Greedy Search in Social Networks

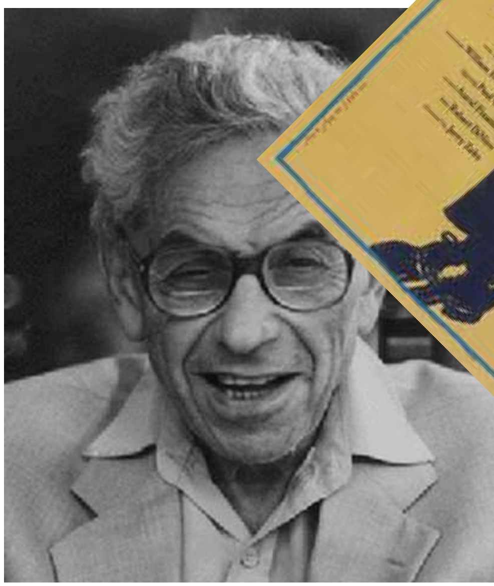
David Liben-Nowell

Carleton College

`dlibenno@carleton.edu`

Joint work with Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins.

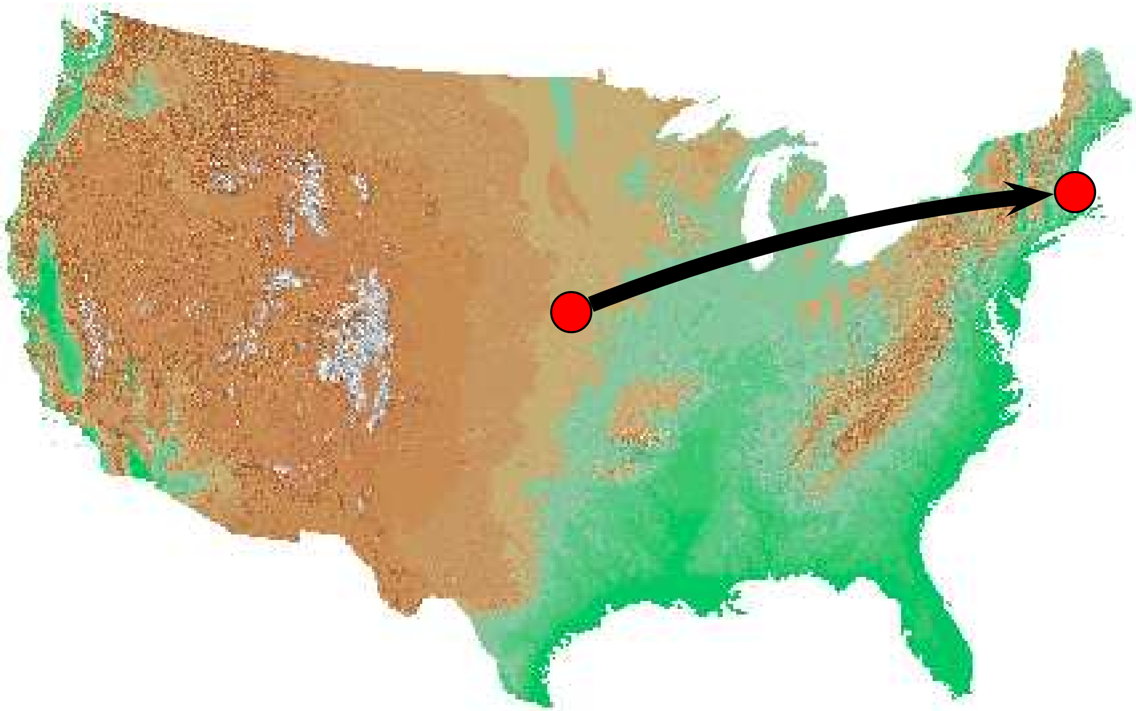
IPAM, Los Angeles | 8 May 2007




Six Degrees of Separation
by
John Guare
Lincoln
Center
Theater

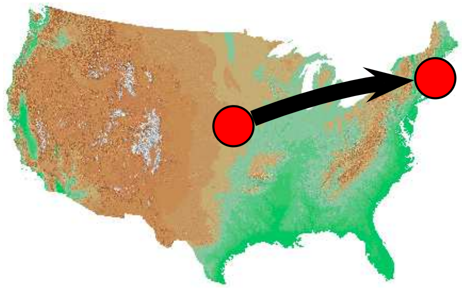
Milgram: Six Degrees of Separation

Social Networks as Networks: [Milgram 1967]



- People given letter, asked to forward to one friend.
- Source: random Omahaians;
  Target: stockbroker in Sharon, MA.
- Of completed chains, averaged six hops to reach target.

Milgram: The Explanation?



"the small-world problem"

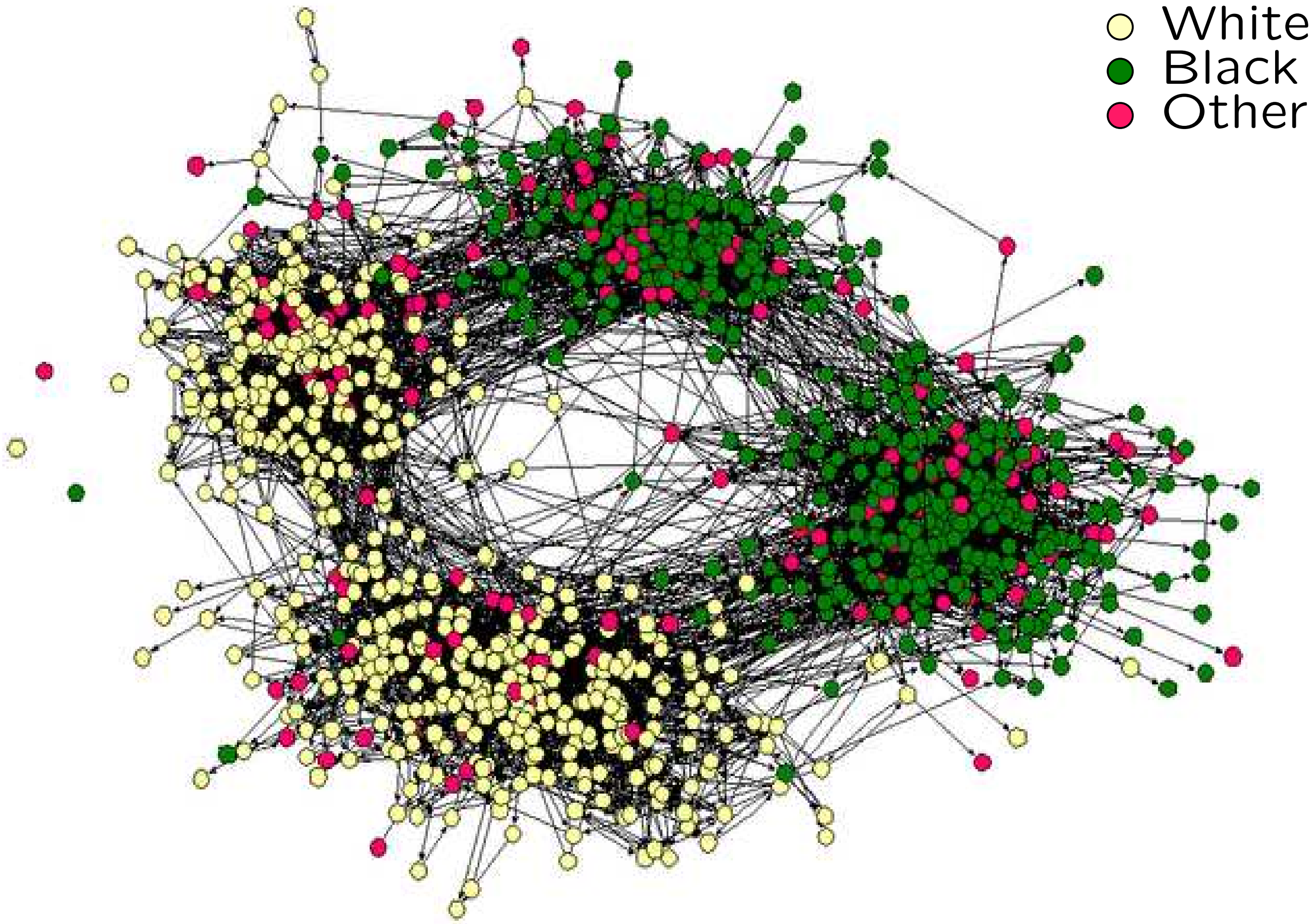
➡ Why is a random Omahaian close to a Sharon stockbroker?

Standard (pseudosociological, pseudomathematical) explanation:
(Erdős/Rényi) random graphs have small diameter.

Bogus! In fact, many bogosities:

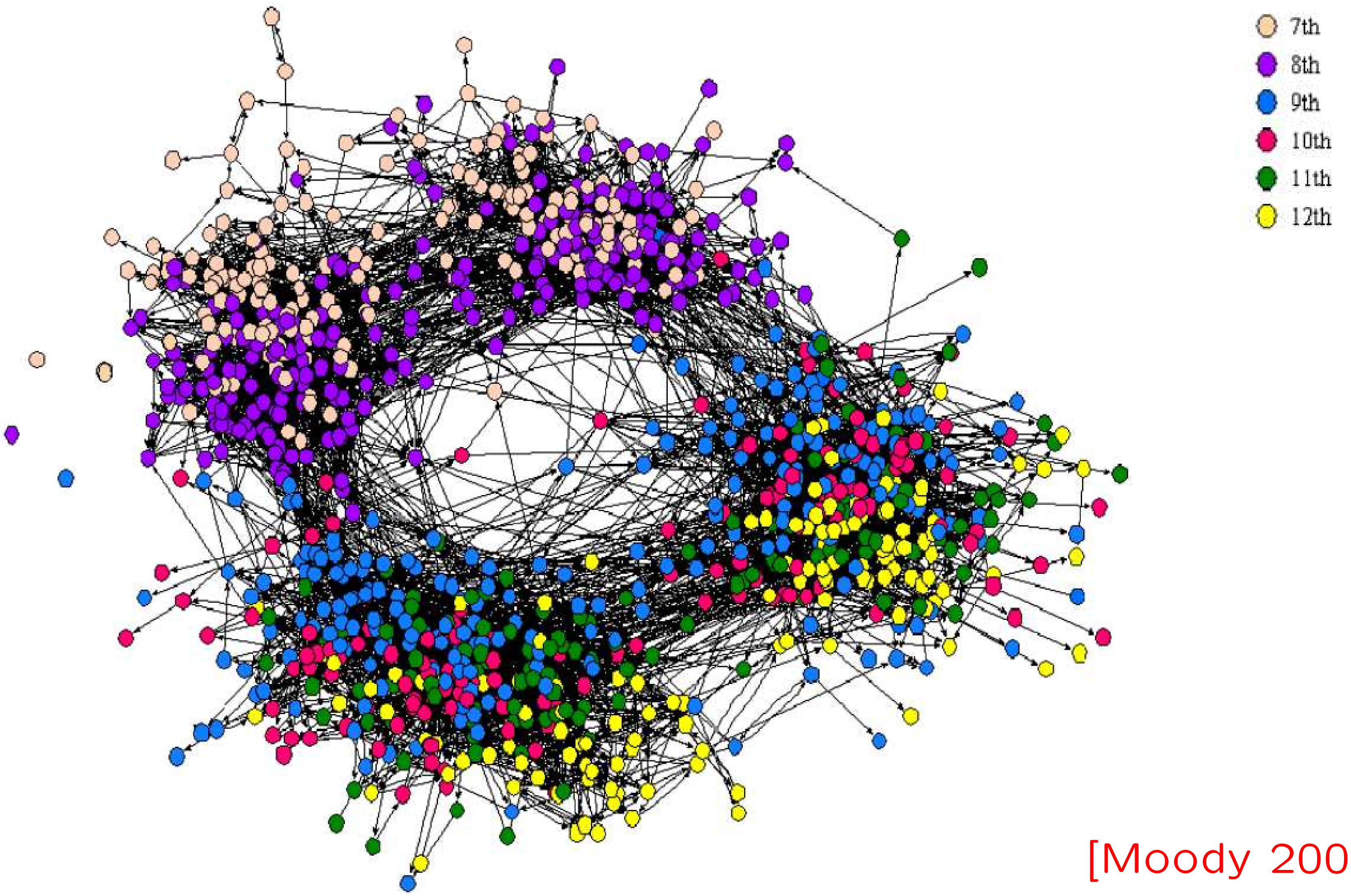
- degree distribution
- clustering coefficients
- ...

High School Friendships



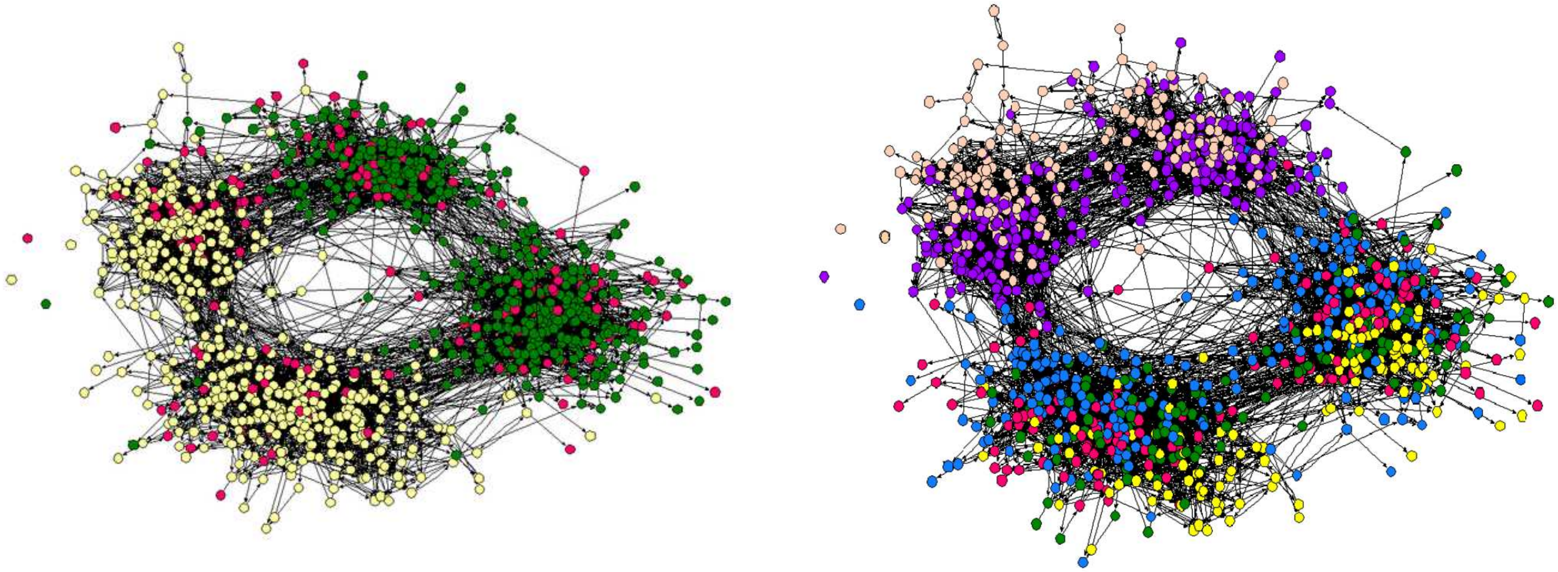
Self-reported high school friendships. [Moody 2001]

High School Friendships



[Moody 2001]

Homophily



homophily: a person x's friends tend to be 'similar' to x.

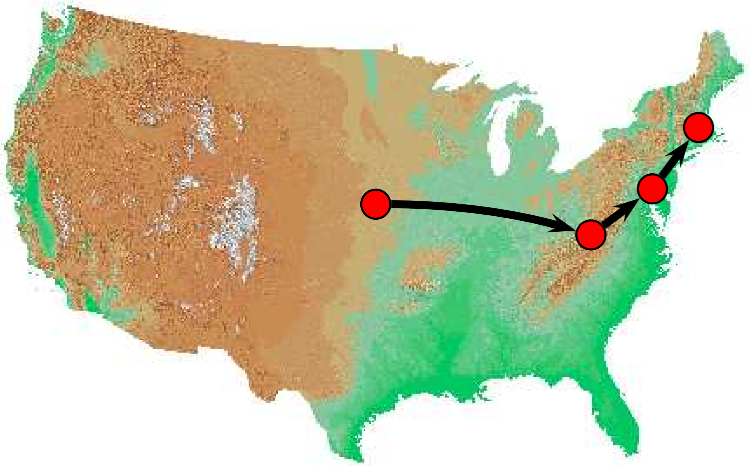
One explanation for high clustering: (semi)transitivity of similarity.

$$x, y \text{ both friends of } u \;\approx\!\!\!\Rightarrow\; x \text{ and } u \text{ similar; } y \text{ and } u \text{ similar}$$
$$\approx\!\!\!\Rightarrow\; x \text{ and } y \text{ similar}$$
$$\approx\!\!\!\Rightarrow\; x \text{ and } y \text{ friends}$$

Navigability of Social Networks

[Kleinberg 2000]
Milgram experiment shows more than small diameter:
People can construct short paths!

Milgram's result is algorithmic, not existential.



Homophily and Greedy Applications

homophily: a person x's friends tend to be similar to x.

Key idea: getting closer in "similarity space"
$\Rightarrow$ getting closer in "graph-distance space"

[Killworth Bernard 1978] ("reverse small-world experiment")
[Dodds Muhamed Watts 2003]
In searching a social network for a target,
most people chose the next step because of
"geographical proximity" or "similarity of occupation"
(more geography early in chains; more occupation late.)

Suggests the greedy algorithm in social-network routing:
if aiming for target t, pick your friend who's 'most like' t.

Greedy Routing

Greedy algorithm:
if aiming for target t, pick your friend who's 'most like' t.

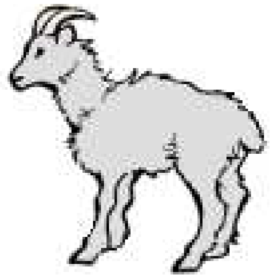
Geography: greedily route based on distance to t.

Occupation: $\approx$ greedily route based on distance in the (implicit) hierarchy of occupations.

Want $\texttt{Pr}[u, v$ friends$]$ to decay smoothly as $d(u, v)$ increases.

(Need social 'cues' to help narrow in on t.
Not just homophily! Can't just have many disjoint cliques.)

The LiveJournal Community



`www.livejournal.com`

"Baaaaah," says Frank.

- Online blogging community.
- Currently 12.8 million users; ∼1.3 million in February 2004.

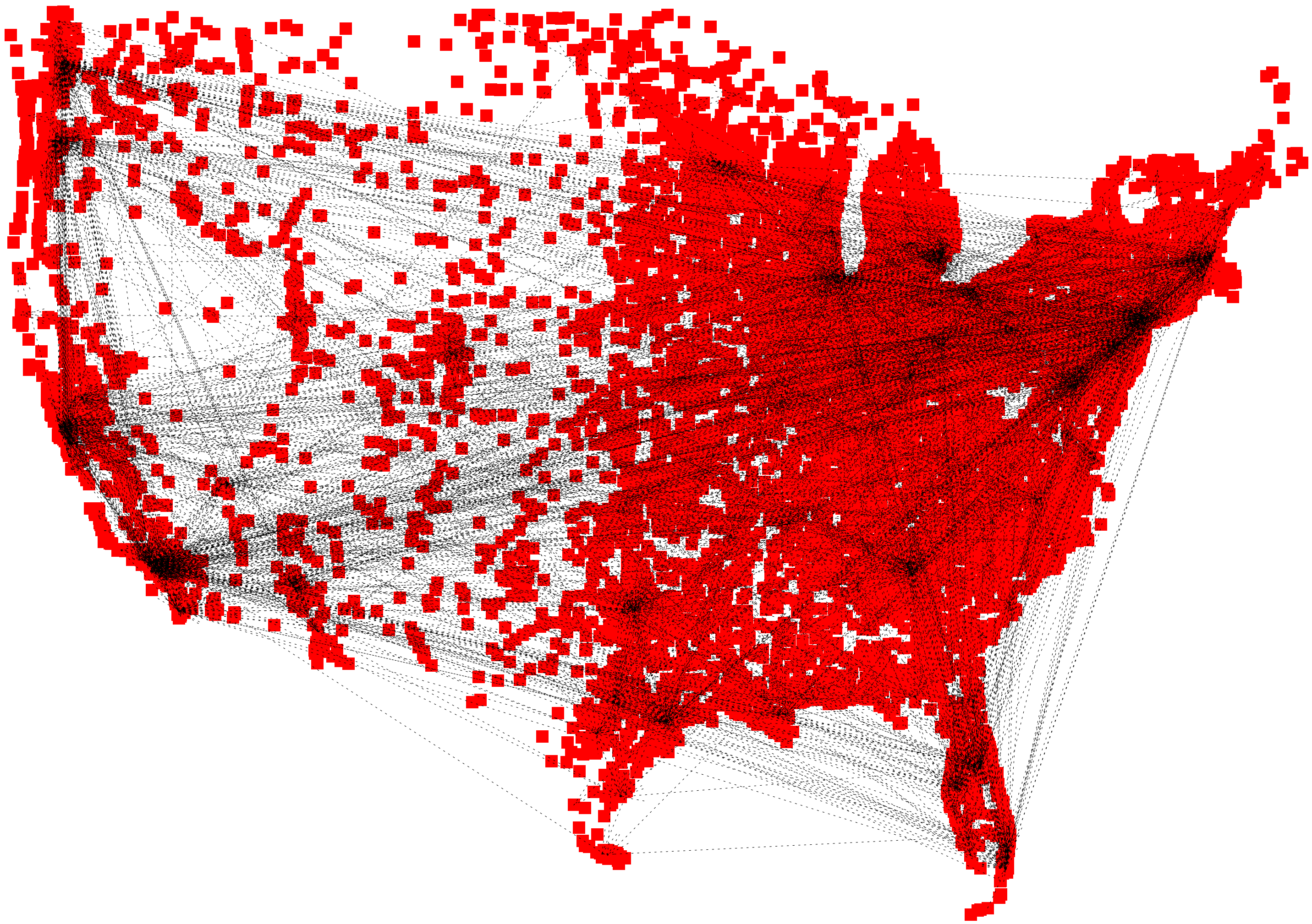
LiveJournal users provide:

- disturbingly detailed accounts of their personal lives.
- profiles (birthday, hometown, explicit list of friends)

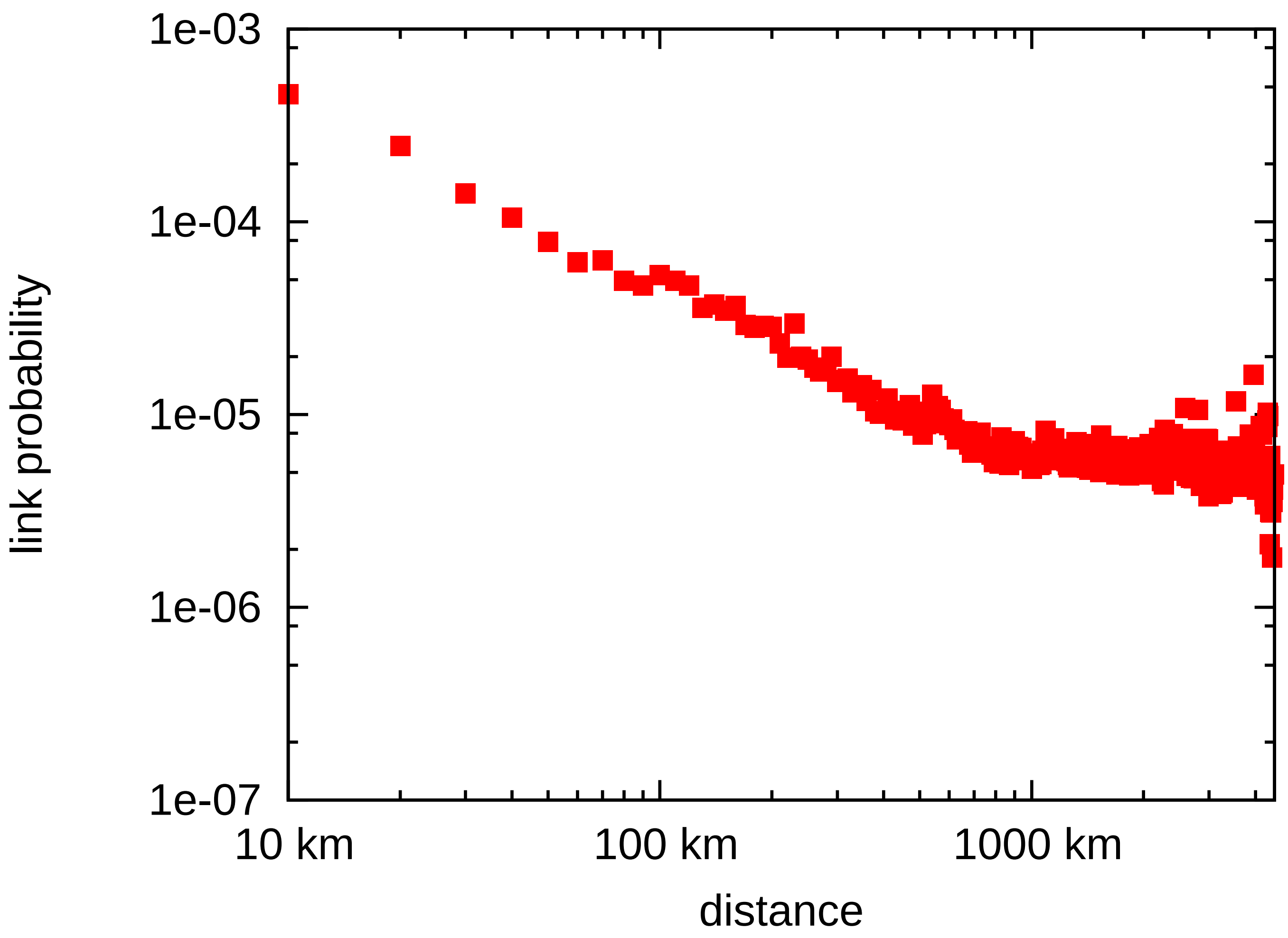
- Yields a social network, with users' geographic locations.
  (∼500K people in the continental US.)

LiveJournal

0.1% of LJ friendships



Distance versus LJ link probability



The Hewlett-Packard Email Community

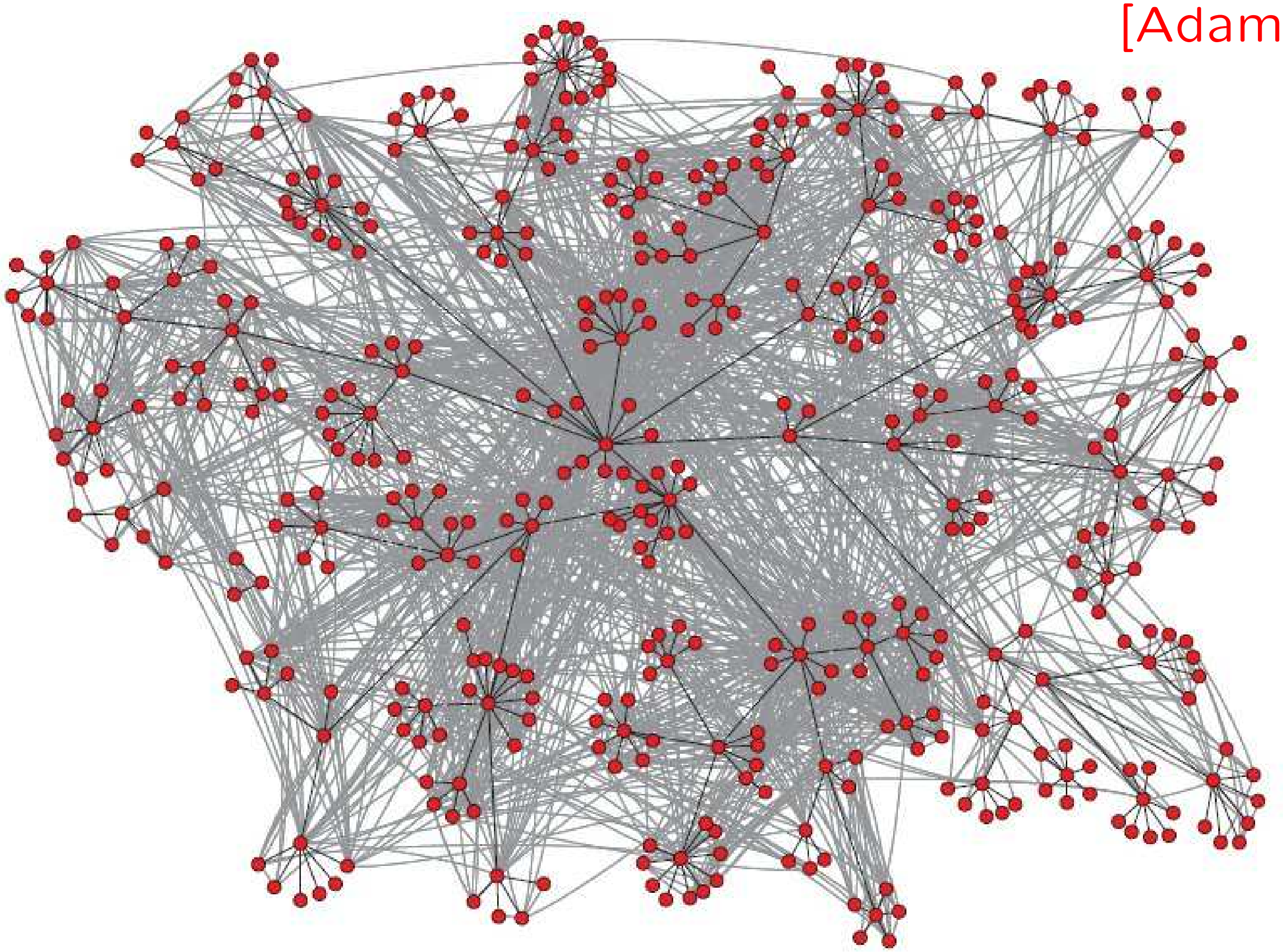
[Adamic Adar 2005]

- Corporate research community.
- Captured email headers over $\sim$3 months.
- Define friendship as ≥ 6 emails $u \to v$ *and* ≥ 6 emails $v \to u$.

- Yields a social network ($n = 430$),
  with positions in the corporate hierarchy.

Emails and the HP Corporate Hierarchy

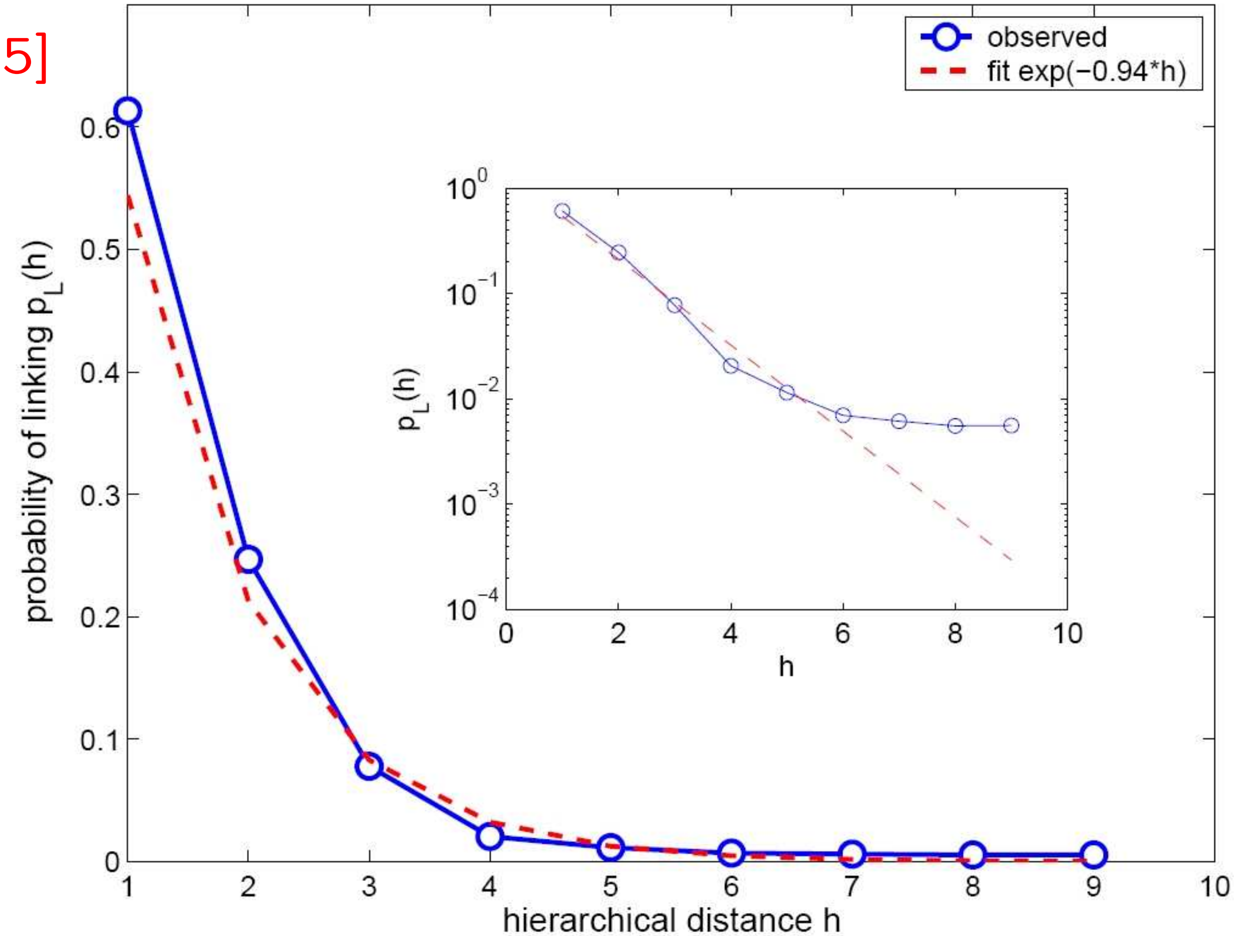
[Adamic Adar 2005]



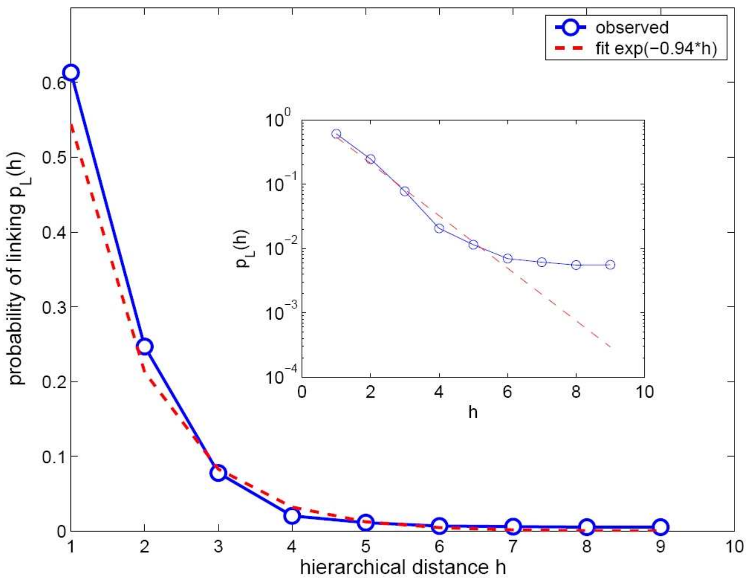
black: HP corporate hierarchy   gray: exchanged emails.

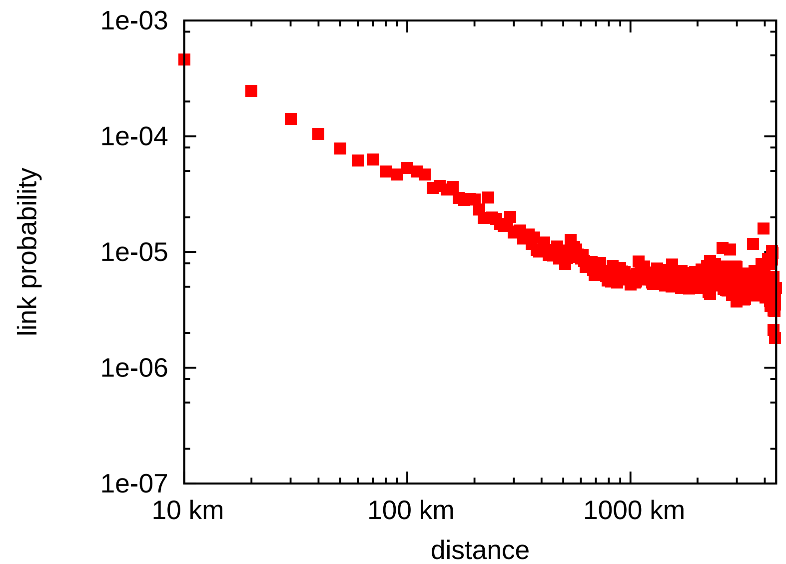
Emails and the HP Corporate Hierarchy

[Adamic Adar 2005]



Requisites for Navigability

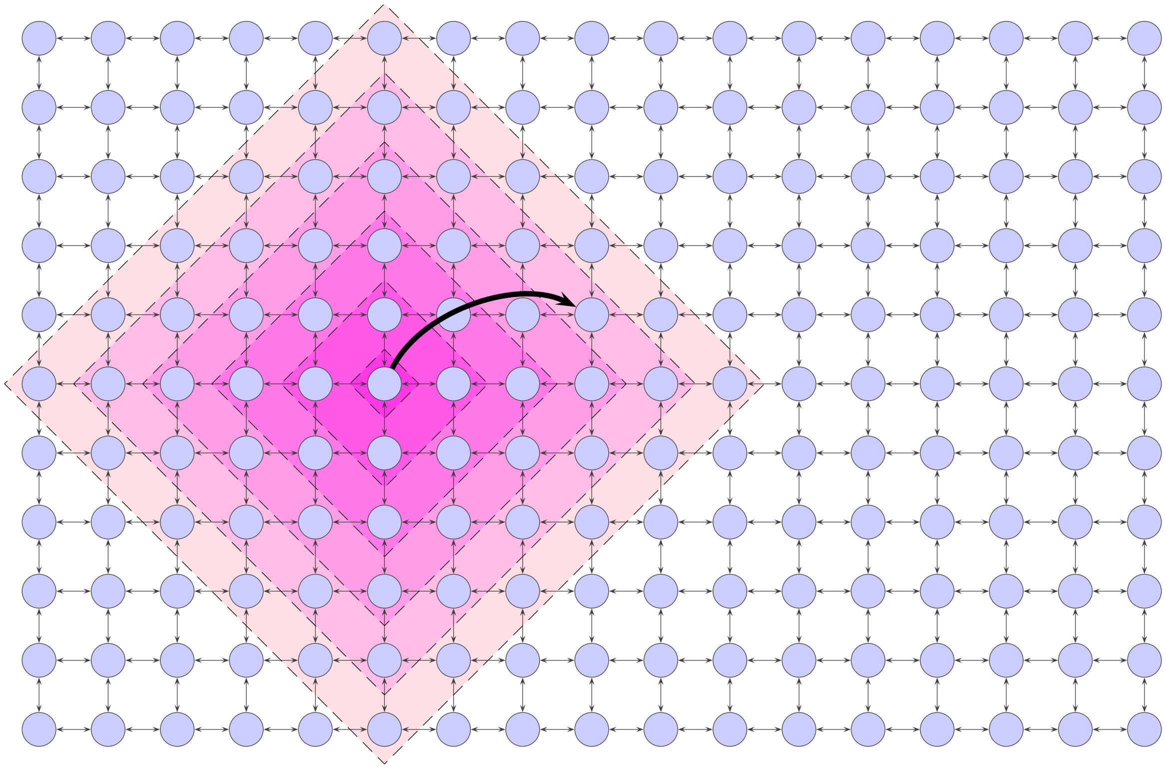




[Kleinberg 2000]:
for a social network to be navigable without global knowledge,
- need 'well-scattered' friends (to reach faraway targets)
- need 'well-localized' friends (to home in on nearby targets)

Kleinberg: Navigable Social Networks

[Kleinberg 2000]



- put n people on a k-dimensional grid
- connect each to its immediate neighbors
- add one long-range link per person; $\mathtt{Pr}[u \to v] \propto \frac{1}{d(u,v)^\alpha}$.

Navigability of Social Networks

- put n people on a k-dimensional grid
- connect each to its immediate neighbors
- add one long-range link per person; $\mathtt{Pr}[u \rightarrow v] \propto \frac{1}{d(u,v)^{\alpha}}$.

Theorem [Kleinberg 2000]: *(short = polylog(n))*

If $\alpha \neq k$
then no local-information algorithm can find short paths.

If $\alpha = k$
then people can find short—$O(\log^2 n)$—paths using the greedy algorithm.

Geography's Role in LiveJournal

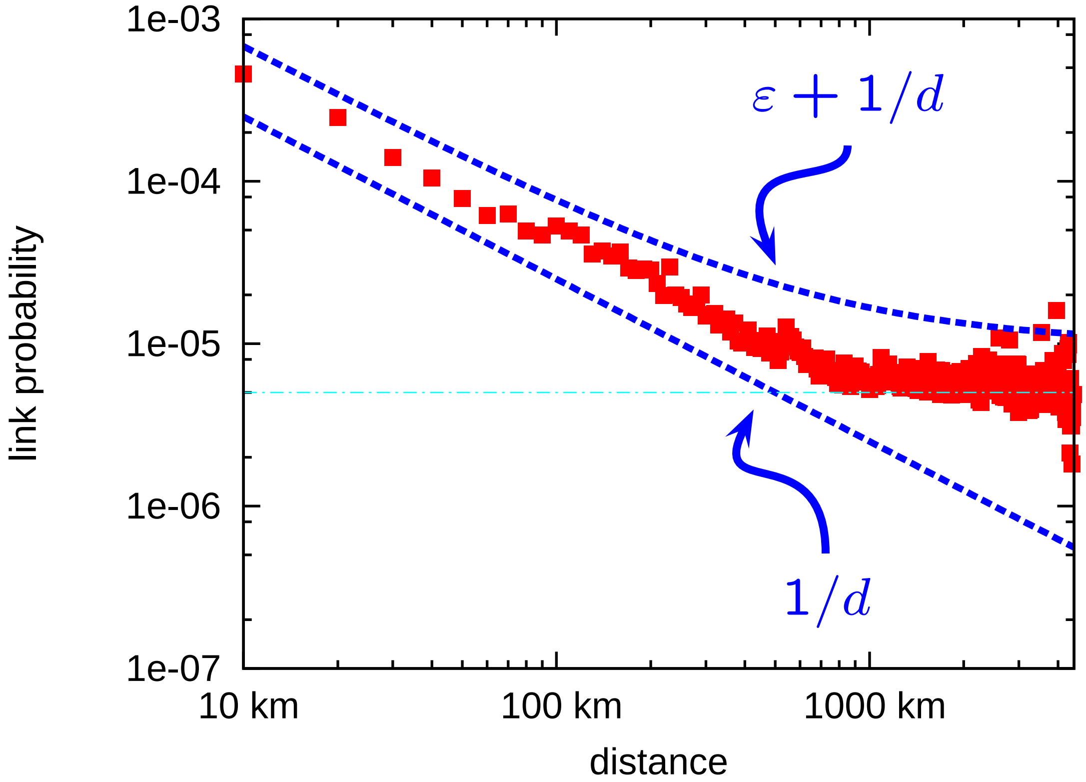
- By simulating the Milgram experiment, we find that LJ is navigable via geographically greedy routing.

- By Kleinberg's theorem, navigable 2-D geographic mesh $\Rightarrow \texttt{Pr}[u \rightarrow v] \propto d(u,v)^{-2}$.

Original goal of this research:

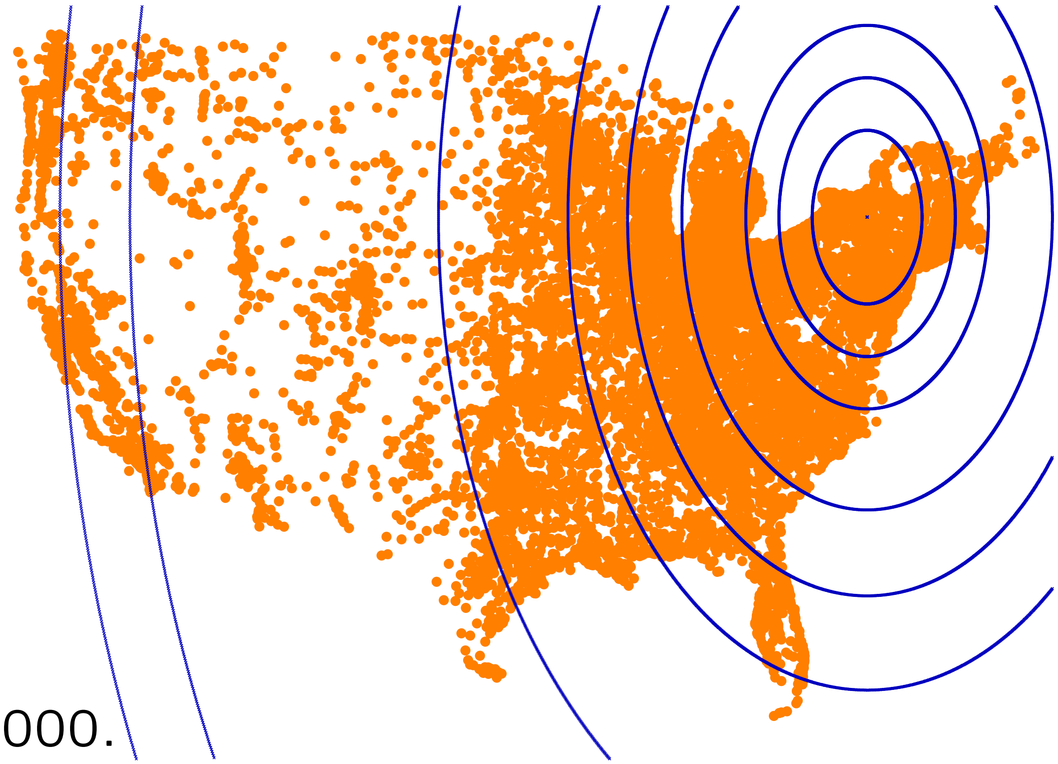
verify that $\texttt{Pr}[u \rightarrow v] \propto d(u,v)^{-2}$ in LiveJournal.

Distance versus link probability



- shows $\texttt{Pr}_{u,v}[u \text{ is friends with } v \mid d(u,v) = d]$
- Kleinberg's $1/d^2$ highly unsupported!
- Not really linear! Link probability levels out to $\sim 5 \times 10^{-6}$.

The LiveJournal Odyssey



Dot shown for every inhabited location in LiveJournal network.

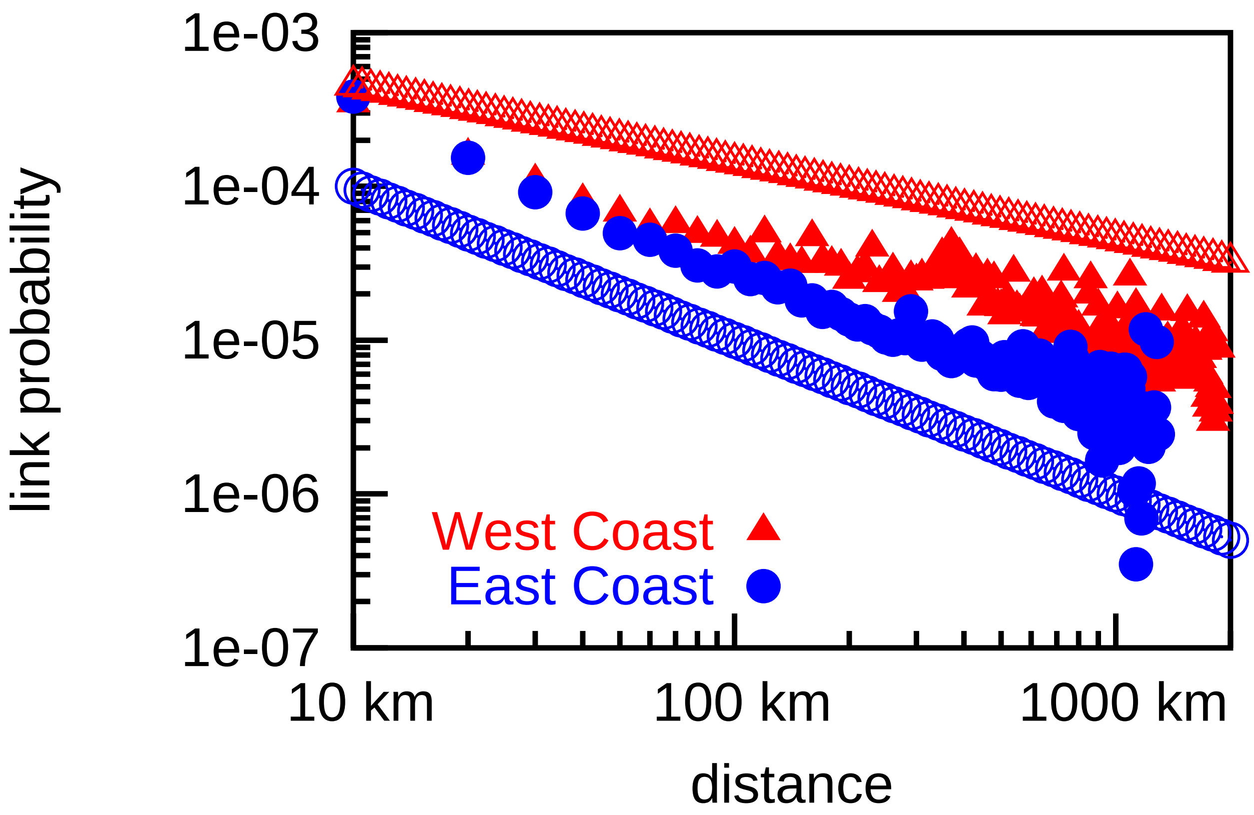
Circles are centered on Ithaca, NY.

Each successive circle's population increases by 50,000.

Uniform population ⇒ radii would decrease quadratically.
(actually mostly increase!)

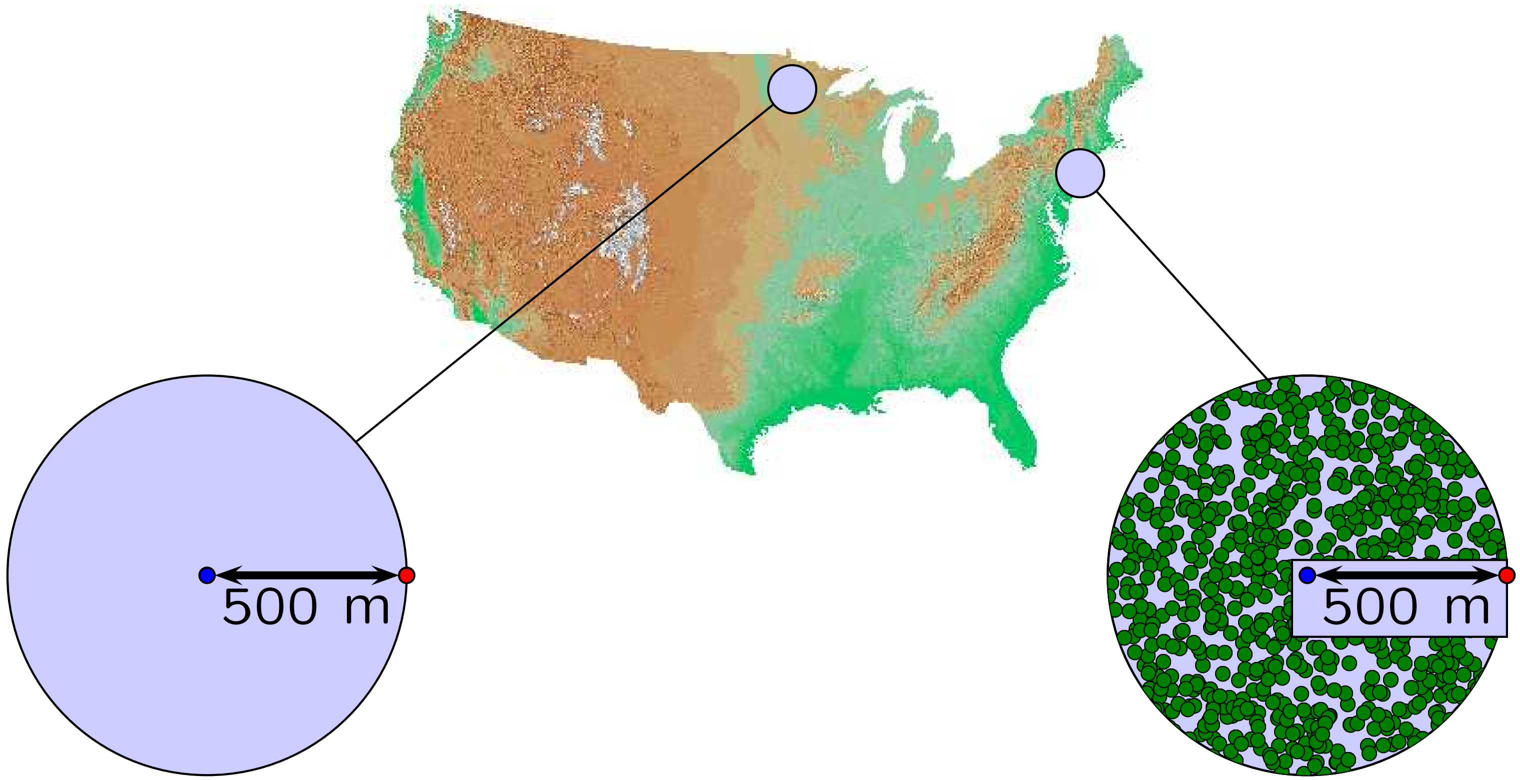
People don't live on a uniform grid!

Coastal Distances and Friendships



- Link probability versus distance.
- Restricted to the two coasts (CA to WA; VA to ME).
- Lines: $P(d) \propto d^{-1.00}$ and $P(d) \propto d^{-0.50}$.

Why does distance fail?



Population density varies widely across the US!

● and ●: best friends in Minnesota, strangers in Manhattan.

Rank-Based Friendship

How do we handle non-uniformly distributed populations?

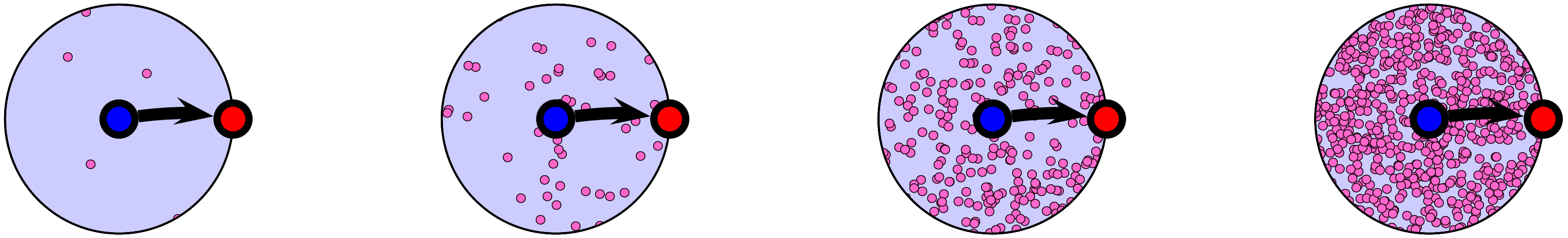
Instead of distance, use rank as fundamental quantity.

$$\text{rank}_A(B) := |\{C : d(A, C) < d(A, B)\}|$$

How many people live closer to A than B does?

Rank-Based Friendship : $\Pr[A \text{ is a friend of } B] \propto 1/\text{rank}_A(B)$.

Probability of friendship $\propto$ 1/(number of closer candidates)



Relating Rank and Distance

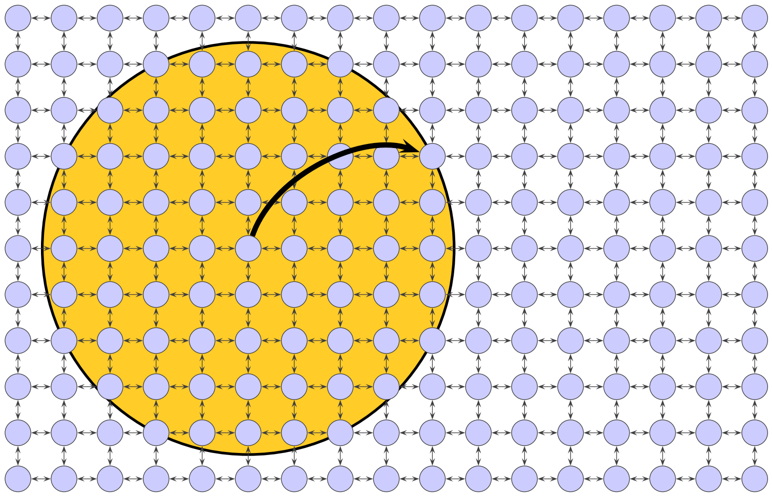
Rank-Based Friendship: $\texttt{Pr}[A \text{ is a friend of } B] \propto 1/\text{rank}_A(B)$.

Kleinberg (k-dim grid): $\texttt{Pr}[A \text{ is a friend of } B] \propto 1/d(A,B)^k$.

Uniform k-dimensional grid:

radius-d ball volume $\approx d^k$

$1/rank \approx 1/d^k$



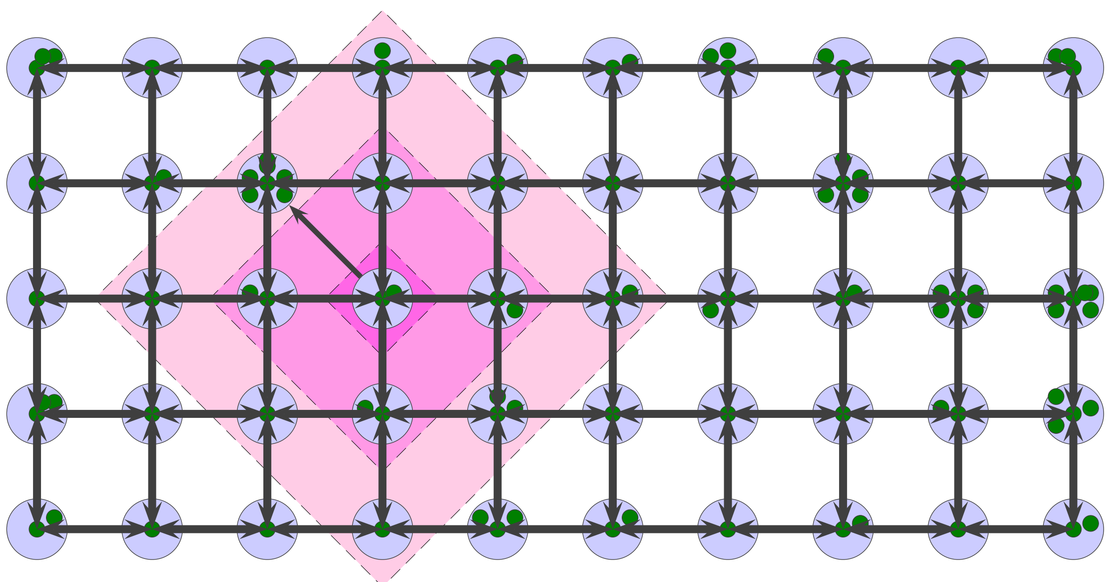
For a uniform grid, rank-based friendship
has (essentially) same link probabilities as Kleinberg.

Population Networks

A rank-based population network consists of:

- a k-dimensional grid L of locations.
- a population P of people, living at points in L $(n := |P|)$.
- a set $E \subseteq P \times P$ of friendships:
  - one edge from each person in each 'direction'
  - one edge from each person, chosen by rank-based friendship

e.g.,
locations rounded
to the nearest
integral point in
longitude/latitude.



Short Paths and Rank-Based Friendships

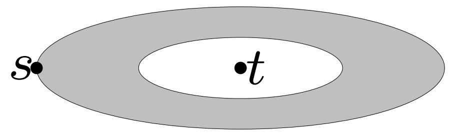
[Kumar DLN Tomkins, ESA'06]

Theorem: For any n-person rank-based population network in a k-dimensional grid, $k = \Theta(1)$, for any source $s \in P$ and for a randomly chosen target $t \in P$, the expected length (over t) of $\mathit{Greedy}(s, \mathsf{loc}(t))$ is $O(\log^3 n)$.

Is this just like all the other proofs?

Typical proof of navigability:

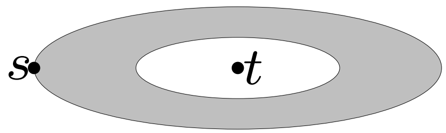
- Claim: $\texttt{Pr}\left[s \text{ friends with } u \text{ within } \frac{d(s,t)}{2} \text{ of } t\right] = \Omega\left(\frac{1}{polylog}\right)$.
- After $\log n$ halvings, done!

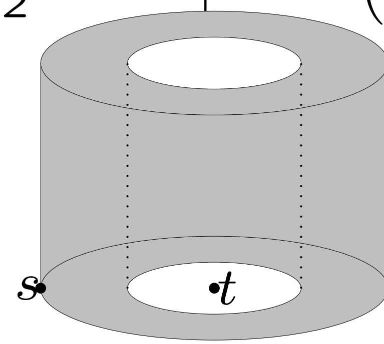


Is this just like all the other proofs?

Typical proof of navigability:

- Claim: $\texttt{Pr}\left[s \text{ friends with } u \text{ within } \frac{d(s,t)}{2} \text{ of } t\right] = \Omega\left(\frac{1}{polylog}\right)$.
- After $\log n$ halvings, done!





- Claim is false if $\{u : d(u,t) < \frac{d(s,t)}{2}\} \ll \{u : d(u,t) < d(s,t)\}$!

Our proof:

- Claim′: $\texttt{Pr}\left[s \text{ friends with } u \text{ within } \frac{d(s,t)}{2} \text{ of } t\right] = \Omega\left(\frac{1}{polylog}\right)$
  for a randomly chosen target t.
- After $\log n$ halvings, done!

The Real Theorem

Theorem: For any n-person rank-based population network in a k-dimensional grid, $k = \Theta(1)$, for any source $s \in P$ and for a randomly chosen target $t \in P$, the expected length (over t) of $\mathit{Greedy}(s, \mathsf{loc}(t))$ is $O(\log^3 n)$.

- Intuition: difficulty of halving distance to isolated target t is canceled by low probability of choosing t.
- Real theorem: not just for grids. (use doubling dimension of metric space instead of k).

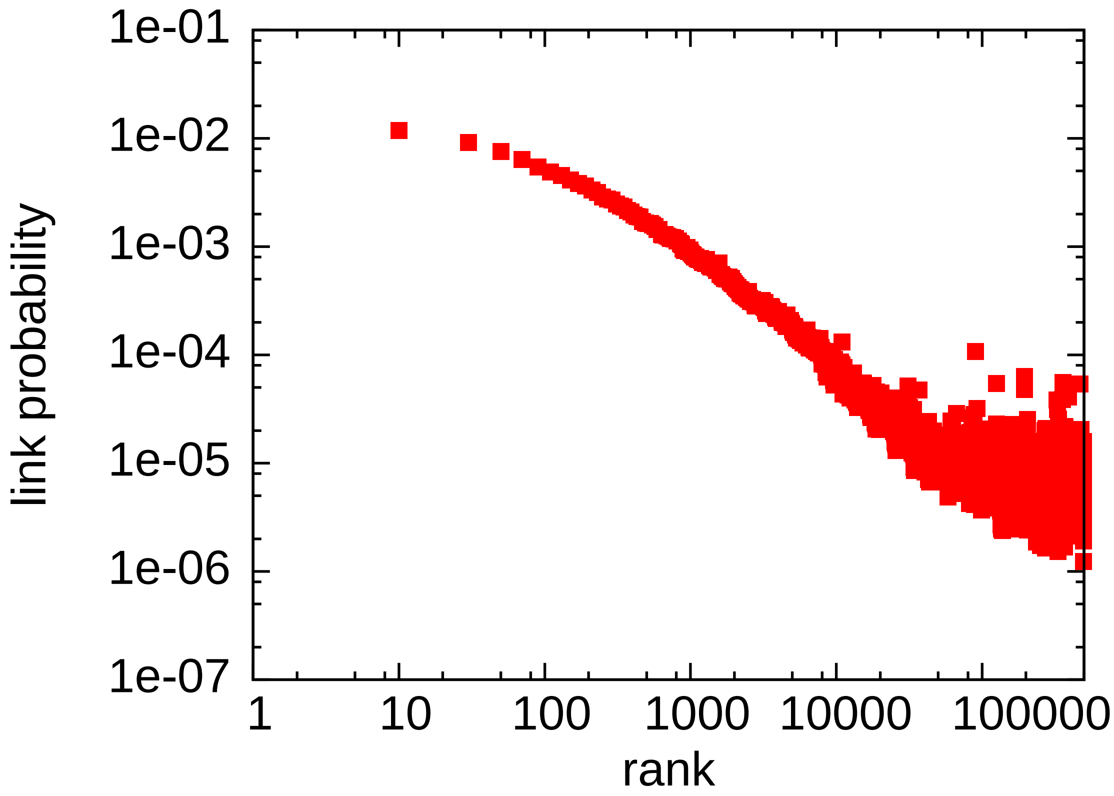
Short Paths and Rank-Based Friendships

Theorem: For any n-person population network in a k-dim grid, for any source $s \in P$ and a randomly chosen target $t \in P$, the expected length (over t) of $\mathit{Greedy}(s,t)$ is $O(\log^3 n)$.

Theorem [Kleinberg 2000]: For any n-person uniform-density population network, any source s, and any target t, the length of $\mathit{Greedy}(s,t)$ is $O(\log^2 n)$ with high probability.

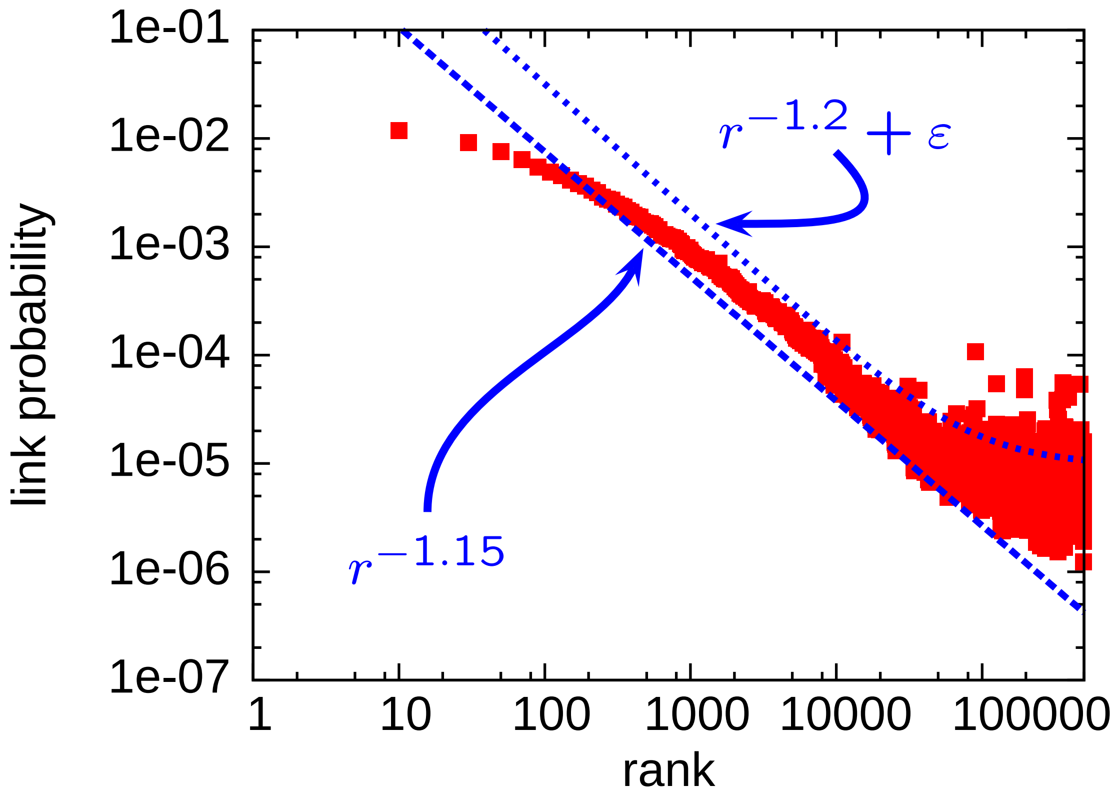
Lose:	expectation (not whp).
Lose:	another log factor.
Gain:	arbitrary population densities.
Gain?	*holds in real networks?*

Ranks and Friendships in LiveJournal



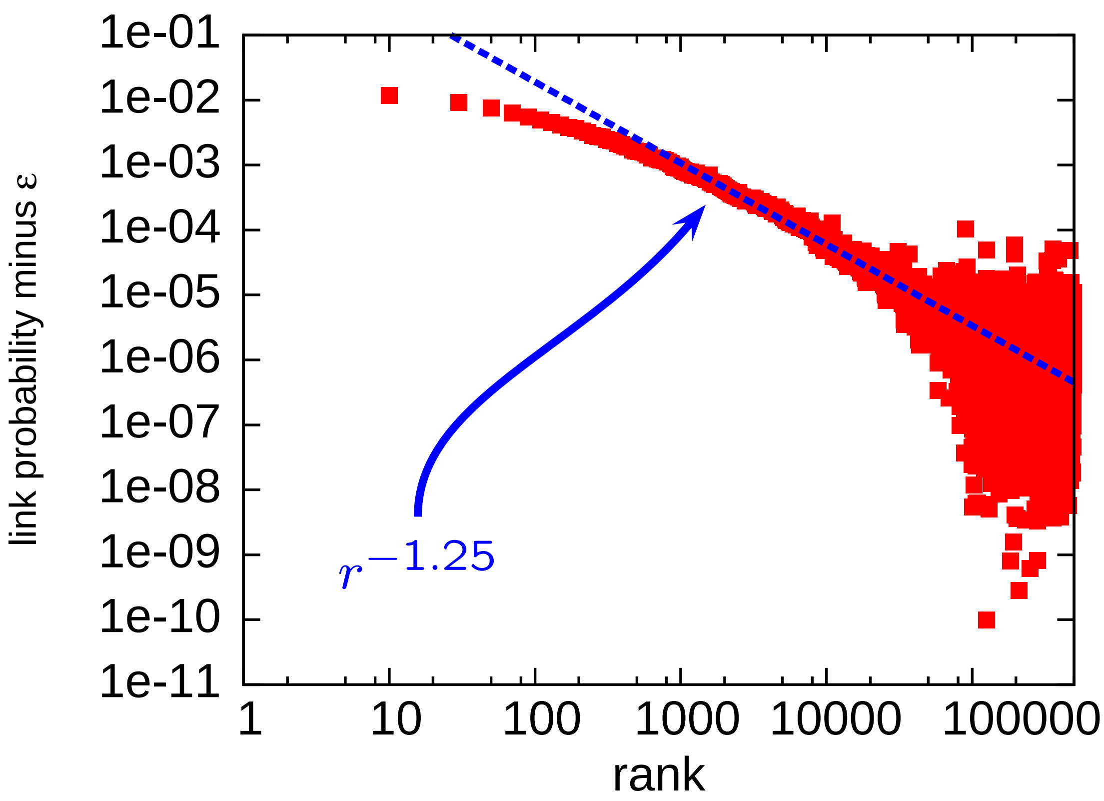
- shows $\mathtt{Pr}_u[u \text{ is friends with the } v : \mathrm{rank}_u(v) = r]$

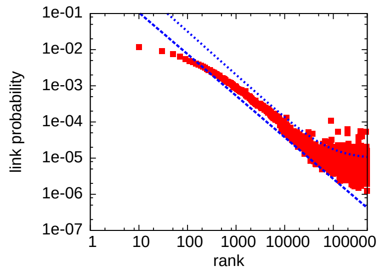
Ranks and Friendships in LiveJournal

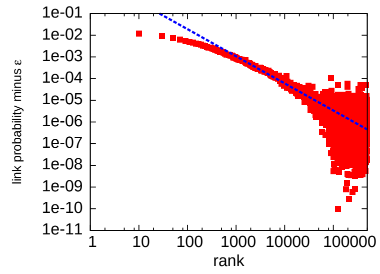


- shows $\mathtt{Pr}_u[u$ is friends with the $v : \mathrm{rank}_u(v) = r]$
- very close to $1/r$, as required for rank-based friendship!
- again, must correct for nongeographic friends.

Ranks and Friendships in LiveJournal

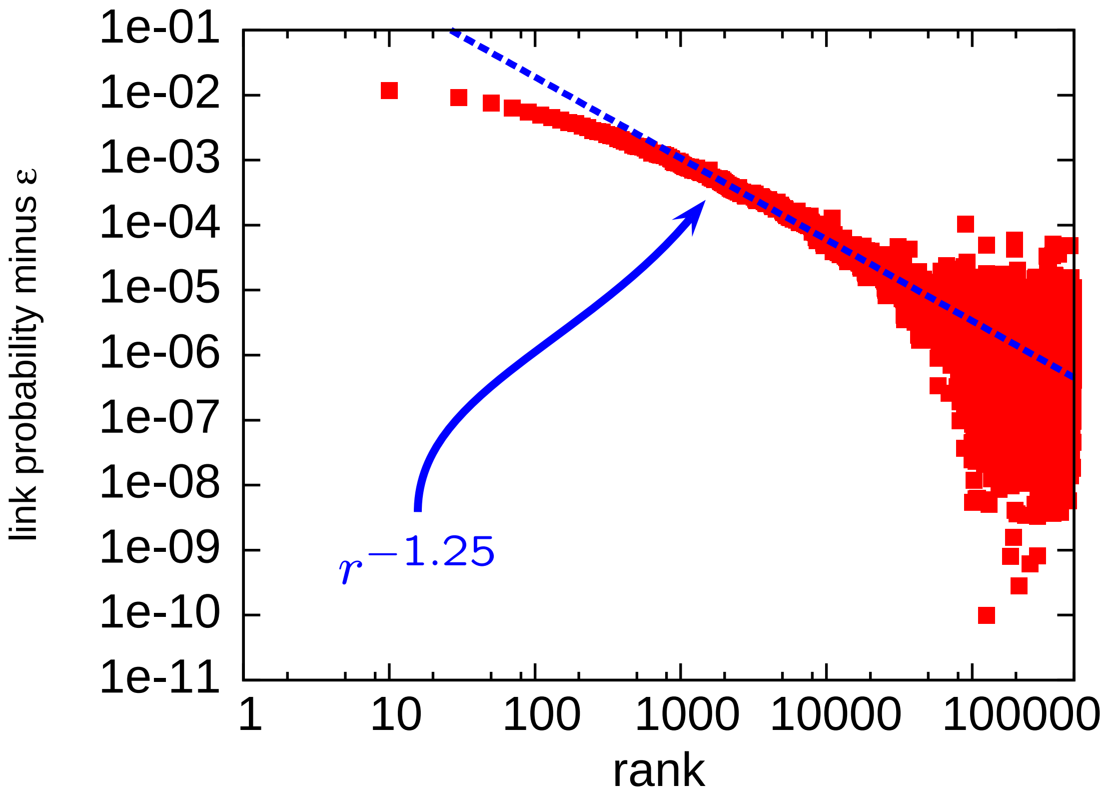


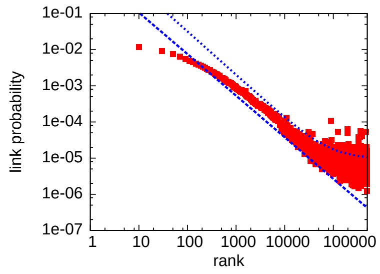


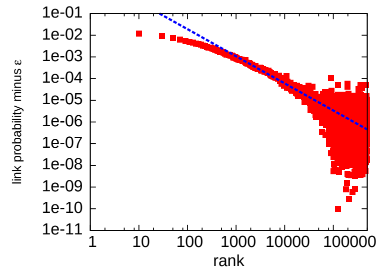


➡ shows probability of rank-r friendship, less $\varepsilon = 5.0 \times 10^{-6}$.

Ranks and Friendships in LiveJournal

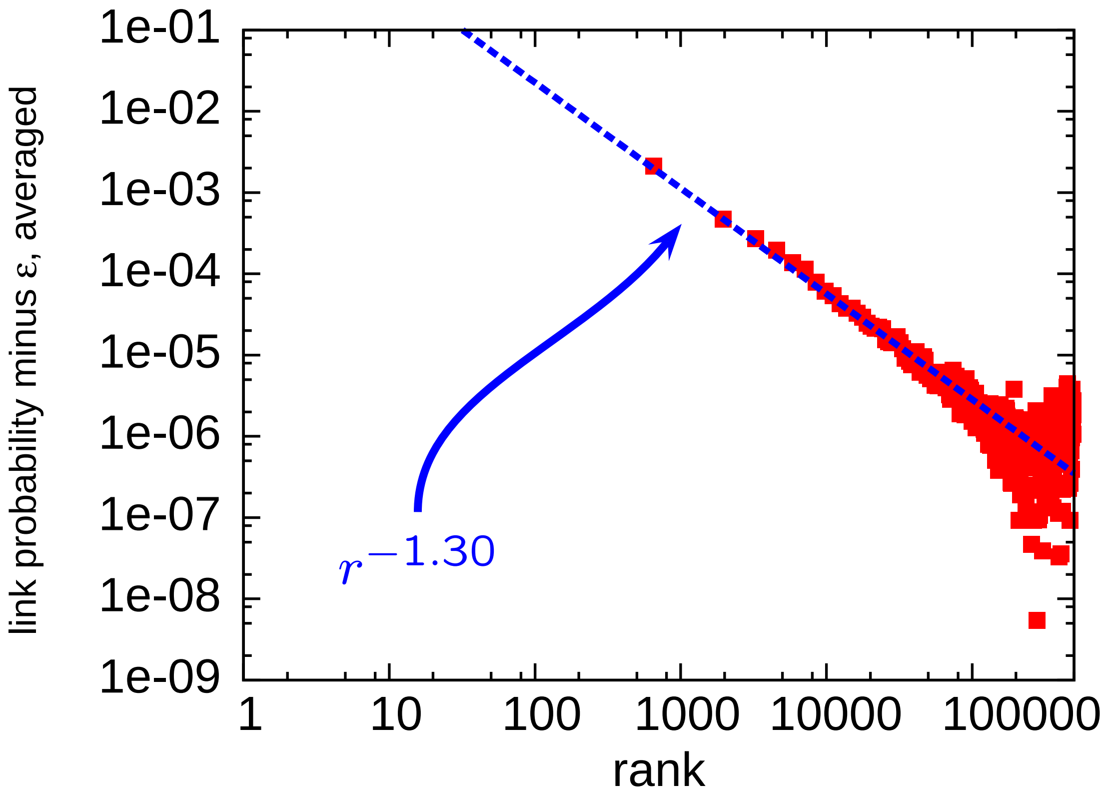


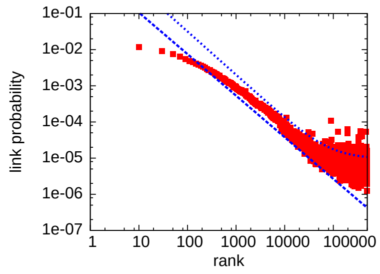


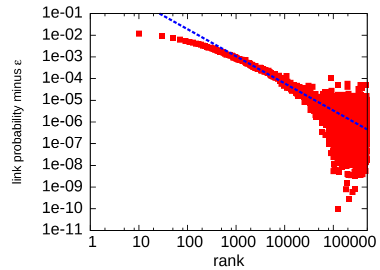


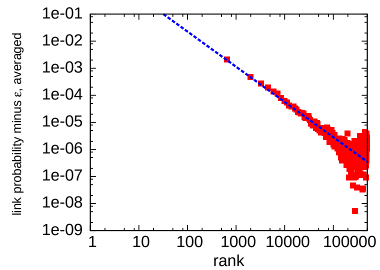
- shows probability of rank-r friendship, less $\varepsilon = 5.0 \times 10^{-6}$.
- LJ "location resolution" is city-only.
  average u's ranks $\{r, \ldots, r + 1300\}$ are in the same city
- $\Rightarrow$ we'll average probabilities over ranks $\{r, \ldots, r + 1300\}$

Ranks and Friendships in LiveJournal



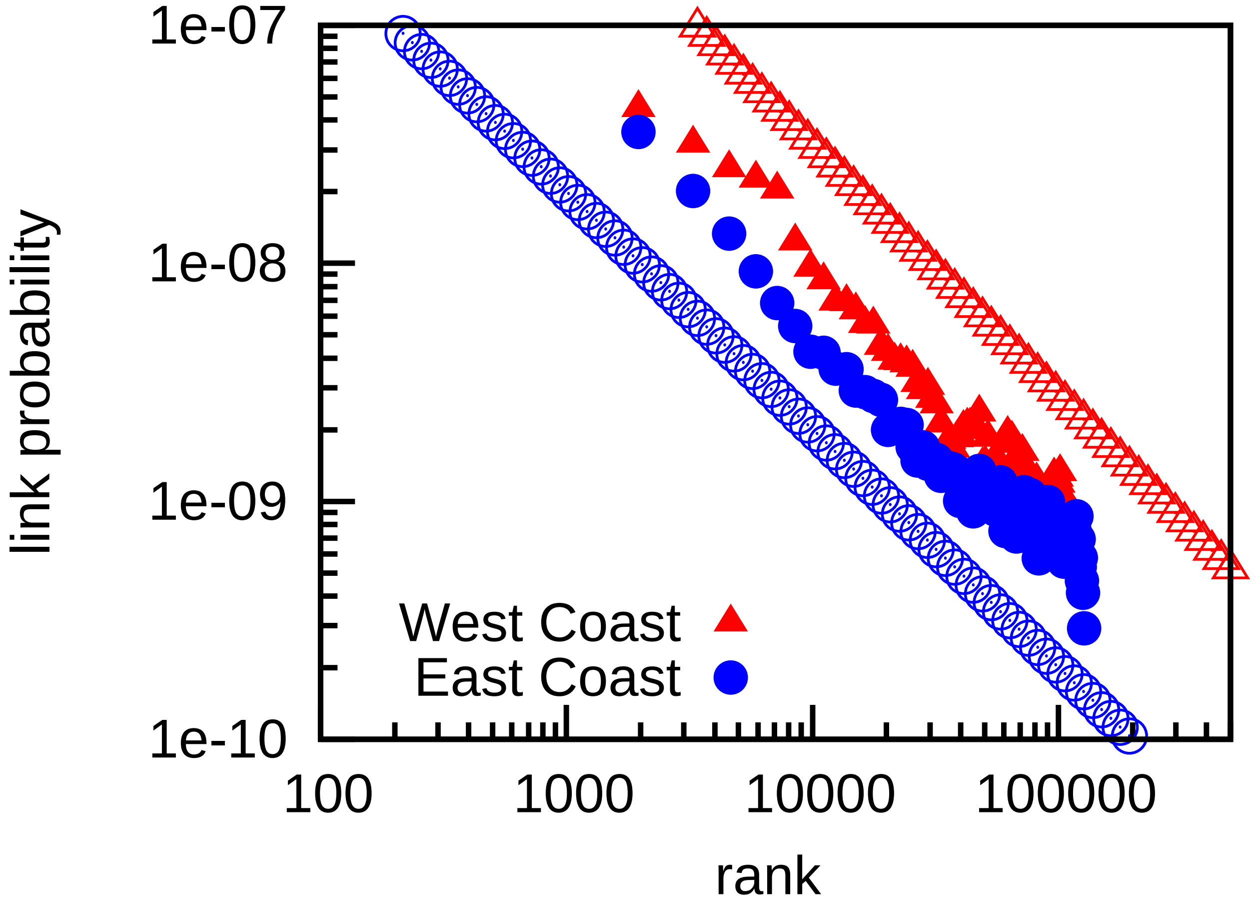






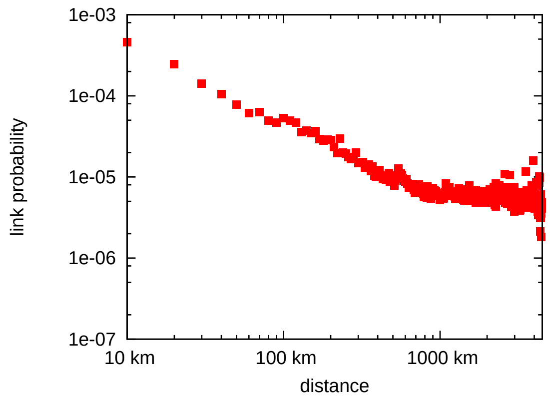
still very close to $1/r$
(though not absolutely perfect)

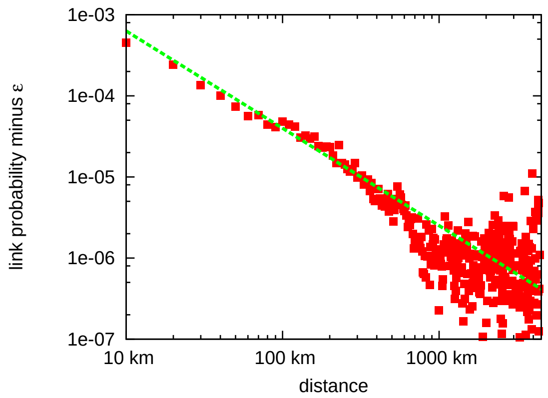
Coastal Ranks and Friendships



- Link probability versus rank.
- Restricted to West (CA to WA) and East (VA to ME).
- Lines: $P(r) \propto r^{-1.00}$ and $P(r) \propto r^{-1.05}$.

Geographic/Nongeographic Friendships





- good estimate of friendship probability:

$$\texttt{Pr}[u \to v] \approx \varepsilon + f(d(u,v)) \text{ for } \varepsilon \approx 5.0 \times 10^{-6}.$$

'ε friends' (nongeographic) 'f(d) friends' (geographic).

- LJ: $\texttt{E}$[number of u's "ε" friends] $= \varepsilon \cdot 500{,}000 \approx 2.5$.
- LJ: average degree ≈ 8.

~5.5/8 ≈ 66% of LJ friendships are "geographic," 33% are not.

Routing Choices

In real life, many ways to choose a next step when searching!

Geography: greedily route based on distance to t.

Occupation: $\approx$ greedily route based on distance in the (implicit) hierarchy of occupations.

Age, hobbies, alma mater, ...

Popularity: choose people with high outdegree. [Adamic Lukose Puniyani Huberman 2001] [Kim Yoon Han Jeong 2002] ...

What does 'closest' mean in real life?
How do you weight various 'proximities'?

minimum over all proximities? [Dodds Watts Newman 2002]
a more complicated combination?

Open Directions

A half-sociological, half-computational question ...

Why should rank-based friendship hold, even approximately?

Are there natural processes that generate it?

E.g., a generative process based on "geographic interests"?

Thank you!

David Liben-Nowell

dlibenno@carleton.edu

http://cs.carleton.edu/faculty/dlibenno