

The Structure of Information Networks

Jon Kleinberg
Cornell University

Information Networks

The information we deal with is taking on a networked character.

- Storage and indexing of massive datasets
 - ▷ Scientific and patent citation databases.
 - ▷ Social networks and collaboration graphs.
 - ▷ The phone call graph.
 - ▷ Consumer preferences, collaborative filtering.
- Creation of content with an intrinsic link structure
 - ▷ The World Wide Web.

A New Continent

The emergence of 'cyberspace' and the
World Wide Web is like the discovery of a new continent.

— Jim Gray

1998 Turing Award address

Complex networks as phenomena — not just technological artifacts.

What structure emerges?

How can we exploit it to understand the underlying content?

Network Structure

- **Unit Structure**

A single node's degree: the number of links to it.

What is the distribution of node degrees?

- **Local Structure**

What does the neighborhood of a node look like?

What small structures recur in the Web's link topology?

- **Global Structure**

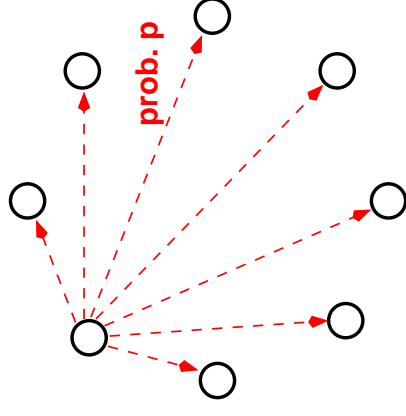
How well connected is the link structure?

What is the typical distance between nodes?

Generative Models

Comparing observed structure with predictions of generative models.

- To what extent can structure be explained by simple models?
- Network models as a testbed for designing heuristics.



Traditional model for randomly generated graphs [Erdős-Rényi 1960]

Start with a fixed set of nodes.

Link v to w independently with probability p .

First observation: degrees strongly concentrated around average.

Very high degrees exponentially unlikely.

Power Laws

Fraction of Web pages with n in-links is $1/n^{2.1}$ (\longrightarrow relatively likely).

[Barabasi et al. 1999], [Broder et al. 2000], [Huberman et al. 1999]

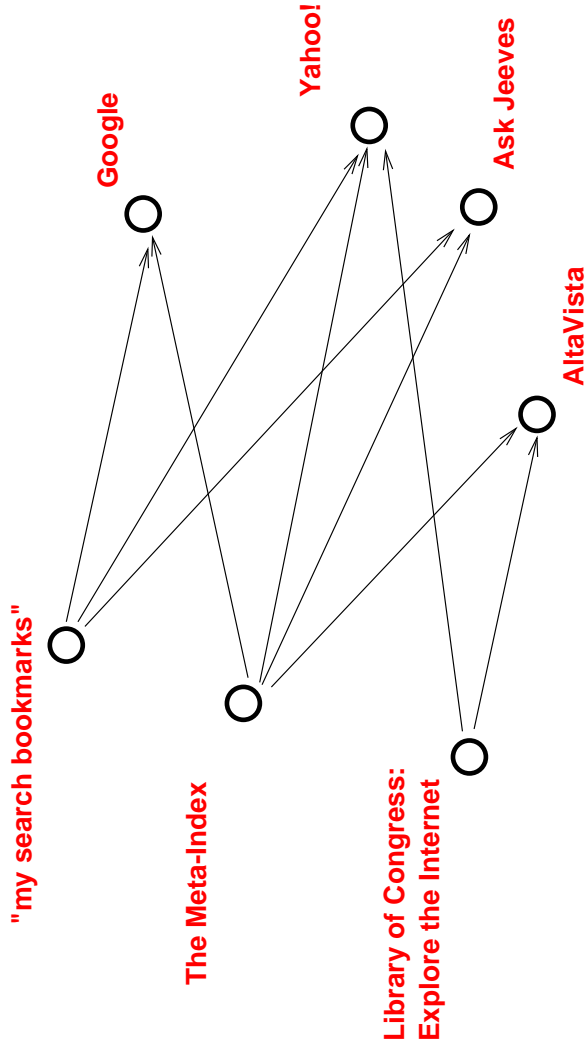
Fraction of Internet domains with n neighbors is $1/n^{2.15}$

[Faloutsos et al. 1999, incl. several other Internet power laws]

Power laws appear throughout socially constructed systems:

- Zipf [1949]: The n^{th} most frequent word occurs at rate $1/n$.
- Lotka [1926]: Fraction of authors with n papers is $1/n^2$.
- Pareto [1897]: Wealth distribution follows a power law.

Local Structure



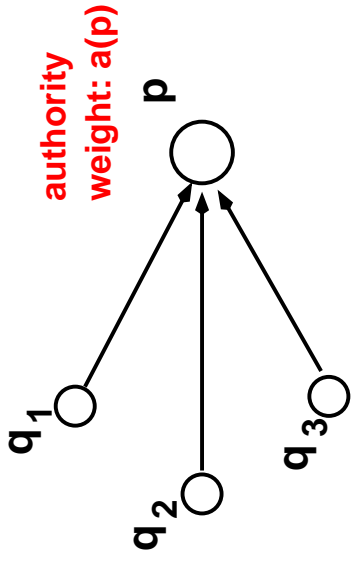
Density of link structure frequently signifies a coherent topic.

The core of a hyperlinked “community” is an inter-linked set of hubs and authorities [Kleinberg 1998].

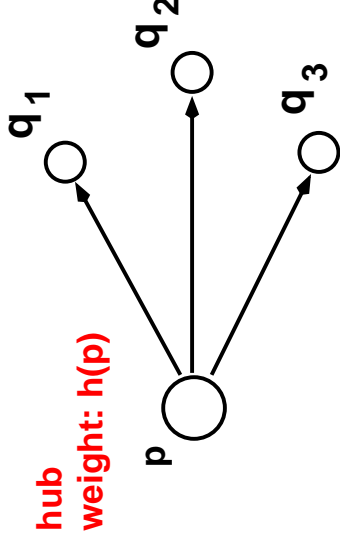
A hub is a page that links to many good authorities.

An authority is a page that is linked to by many good hubs.

Finding Hubs and Authorities



$$a(p) \leftarrow \sum_{q \rightarrow p} h(q)$$

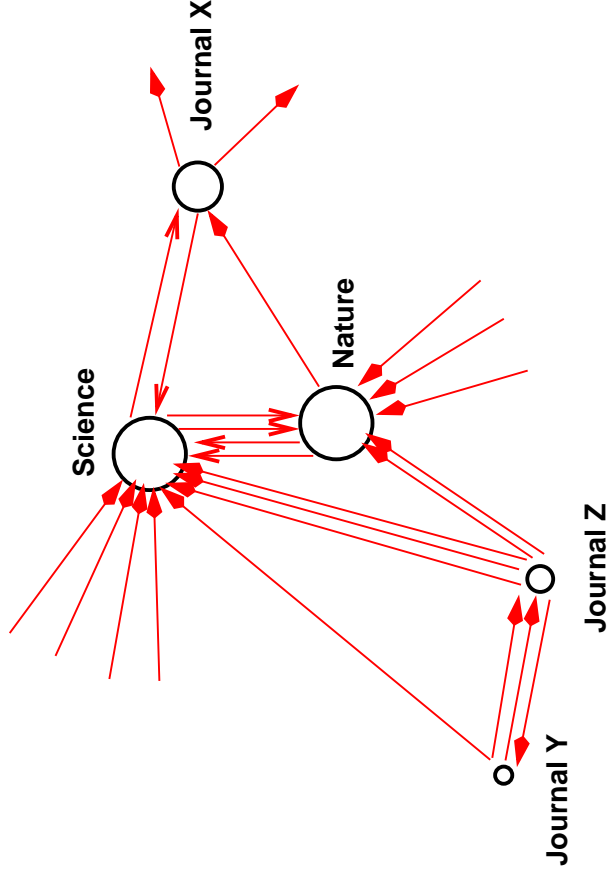


$$h(p) \leftarrow \sum_{p \rightarrow q} a(q)$$

Theorem: Repeated updating operations converge to a fixed point.

Limiting vectors of weights a^* and h^* are eigenvectors of “symmetrized” adjacency matrices $A^T A$ and AA^T .

Bibliometrics, . . .



Measures of centrality in social networks [Katz 1953], [Hubbell 1965].

Garfield, 1972: The Impact Factor. **Normalized citation counting.**

Pinski-Narin, 1976: Influence Weights

- **Influential journals are cited, recursively, by other influential journals.**
- **Weight-propagation: “authorities” directly reinforce other “authorities.”**

Random Walks, . . . and Web Search

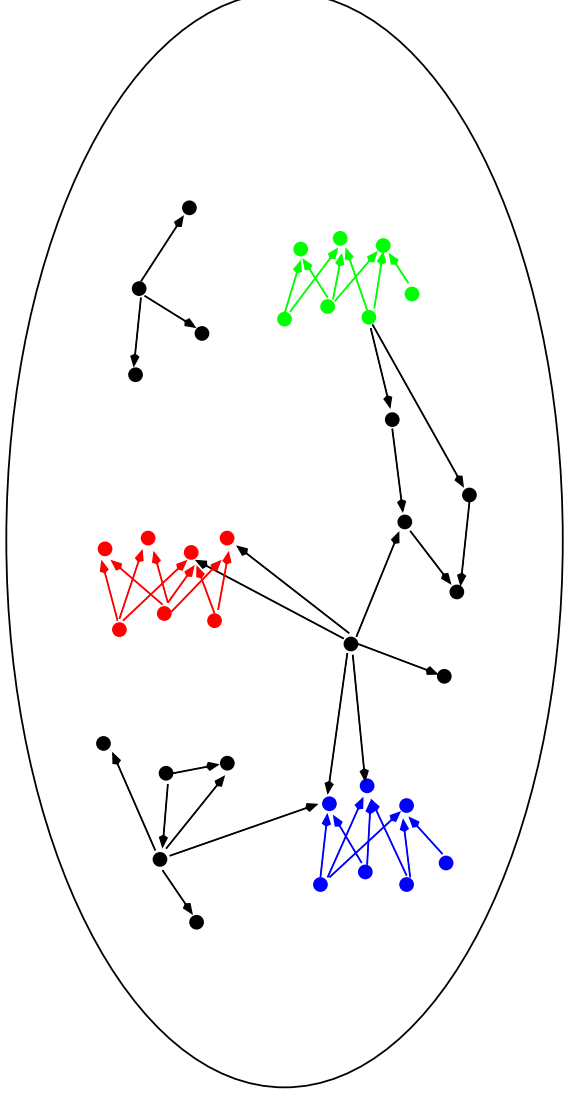
Brin and Page, 1998.

- How should we model Web browsing?
- User who repeatedly follows six random links, then takes random jump.
- Equilibrium of random walk is “one-level” weight-propagation scheme.
 - Simplified variation of hub-authority computation, with no hubs.
- Can be used to rank Web pages.
- Formulation related to Pinski-Narin influence weights.

. . . 

How is authority conferred on the Web?

- Authority-to-authority? Through a layer of hub pages?
(CiteSeer, Teoma, . . .)



Scan a large crawl for “signatures” of Web communities. [Kumar et al. 1999]:

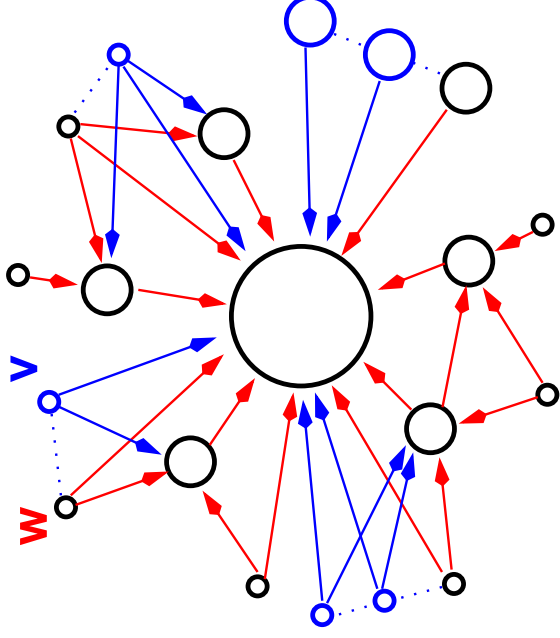
- Expand small complete bipartite subgraphs to hubs/authorities.
- Can discover emerging cyber-communities (over 100,000 to date).

E.g.: Alumni of Delta Sigma Phi fraternity.

Resources on Australian fire brigades.

People concerned with oil spills off the Japanese coast.

- How do we represent and navigate such a structure?



Evolving network model [Kumar et al.], [Barabasi-Albert], [Aiello-Chung-Lu]

- Nodes join the network one at a time.
- v arrives: chooses a node w at random; copies a subset of its out-links.
- The rich-get-richer dynamics of imitation.

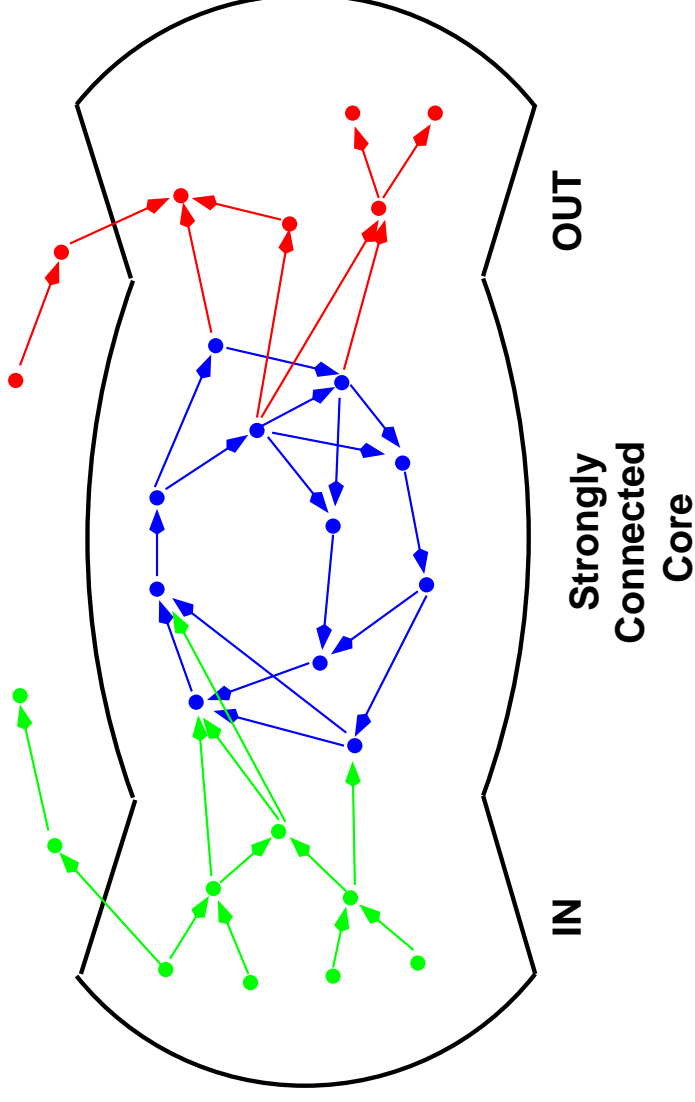
Theorem:

- Fraction of nodes with n in-links is $1/n^\alpha$.
- With high probability, many complete bipartite subgraphs.

Further Analysis

- [Achlioptas-Azar-Fiat-Karlin-McSherry-Saia, 2000 and 2001]
Generative model for networks, where each node has hidden “hub importance” and “authority importance” that determine link formation.
Can prove that hub/authority computation will approximately recover these hidden values.
- [Borodin-Roberts-Rosenthal-Tsaparas 2001]
Identification of principles underlying page ranking algorithms.
Evaluation of algorithms according to these principles.
- [Ng-Zheng-Jordan 2001]
Stability as a goal in the design of page ranking algorithms.
- Consensus on a generative model for networks?
Other meaningful graph-theoretic structures in the Web’s hyperlinks?

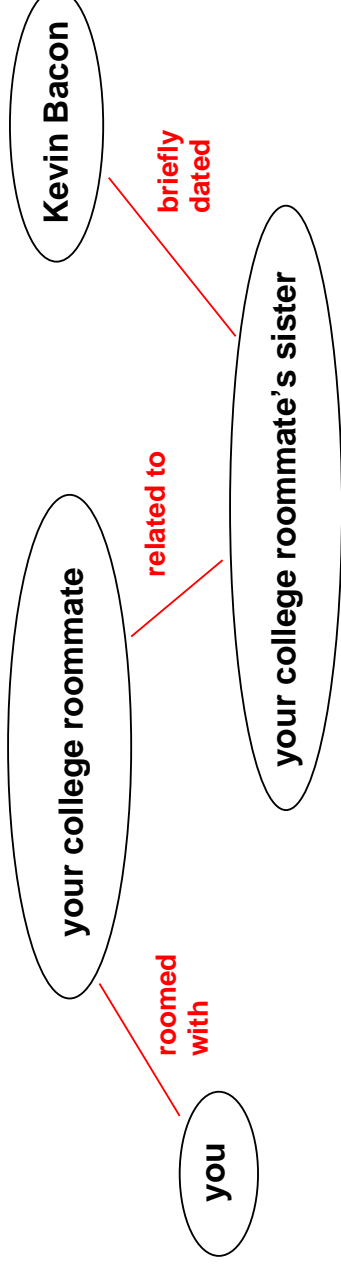
Global Structure



[Broder et al. 2000]

- Decompose Web graph into giant strongly connected component, in-set, out-set, and a collection of “tendrils.”
- Distances in strongly connected core very small: avg. 16-20 links. A “small-world” property [Albert-Jeong-Barabasi 1999].

The Small-World Phenomenon



“It’s a small world”

- Why do strangers so often discover they have a friend in common?
- Or a short chain of friends “connecting” them?
- What does it say about our network of acquaintances?

Network whose nodes are people, and links joining people who know each other on a first-name basis.

Are most pairs of people really linked by short paths?

Seems to require experimental validation. But how?

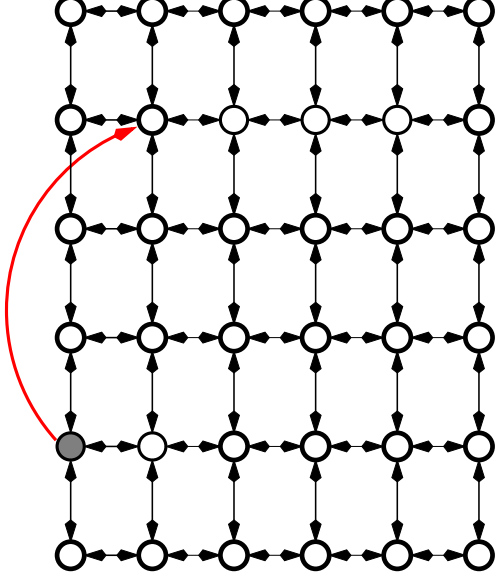
Stanley Milgram's Experiment (1967)

- (1) Pick a **source** person in Nebraska and a **target** person in Massachusetts.
- (2) Tell the source basic information about the target:
name, address, occupation.
- (3) Rules for the source person:
Send the letter to someone you know on a first-name basis, with the goal of reaching the target in as few steps as possible.
- (4) All future recipients in the chain get same information and instructions, plus history.
- (5) Continue until the target receives the letter.

Over chains that completed, average number of intermediate steps was between 5 and 6

→ “six degrees of separation.”

A Small-World Network Model



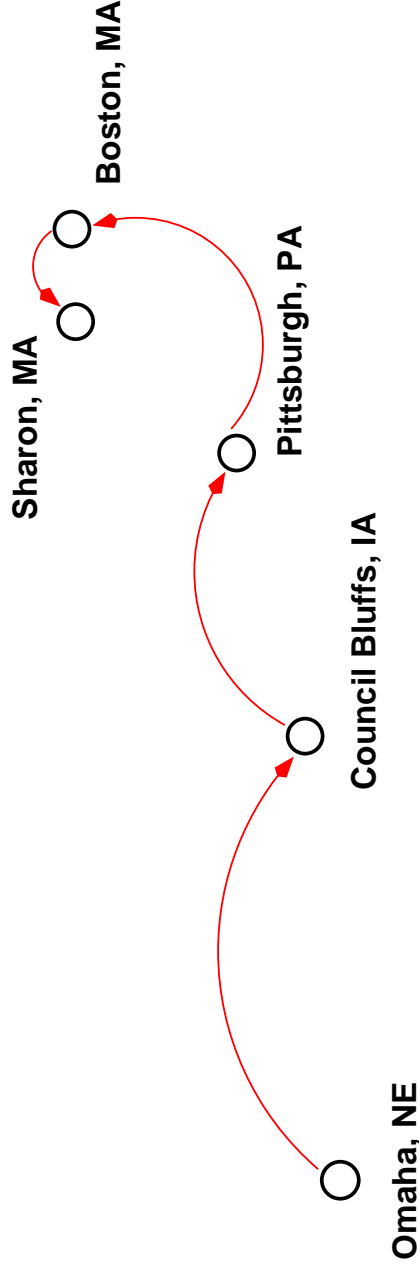
A class of networks with orderly local structure and small diameter

[Watts-Strogatz 1998].

- Start with structured lattice network (e.g. 2-dimensional).
- Add a small number of random links (e.g. 1 per node).
- Diameter drops very quickly, while local neighborhoods remain “clustered.” (See [Bollobás-Chung 1988])

Modeling low-diameter networks as a superposition of two networks.

An Algorithmic Question



- Why should pairs of strangers be able to **find** short chains of acquaintances linking them together?
- What does it say about our network of acquaintances?

What navigational cues do people use in an exponentially branching world?
The network must contain a “gradient” that guides you toward a target.

- ▷ **Goal: Characterize what is possible for algorithms with local information.**
[Kleinberg 2000].

Decentralized Algorithms

Current message holder knows grid structure, destination, and path so far.

Long-range contacts of node v only known if v has touched the message.

Can a decentralized algorithm achieve delivery time polynomial in $\log n$?

(Delivery time = expected number of steps in random network.)

People were very successful at finding short paths.

Watts-Strogatz framework ideal as simple model for asking why:

- Need a model of a network that is partially known and partially unknown.
- Known part has high diameter.
- Full network (known + unknown) has low diameter.
- Structure of known part provides cues for using unknown part.

An Impossibility Result

Theorem:

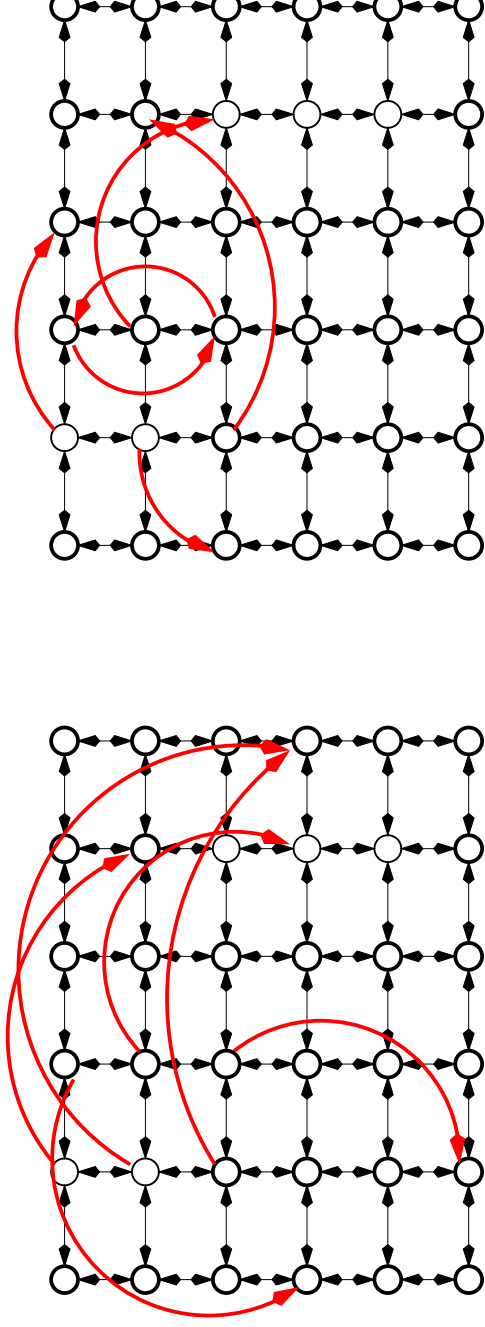
The delivery time of every decentralized algorithm in the Watts-Strogatz network is at least $\varepsilon n^{2/3}$ (for a constant ε).

In contrast, network has diameter polynomial in $\log n$ with high probability...

Despite the low diameter, there is no “gradient” to guide the message.

Can we find a model that has such a gradient?

Generalizing the Network Model



$n \times n$ grid and nearest-neighbor links as before.

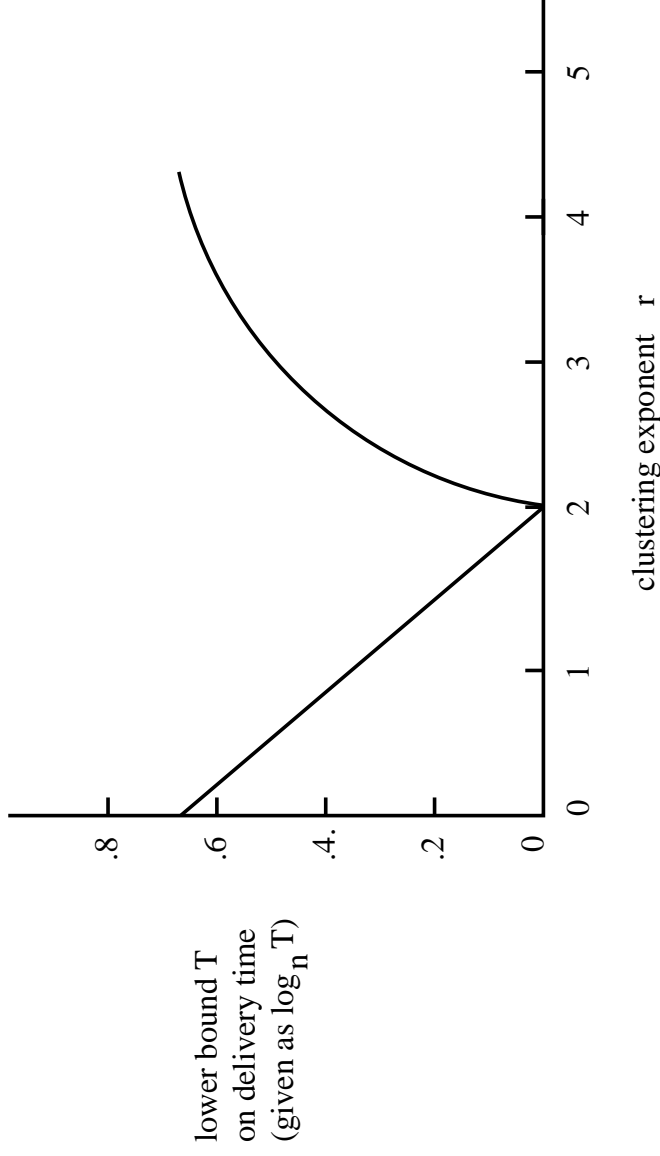
Clustering exponent r .

- For each node v , add a directed link to a random long-range contact.
- Choose w as the contact with probability proportional to $d(v, w)^{-r}$ where $d(v, w)$ is the lattice distance from v to w .

The Inverse r^{th} -power model:

People are more likely to have a nearby contact.

A Full Characterization



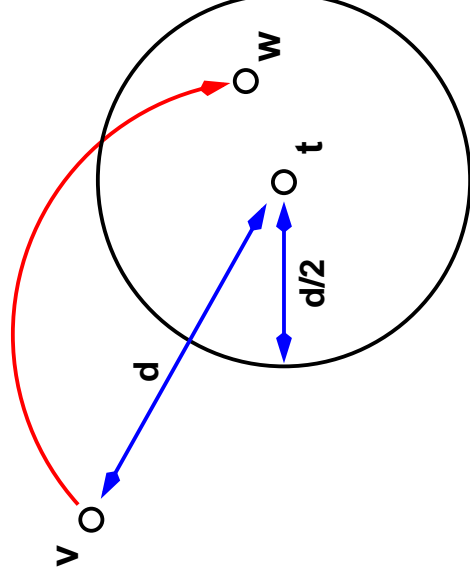
(a) $r = 2$: There is a decentralized algorithm with delivery time proportional to $\log^2 n$.

(b) $r < 2$: Any decentralized algorithm has delivery time $\geq \varepsilon_r n^{(2-r)/3}$.

(c) $r > 2$: Any decentralized algorithm has delivery time $\geq \varepsilon_r n^{(r-2)/(r-1)}$.

The Inverse-Square Network

When $r = 2$, the greedy algorithm is effective:
always forward to contact closest to target.



Key idea in the analysis:

In any step, there is a probability $\geq \epsilon / \log n$
that the message's distance to the target will be halved.

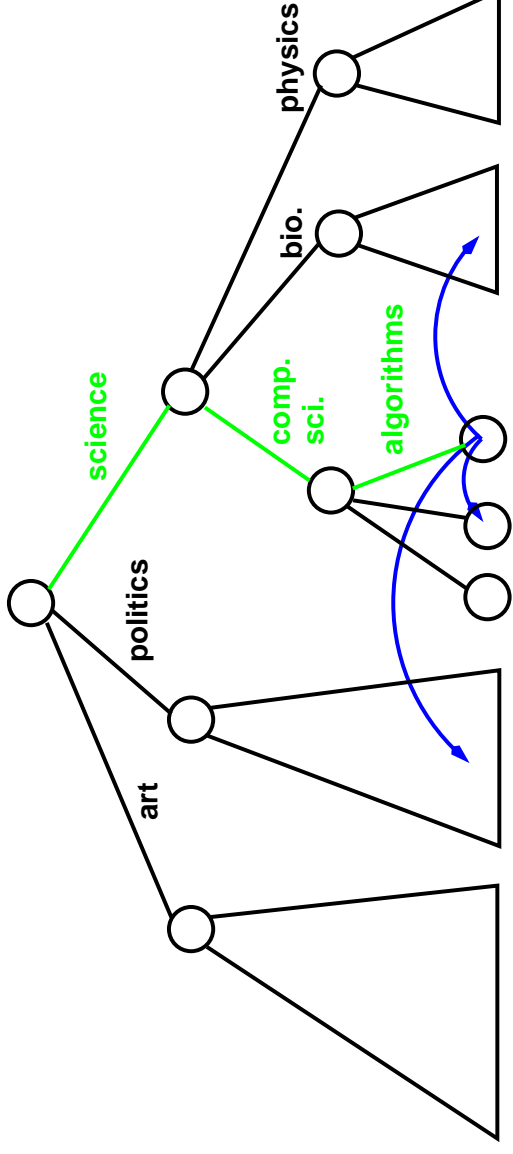
Why an Inverse-Square Law?

- Exponentially layered “distance scales” around nodes.

$$[1, 2], [2, 4], \dots [2^j, 2^{j+1}], \dots$$

- When $r = 2$, nodes have roughly same proportion of links to each distance scale.
(e.g. [Berners-Lee 1999])
- $r < 2$: too long-range.
- $r > 2$: too short-range.

An Analogue for Links on the Web



Instead of a lattice, suppose underlying model is a topic hierarchy.

- Pages reside at leaf nodes. Links created by random process:
 - Each page chooses ancestor at random (i.e. level of generalization)
 - then chooses a leaf at random in that ancestor's subtree.
- Analogue of small-world result:
 - Navigation is most efficient when ancestor chosen uniformly at random.

Reflections

- Decentralized peer-to-peer networks: a small-world navigation problem [Gnutella, Freenet].
- Consensus on generative models for networks?
- Usage, traffic, and temporal phenomena?
- Visualization and partial visualization of complex networks?
- The structures that grow up around information:
Many of these networks were already there.
We just had to start exploring.