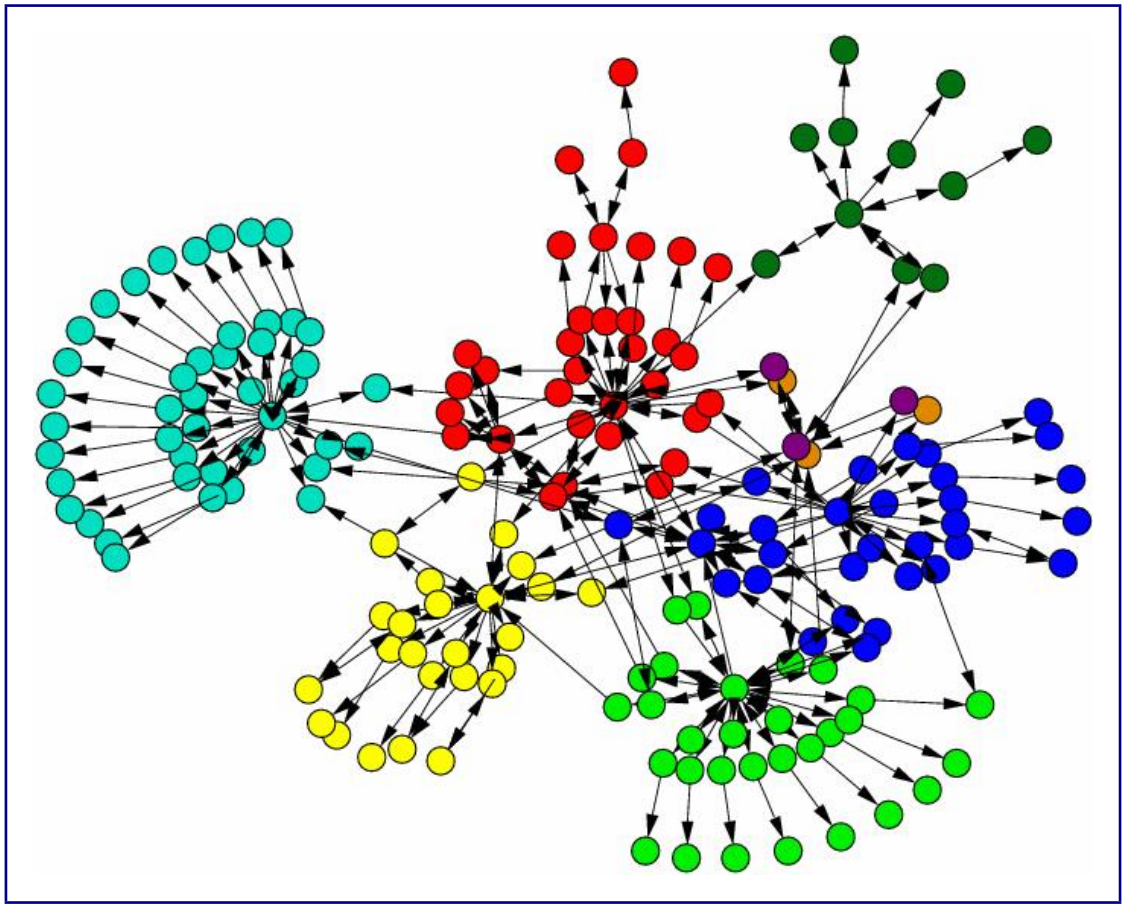
MAE 298, Lecture 9

May 2, 2006



"Web search and decentralized search on small-worlds"

Search for information

Assume some resource of interest is stored at the vertices of a network:

- Web pages
- Files in a file-sharing network

Would like to determine rapidly where in the network a particular item of interest can be found.

To warehouse data or search on demand?

- Centralized : Catalogue data in one central place.

  – Makes most sense when high cost to search network in real time.

  – Requires resources for learning the data and storing it.

- Decentralized : Data is spread out in a distributed data base.

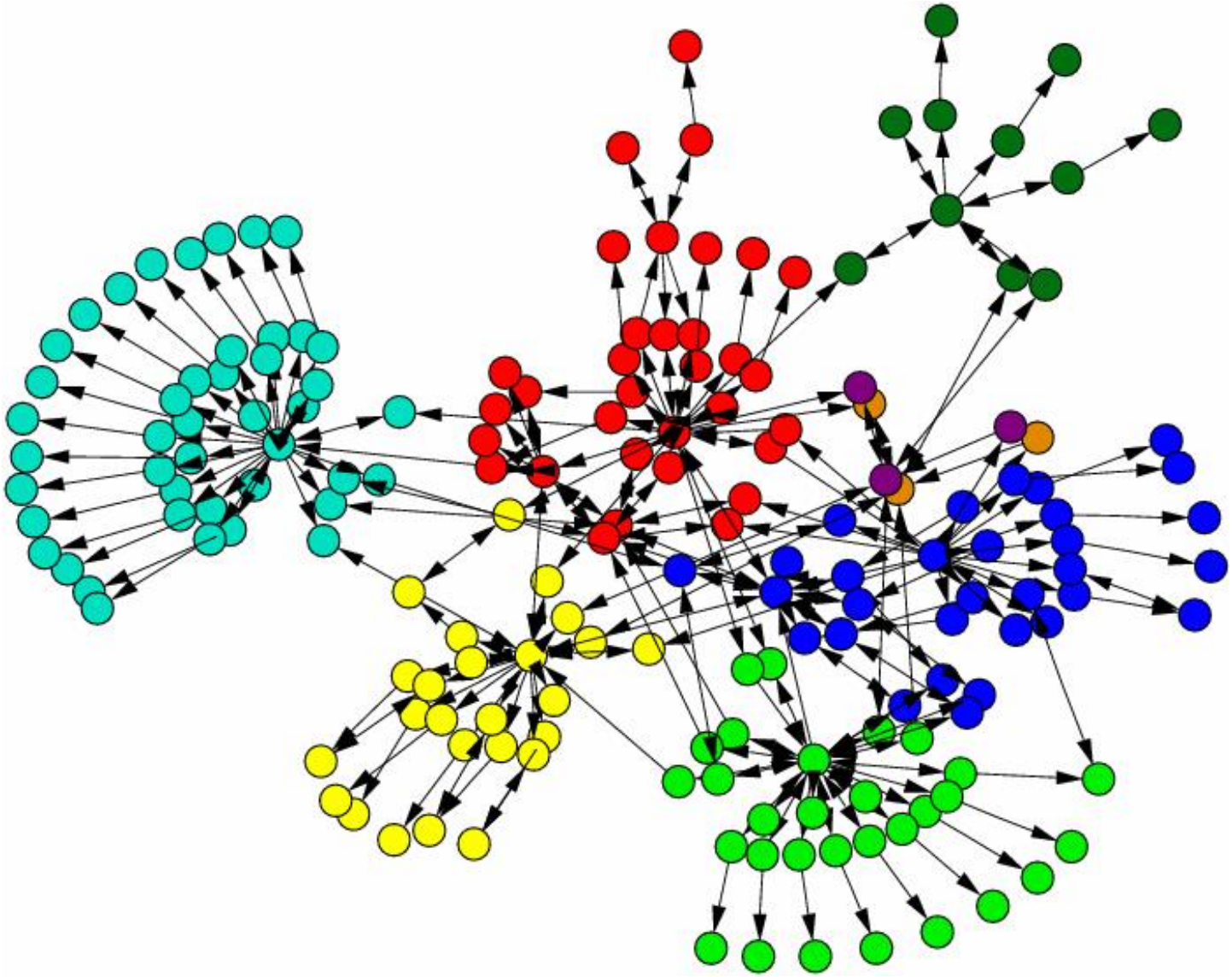
  – Can be a very slow process to search.

  – But dependent on network topology may be able to devise "quick" algorithms.

Web search

- Centralized warehousing of information (need results as quickly as possible)
- Key: Use information contained in the edges as well as the vertices! (Assumes edges contain information about relevance).
- Process: Query arrives, select subset of pages which match, order that subset by ranking based on link structure.

Typical web domain



M. E. J. Newman

Ranking pages in a connected component

- Each site starts with unit "rank" (i.e., weight).
- Transfers fraction of this rank equally to each connected site.
- So the rank of vertex i, r_i:

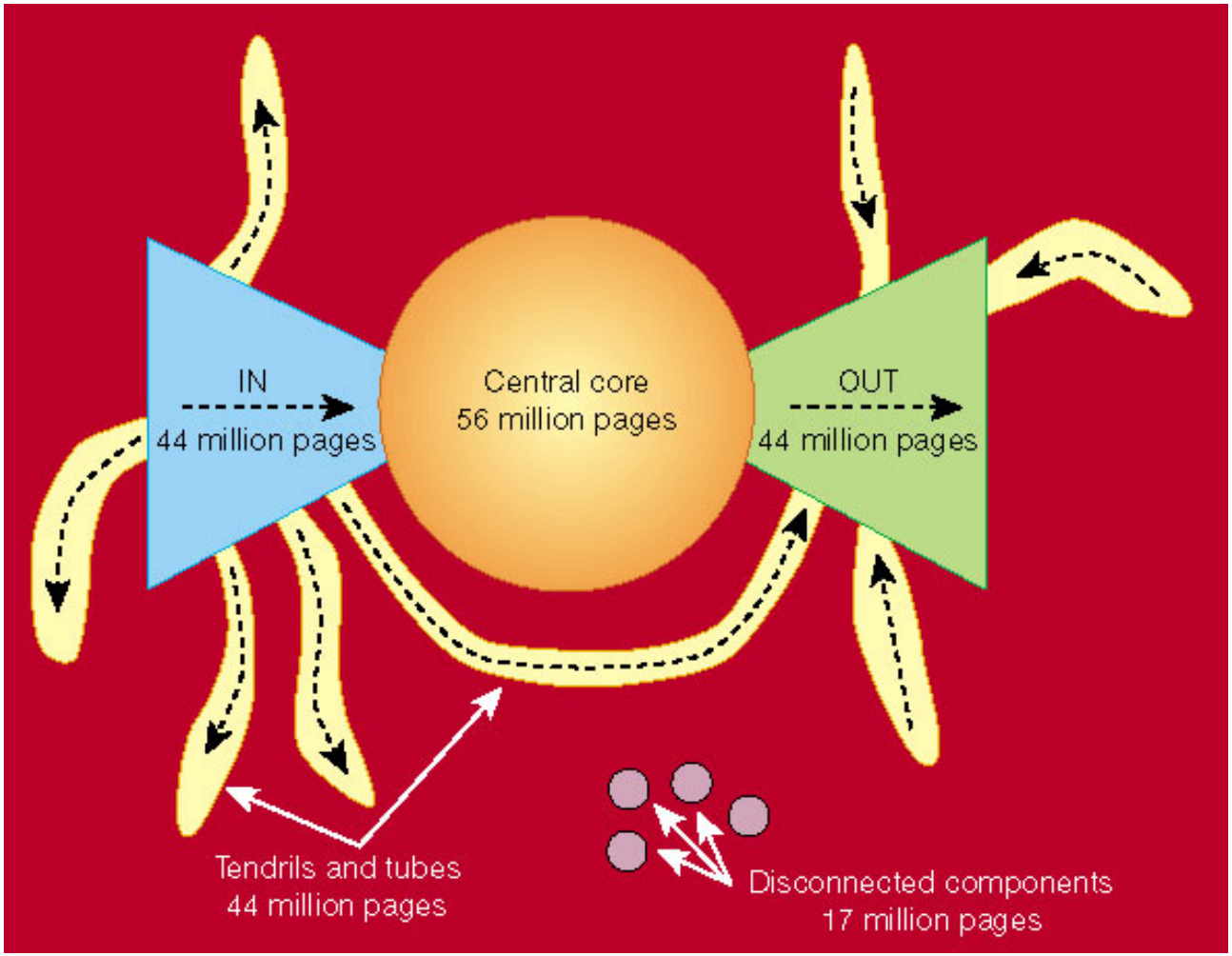
$$r_i \propto \sum_j A_{ij} r_j,$$

  where A is the adjacency matrix.

- Using the random walk analogue, the occupancy probabilities in steady-state (i.e., the vector corresponding the $\lambda = 1$):

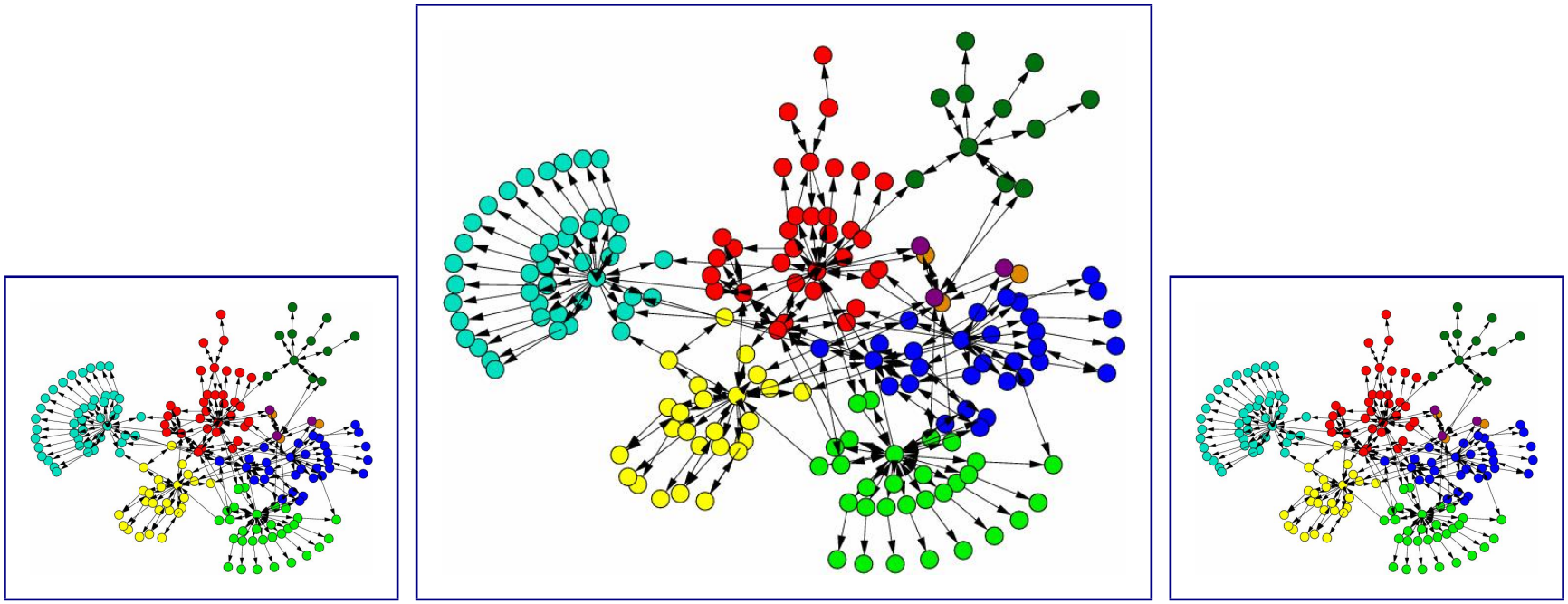
  $\boxed{\vec{r} = M\vec{r}}$, where M is state transition matrix.

The Web as a whole



(This is an old picture, circa 2000, currently Google “8 billion pages served...”, Yahoo claims 20 billion .... Is there really that much valid content out there?)

Disconnected components



We understand how to deal with each components. How do we deal with getting a consistent rank across the whole web?

The "Random Surfer" model

[Brin and Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks*, 30 (1998)]

[L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing Order to the web", technical report, Stanford University, Stanford, CA, 1998.]

- With probability ϵ follow an out-link of current page.
- With probability $[1 - \epsilon]$ jump at random to some other web page. (Usually assume jump is random, so land at any site with prob $1/N$).

The "Random Surfer" model

The weight of a page j is the sum over all the in-links pointing to it, including those gained by the random jump:

$$r_j = \sum_{i \to j} \{ r_i \epsilon(i) + [1 - \epsilon(i)] J_{ij} \} .$$

Rules of thumb:

$$\epsilon(i) = \epsilon$$

$$J_{ij} = 1/N$$

$$\epsilon = 0.8.$$

Real-world complications: Page Rank

- Newer pages have less rank, even though they may be extremely relevant.
- Calculate eigenvalues of 10^9 by 10^9 matrix!!!
- spam, spam, spam
- “Search engine optimizers” (reverse engineer search engines)
- Link farms (both invisible and visible)
- Selling highly ranked domain names
- delisting web sites

Real-world complications in general

- Stop words typically removed: and, of, the to, be, or.
  So how to handle query “to be or not to be”?

- Dealing with complications of multiple languages.

Alternate approaches – topology based

[Kleinberg and Lawrence, "The structure of the Web", *Science*, 294 (2001)]

[Kleinberg, "Authoritative sources in a hyperlinked environment", *J. ACM*, 46 (1999)]

- Slightly more sophisticated. Kleinberg proposes to use in-links and out-Links.
- Google assumes a page is important if other important pages point to it.
- Kleinberg identifies two kinds of importance: "hubs" and "authorities"

Hubs and authorities

- A page pointed to by highly ranked pages in an authority
- A page that points to highly ranked pages is a hub . (May not contain the information, but will tell you where to find it).

In use:

- Teoma search engine (http://search.ask.com/)
- Citeseer literature search engine.

Alternate approaches

- Usage/click based.
- Negative edge weights: (Penalize spam linking to you.)
- Genetic algorithms (Multiple agents searching simultaneously. The "least-fit" killed-off and the "most fit" duplicated. Relies on assumption that pages on a particular topic clustered together). [Menczer and Belew, "Adaptive retrieval agents: Internalizing local context and scaling up to the Web", *Machine Learning*, 39 (2000)]
- Clustering results by subject, e.g., http://clusty.com/
- . . .

Distributed search

Some resource of interest is stored at the vertices of a network: i.e., Files in a file-sharing network

Search on arbitrary networks: $O(N)$

- Depth-first
- Breadth-first

Search on power-law random graphs

- Breadth-first passing always to highest-degree node possible. [Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks", Phys. Rev. E, 64 2001], find between $O(N^{2/3})$ and $O(N^{1/2})$.

Decentralized search on small-world networks

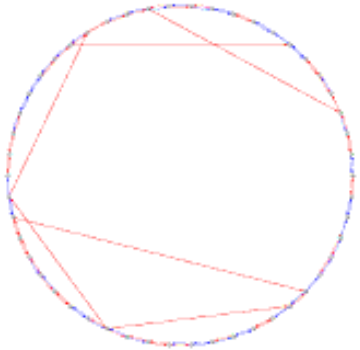
Consider a network with small diameter.

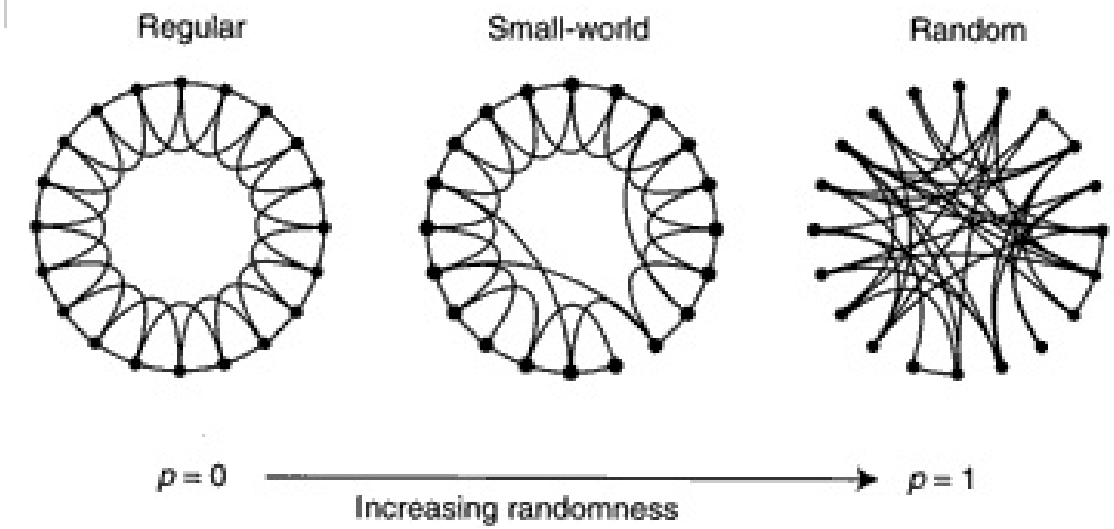
Will this help with decentralized search? (i.e. Find a local algorithm with performance less than $O(N)$?)

What is a small-world

[Watts and Strogatz, “Collective dynamics of ‘small-world’ networks”, *Nature*, 393 (1998)]

Start with regular 1D lattice, add links uniformly at random:





Watts-Strogatz small-world model

- Together with Barabasi-Albert launched the flurry of activity on networks.
- Watts and Strogatz showed that networks from both the natural and manmade world, such as the neural network of C. elegans and power grids, exhibit the small-world property.
- Originally they wanted to understand the synchronization of cricket chirps.
- Introduced the mathematical formalism, which *interpolates between lattices and networks .*

A new paradigm

”I think I’ve been contacted by someone from just about every field outside of English literature. I’ve had letters from mathematicians, physicists, biochemists, neurophysiologists, epidemiologists, economists, sociologists; from people in marketing, information systems, civil engineering, and from a business enterprise that uses the concept of the small world for networking purposes on the Internet.” – Duncan Watts

Navigation

Clearly if central coordination, can use short paths to deliver info quickly.

But, can someone living in a small world actually make use of this info and do efficient decentralized routing?

- Instead of designing search algorithms, given a local greedy algorithm, are there any topologies that enable $O(\log N)$ delivery times?

Precise topologies required

[J. M. Kleinberg, “Navigation in a small world”, *Nature*, 406 (2000)]

- Start with a regular 2D square lattice (consider vertices and edges).
- Add random long links, with bias proportional to distance between two nodes:

$$\boxed{p(e_{ij}) \propto 1/d_{ij}^{\alpha}}$$

- Find mean delivery time $t \sim N^{\beta}$, with $\beta > 1$ unless $\alpha = 2$.
- Only for $\alpha = 2$ will decentralized routing work, and packet can go from source to destination in $O(\log N)$ steps.
- For d-dimensional lattice need $\alpha = d$.

But we know greedy decentralized routing works for human networks (c.f. Milgram's experiments "six-degrees of separation" [S. Milgram, "The small world problem", *Psych. Today*, 2, 1967.]

So how do we get beyond a lattice model?

Navigating social networks

[Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks", Science, 296 (2002)]

[Kleinberg, "Small world phenomena and the dynamics of information", in Proceedings of NIPS 2001].

- Premise: people navigate social networks by looking for common features between there acquaintances and the targets (occupation, city inhabited, age, ...)
- Brings in DATA!

Hierarchical “social distance” tree

- Individuals are grouped into categories along many attributes.
- One tree for each attribute.
- Trees are not the network, but complementary mental constructs believed to be at work.
- Assume likelyhood of acquaintance falls of exponentially with “social distance”.

Summary

- Web search
- Decentralized search

Next time: Epidemiology on networks. (Flow of diseases).