.
  - e.g "automobile manufactures". Honda Toyota may not use these terms on their web pages.

The Web as a Directed Graph



- nodes: Pages
- directed edges: Links

The Web as a Directed Graph



- nodes: Pages
- directed edges: Links
- Assumptions
  - The hyperlink from A to B connotes a conferral of authority on page B, by the creator of page A.
  - The anchor text describes the page B.

HITS [Kleinberg 98]

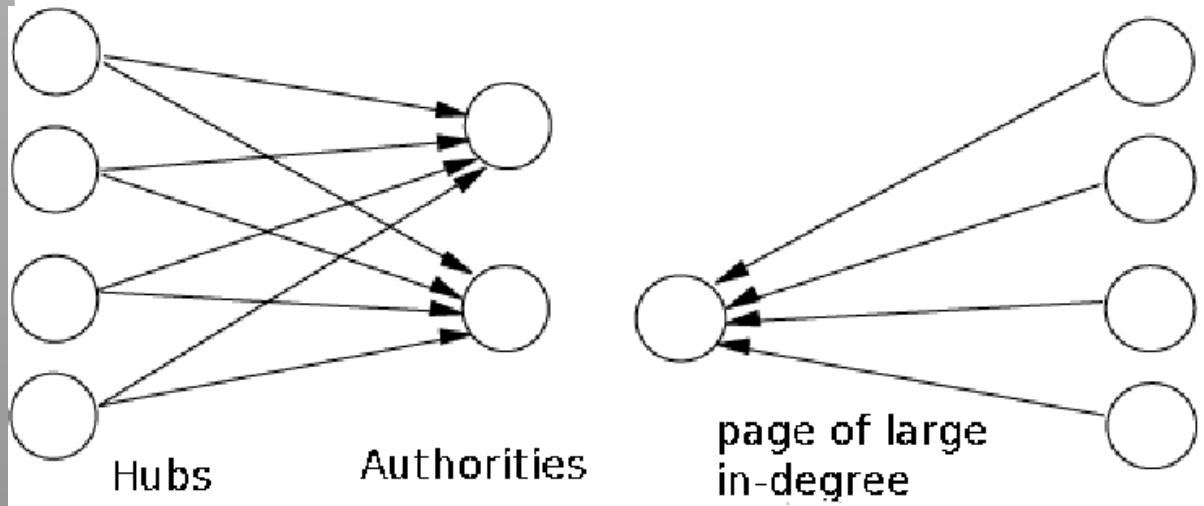
Hyperlink Induced Topic Search (HITS)

- In response to a query, instead of an ordered list of pages each meeting the query, find two sets of inter-related pages:
  - *Hub* pages are good lists of links on a subject.
    - e.g., "Bob's list of cancer-related links".
  - *Authority* pages occur recurrently on good hubs for the subject.
    - `www.cancer.org/` American Cancer Society.

Hubs and Authorities

- Thus, a good hub page for a topic *points* to many authoritative pages for that topic.
- A good authority page for a topic is *pointed* to by many good hubs for that topic.
- Circular definition will turn this into an iterative computation.

Idealized View



- Authoritative pages: not only large in-degree, also
- overlap in the *set* of pages that point to them.
- Hubs pages: link to multiple authoritative pages.
- Large in-degree pages: popular pages like yahoo! or google.

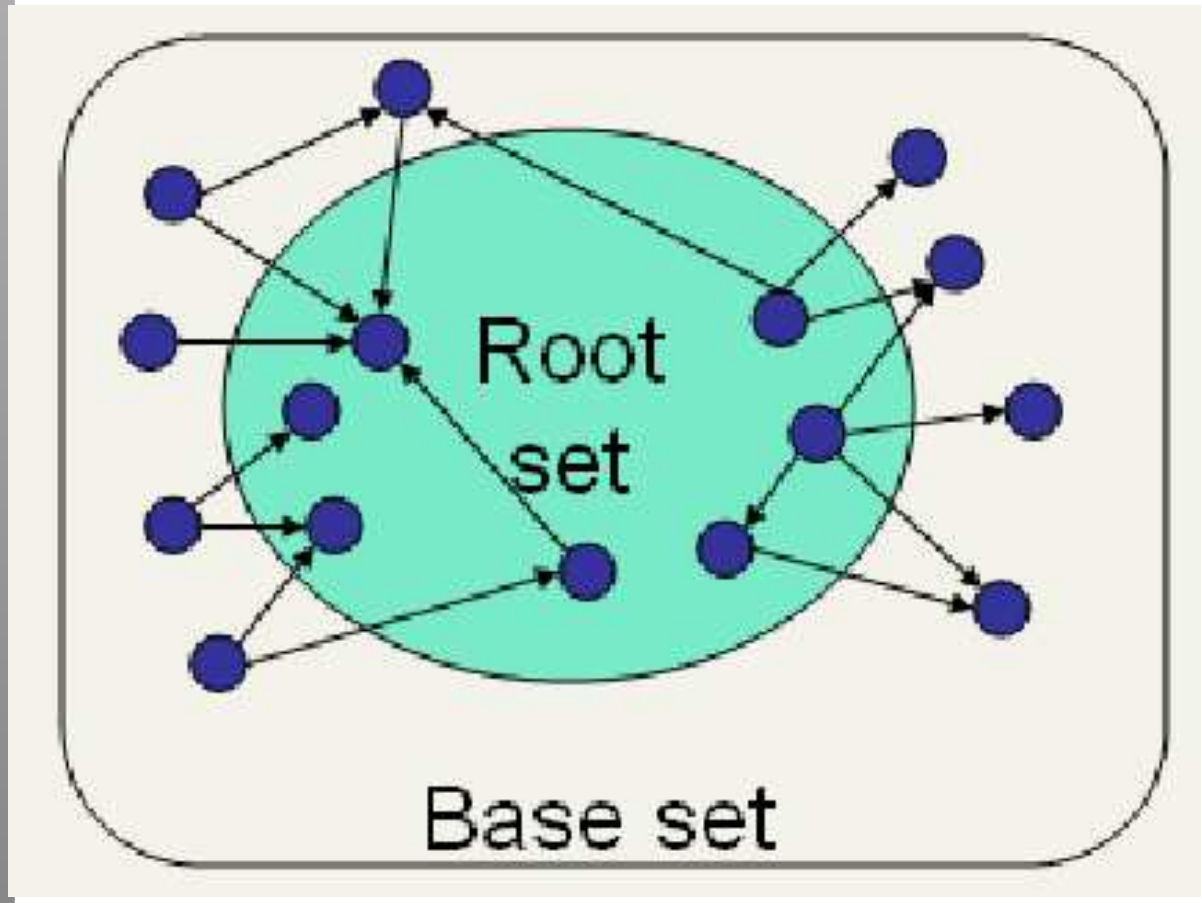
High-Level Scheme

- Extract from the web a base set of pages that *could* be good hubs or authorities.
- From these, identify a small set of top hub and authority pages.
- Note the base set is query dependent (as opposed to the PageRank).

Base Set

- Given text query (say browser), use a text index to get all pages containing browser.
  - Call this the root set of pages
- Add in any page that either
  - points to a page in a root set, or
  - it is pointed to by a page on the root set.
- Call this the base set.

Visualization



Assembling the Base Set

- Root set typically 200-1000 nodes
- Base set may have up to 5000 nodes.

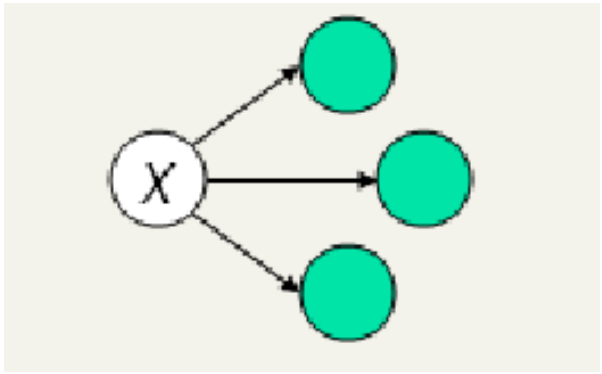
Computing Hubs and Authorities

- Compute for each page x in the base set a hub score $h(x)$ and an authority score $a(x)$.
- Initialize: for all x, $h(x) = 1$, $a(x) = 1$
- Iteratively update all $h(x)$, $a(x)$ (next slide).
- After Iterations
  - Output pages with highest $a()$ scores as top authorities.
  - Pages with highest $h()$ scores as top hubs.

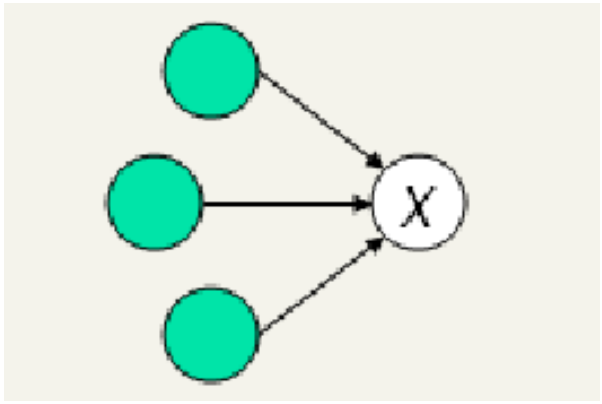
Iterative update

Repeat the following updates, for all x:

$$h(x) \leftarrow \sum_{x \to y} a(y)$$



$$a(x) \leftarrow \sum_{y \to x} h(y)$$



Scaling

- To prevent the $a()$ and $h()$ values from getting too big, scale down after each iteration.
- Scaling factor doesn't matter
- We care about the *relative* values of the scores.

How Many Iterations?

- Scores will converge after a few iterations:
  - Suitable scaled, $h()$ and $a()$ scores converge to a steady state.
  - Proof later in the talk.
- In practice, 5 iterations get you close to stability.

Japan Elementary Schools

Hubs

- schools
- LINK Page-13
- “¨–{‚ÃŠwçZ
- ça‰„è¨ŠwçZƒzÅ[ƒÄƒyÅ[ƒW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...rnet and Education)
- http://www...iglobe.ne.jp/~IKESAN
- ‚l‚f‚jè¨ŠwçZ‚U”N‚P‘g•®ŒÍ
- è“Š—’¨—ßè“Š—“Œè¨ŠwçZ
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- –y“Ïè¨ŠwçZ‚Ãƒ‚zÅ[ƒÄƒyÅ[ƒW
- UNIVERSITY
- ‰J—≥è¨ŠwçZ DRAGON97-TOP
- é¬‰ªè¨ŠwçZ‚T”N‚P‘gƒzÅ[ƒÄƒyÅ[ƒW
- ∂µ∞È°¬¡© •·•À•Â°º •·•À•Â°º

Authorities

- The American School in Japan
- The Link Page
- ‰ªçËés—ßˆ‰“cè¨ŠwçZƒzÅ[ƒÄƒyÅ[ƒW
- Kids' Space
- ˆ¿èÈés—ßˆ¿èÈê°•”è¨ŠwçZ
- ‹{èÈ‹≥ˆÁ‘ÂŠw•ç‘Æè¨ŠwçZ
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- ê_“fiêÏŒßÅE‰°•lés—ß’†êÏê°è¨ŠwçZ‚Ãƒy
- http://www...p/~m_maru/index.html
- fukui haruyama-es HomePage
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

Things to Note

- Pulled together good pages regardless of language of page content
- Use only link analysis after base set assembled.
  - Iterative scoring is query-independent.

Proof of Convergence

- $n \times n$ Adjacency Matrix A:
  - Each of the n pages in the base set has a row an a column in the matrix.
  - Entry $A_{ij} = 1$ if page i links to page j, else 0.
- View the hubs scores and the authority scores as vectors with n components
  - $h(x) \leftarrow \sum_{x \to y} a(y)$ equivalent to $h \leftarrow Aa$
  - $a(x) \leftarrow \sum_{y \to x} h(y)$ equivalent to $a \leftarrow A^t h$
  - We get $h \leftarrow AA^t h$ and $a \leftarrow A^t Aa$
  - If v is the initial vector
  - Iteration k, h_k is a unit vector in direction $(AA^t)^k v$
  - a_k is a unit vector in direction $(A^t A)^k v$

Proof of Convergence (Cont)

Let $\lambda_1(M), \ldots, \lambda_n(M)$ the eigenvalues of a $n \times n$ matrix M in decreasing order of absolute value. Suppose $|\lambda_1(M)| > |\lambda_2(M)|$. Then

- If M is symmetric, v a vector not orthogonal to the principal eigenvector $w_1(M)$, then
  - Unit vector in the direction $M^k v$ converges to $w_1(M)$.
  - If M has non-negative entries then $w_1(M) \geq 0$.
- Applying these facts to AA^t and (A^tA)
  - The initial vector $v = (1, 1, \ldots, 1)$ is non-orthogonal to $w_1(AA^t)$ and to $w_1(A^tA)$.
  - $(A^tA)^k v$ normalized converges to $w_1(A^tA) = a$.
  - $(AA^t)^k v$ normalized converges to $w_1(AA^t) = h$.

Issues

- Topic Drift
  - Off-topic pages can cause off-topic "authorities" to be returned.
  - E.g., the neighborhood graph can be about a "super topic".
  - HITS favors the most most dense bipartite component of the hub/authority graph.
- Mutually Reinforcing Affiliates
  - Affiliated pages/sites can boost each others' scores.
  - Linkage between affiliated pages is not useful signal.

More Issues, Chakrabarti et al 2001

- Pages are more complex, have more links
- More noise: banners, navigation panels, and advertisements.
- Topic distillation algorithms treat whole pages as atomic, indivisible nodes with no internal structure.
  - Vulnerable to “clique attacks”, a collection of sites linking to each other without semantic reason.

Clique Attacks

from Chakrabarti et al 2001

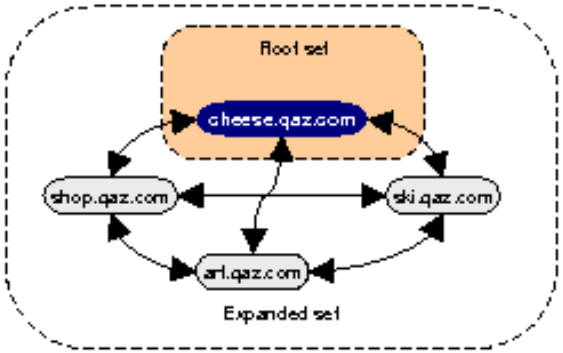
Clique Attacks

Teddington Cheese - **Buy online.** Great selection of cheeses to purchase from this English emporium. Site Guide

All about cheese - **Cheese info.** Search the database of 652 cheeses by country, texture, name or milk. Includes facts and history, and online bookstore.

Cheesiest Site on the Net - **Cheese and more cheese.** Guide to cheese, with tips for making your own, and top cheese retailers.

bet
book
broadcasting
car
card
carhire
cd
charity
chat
cheese
chocolate
christmas
clothing



"Clique attack" on HITS for the query *cheese*.

from Chakrabarti et al 2001

Resources

- Course "The Structure of Information Networks". Jon Kleinberg `http://www.cs.cornell.edu/Courses/cs685/2007f` Lots of papers.
- Book: “Introduction to Information Retrieval”. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. `http://www.informationretrieval.org` Online book and slides!!!
- Course “Data Mining”. Anand Rajaraman, Jeffrey D. Ullman `http://www.stanford.edu/class/cs345a/` Online slides!!!

References

- J. Kleinberg. Authoritative sources in a hyperlinked environment. In Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pages 668-677, ACM Press, New York, 1998
- Lawrence Page and Sergey Brin and Rajeev Motwani and Terry Winograd.The PageRank Citation Ranking: Bringing Order to the Web.
- Link Analysis Ranking Algorithms Theory And Experiments (2004) Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, Panayiotis Tsaparas.
- Chakrabarti S., Mukul M. Joshi and Vivek B. Tawde. Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks.SIGIR 2001