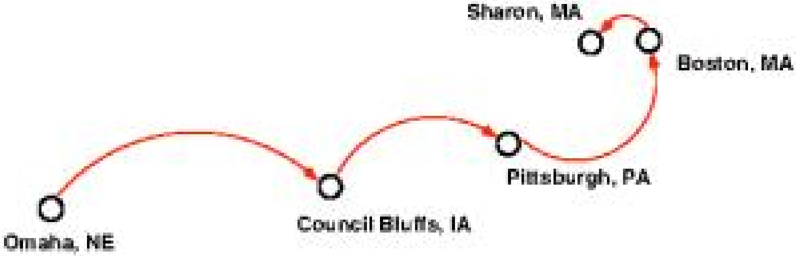
The Structure of Information Networks

Jon Kleinberg

Cornell University

An Algorithmic Perspective

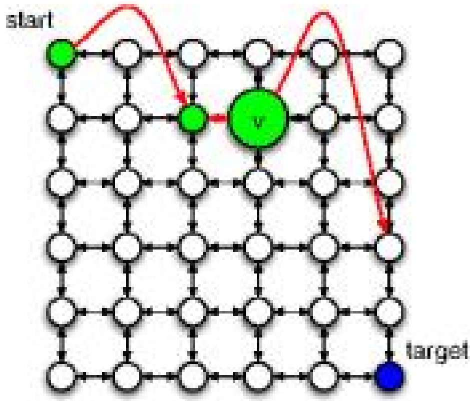


- Algorithmic question: why should pairs of strangers be able to find short chains of acquaintances linking them together?
- How do people navigate in an unknown social network?
- Need models in which local information is sufficient.

Decentralized Algorithms

Decentralized algorithm:

- Given: long-range links(s) of current node v, path so far, location of target.
- Produce: Choice of which neighbor to contact.

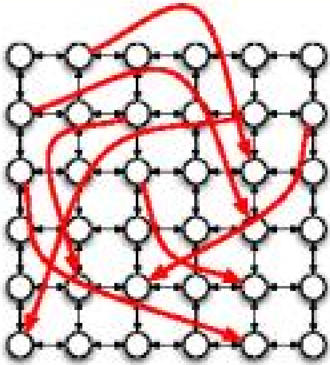


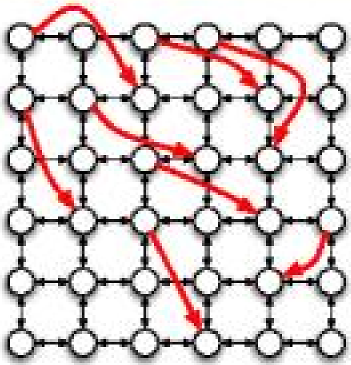
- Delivery time: expected number of steps, over random generation of graph and random start and target.
- Gold standard: network has $n \times n$ nodes, but we want an algorithm with exponentially better delivery time: polynomial function of $\log n$, not n (polylogarithmic).

Impossibility Result

- Theorem [Kleinberg 2000]: There is a constant $c > 0$ such that the delivery time of any decentralized algorithm in the Watts-Strogatz model is at least $cn^{2/3}$.
- Since the diameter is $\leq c' \log n$ for a constant c', this is an exponential gap between the length of the shortest path and the length of the shortest "findable" path.
- Is there a (mild) generalization of the Watts-Strogatz model where decentralized algorithms succeed?

Generalizing the Network Model





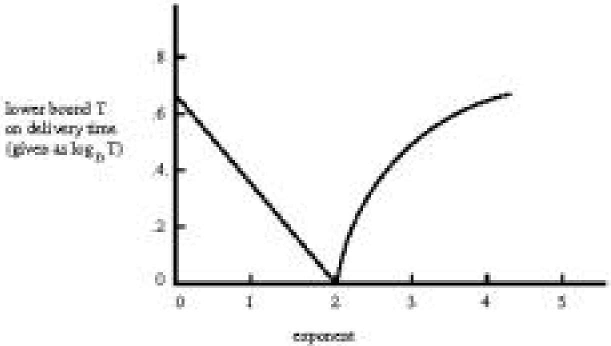
$n \times n$ grid and nearest-neighbor links as before.

Add further parameter α.

- For each v, add directed link to random node.
- Choose w as other end of link with probability proportional to $d(v, w)^{-\alpha}$ where $d(v, w)$ is the lattice distance from v to w.

A type of long-range percolation model [Schulman'83, Newman-Schulman'86, Aizenman-Newman'86, Aizenman et al'88]

Dependence on α



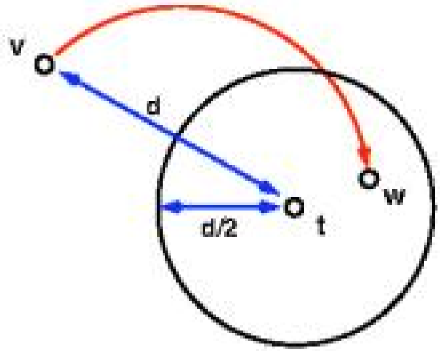
Thm [Kleinberg 2000]: There exist constants c_α ($\alpha \geq 0$) such that

(a) for $\alpha = 2$, there is a decentralized algorithm with delivery time $\leq c_2(\log n)^2$;

(b) for any $\alpha < 2$, the delivery time of any decentralized algorithm is $\geq c_\alpha n^{(2-\alpha)/3}$; and

(c) for any $\alpha > 2$, the delivery time of any decentralized algorithm is $\geq c_\alpha n^{(\alpha-2)/(\alpha-1)}$.

High-level view of analysis



Greedy algorithm:

- Always aim as close to the target as possible.

- Suppose message is at v, distance d from target.
- With probability roughly $1/\log n$, message will enter ball of radius $d/2$ around target.
- Distance to target is halved roughly every $\log n$ steps.
- Distance can only be halved $\log n$ times, so delivery time is bounded by $c(\log n)^2$.

Why an Inverse-Square Law?

- Exponentially layered "distance scales" around nodes.

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

- When $\alpha = 2$, nodes have same proportion of links to each distance scale.
- The right exponent scales with dimension.

Long-Range Percolation

Close connections between searchable small-world networks and long-range percolation models.

- Original model: on D-dimensional integer lattice $\mathbf{Z}^D$, include undirected edge for each pair (v, w) independently with probability $\rho(v, w)^{-\alpha}$, where $\rho(v, w)$ is lattice distance.
- Note (small) differences with models thus far:
  - Graph is undirected and infinite.
  - Node degrees take different values.
- Initial questions concerned existence of infinite connected component [e.g. Schulman'83, Newman-Schulman'86, Aizenman-Newman'86, Aizenman-Chayes-Chayes-Newman'88]

Motivated by small-world model, recent long-range percolation work has considered graph diameter, restricted to finite graphs on $\{1, 2, \ldots, n\}^D$ [Benjamini-Berger '01, Coppersmith et al '02, Biskup '04, Berger '06].

Long-Range Percolation

Diameter results for long-range percolation on $\{1, 2, \ldots, n\}^D$.
(Note: concerned here with existence of paths, not finding paths.)

- $\alpha < D$: Constant diameter (note: very large degrees) [via Benjamini-Kesten-Peres '04].
- $\alpha = D$: Diameter proportional to $\left(\frac{\log n}{\log \log n}\right)$ [Coppersmith et al '02].
- $D < \alpha < 2D$: Diameter is polylogarithmic in n [tight bound due to Biskup '04, '06].
- $\alpha = 2D$: Mainly an open question. (Transition between "small world" and "large world.")
- $\alpha > 2D$: Diameter is linear in n [Berger '06].

Further Algorithmic Aspects

Decentralized search when $\alpha = D$ (redux)

- Since degrees are now logarithmic, greedy algorithm finds paths of length $\leq c \log n$.
- Theorem [Manku-Naor-Wieder 2004]: The following "2-step" algorithm finds paths of length $\leq c' \log n / \log \log n$:
  - Examine all neighbors of your neighbors, and send message to the one closest to target.
- The power of lookahead — and two steps is enough to match the diameter within constant factors.

"Epidemic algorithms" for spreading info. in distributed systems.

- When a node has information, it tells a random other node. Random choice made with probability $\rho^{-\alpha}$. [van Renesse-Birman-Vogels'03, Kempe-Kleinberg-Demers'01]
- $\alpha = 2d$ is the main value used, due to scalability. Lack of understanding of $\alpha = 2d$ is an obstacle to full analysis.

Computing and Information Systems

Rich framework for posing computer science questions:
Computer receives input; produces output.

- How efficiently can output be computed?
- For certain problems, can we prove there is no fast algorithm?

A large and growing aspect of computer use:
Computers as mediators, connecting people to information and to other people.

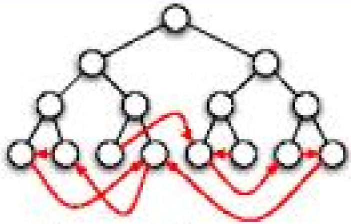
- Information: Web pages, personal digital archives.
- Other people: e-mail, instant messaging, electronic markets.
- In between: blogging, on-line discussion, p2p file-sharing.

How should computer scientists combine these threads?

Searchable Networks on Different "Scaffolds"

- Nodes reside at leaves of a complete b-ary tree [Kleinberg 2002, Watts-Dodds-Newman '02].

  Prob[$v \to w$] decreases in least common ancestor height.

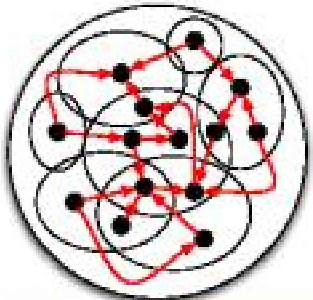


- Nodes reside in a metric space with low combinatorial dimension [Slivkins '05, Fraigniaud-Lebhar-Lotker '06].
- Nodes belong to a graph of low tree-width, or with fixed excluded minor [Fraigniaud '05, Abraham-Gavoille '06].

General model based on set systems.

- Consider a set system C on the collection of nodes.

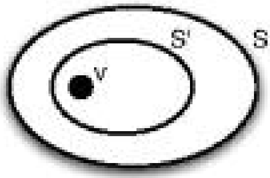
  $g(v, w)$ = size of smallest set containing v, w.
  Link probability decreases in $g(v, w)$.



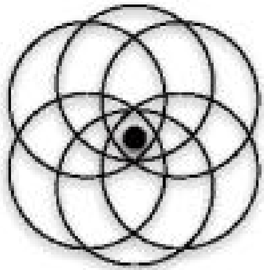
Building a Network on a Set System

Consider a set system C over a ground set of nodes V, such that $V \in C$, and satisfying the following two properties (for parameters $\lambda < 1$ and $\kappa > 1$).

(i) If $v \in S \in C$, then there exists $S' \in C$ such that $v \in S' \subseteq S$ and $\min(\lambda|S|, |S| - 1) \leq |S'| < |S|$.



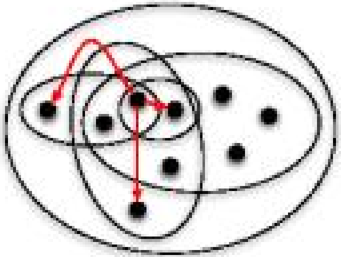
(ii) If $\cap_i S_i$ is non-empty, then $|\cup_i S_i| \leq \kappa \max_i |S_i|$.



Building a Network on a Set System

Random graph model with out-degree $k(n)$ and exponent γ:

- Each node v generates $k(n)$ out-links, choosing w as endpoint of i^{th} link independently with probability prop. to $g(v, w)^{-\gamma}$.



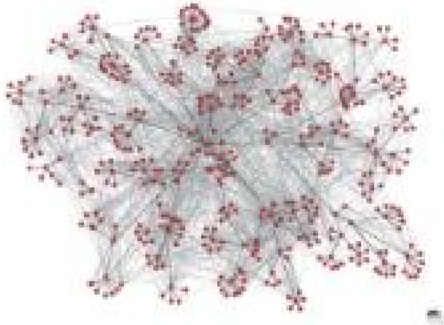
Theorem[Kleinberg 2002]: For arbitrary $\mathcal{C}$ satisfying (i), (ii):

(a) There is a decentralized algorithm with polylogarithmic delivery time in the random graph model with set system $\mathcal{C}$, exponent $\gamma = 1$, and out-degree $k = c(\log n)^2$ (suff. large c).

(b) For every $\gamma < 1$, and every polylogarithmic function $k(n)$, there is no decentralized algorithm achieving polylogarithmic delivery time in the random graph model with set system $\mathcal{C}$, exponent γ and out-degree $k(n)$.

Social Network Data

- [Adamic-Adar 2003]: social network on 436 HP Labs researchers.
- Joined pairs who exchanged ≥ 6 e-mails (each way).



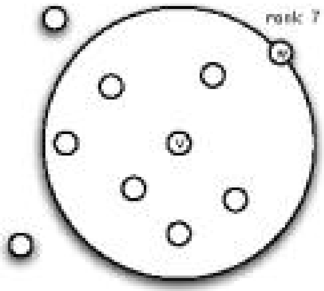
- Adamic-Adar compared to set system model.
  - Probability of link (v, w) prop. to $g(v, w)^{-\gamma}$, where $g(v, w)$ is size of smallest group containing v and w.
  - $\gamma = 1$ gives optimal search performance.
- In HP Labs, groups defined by sub-trees of hierarchy.
- Links scaled as $g^{-3/4}$.

Geographic Data: LiveJournal

Liben-Nowell, Kumar, Novak, Raghavan, Tomkins (2005) studied LiveJournal, an on-line blogging community with friendship links.

- Large-scale social network with geographical embedding:
  - 500,000 members with U.S. Zip codes, 4 million links.
- Analyzed how friendship probability decreases with distance.
- Difficulty: non-uniform population density makes simple lattice models hard to apply.

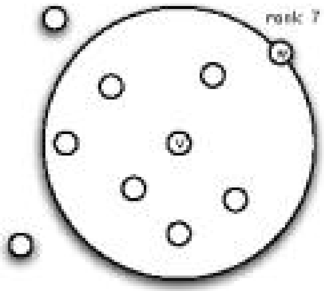
LiveJournal: Rank-Based Friendship

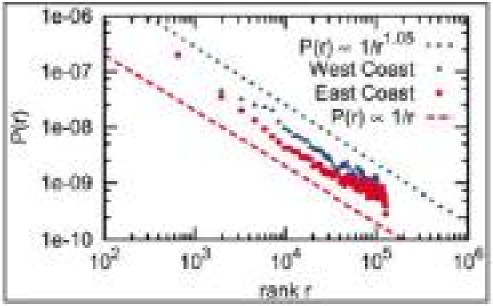


Rank-based friendship: rank of w with respect to v is number of people x such that $d(v, x) < d(v, w)$.

- Result of [LKNRT'05]: Efficient routing for (nearly) arbitrary population density, if link probability proportional to $1/\text{rank}$.
- Generalization of lattice result (diff. from set systems).

LiveJournal: Rank-Based Friendship





Rank-based friendship: <u>rank</u> of w with respect to v is number of people x such that $d(v, x) < d(v, w)$.

- Result of [LKNRT'05]: Efficient routing for (nearly) arbitrary population density, if link probability proportional to $1/\mathrm{rank}$.
- Generalization of lattice result (diff. from set systems).

Punchline: LiveJournal friendships approximate $1/\mathrm{rank}$.

Peer-to-Peer File-Sharing

Internet file-sharing (Napster, Freenet, Kazaa, Morpheus, ...)

- After demise of Napster, centralized solutions not feasible: File-sharing becomes a small-world search problem.
- Each node has some files and some neighbors it knows. When file request arrives, node asks neighbors to help locate.

Decentralized peer-to-peer systems

- Gnutella: brute-force flooding of network.
- Freenet [Clarke et al. '00]: small-world-style directed search.
- Research prototypes place nodes in "virtual" Cartesian space and perform search relative to these coordinates.
  Chord [Stoica et al. '01], CAN [Ratnasamy et al. '01], Tapestry [Zhao et al. '01], Pastry [Rowstron et al. '01], Viceroy [Malkhi et al. '02], Symphony [Manku et al '03].

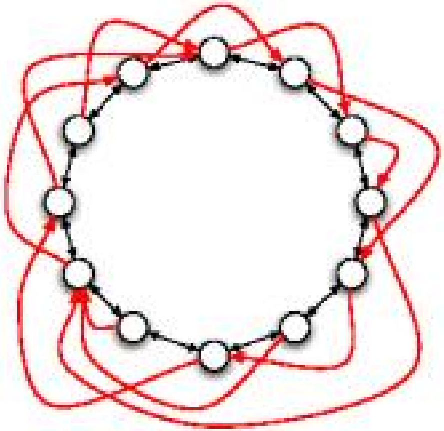
Reflections

- Computer science research questions arising from
  - large volumes of networked information, and
  - large collections of people interacting on-line.
- Rich data lets us study questions that would have been impossible to formulate 20 years ago.
- Need more powerful frameworks for modeling aggregate properties in these networks.
  - Algorithmic and probabilistic models of network phenomena.
  - Algorithmic game theory to model network interactions [Papadimitriou 2001, Tardos 2004, Roughgarden 2005].
- Research arising from the tension between privacy and massive datasets on human activity.

Open Question: Network Evolution

What causes a network to evolve toward searchability?

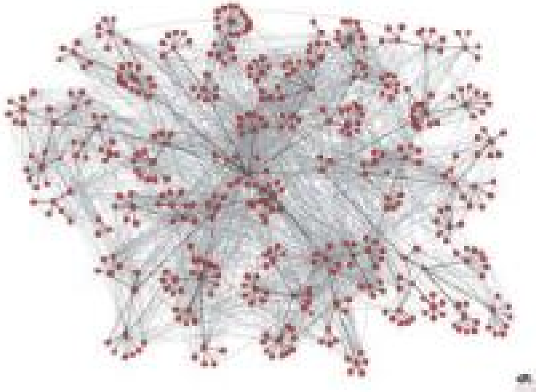
- A proposal by Sandberg and Clarke 2006, based on their work on Freenet:



- n nodes on a ring, each with neighbor links and a long link.
- At each time $j = 1, 2, 3, \ldots$, choose random start s, target t, and perform greedy routing from s to t.
- Each node on resulting path updates long-range link to point to t, independently with (small) probability p.

The Role of Networks

- Networks play a central role in studying large-scale information systems.
- Model as undirected or directed graphs $G = (V, E)$

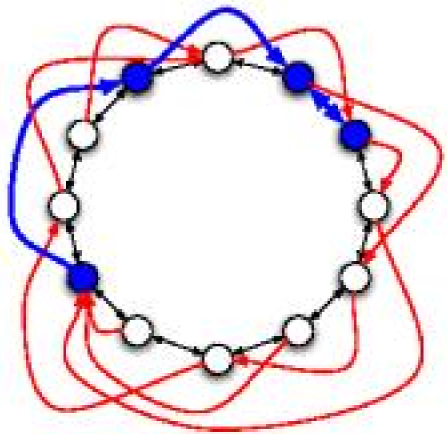


- Communication networks: Internet (routers, links)
- Information networks: World Wide Web (pages, hyperlinks)
- Social networks: e-mail, instant messaging (people, message exchange).

Open Question: Network Evolution

What causes a network to evolve toward searchability?

- A proposal by Sandberg and Clarke 2006, based on their work on Freenet:

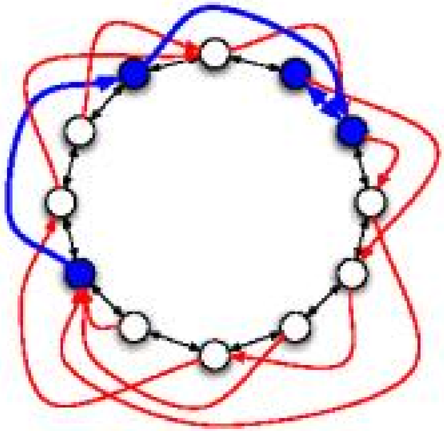


- n nodes on a ring, each with neighbor links and a long link.
- At each time $j = 1, 2, 3, \ldots$, choose random start s, target t, and perform greedy routing from s to t.
- Each node on resulting path updates long-range link to point to t, independently with (small) probability p.

Open Question: Network Evolution

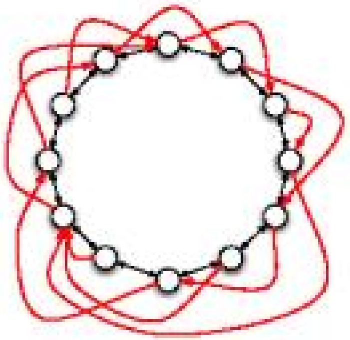
What causes a network to evolve toward searchability?

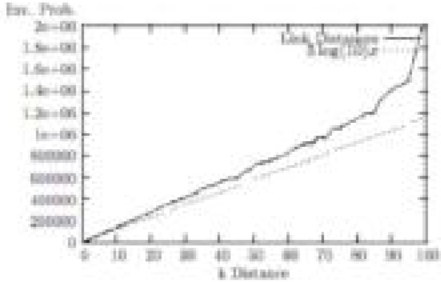
- A proposal by Sandberg and Clarke 2006, based on their work on Freenet:



- n nodes on a ring, each with neighbor links and a long link.
- At each time $j = 1, 2, 3, \ldots$, choose random start s, target t, and perform greedy routing from s to t.
- Each node on resulting path updates long-range link to point to t, independently with (small) probability p.

Open Question: Network Evolution





This defines a Markov chain on labeled graphs.

Conjecture [Sandberg-Clarke 2006]:

- At stationarity, distribution of distances spanned by long-range links is (close to) theoretical optimum for search.
- At stationarity, expected length of searches is polylogarithmic.
- Conjectures are supported by simulation.

Reflections

- Computer science research questions arising from
  - large volumes of networked information, and
  - large collections of people interacting on-line.
- Rich data lets us study questions that would have been impossible to formulate 20 years ago.
- Need more powerful frameworks for modeling aggregate properties in these networks.
  - Algorithmic and probabilistic models of network phenomena.
  - Algorithmic game theory to model network interactions [Papadimitriou 2001, Tardos 2004, Roughgarden 2005].
- Research arising from the tension between privacy and massive datasets on human activity.

Networks as Phenomena

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 designed artifacts.
- Models rooted in graph theory and discrete probability: random graphs, random walks, percolation theory.
- Algorithmic models in the social sciences.

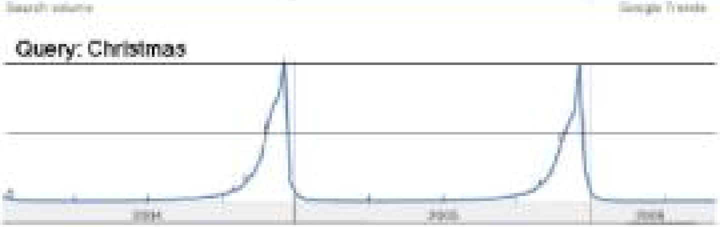
Models as Design Principles

Models become principles for analyzing on-line data and designing systems.

A first example: link analysis for Web search.

- Spectral properties of the hyperlink graph form the basis for modern search engine ranking functions.
- Novel types of matrix factorization, applied to the adjacency matrix, expose tightly-knit communities.
- The "search economy": bidding for ads on search keywords. New game-theoretic analysis and mechanism design issues.

Second example: Aggregate behavior



- Is there a basic vocabulary of usage patterns?
- Data streams: Analysis based on probabilistic models, property testing, communication complexity, harmonic analysis.
- How do we link individual behavior to aggregate properties?

Small-World Networks and Decentralized Search

Exploring an information network using only local information.

- Origins in research in social psychology
  - The small-world experiment [Milgram 1967]
- Initial models
  - Graphs based on superposition of structure and randomness [Watts-Strogatz 1998, Kleinberg 2000]
- Abstracting a general pattern
- Identifying the pattern in large-scale network data
  - Web hyperlinks [Menczer 2002]
  - E-mail communication in an organization [Adamic-Adar 2003]
  - Friendships in on-line communities [Liben-Nowell et al. 2005]
- The models as design principles
  - Decentralized peer-to-peer file-sharing systems
- Results and open questions in long-range percolation

Milgram's Small-World Experiment (1967)

(1) Pick a source person in Nebraska and target 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, to try reaching the target in as few steps as possible.

(4) All future recipients in the chain get same information and instructions, plus history.

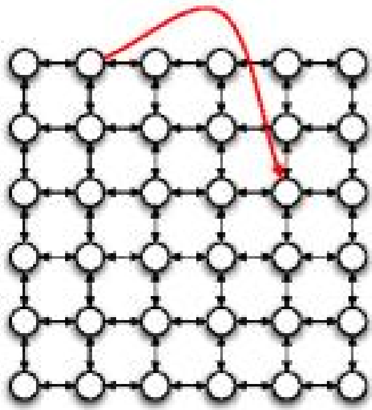
(5) Continue until the target receives the letter.

Over completed chains, median number of steps was 6

$\longrightarrow$ "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 grid 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." (cf. [Bollobás-Chung 1988])

Modeling low-diameter networks as a superposition of two.