Navigation and Evolution of Social Networks

David Liben-Nowell

Carleton College

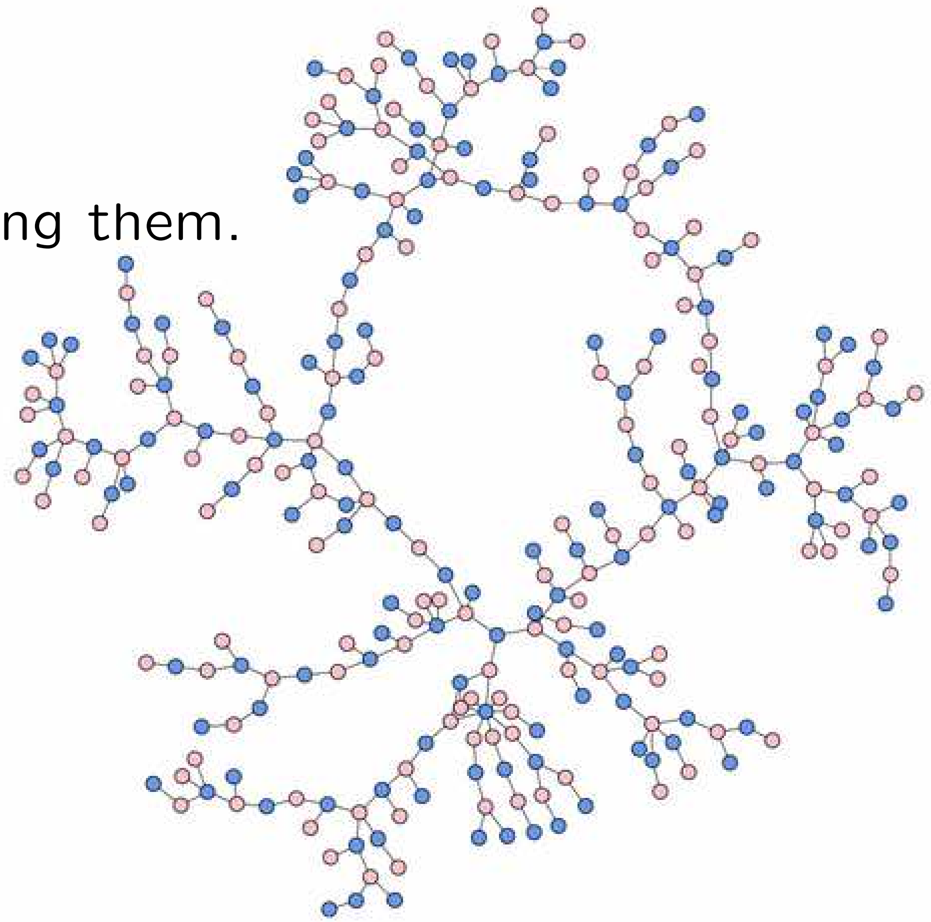
`dlibenno@carleton.edu`

WAW'06 Winter School | 28 November 2006

Social Networks

A social network:

- a set of people.
- a social relationship linking them.

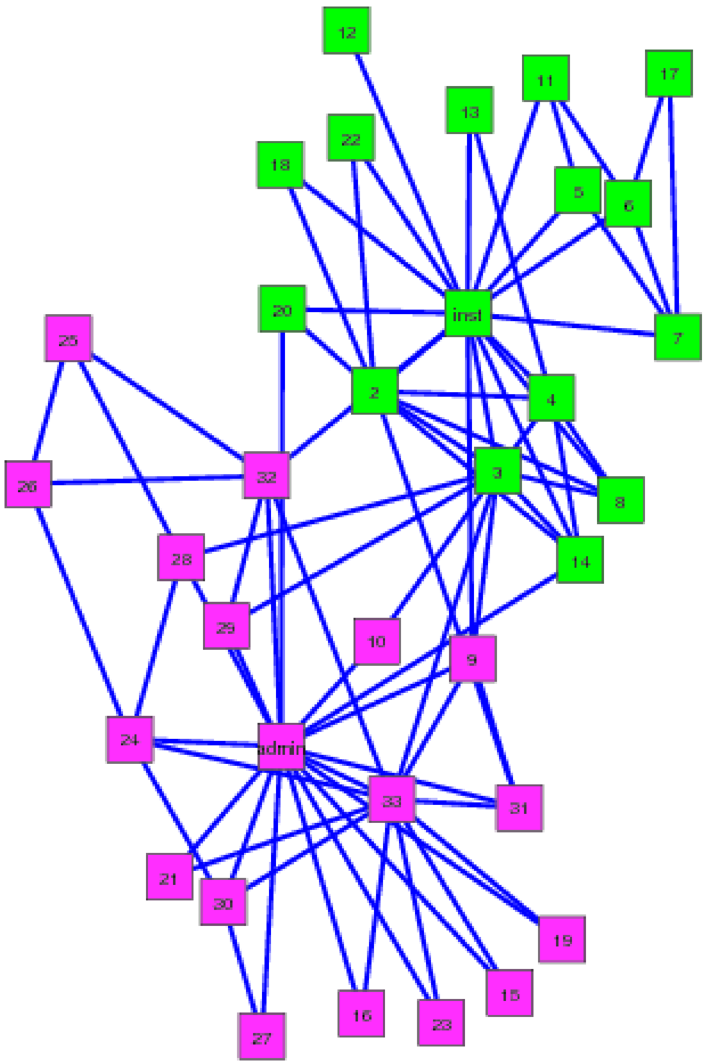


[Bearman Moody Stovel 2002; image by Mark Newman]

Social Network Analysis: Old School

- social networks have been around for 100K+ years!
- before the web, hard to acquire (surveys, interviews, ...).
- but many interesting, relevant, generalizable observations!

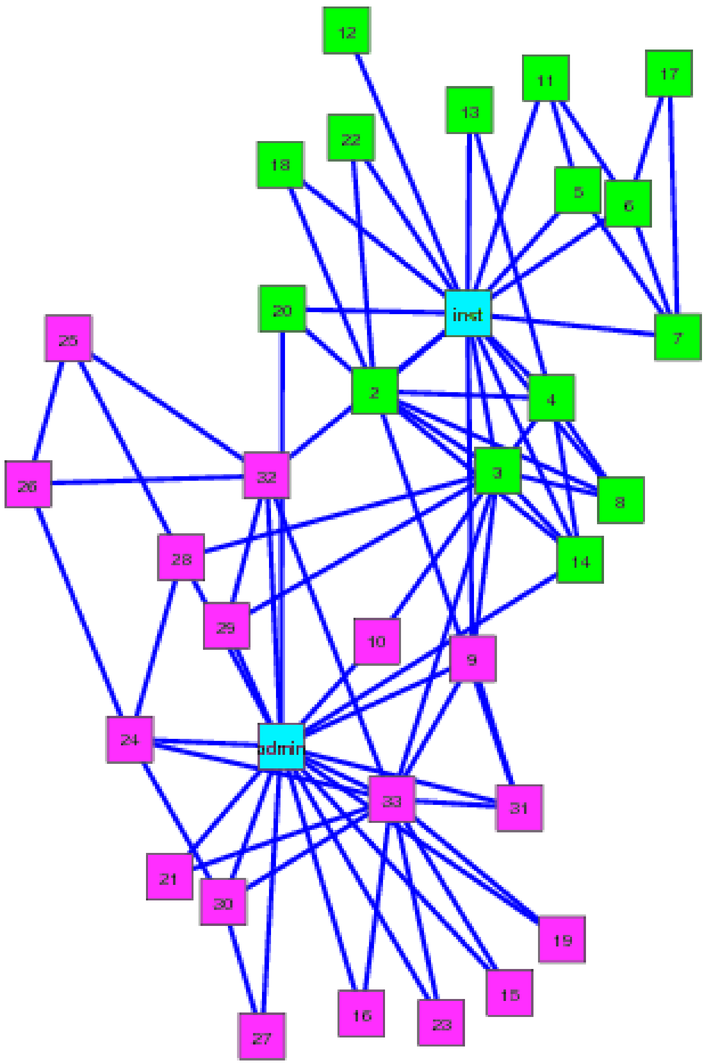
Zachary's Karate Club



[Zachary 1977]

Recorded interactions in a karate club for 2 years.

Zachary's Karate Club



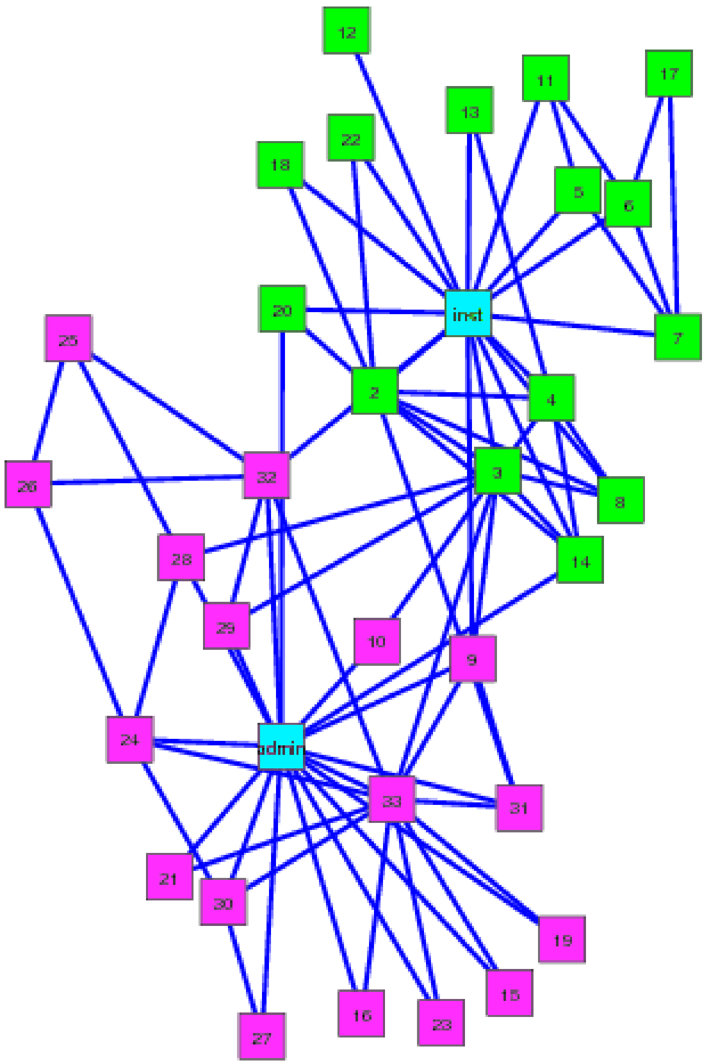
[Zachary 1977]

Recorded interactions in a karate club for 2 years.

During observation, adminstrator/instructor conflict developed

⇒ *broke into two clubs.*

Zachary's Karate Club



[Zachary 1977]

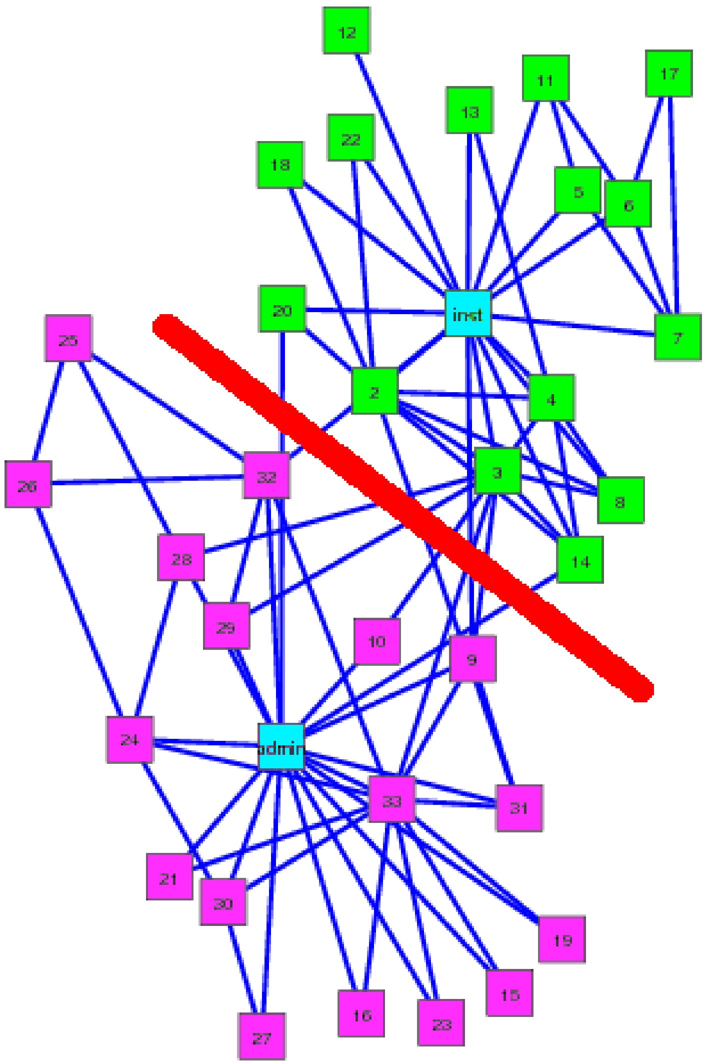
Recorded interactions in a karate club for 2 years.

During observation, adminstrator/instructor conflict developed

⇒ *broke into two clubs.*

Who joins which club?

Zachary's Karate Club



[Zachary 1977]
Recorded interactions in a karate club for 2 years.

During observation, adminstrator/instructor conflict developed
⇒ *broke into two clubs.*

Who joins which club?

Split along administrator/instructor minimum cut (!)

Part I:

Search in Social Networks

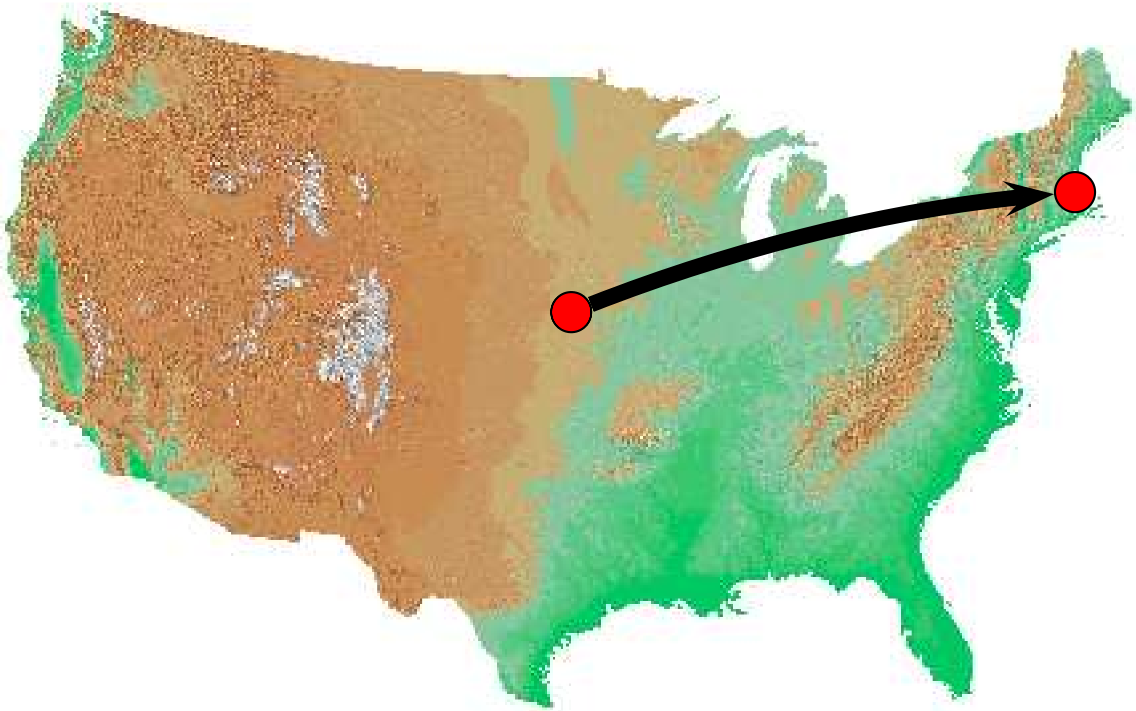
Part I:

Search in Social Networks*

* with a somewhat biased DLN-centric perspective.

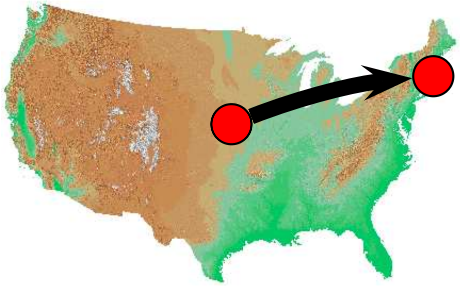
Milgram: Six Degrees of Separation

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?

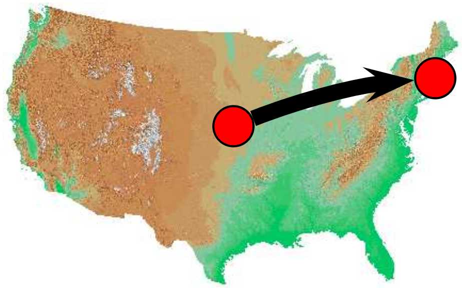


"the small-world problem"

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

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

Milgram: The Explanation?



"the small-world problem"

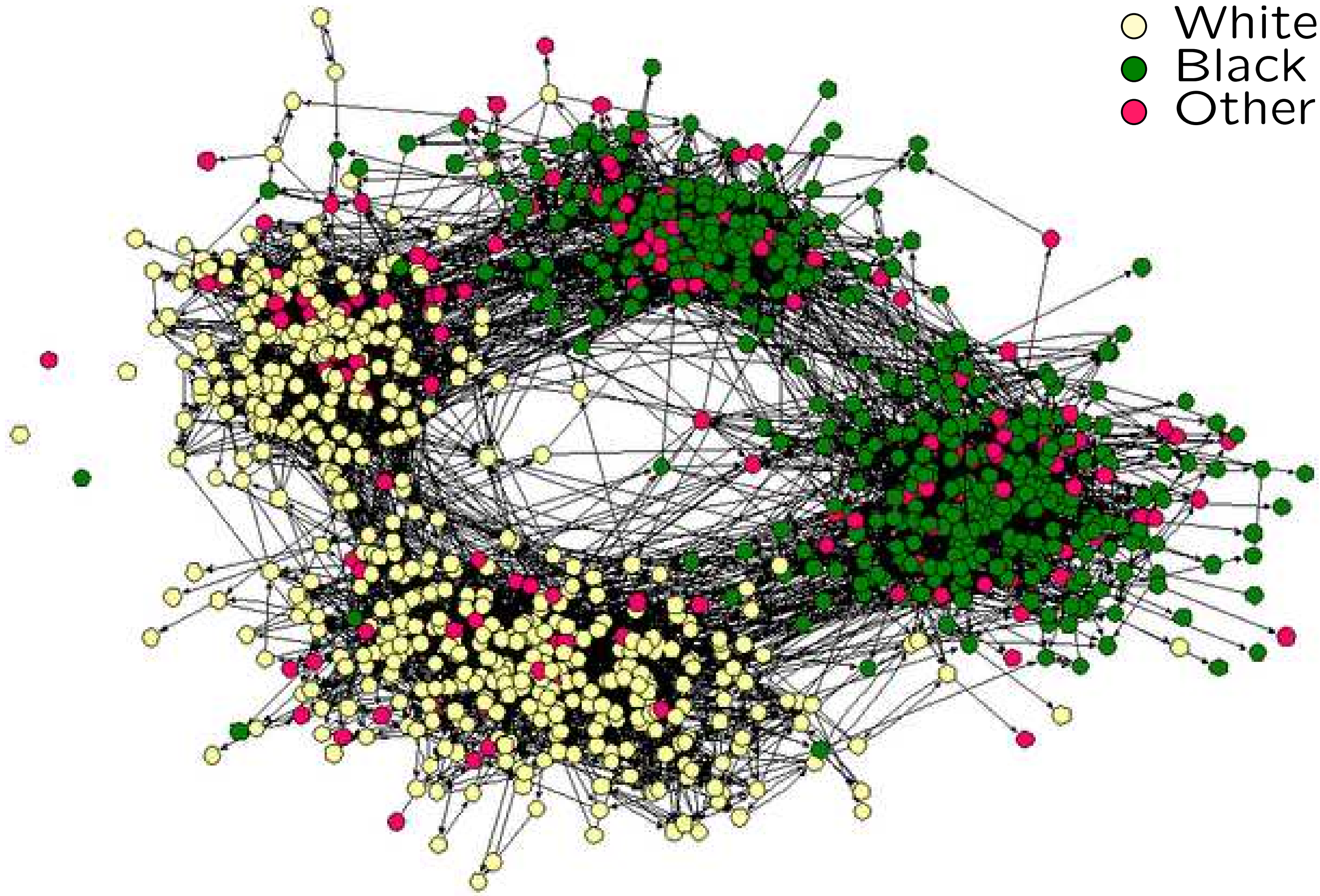
➡ Why is a random Omahaian close to a Sharon stockbroker?

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

Bogus! In fact, many bogosities:

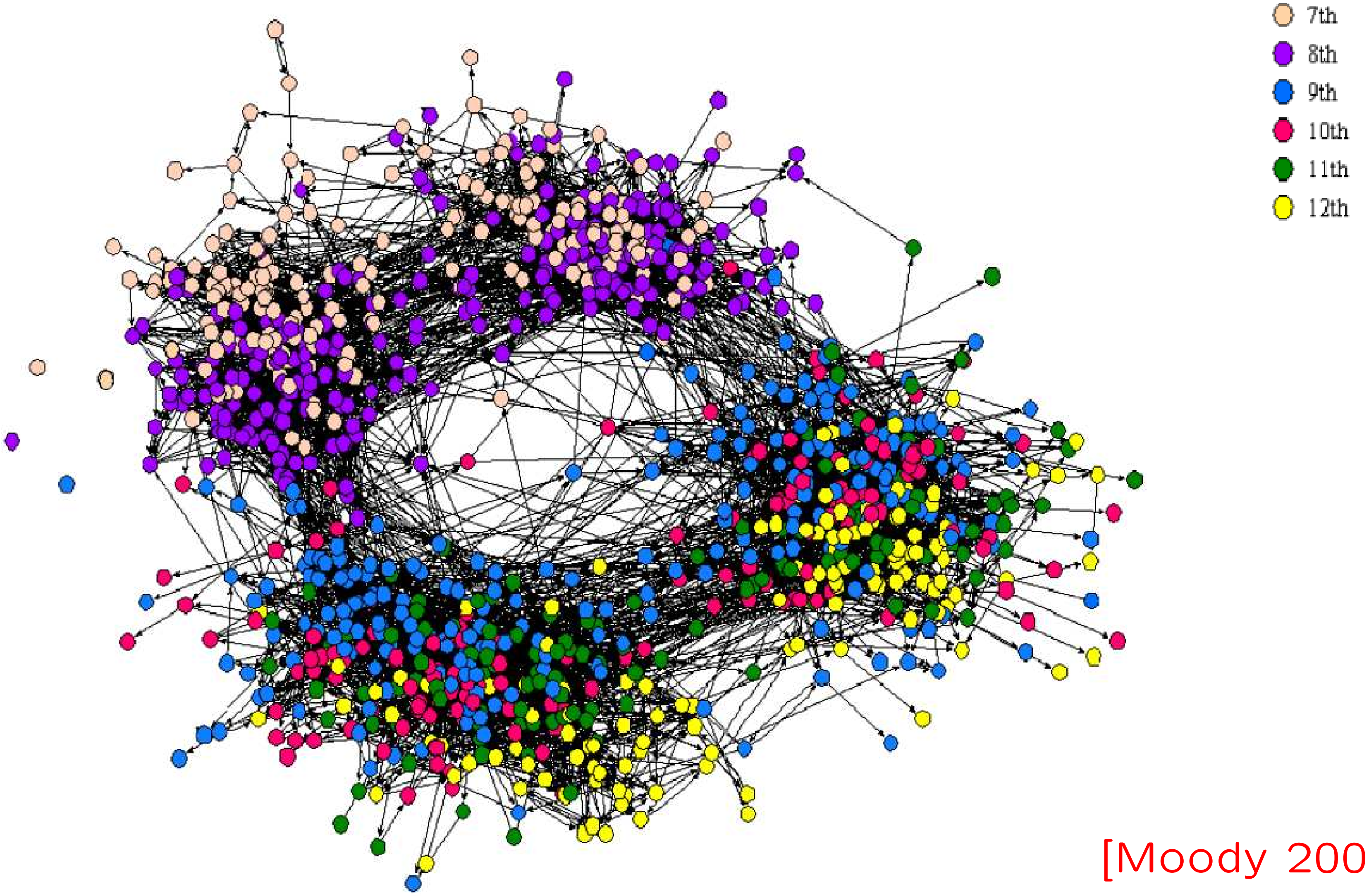
- degree distribution
- clustering coefficients
- ...

High School Friendships



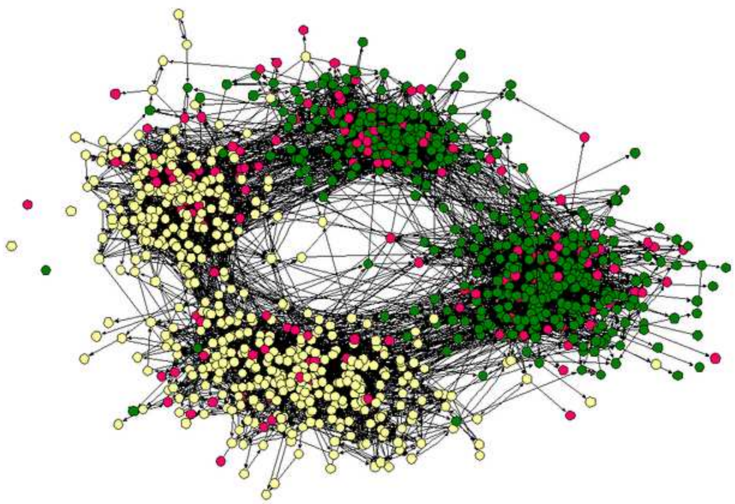
Self-reported high school friendships. [Moody 2001]

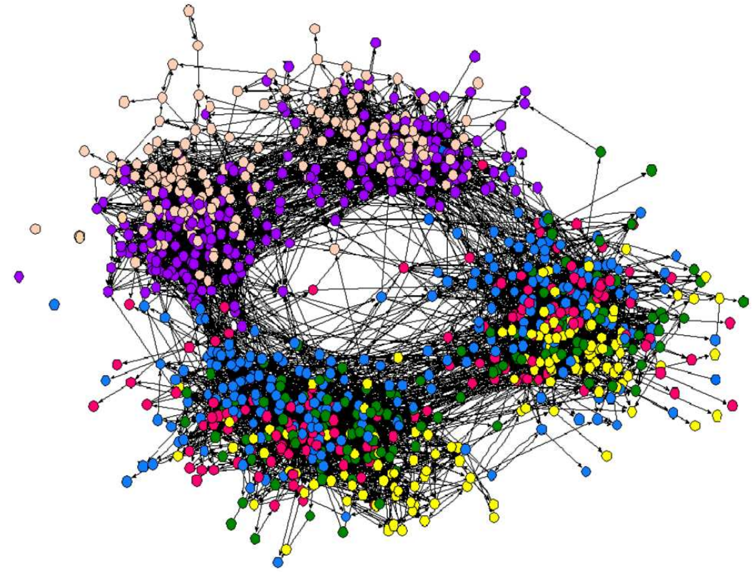
High School Friendships



[Moody 2001]

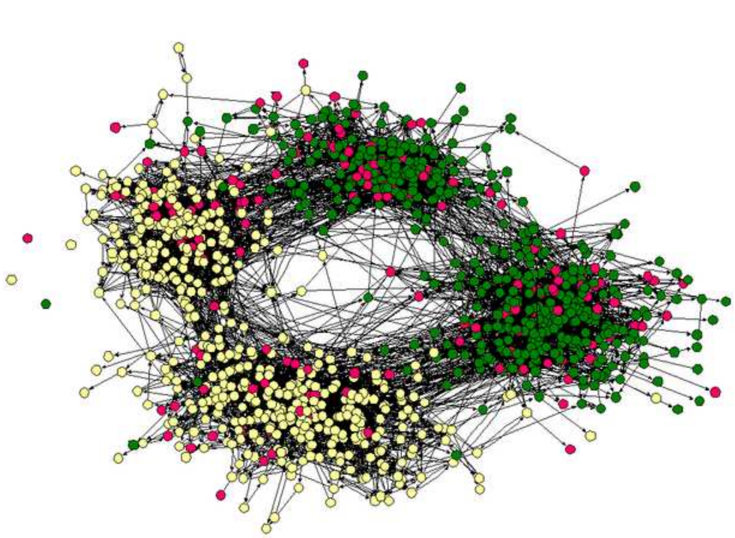
Homophily

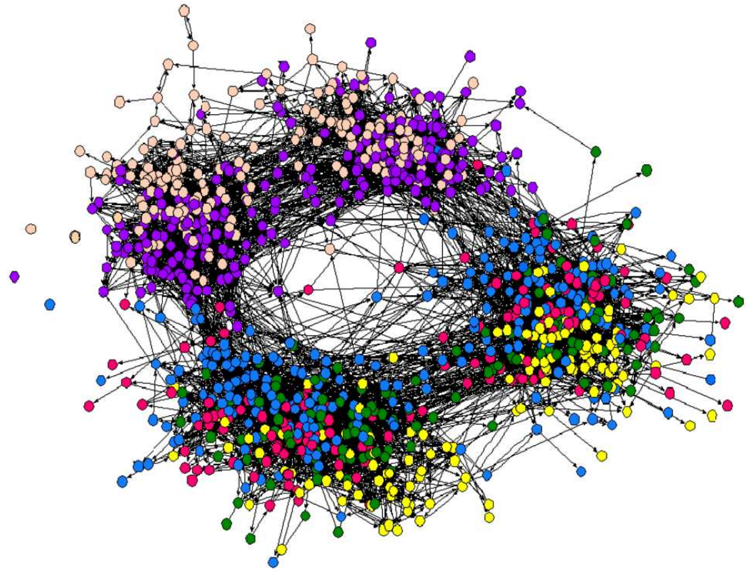




homophily: a person x's friends tend to be 'similar' to x.

Homophily





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

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

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

Watts/Strogatz: Rewired Ring Lattice

A model with small diameter and large clustering coefficient?

Some well-studied models in graph theory:

e.g., [Bollabas Chung 1988]

e.g., later today (13:30–15:45)!

Watts/Strogatz: Rewired Ring Lattice

A model with small diameter and large clustering coefficient?

Some well-studied models in graph theory:

e.g., [Bollabas Chung 1988]

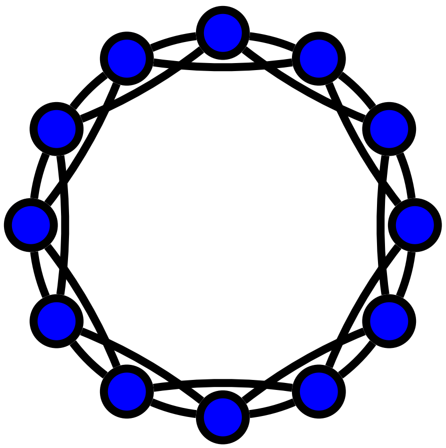
e.g., later today (13:30–15:45)!

[Watts Strogatz 1998]

Put people on circle; connect each to δ closest neighbors.

With probability p, rewire each connection randomly.

$\boxed{p = 0.0}$



Watts/Strogatz: Rewired Ring Lattice

A model with small diameter and large clustering coefficient?

Some well-studied models in graph theory:

e.g., [Bollabas Chung 1988]

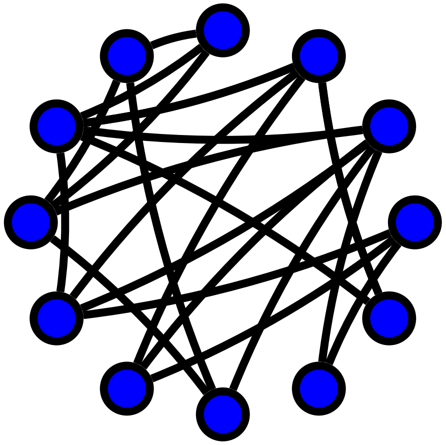
e.g., later today (13:30–15:45)!

[Watts Strogatz 1998]

Put people on circle; connect each to δ closest neighbors.

With probability p, rewire each connection randomly.

$\boxed{p = 1.0}$



Watts/Strogatz: Rewired Ring Lattice

A model with small diameter and large clustering coefficient?

Some well-studied models in graph theory:

e.g., [Bollabas Chung 1988]

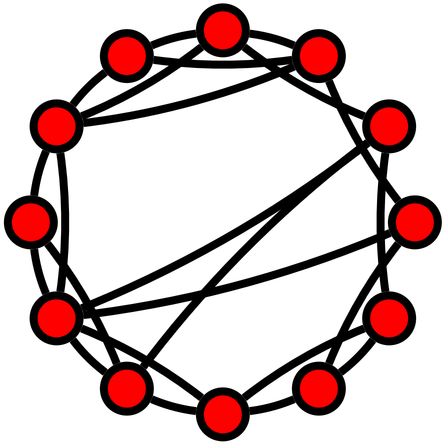
e.g., later today (13:30–15:45)!

[Watts Strogatz 1998]

Put people on circle; connect each to δ closest neighbors.

With probability p, rewire each connection randomly.

$\boxed{p = 0.1}$



Watts/Strogatz: Rewired Ring Lattice

A model with small diameter and large clustering coefficient?

Some well-studied models in graph theory:

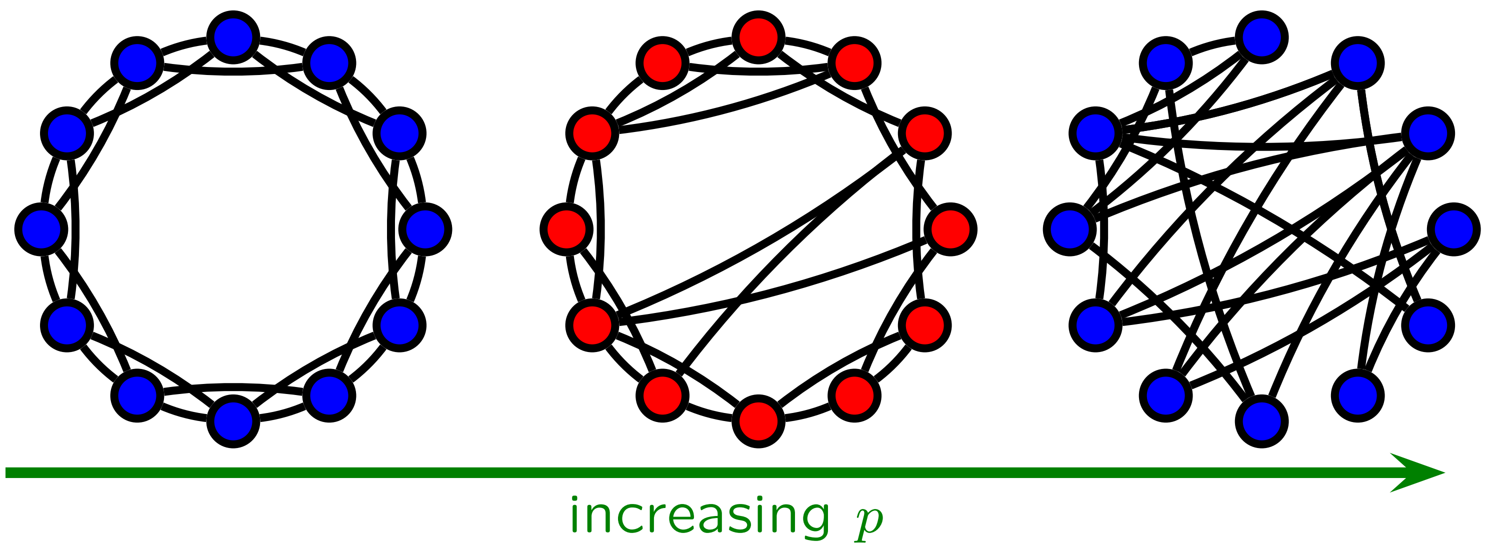
e.g., [Bollabas Chung 1988]

e.g., later today (13:30–15:45)!

[Watts Strogatz 1998]

Put people on circle; connect each to δ closest neighbors.

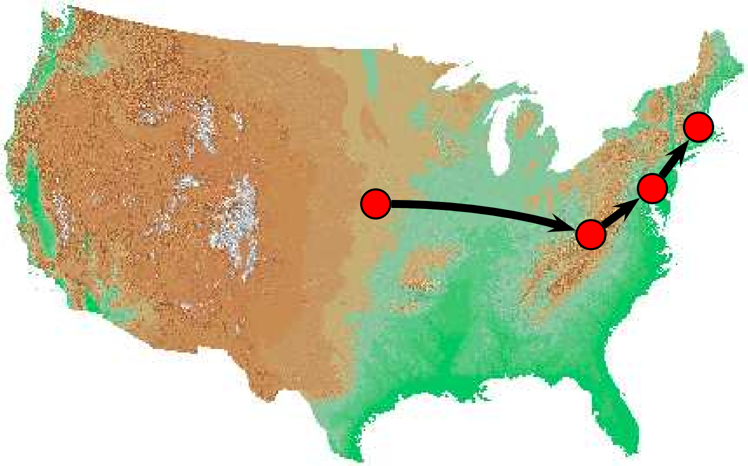
With probability p, rewire each connection randomly.



Navigability of Social Networks

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

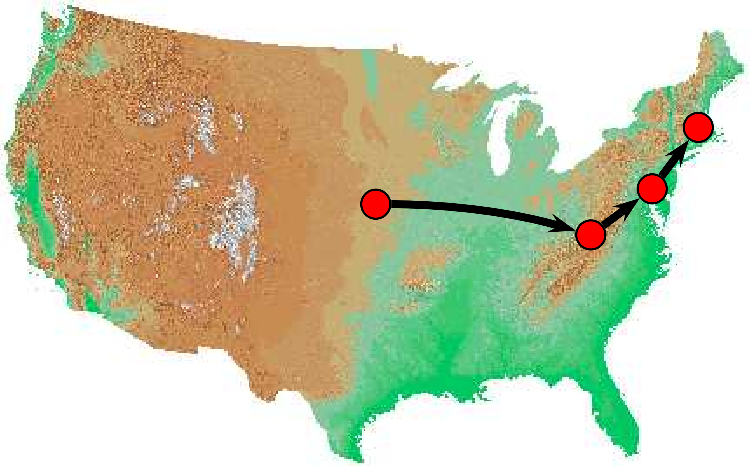
Milgram's result is algorithmic, not existential.



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.



Theorem [Kleinberg 2000]: No local-information algorithm can find short paths in Watts/Strogatz networks.

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"

Columbia Small-World Experiment

[Dodds Muhamad Watts 2003]

```
Date:  Mon, 25 Mar 2002 02:17:33 -0500
From:  Parviz Parvizi <parviz_parvizi@mckinsey.com>
To:  dln@mit.edu
Subject:  Small World Research Project

Dear David Liben-Nowell,
Here's a message from Parviz Parvizi
who chose you as the next person in this experiment:

sir, i assume that you *must* know this guy personally or know someone who
knows him.  (my random guess would be that since this duncan watts character
got his ph.d. at cornell, the person we are trying to reach--steven
strogatz--was one of his advisers/mentors & therefore somehow possibly knows
your pal kleinberg since watts' work relates to the 6 degrees phenomenon.  so,
i would say, forward this to kleinberg unless you happen to know our good man
strogatz yourself.)


We request your assistance with a Columbia University research project.
(An article about this project has just appeared in the New York Times
http://www.nytimes.com/2001/12/20/technology/circuits/20STUD.html)

Our request is simple and will only take a minute or two of your time.

We are a team of sociologists interested in what is known as
the "Small World Phenomenon".  This is the idea that
everyone in the world can be reached through
```

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.)

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.

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.

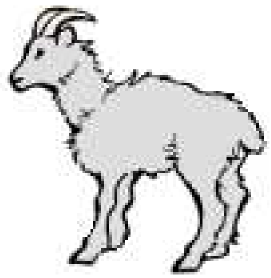
Occupation: $\approx$ greedily route based on distance in the (implicit) hierarchy of occupations.

Want $\texttt{Pr}[u, v \text{ friends}]$ 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

[DLN Novak Kumar Raghavan Tomkins 2005]



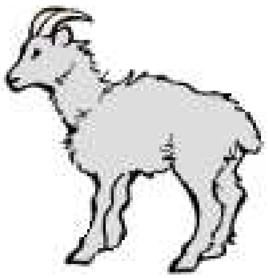
`www.livejournal.com`

"Baaaaah," says Frank.

- Online blogging community.
- Currently 11.6 million users; ~1.3 million in February 2004.

The LiveJournal Community

[DLN Novak Kumar Raghavan Tomkins 2005]



`www.livejournal.com`

"Baaaaah," says Frank.

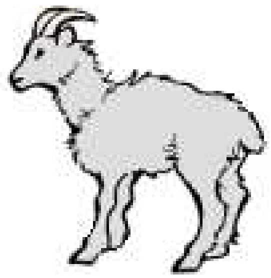
- Online blogging community.
- Currently 11.6 million users; ~1.3 million in February 2004.

LiveJournal users provide:

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

The LiveJournal Community

[DLN Novak Kumar Raghavan Tomkins 2005]



`www.livejournal.com`

"Baaaaah," says Frank.

- Online blogging community.
- Currently 11.6 million users; $\sim$1.3 million in February 2004.

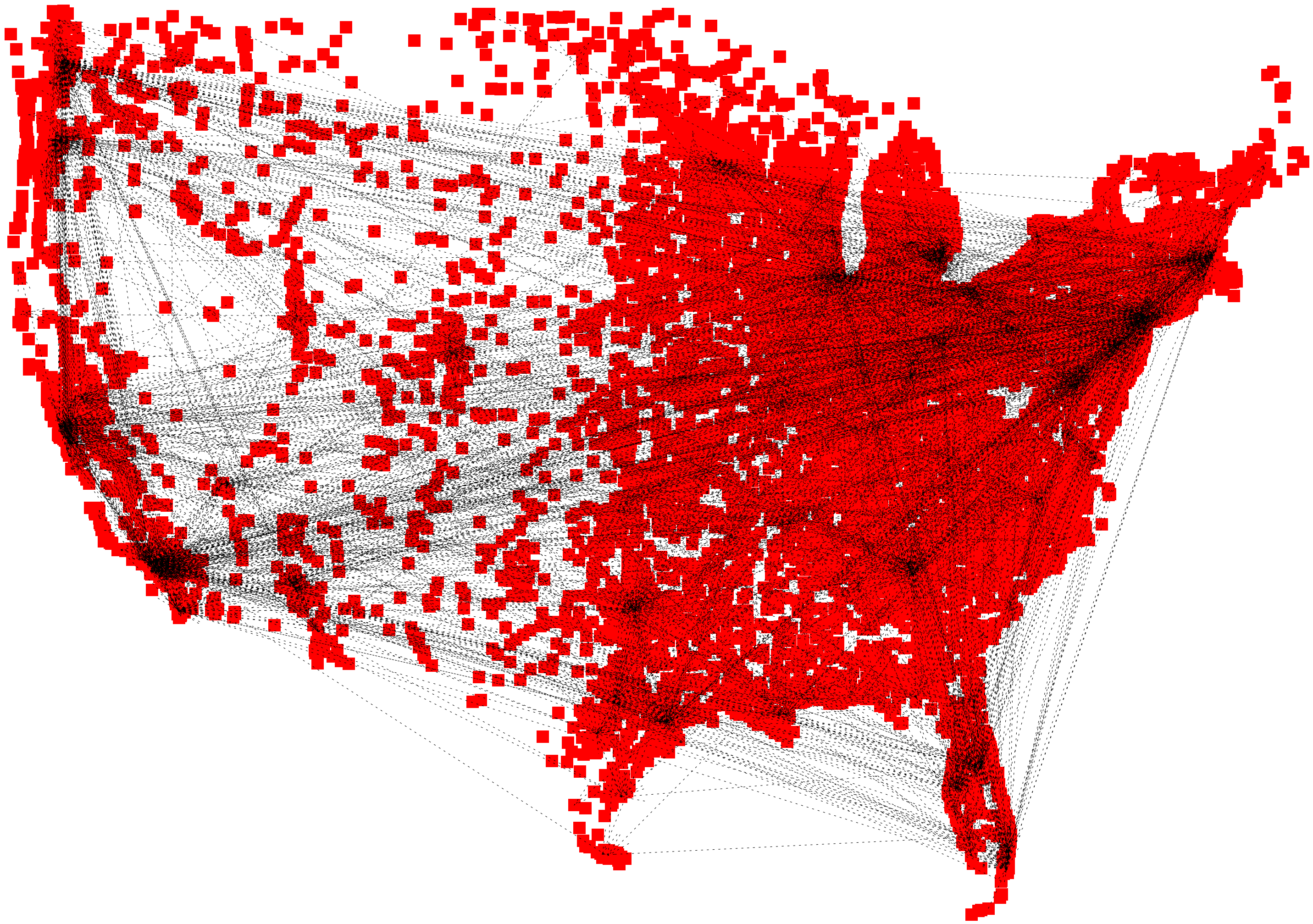
LiveJournal users provide:

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

- Yields a social network, with users' geographic locations.

LiveJournal

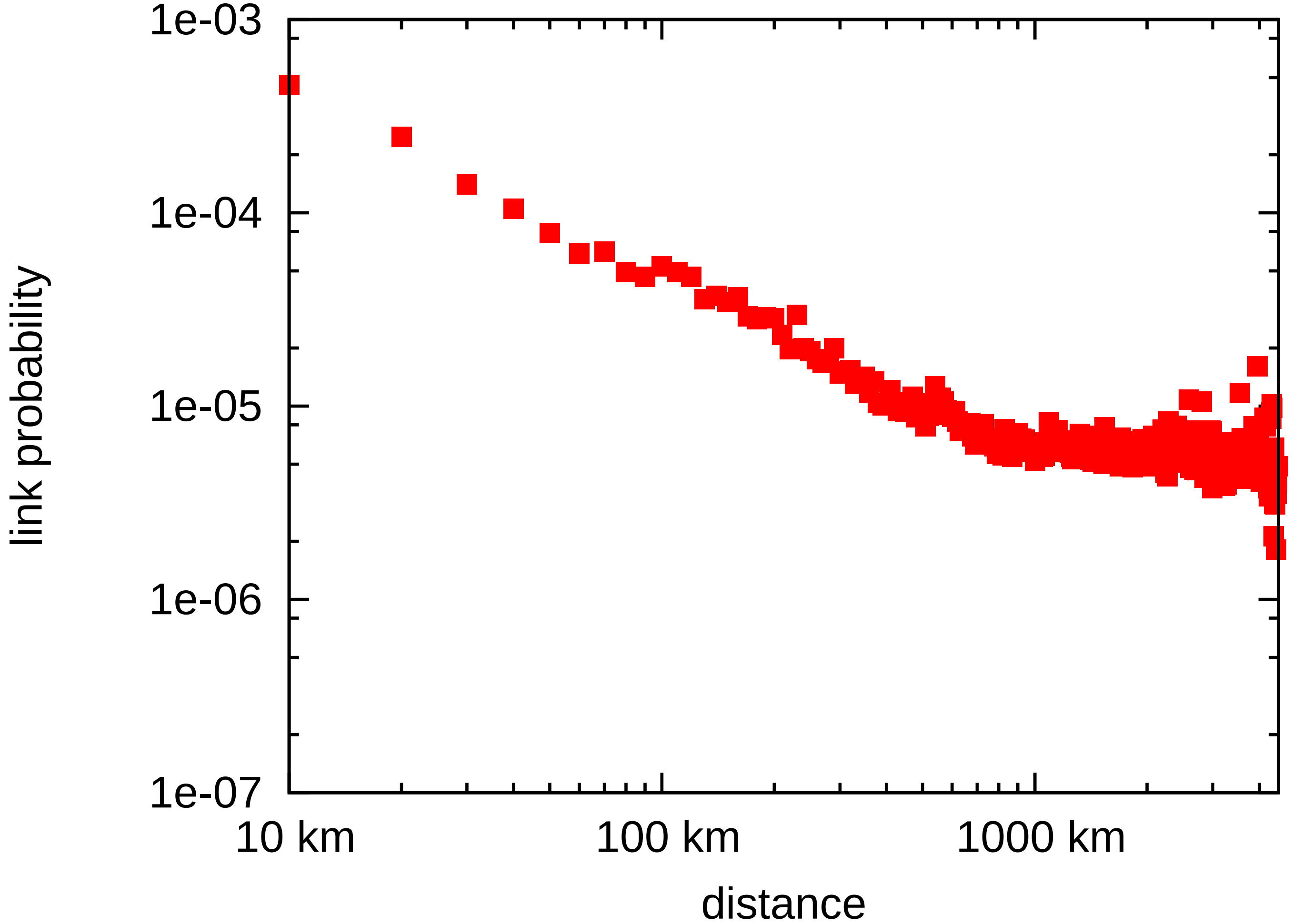
0.1% of LJ friendships



[DLN Novak Kumar Raghavan Tomkins 2005]

Distance versus LJ link probability

[DLN Novak Kumar Raghavan Tomkins 2005]



Analyzing Social Networks, pre-1995

Social Network Analysis: Old School

- social networks have been around for 100K+ years!
- before the web, hard to acquire (surveys, interviews, ...).
- but many interesting, relevant, generalizable observations!

Social Network Analysis: `New Sch00l`

- automatically extract networks without having to ask.
- phone calls, emails, online communities, ...
- big! but are these really social networks?

The Hewlett-Packard Email Community

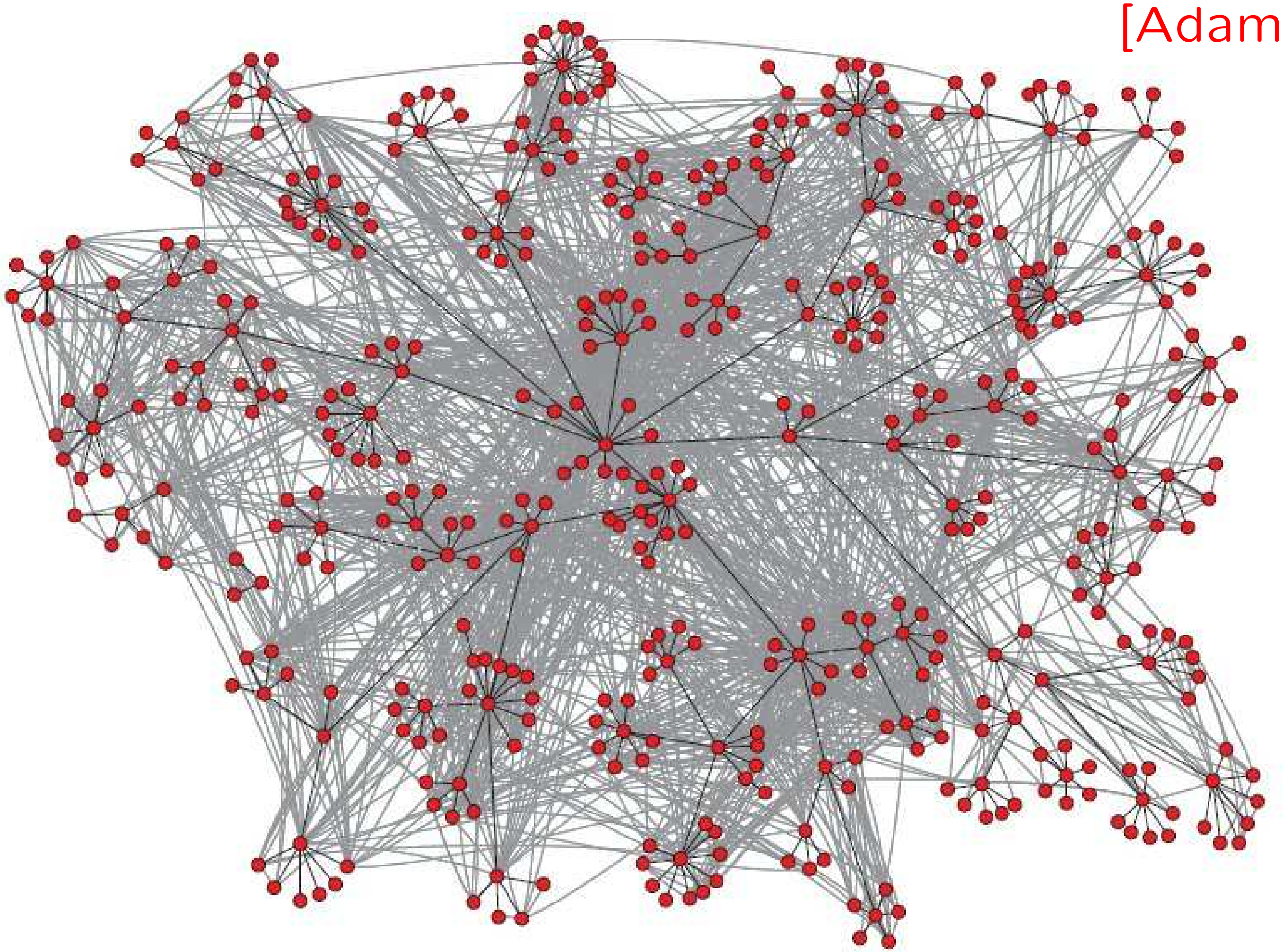
[Adamic Adar 2005]

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

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

Emails and the HP Corporate Hierarchy

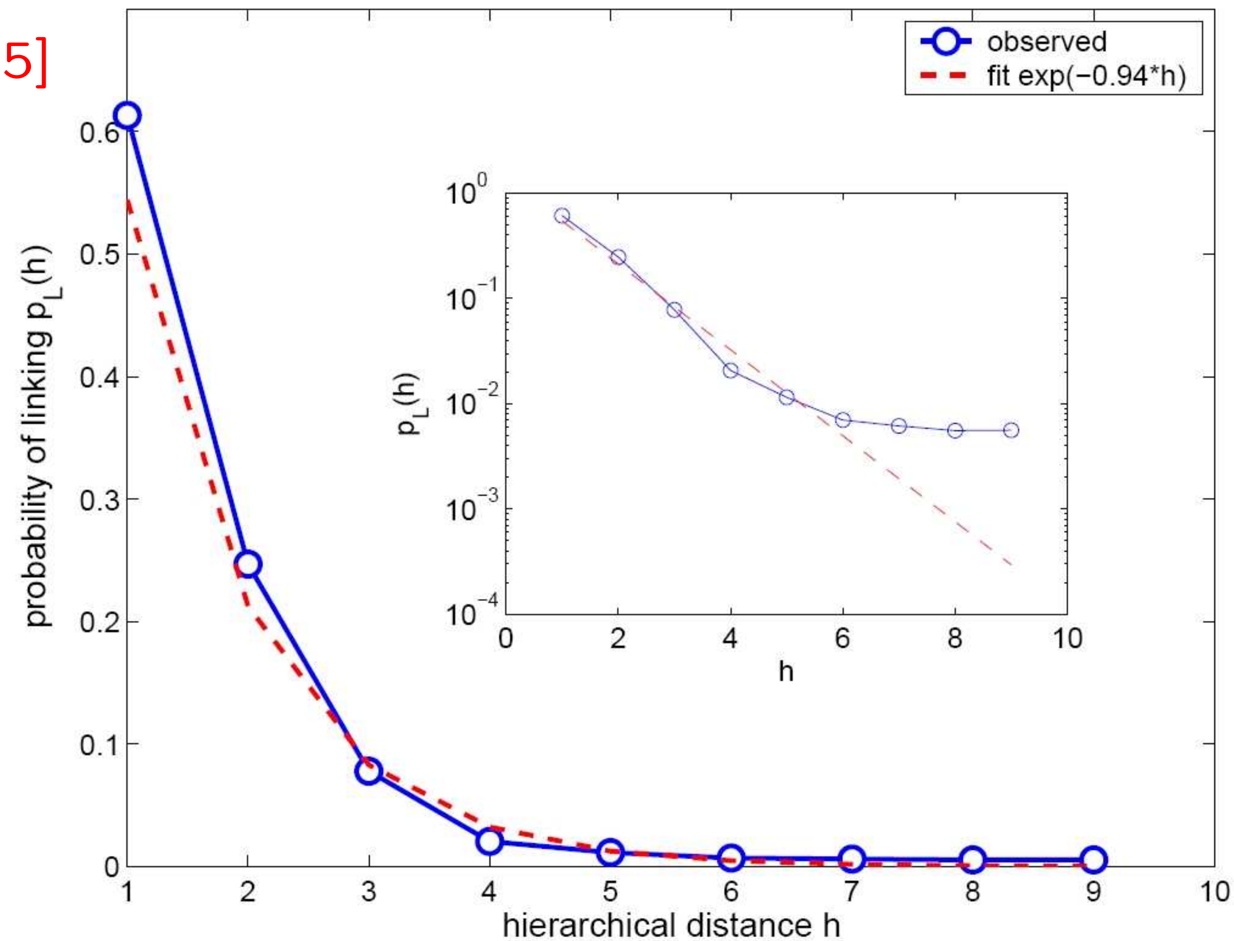
[Adamic Adar 2005]



black: HP corporate hierarchy gray: exchanged emails.

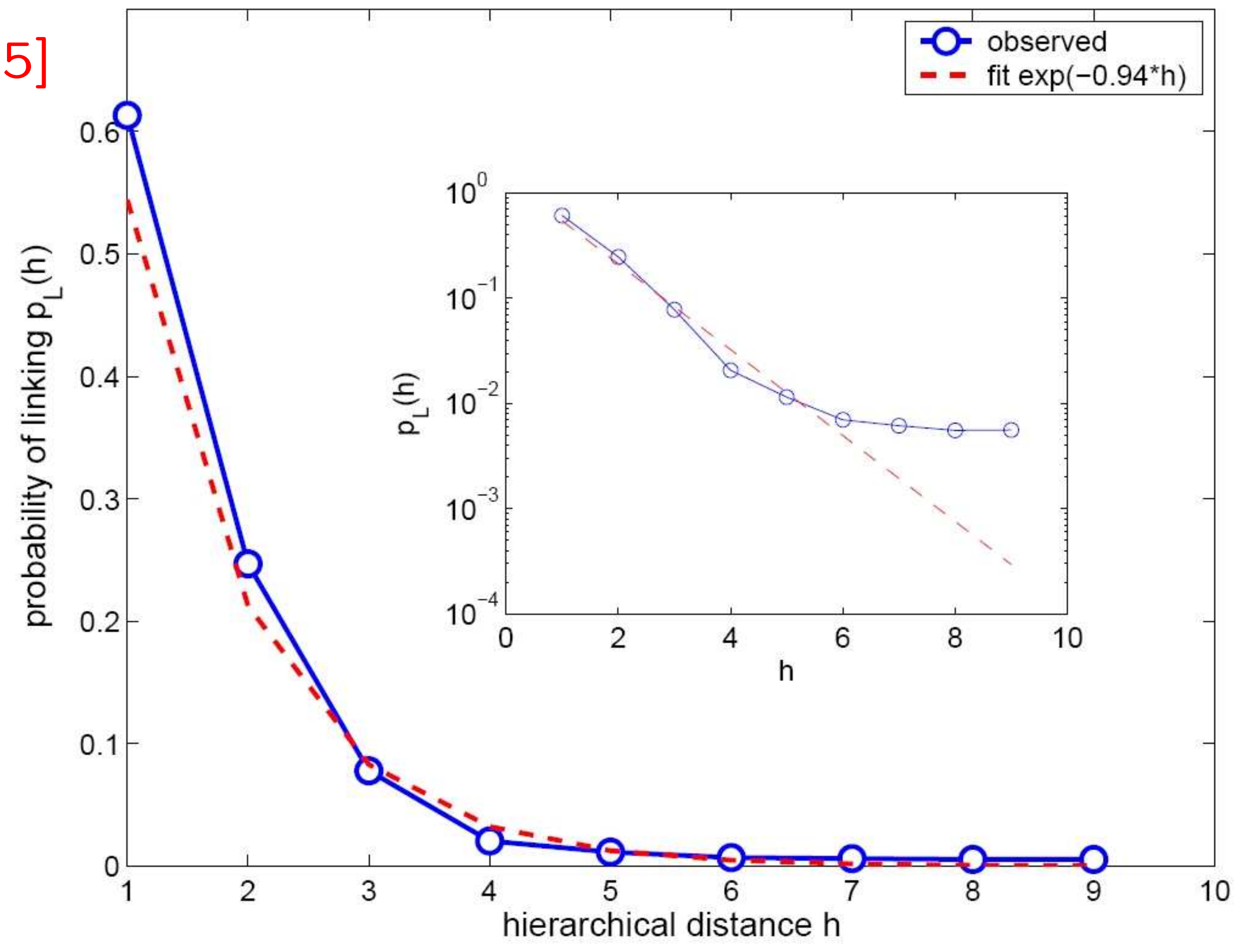
Emails and the HP Corporate Hierarchy

[Adamic Adar 2005]



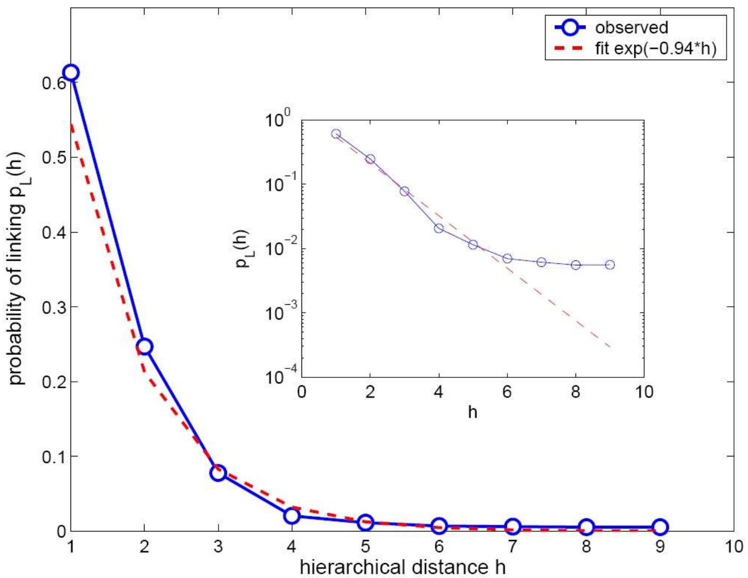
Emails and the HP Corporate Hierarchy

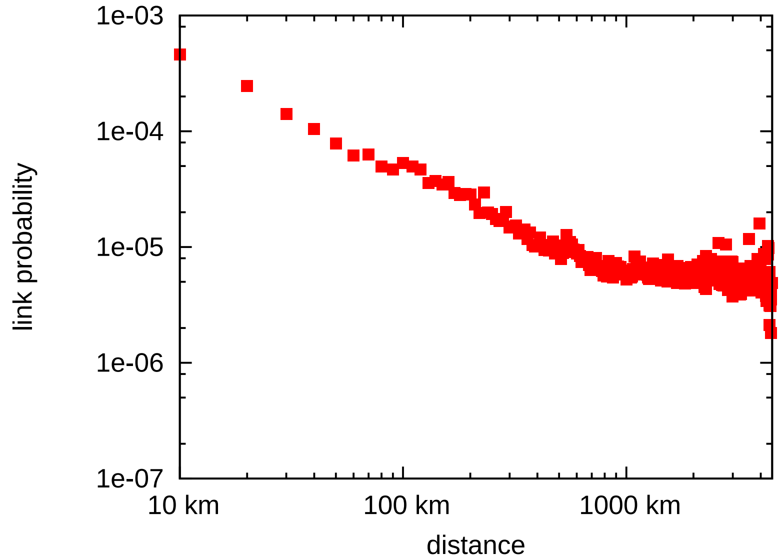
[Adamic Adar 2005]



So what?

Requisites for Navigability





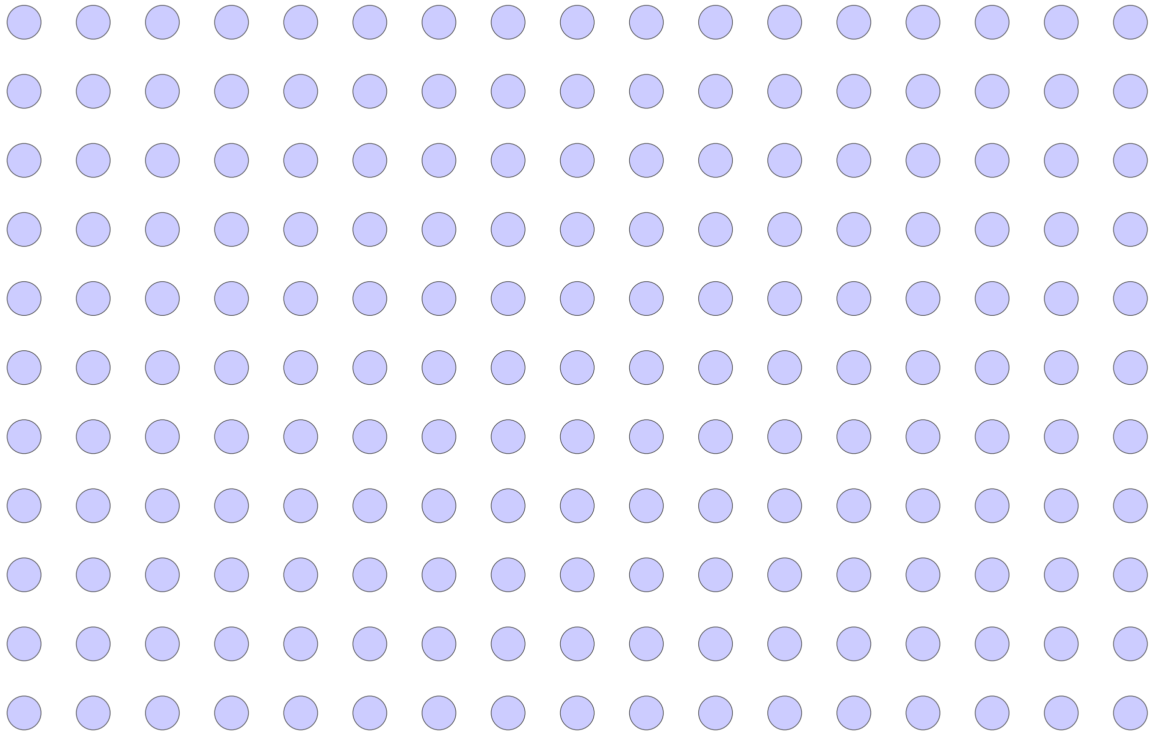
[Kleinberg 2000]:

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

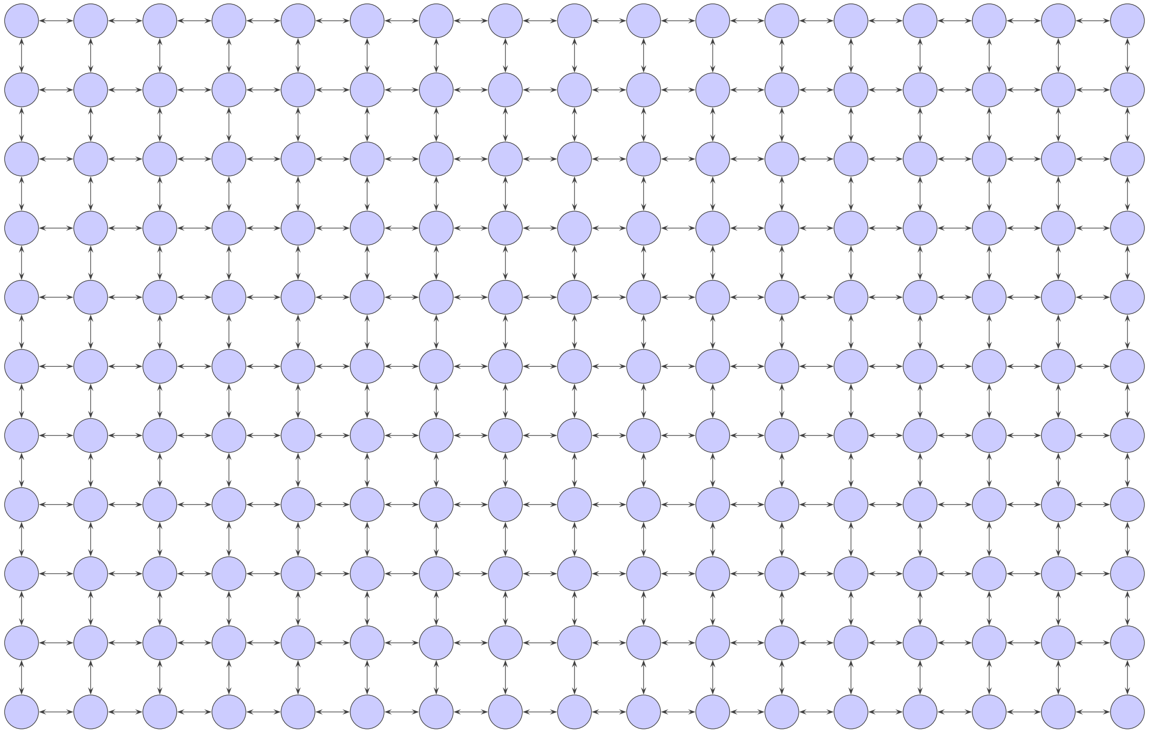
[Kleinberg 2000]



- put n people on a k-dimensional grid

Kleinberg: Navigable Social Networks

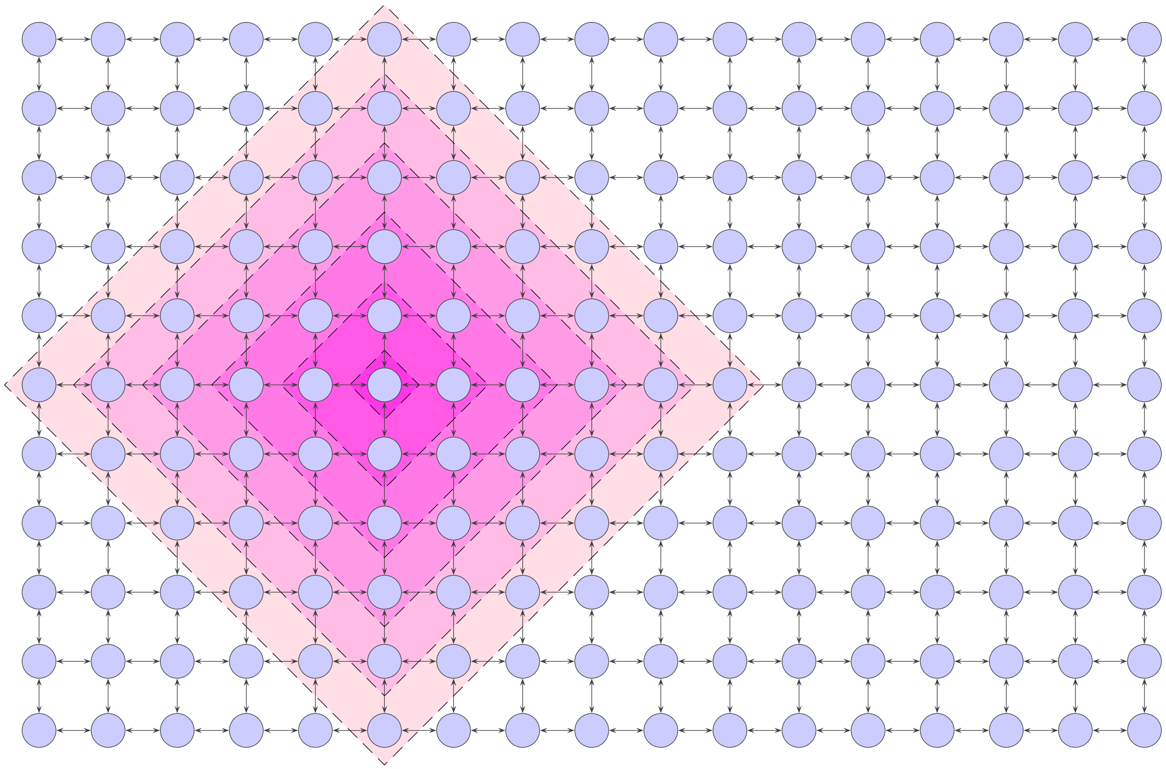
[Kleinberg 2000]



- put n people on a k-dimensional grid
- connect each to its immediate neighbors

Kleinberg: Navigable Social Networks

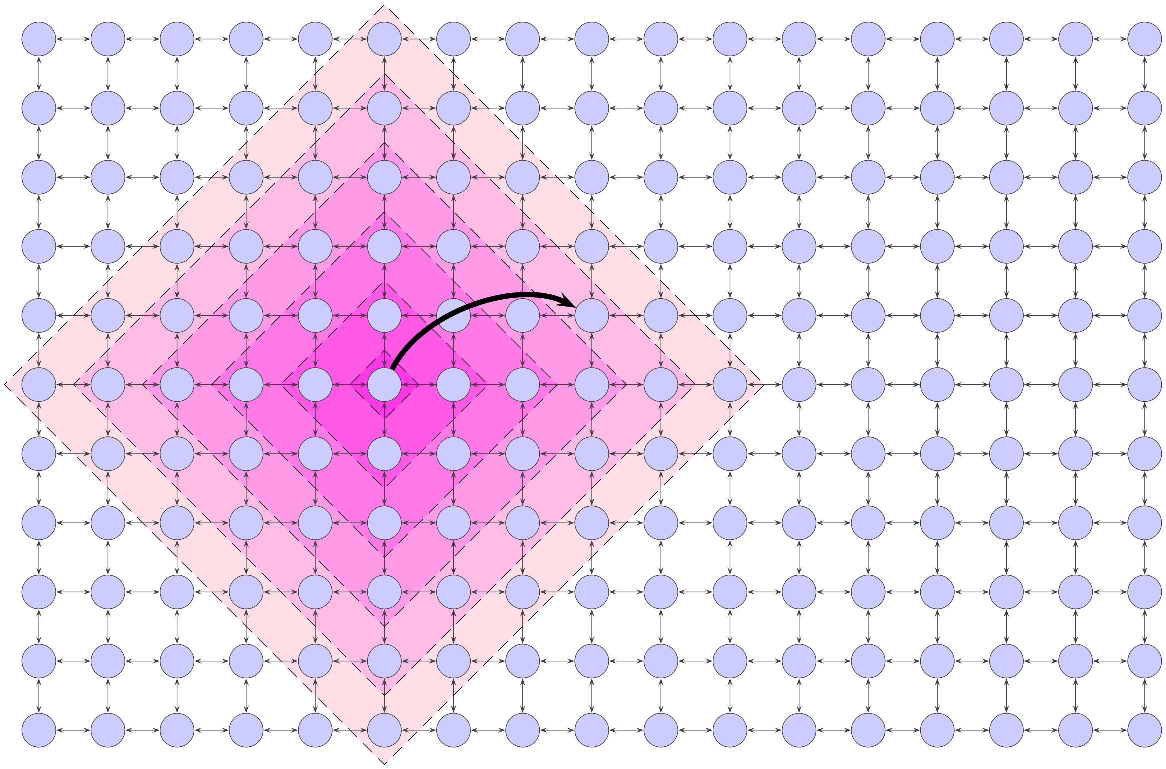
[Kleinberg 2000]



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

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; $\mathtt{Pr}[u \rightarrow v] \propto \frac{1}{d(u,v)^{\alpha}}$.

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; $\texttt{Pr}[u \to v] \propto \frac{1}{d(u,v)^\alpha}$.

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

If $\alpha \neq k$ (Watts/Strogatz: $\alpha = 0$)

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.

Sketch of Kleinberg's Proof

Kleinberg:

- n people on k-dimensional grid, with local links.
- $\texttt{Pr}[u \rightarrow v] \propto d(u,v)^{-k}$.

Sketch of Kleinberg's Proof

Kleinberg:
- n people on k-dimensional grid, with local links.
- $\mathtt{Pr}[u \to v] \propto d(u,v)^{-k}$.

$$\sum_{v \neq u} d(u,v)^{-k} \leq \sum_{d=1}^{n} d^{k-1} \cdot d^{-k} = O(\log n)$$

Sketch of Kleinberg's Proof

Kleinberg:
- n people on k-dimensional grid, with local links.
- $\texttt{Pr}[u \to v] \propto d(u,v)^{-k}$.

$$\sum_{v \neq u} d(u,v)^{-k} \leq \sum_{d=1}^{n} d^{k-1} \cdot d^{-k} = O(\log n)$$

$$\Rightarrow \boxed{\texttt{Pr}[u \to v] = \Omega\left(\frac{1}{d(u,v)^k \cdot \log n}\right)}.$$

Sketch of Kleinberg's Proof

Kleinberg:
- n people on k-dimensional grid, with local links.
- $\texttt{Pr}[u \to v] \propto d(u,v)^{-k}$.

$$\sum_{v \neq u} d(u,v)^{-k} \leq \sum_{d=1}^{n} d^{k-1} \cdot d^{-k} = O(\log n)$$

$$\Rightarrow \boxed{\texttt{Pr}[u \to v] = \Omega\left(\frac{1}{d(u,v)^k \cdot \log n}\right)}.$$

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

- After $\log n$ halvings, done!

Sketch of Kleinberg's Proof

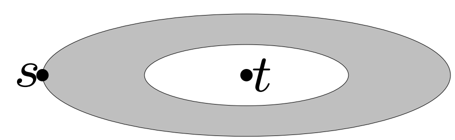
Kleinberg:
- n people on k-dimensional grid, with local links.
- $\mathtt{Pr}[u \to v] \propto d(u,v)^{-k}$.

$$\sum_{v \neq u} d(u,v)^{-k} \leq \sum_{d=1}^{n} d^{k-1} \cdot d^{-k} = O(\log n)$$

$$\Rightarrow \boxed{\mathtt{Pr}[u \to v] = \Omega\left(\frac{1}{d(u,v)^k \cdot \log n}\right)}.$$

Claim: $\mathtt{Pr}\left[s \text{ friends with } u \text{ within } \frac{d(s,t)}{2} \text{ of } t\right] = \Omega\left(\frac{1}{\log n}\right)$.

- After $\log n$ halvings, done!



- Proof of claim: $\quad d = d(s,t)$
  Number of people within distance $d/2$ of t is $\Theta(d^k)$.
  Distance from s to any of them is $\leq 3d/2$.

Sketch of Kleinberg's Proof

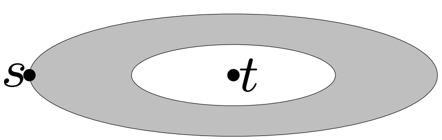
Kleinberg:
- n people on k-dimensional grid, with local links.
- $\mathtt{Pr}[u \to v] \propto d(u,v)^{-k}$.

$$\sum_{v \neq u} d(u,v)^{-k} \leq \sum_{d=1}^{n} d^{k-1} \cdot d^{-k} = O(\log n)$$

$$\Rightarrow \boxed{\mathtt{Pr}[u \to v] = \Omega\left(\frac{1}{d(u,v)^k \cdot \log n}\right)}.$$

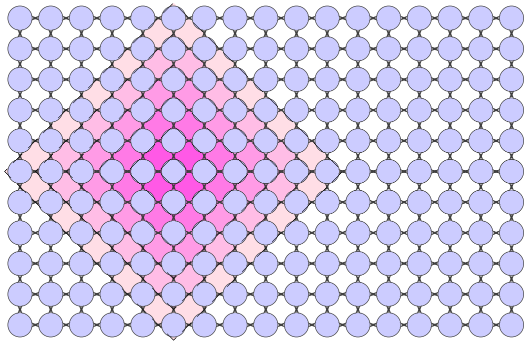
Claim: $\mathtt{Pr}\left[s \text{ friends with } u \text{ within } \frac{d(s,t)}{2} \text{ of } t\right] = \Omega\left(\frac{1}{\log n}\right)$.

- After $\log n$ halvings, done!

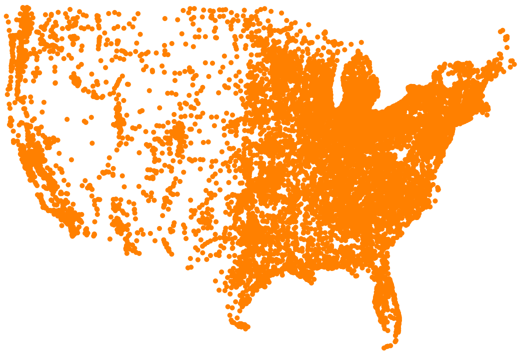


- Proof of claim: $\quad d = d(s,t)$
  Number of people within distance $d/2$ of t is $\Theta(d^k)$.
  Distance from s to any of them is $\leq 3d/2$.
  Probability of s linking to one of them is $\Omega(d^{-k}/\log n)$.
  Probability of s linking to any one of them is $\Omega(1/\log n)$.

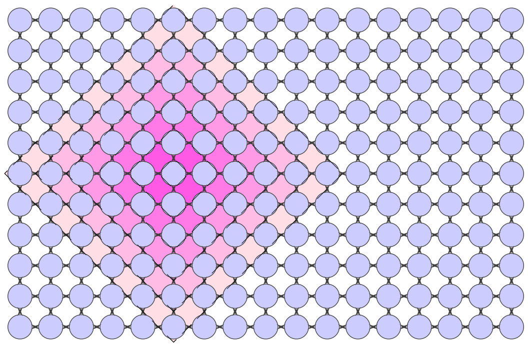
Going off the Grid



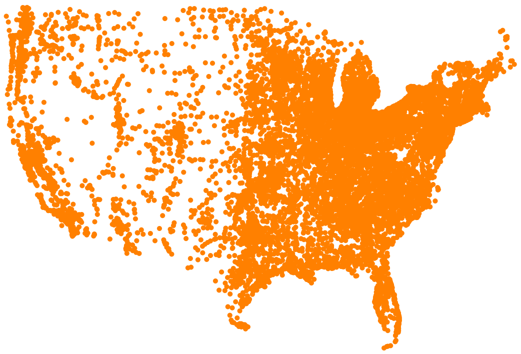
Even for geography, the uniform grid is a poor model of real populations.



Going off the Grid

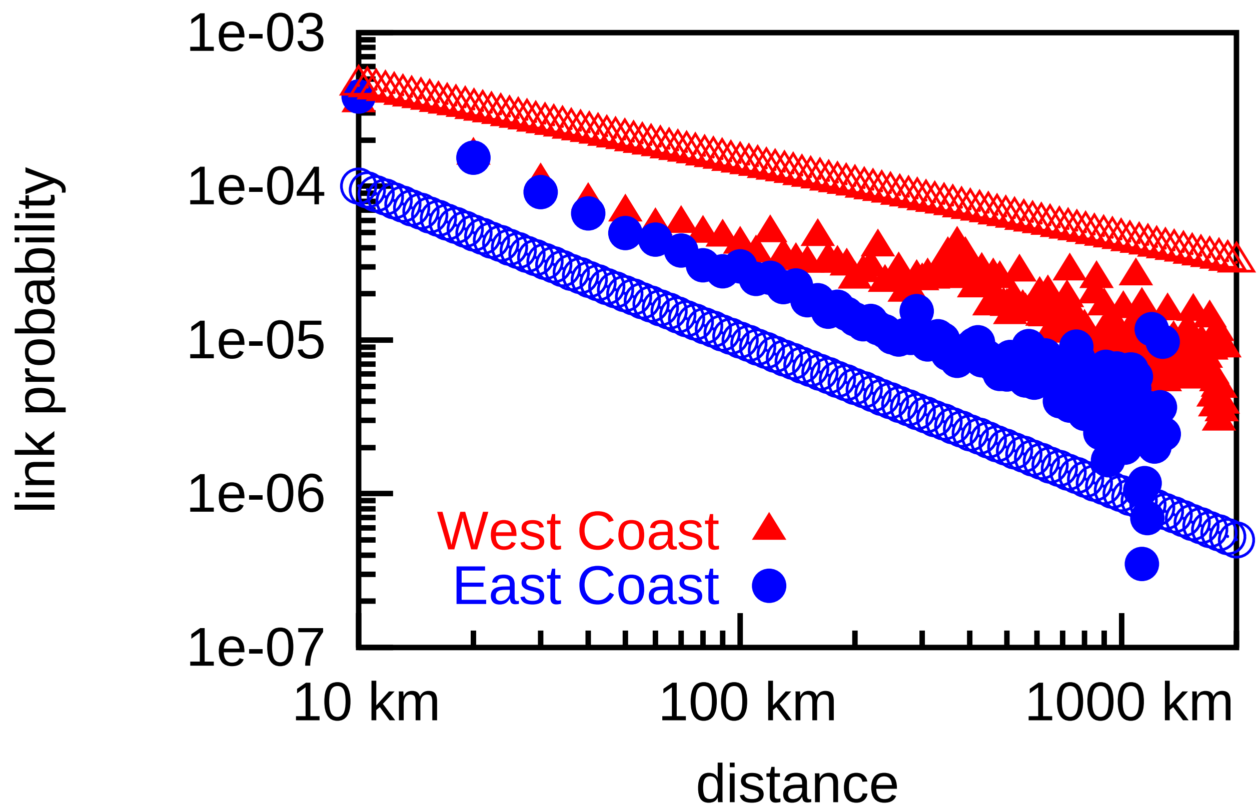


Even for geography, the uniform grid is a poor model of real populations.



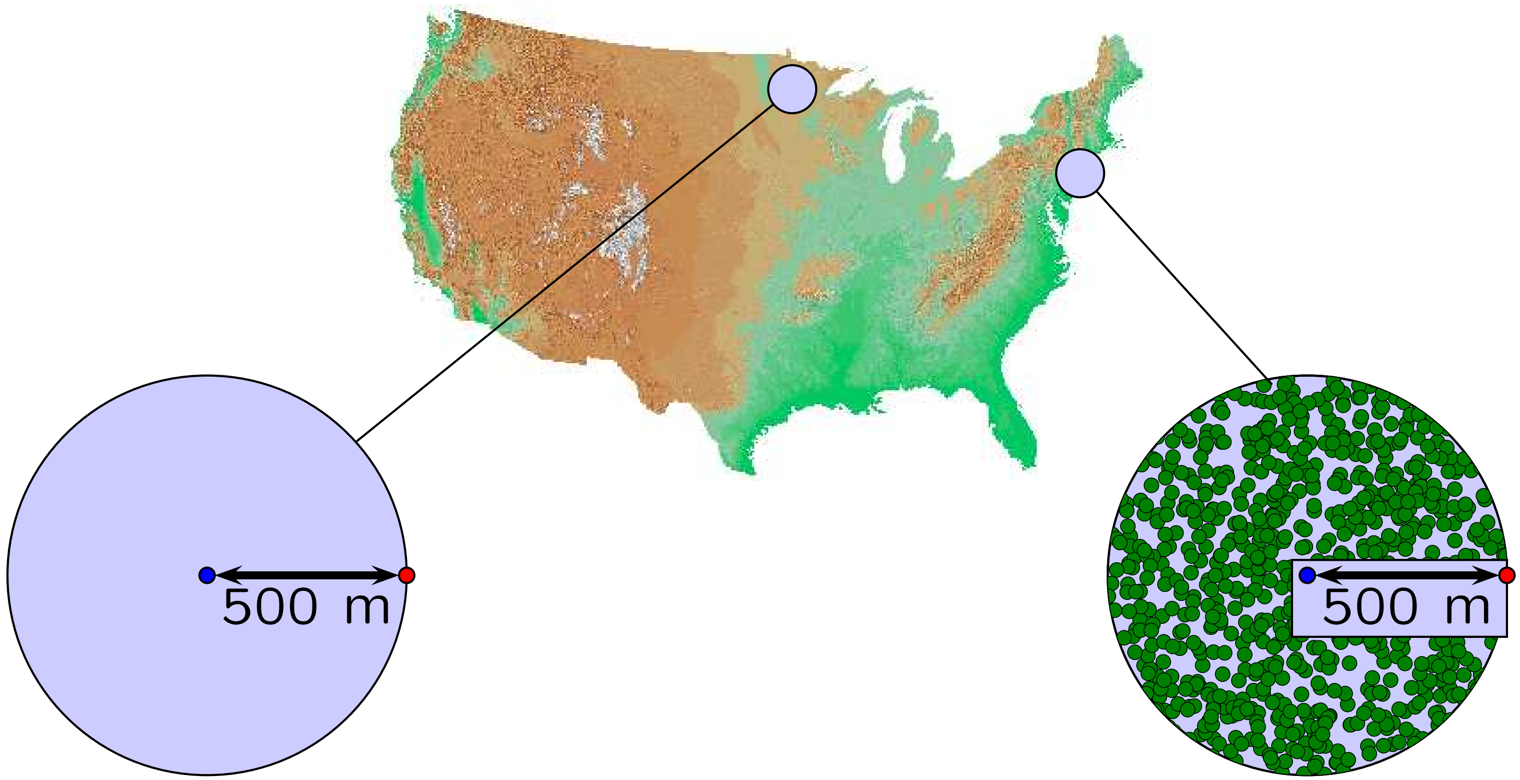
- Hierarchical models of proximity.
  [Kleinberg 2001] [Watts Dodds Newman 2002] ...
- An arbitrary metric space of points ...
  [Slivkins 2005] [Duchon Hanusse Lebhar Schabanel 2005] [Fraigniaud Lebhar Lotker 2006] ...
- ... with arbitrary population distributions.
  [DLN Novak Kumar Raghavan Tomkins 2005] [Kumar DLN Tomkins 2006] ...

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.

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

How many people live closer to A than B does?

Rank-Based Friendship

How do we handle non-uniformly distributed populations?

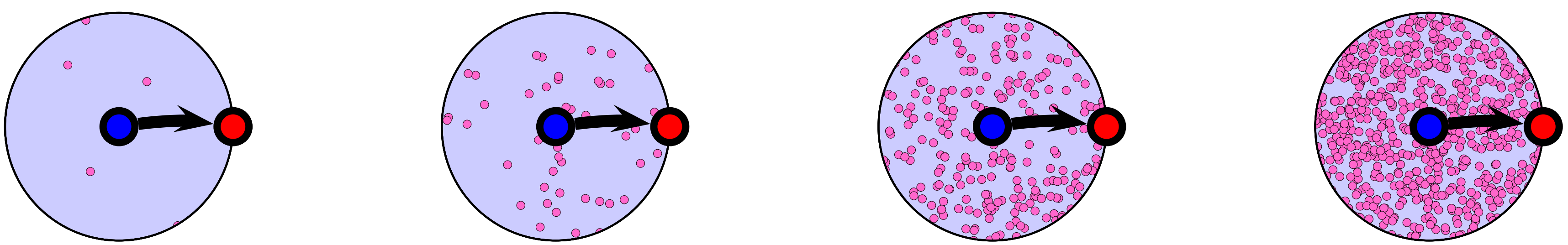
Instead of distance, use rank as fundamental quantity.

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

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

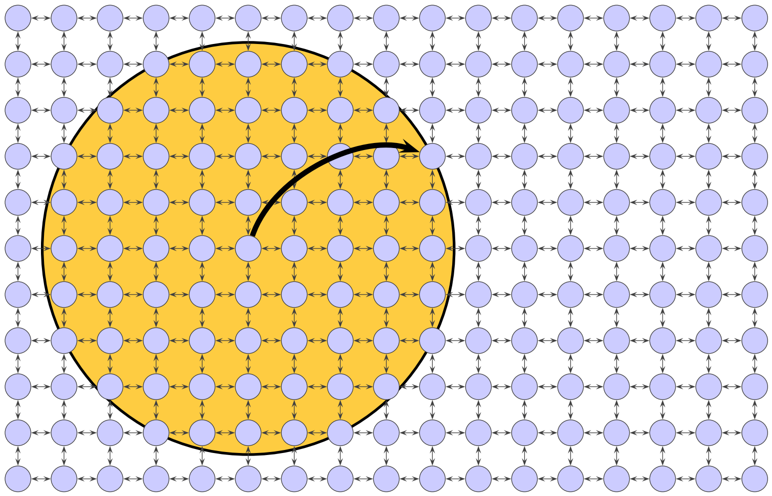
Kleinberg (k-dim grid): $\texttt{Pr}[A \text{ is a friend of } B] \propto 1/d(A,B)^k$.

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 $\approx d^k$

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

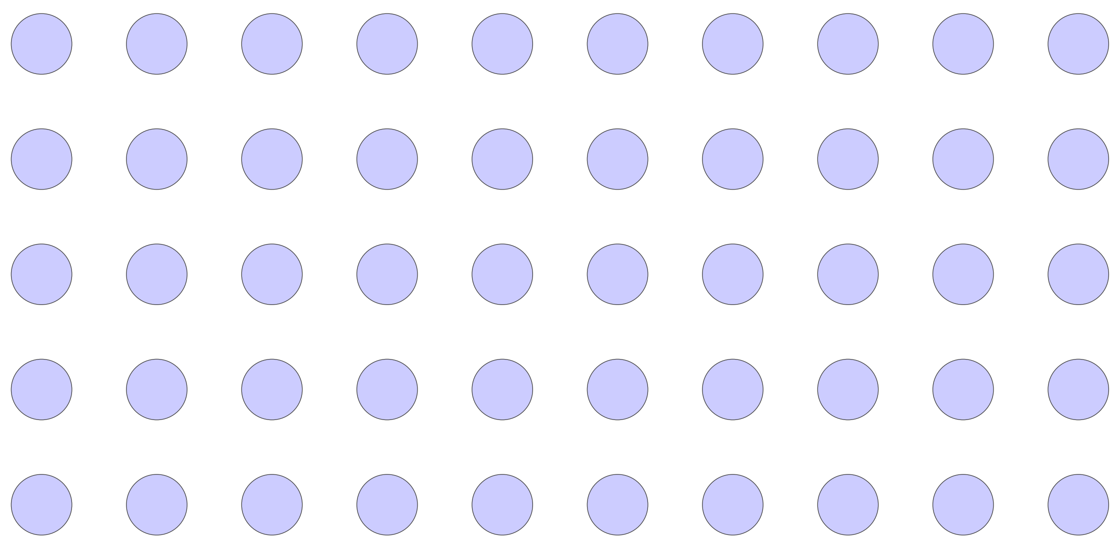
For a uniform grid, rank-based friendship has (essentially) same link probabilities as Kleinberg.

Population Networks

A rank-based population network consists of:

➡ a k-dimensional grid L of locations.

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

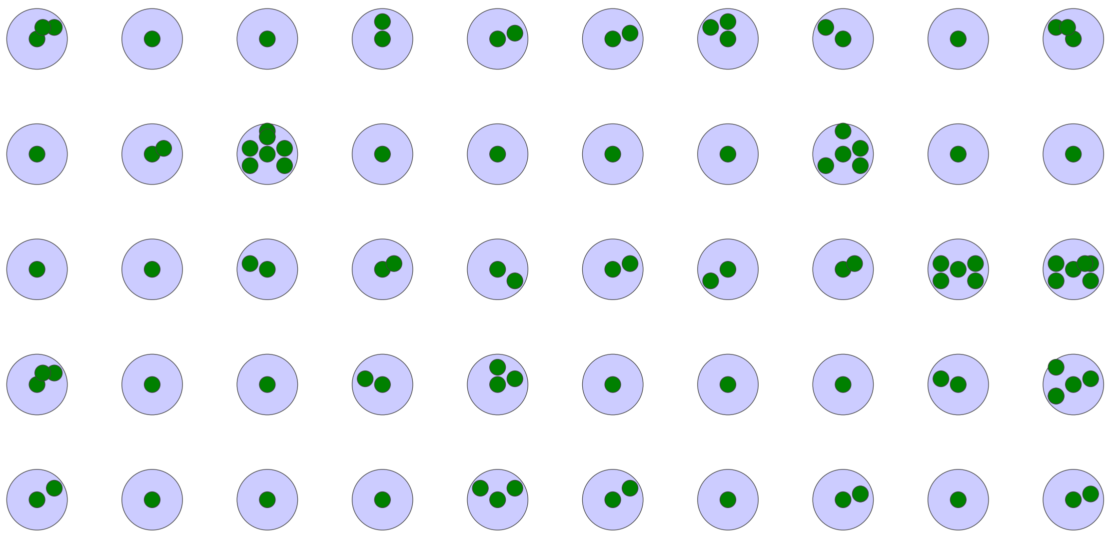


Population Networks

A rank-based population network consists of:

- a k-dimensional grid L of locations.
- a population P of people, living at points in L ($n := |P|$).

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

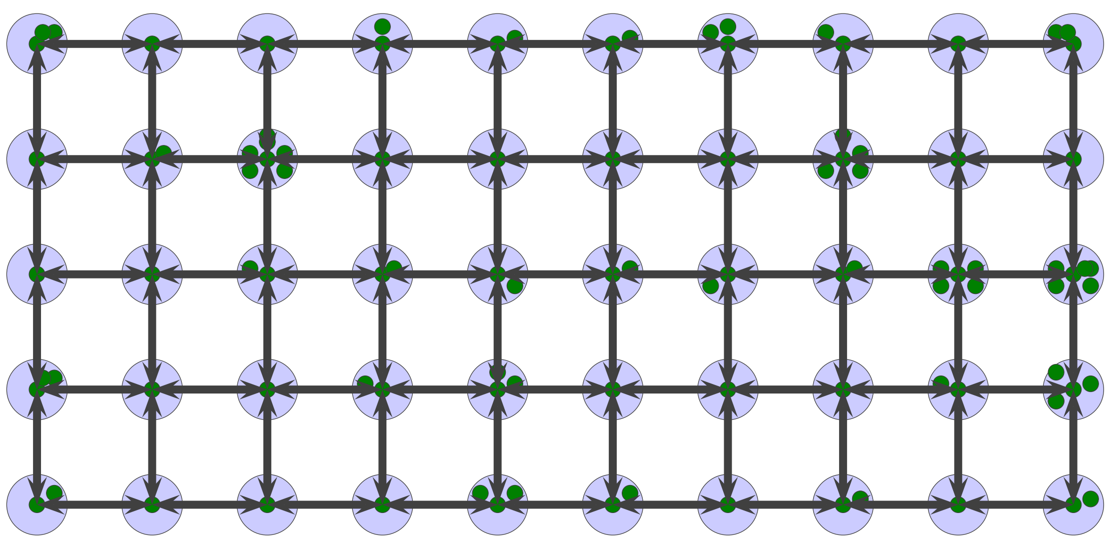


Population Networks

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'

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

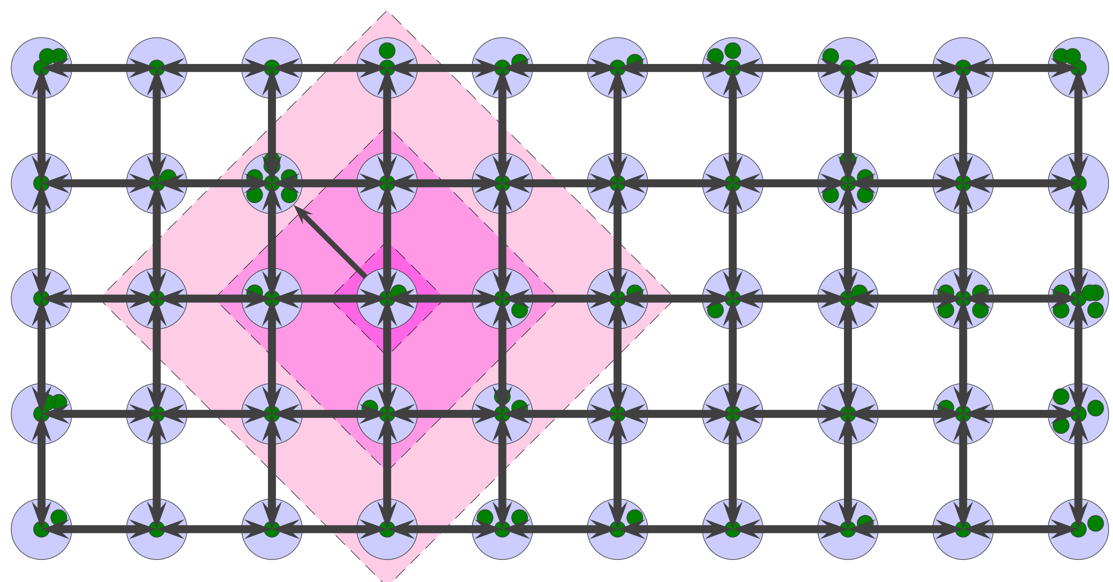


Population Networks

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

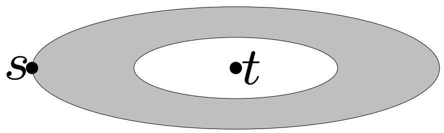
[Kumar DLN Tomkins 2006]

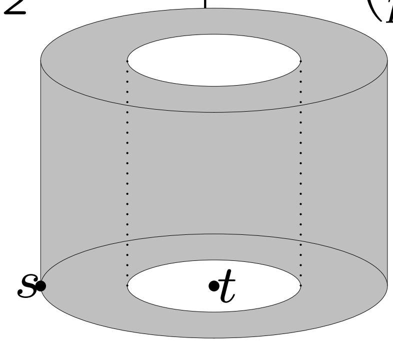
Theorem: For any n-person rank-based population network in a k-dimensional grid, $k = \Theta(1)$, for any source $s \in P$ and for a randomly chosen target $t \in P$, the expected length (over t) of $\mathit{Greedy}(s, \mathsf{loc}(t))$ is $O(\log^3 n)$.

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!

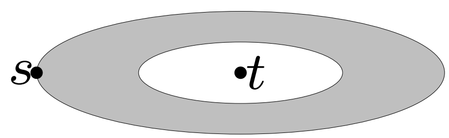


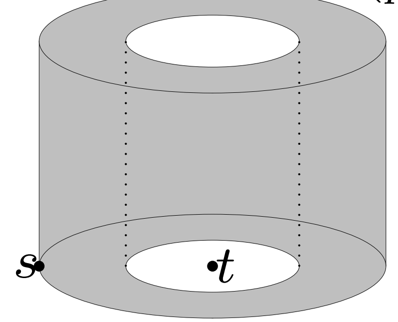


Is this just like all the other proofs?

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!



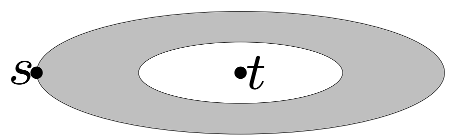


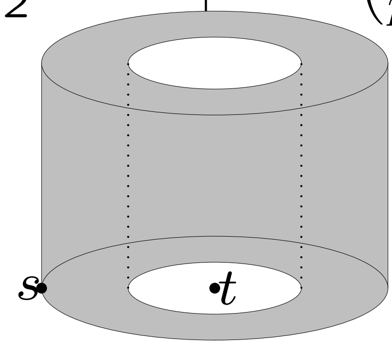
- Claim is false if $\{u : d(u,t) < \frac{d(s,t)}{2}\} \ll \{u : d(u,t) < d(s,t)\}$!

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!

Short Paths and Rank-Based Friendships

[Kumar DLN Tomkins 2006]

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 $O(\log^3 n)$.

High-level proof idea:

- Intuition: difficulty of halving distance to isolated target t is canceled by low probability of choosing t.
- Fix t; let $\beta_t := \min\{\mathtt{Pr}[u \text{ is 'halving}_t\text{'}] : u \text{ found by Greedy}\}$.
  Claim: $\mathtt{E}_t[1/\beta_t] = O(\log^2 n)$.
  Then total path length is $O(\log^3 n)$.

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, ...

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.
[Kim Yoon Han Jeong 2002]
[Adamic Lukose Puniyani Huberman 2001]

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.
[Kim Yoon Han Jeong 2002]
[Adamic Lukose Puniyani Huberman 2001]

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

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.

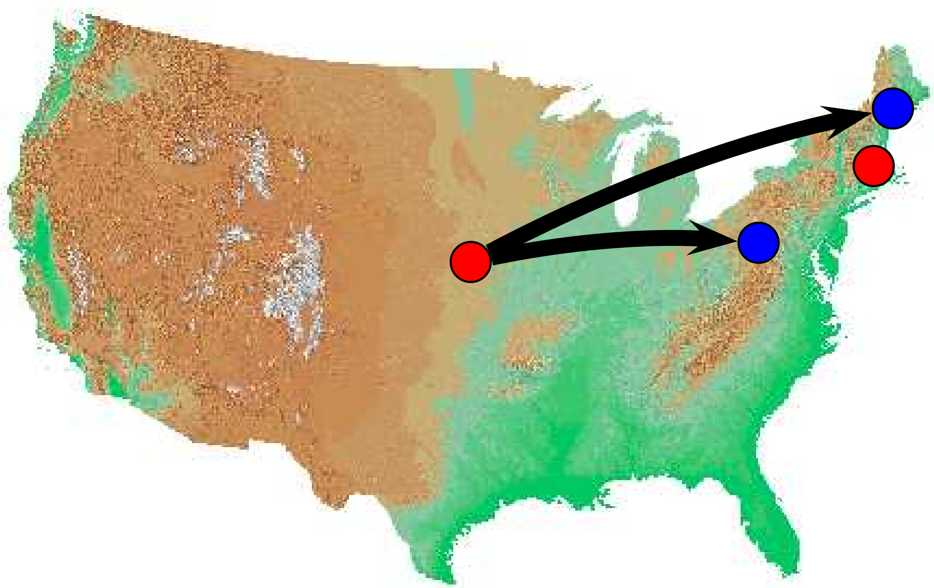
Age, hobbies, alma mater, ...

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

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?

Expected-Value Navigation

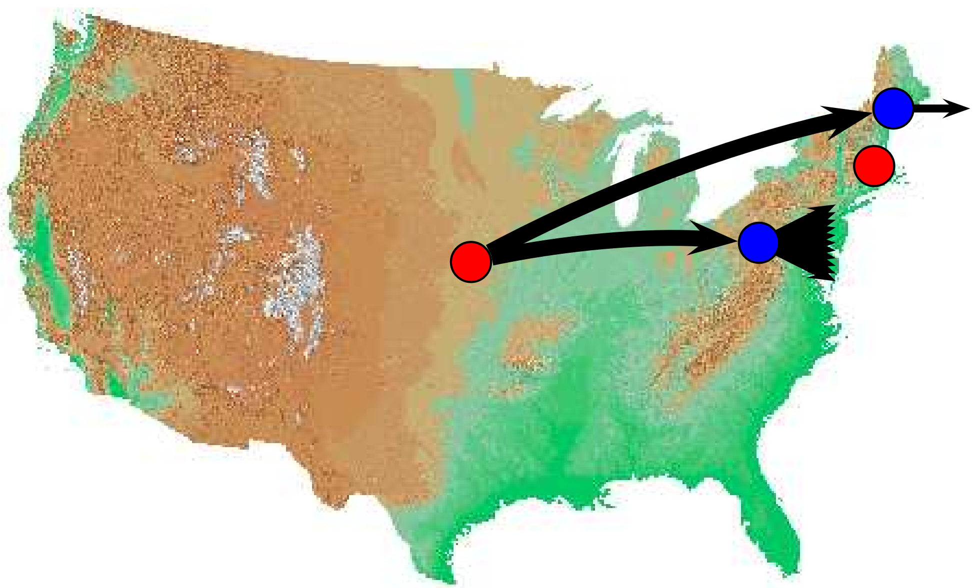


[Şimşek Jensen 2005]

Obviously 'should' choose as next step in chain

$u \in$ Friends minimizing $\quad \mathbb{E}$[length of $u \to t$ path]

Expected-Value Navigation

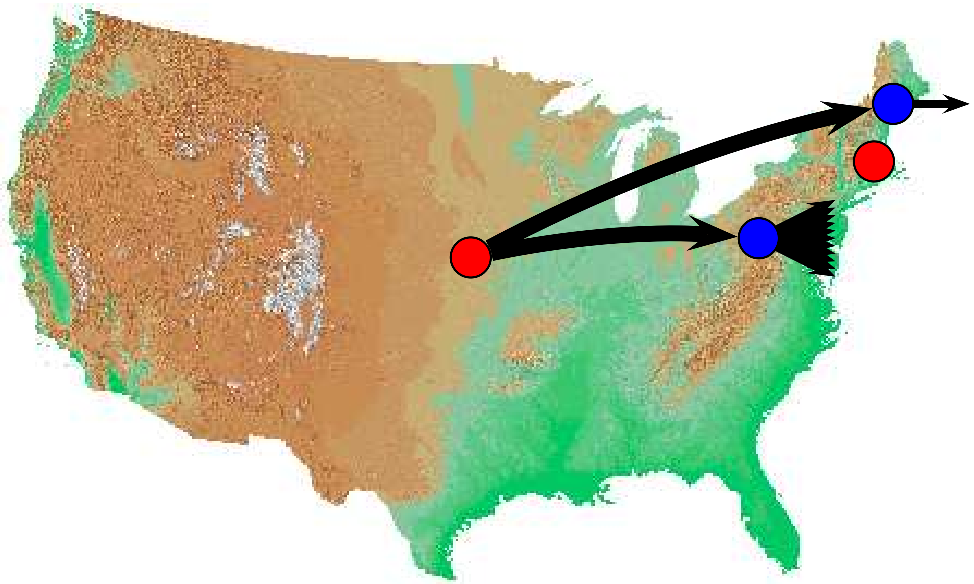


[Şimşek Jensen 2005]

Obviously 'should' choose as next step in chain

$u \in$ Friends minimizing $\quad \mathtt{E}$[length of $u \to t$ path]

Expected-Value Navigation



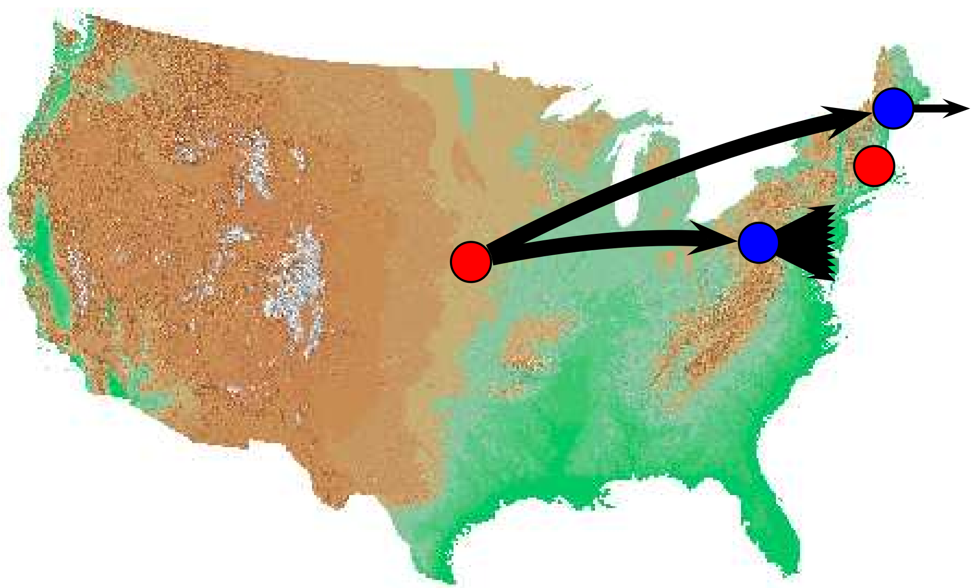
[Şimşek Jensen 2005]

Obviously 'should' choose as next step in chain

$$u \in \text{Friends minimizing} \quad \mathtt{E}[\text{length of } u \to t \text{ path}]$$

$$= \quad u \in \text{Friends minimizing} \quad \sum_i i \cdot \mathtt{Pr}[\exists u \to t \text{ path of length } i]$$

Expected-Value Navigation

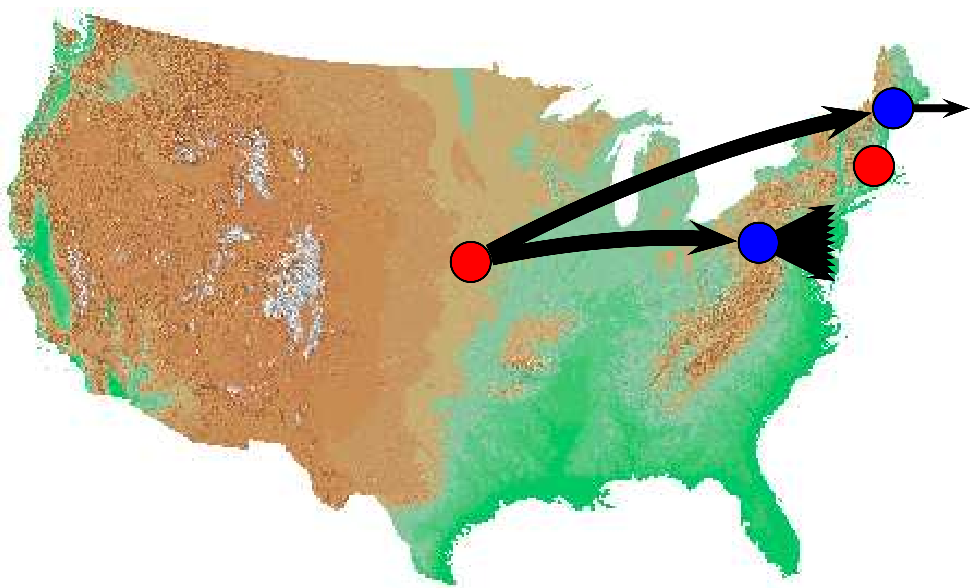


[Şimşek Jensen 2005]

Obviously 'should' choose as next step in chain

$$
\begin{array}{rll}
 & u \in \text{Friends minimizing} & \mathtt{E}[\text{length of } u \to t \text{ path}] \\
= & u \in \text{Friends minimizing} & \sum_i i \cdot \mathtt{Pr}[\exists u \to t \text{ path of length } i] \\
\approx & u \in \text{Friends maximizing} & \mathtt{Pr}[\exists u \to t \text{ path of length 1}]
\end{array}
$$

Expected-Value Navigation

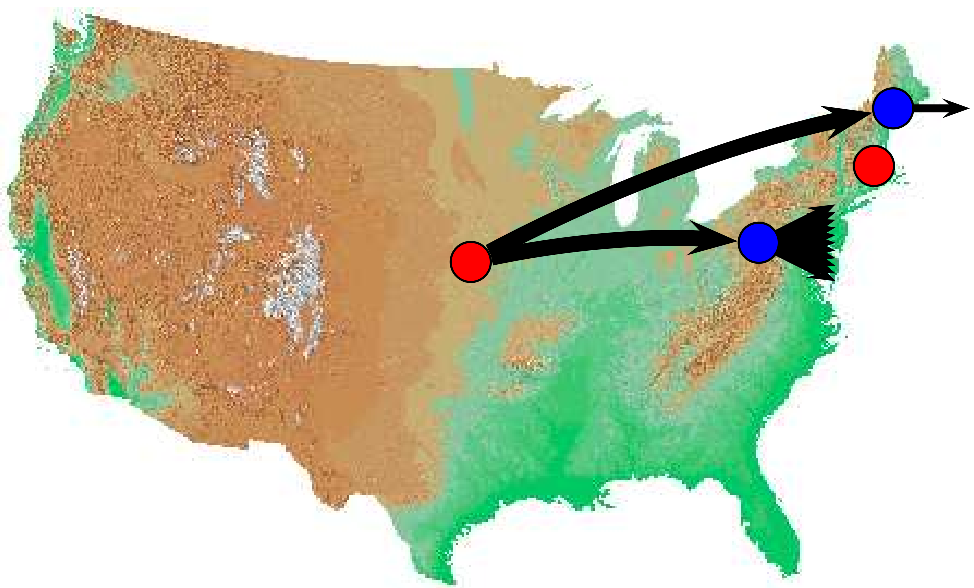


[Şimşek Jensen 2005]

Obviously 'should' choose as next step in chain

$$
\begin{aligned}
& u \in \text{Friends minimizing} && \mathtt{E}[\text{length of } u \to t \text{ path}] \\
= \;& u \in \text{Friends minimizing} && \sum_i i \cdot \mathtt{Pr}[\exists u \to t \text{ path of length } i] \\
\approx \;& u \in \text{Friends maximizing} && \mathtt{Pr}[\exists u \to t \text{ path of length } 1] \\
= \;& u \in \text{Friends minimizing} && (1 - \mathtt{Pr}[\text{one link from } u \text{ goes to } t])^{\text{degree}(u)}
\end{aligned}
$$

Expected-Value Navigation



Combines (one) proximity with a high-degree seeking strategy for search.
(What about > 1?)

[Şimşek Jensen 2005]

Obviously 'should' choose as next step in chain

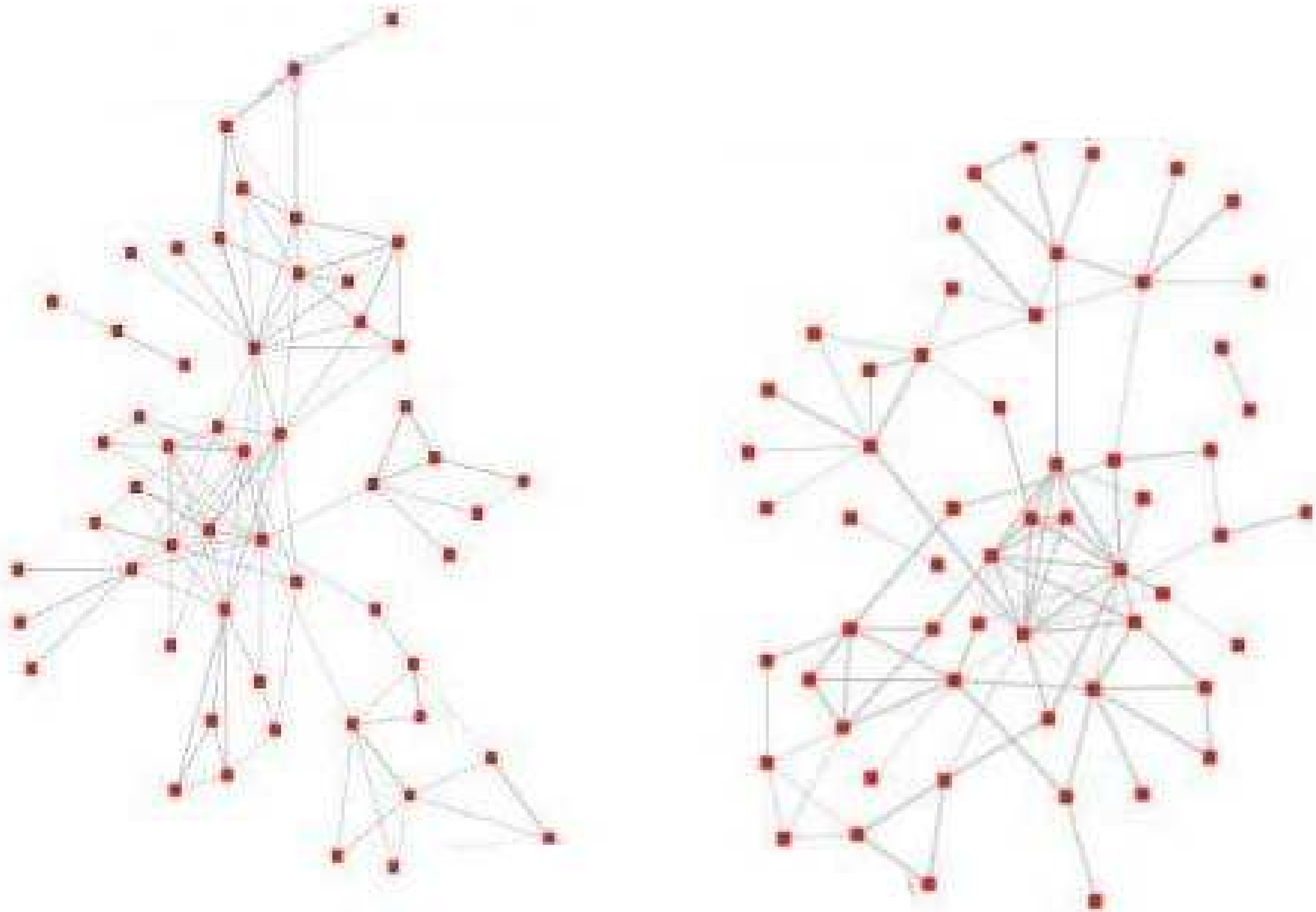
$$
\begin{aligned}
& u \in \text{Friends minimizing} \quad \mathtt{E}[\text{length of } u \to t \text{ path}] \\
= \; & u \in \text{Friends minimizing} \quad \sum_i i \cdot \mathtt{Pr}[\exists u \to t \text{ path of length } i] \\
\approx \; & u \in \text{Friends maximizing} \quad \mathtt{Pr}[\exists u \to t \text{ path of length } 1] \\
= \; & u \in \text{Friends minimizing} \quad (1 - \mathtt{Pr}[\text{one link from } u \text{ goes to } t])^{\text{degree}(u)}
\end{aligned}
$$

Part II:

The Information Content of Social Networks*

* the biased DLN-centric perspective continues.

Two Social Networks



One Fortune 500 executives; one terrorist network.
[Krebs 2002]

Milgram: Some Doubts

"6 degrees of separation between any 2 people!"

"the social network's diameter is 6."

Milgram: Some Doubts

"6 degrees of separation between any 2 people!"

"the social network's diameter is 6."

Some reasons for skepticism: [Kleinfeld 2002]

- only $n = 96$ chains ... and only 18 completed!

Milgram: Some Doubts

“6 degrees of separation between any 2 people!”

“the social network’s diameter is 6.”

Some reasons for skepticism: [Kleinfeld 2002]

- only $n = 96$ chains ... and only 18 completed!
- socially prominent target (and socially ‘active’ sources?)
- subsequent experiments:
  poor black target person significantly harder to reach.

Milgram: Some Doubts

“6 degrees of separation between any 2 people!”

“the social network’s diameter is 6.”

Some reasons for skepticism: [Kleinfeld 2002]

- only $n = 96$ chains ... and only 18 completed!
- socially prominent target (and socially ‘active’ sources?)
- subsequent experiments:
  poor black target person significantly harder to reach.
- little data on why failed chains failed
  (maybe they got badly stuck and people gave up?)

Email-based Small-World Experiment

A recent email-based retrial. [Dodds Muhamad Watts 2003]

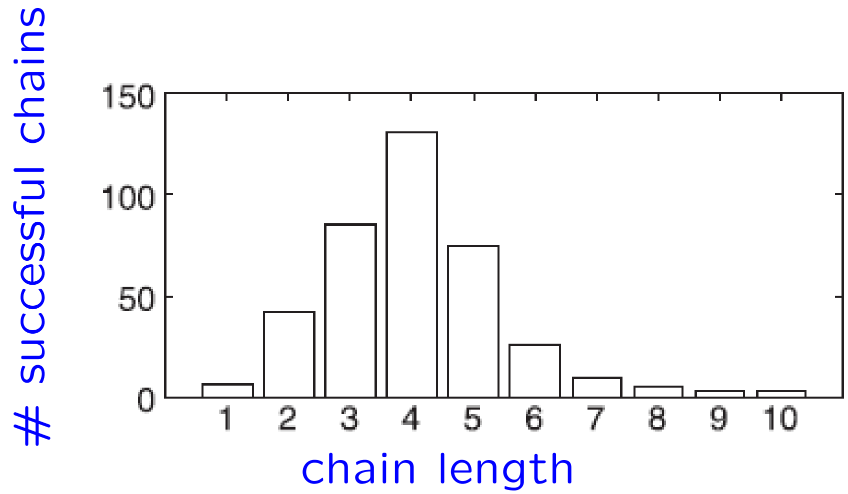
- 18 targets, 24K message chains, 61K participants.
- Success rate: 1.59% (384/24K chains).

Email-based Small-World Experiment

A recent email-based retrial. [Dodds Muhamad Watts 2003]

- 18 targets, 24K message chains, 61K participants.
- Success rate: 1.59% (384/24K chains).

Attrition rate $\approx 2/3 \Rightarrow (1/3)^4 = 1.2\%$ of chains survive 4 steps.



Failure is not not an Option!

In small-world experiments, most chains fail!

What do we conclude?

- It's a big world after all?

Failure is not not an Option!

In small-world experiments, most chains fail!

What do we conclude?

- It's a big world after all? (But it isn't.)
- It's a bit messier than I've admitted so far. (It is.)

Failure is not not an Option!

In small-world experiments, most chains fail!

What do we conclude?

- It's a big world after all? (But it isn't.)
- It's a bit messier than I've admitted so far. (It is.)
  - some friendships form because of geographic proximity.
  - some form because of occupational proximity.
  - but some form because you sit next to someone interesting on a flight to YYC.

Failure is not not an Option!

In small-world experiments, most chains fail!

What do we conclude?

- It's a big world after all? (But it isn't.)
- It's a bit messier than I've admitted so far. (It is.)
  - some friendships form because of geographic proximity.
  - some form because of occupational proximity.
  - but some form because you sit next to someone interesting on a flight to YYC.

  - some 'systematic' friendships, some 'random.'

Failure is not not an Option!

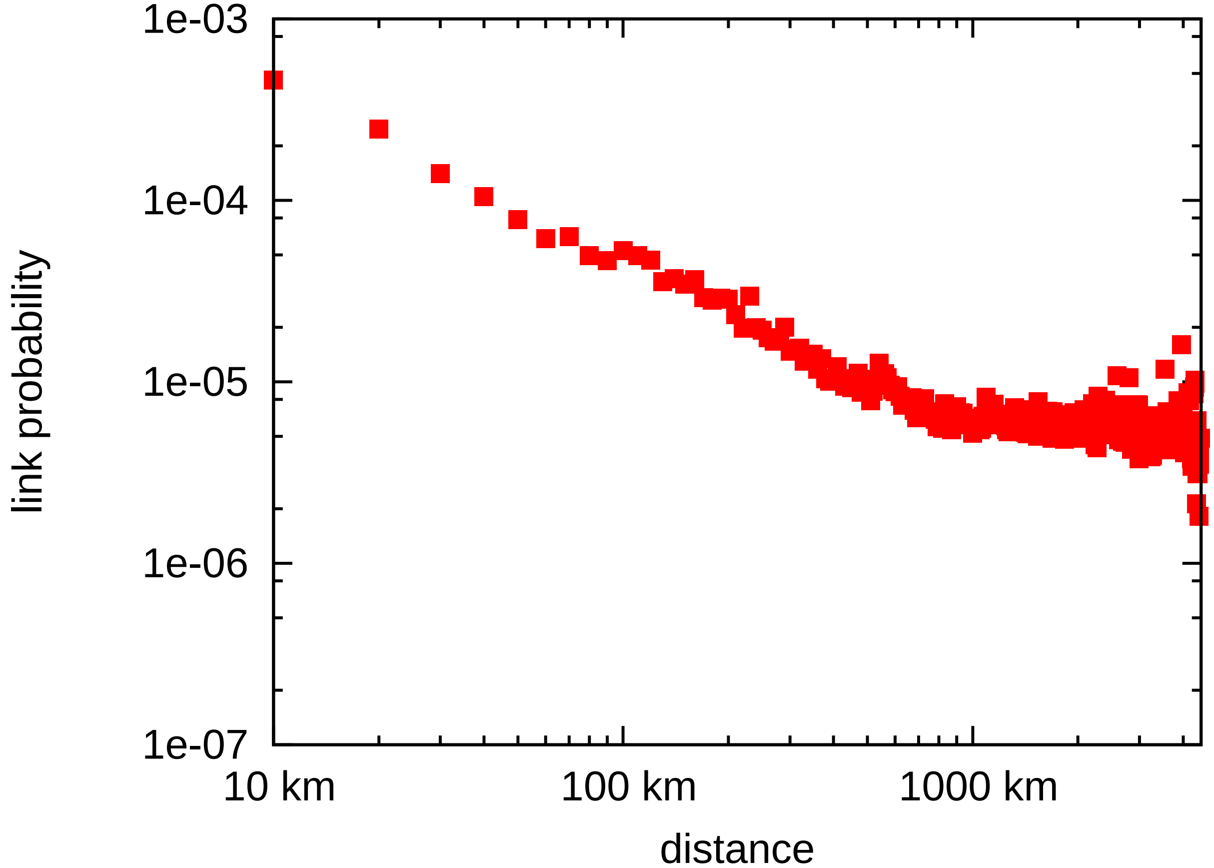
In small-world experiments, most chains fail!

What do we conclude?

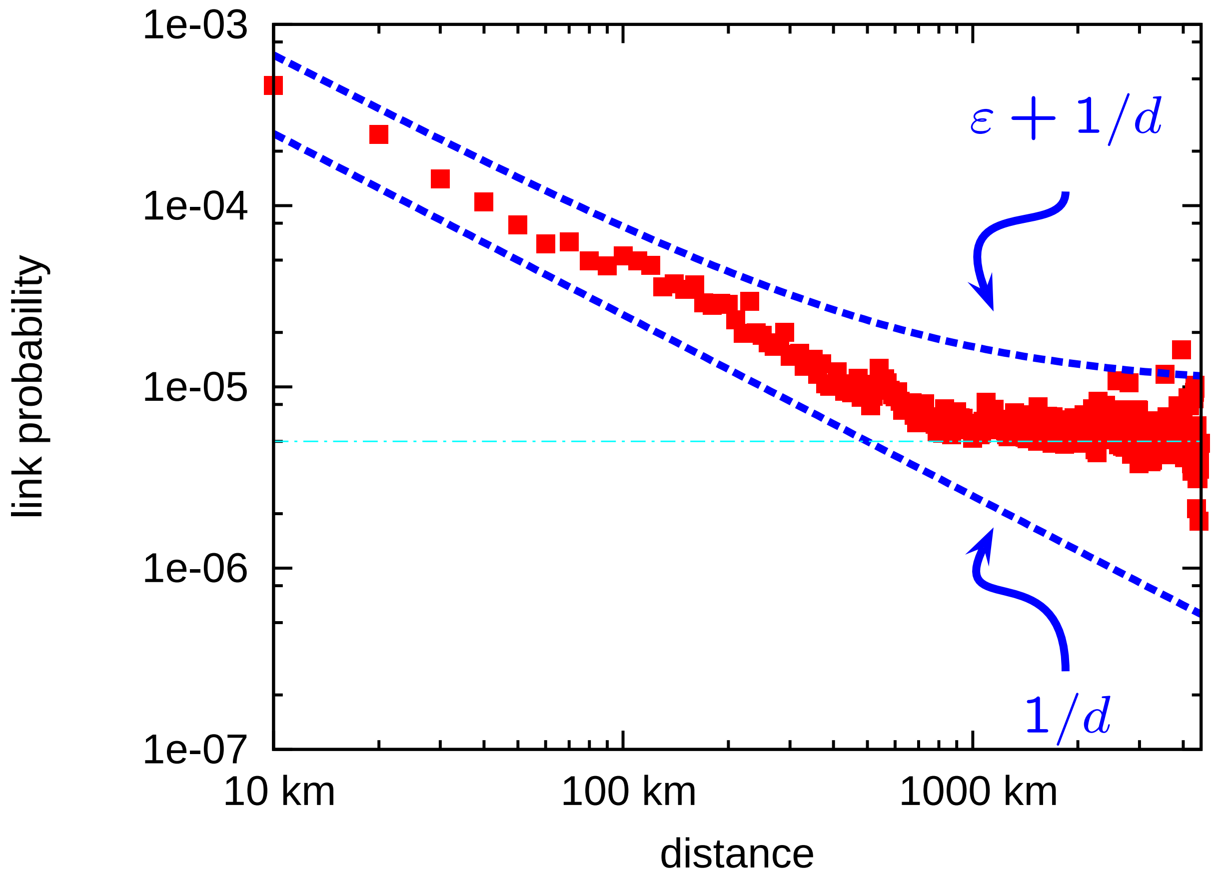
- It's a big world after all? (But it isn't.)
- It's a bit messier than I've admitted so far. (It is.)
  - some friendships form because of geographic proximity.
  - some form because of occupational proximity.
  - but some form because you sit next to someone interesting on a flight to YYC.

  - some 'systematic' friendships, some 'random.'
  - how much of each?

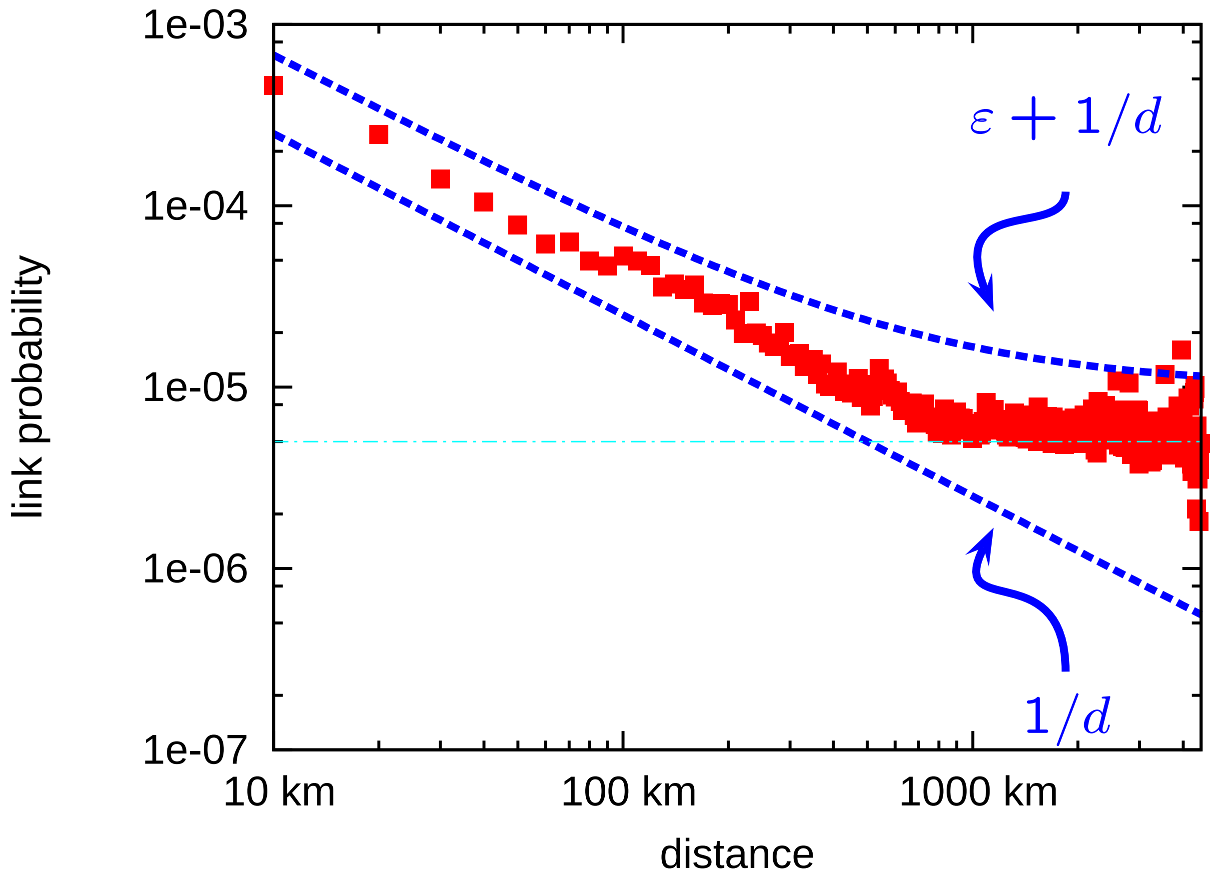
Distance versus link probability



Distance versus link probability

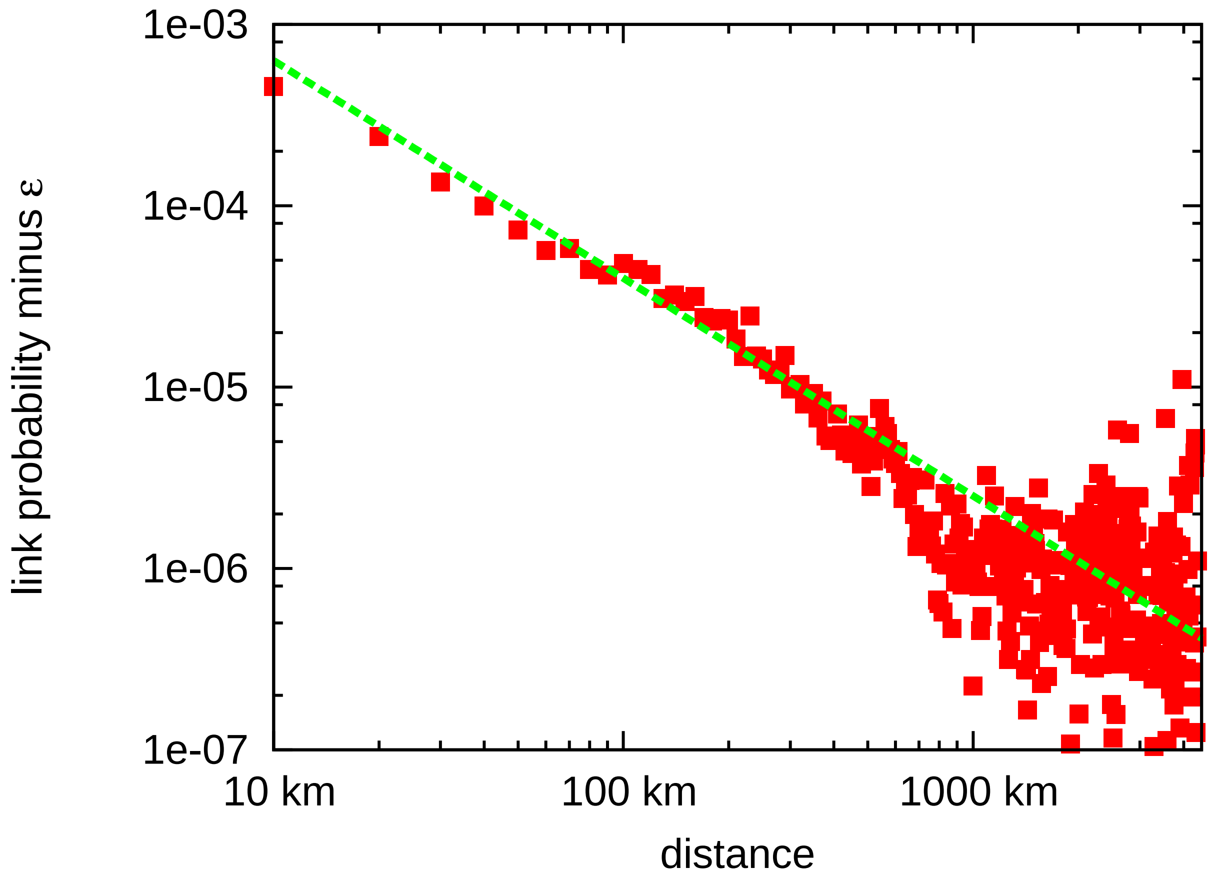


Distance versus link probability



- Not linear: link probability levels out to $\sim 5 \times 10^{-6}$.

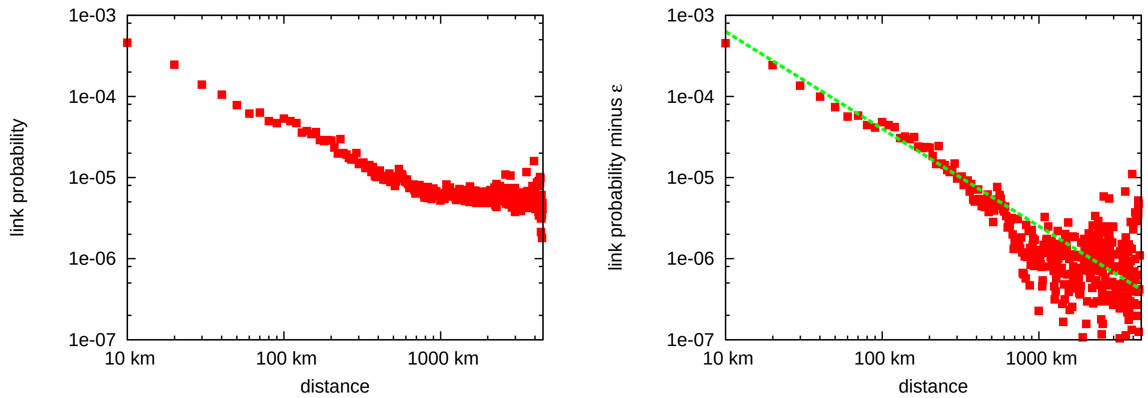
Distance versus link probability



Not linear: link probability levels out to $\sim 5 \times 10^{-6}$.

Geographic/Nongeographic Friendships

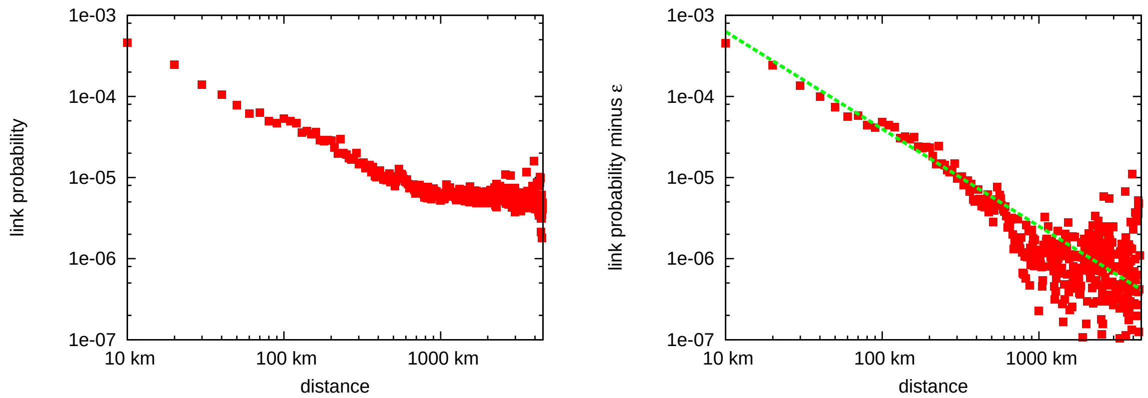
[DLN Novak Kumar Raghavan Tomkins 2005]



- good estimate of friendship probability:
  $\mathtt{Pr}[u \to v] \approx \varepsilon + f(d(u,v))$ for $\varepsilon \approx 5.0 \times 10^{-6}$.

Geographic/Nongeographic Friendships

[DLN Novak Kumar Raghavan Tomkins 2005]



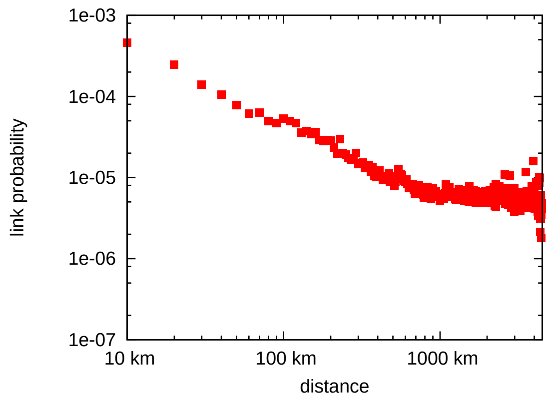
good estimate of friendship probability:

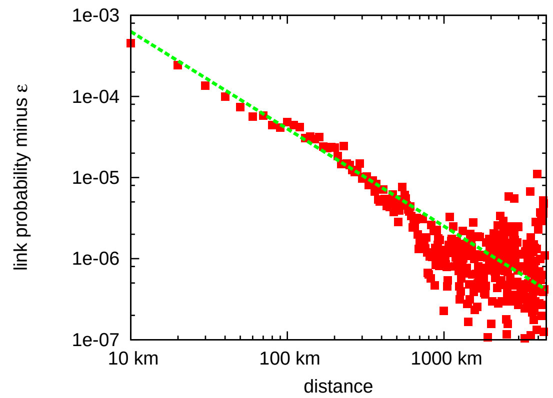
$$\texttt{Pr}[u \to v] \approx \varepsilon + f(d(u,v)) \text{ for } \varepsilon \approx 5.0 \times 10^{-6}.$$

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

Geographic/Nongeographic Friendships

[DLN Novak Kumar Raghavan Tomkins 2005]





- good estimate of friendship probability:

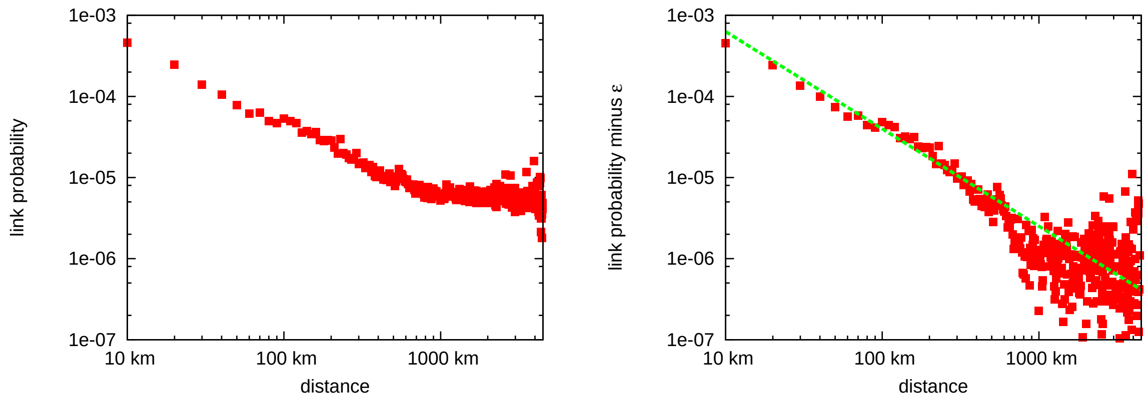
$$\texttt{Pr}[u \longrightarrow v] \approx \varepsilon + f(d(u,v)) \text{ for } \varepsilon \approx 5.0 \times 10^{-6}.$$

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

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

Geographic/Nongeographic Friendships

[DLN Novak Kumar Raghavan Tomkins 2005]



- good estimate of friendship probability:

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

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

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

$\sim 5.5/8 \approx 66\%$ of LJ friendships are "geographic," 33% are not.

Evolution of Social Networks

Implicitly, this is a model of the evolution of social networks.

I decide to make a new friend u. *How do I pick u?*

➡ I choose u according to (rank-based) geography.

With probability $1/3$, I choose u uniformly at random.

With probability $2/3$, I choose u according to (rank-based) geography.

Evolution of Social Networks

Implicitly, this is a model of the evolution of social networks.
I decide to make a new friend u. *How do I pick u?*

- I choose u according to (rank-based) geography.

  With probability 1/3, I choose u uniformly at random.

  With probability 2/3, I choose u according to (rank-based) geography.

- OR: I choose u according to occupational proximity.

Evolution of Social Networks

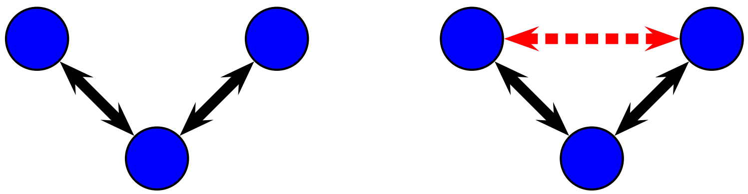
Implicitly, this is a model of the evolution of social networks.
I decide to make a new friend u. *How do I pick u?*

- I choose u according to (rank-based) geography.
  With probability $1/3$, I choose u uniformly at random.
  With probability $2/3$, I choose u according to (rank-based) geography.

- OR: I choose u according to occupational proximity.

- OR:

Evolution through Common Friends



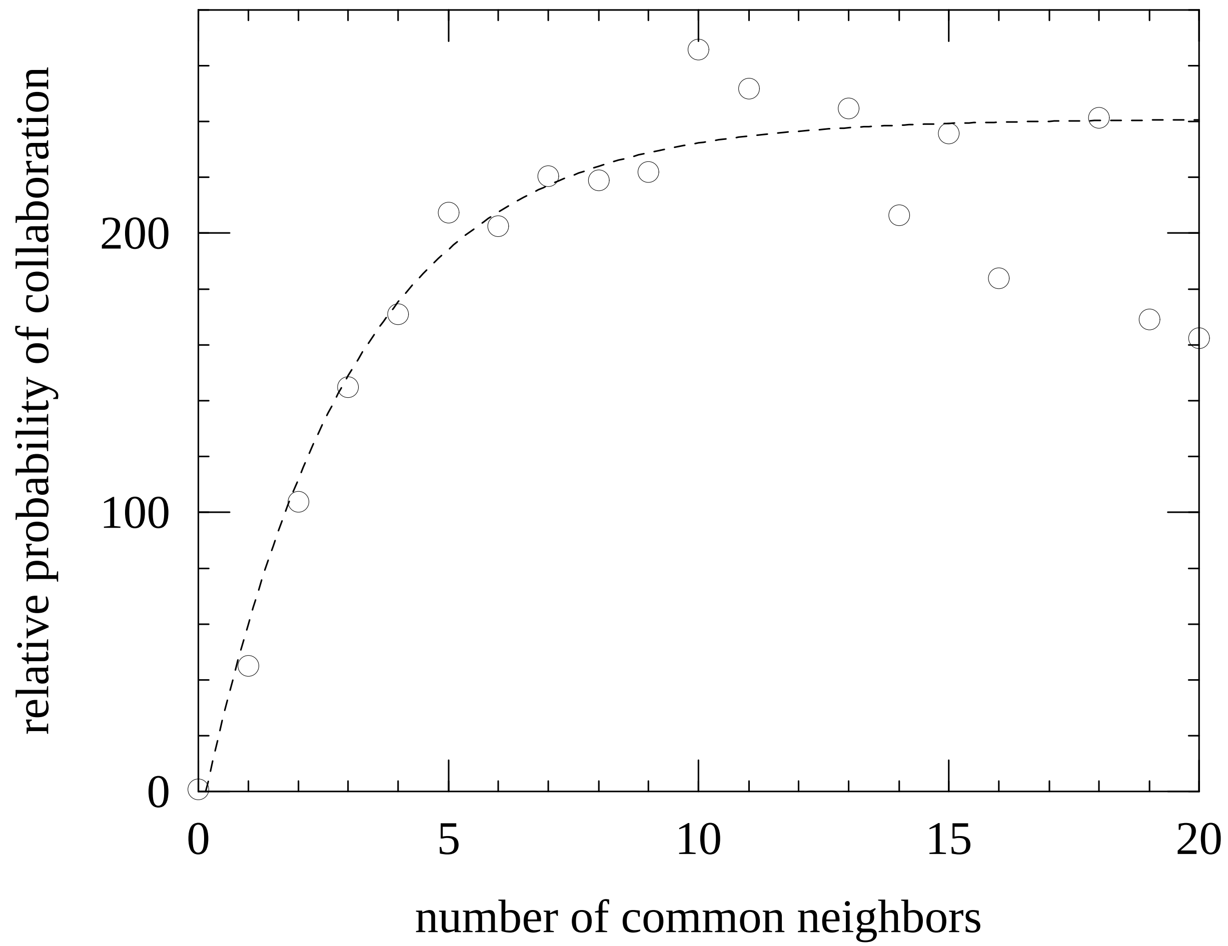
If you and I both know Will Shortz,
then maybe he'll introduce us to each other.

"closing a triangle"

- A (direct) explanation for high clustering coefficients.

Statistics on Triangle Closing

By what factor does Pr*[friendship] increase if* $\exists$ *common friends?*



Collaboration network among physicists. [Newman 2001]

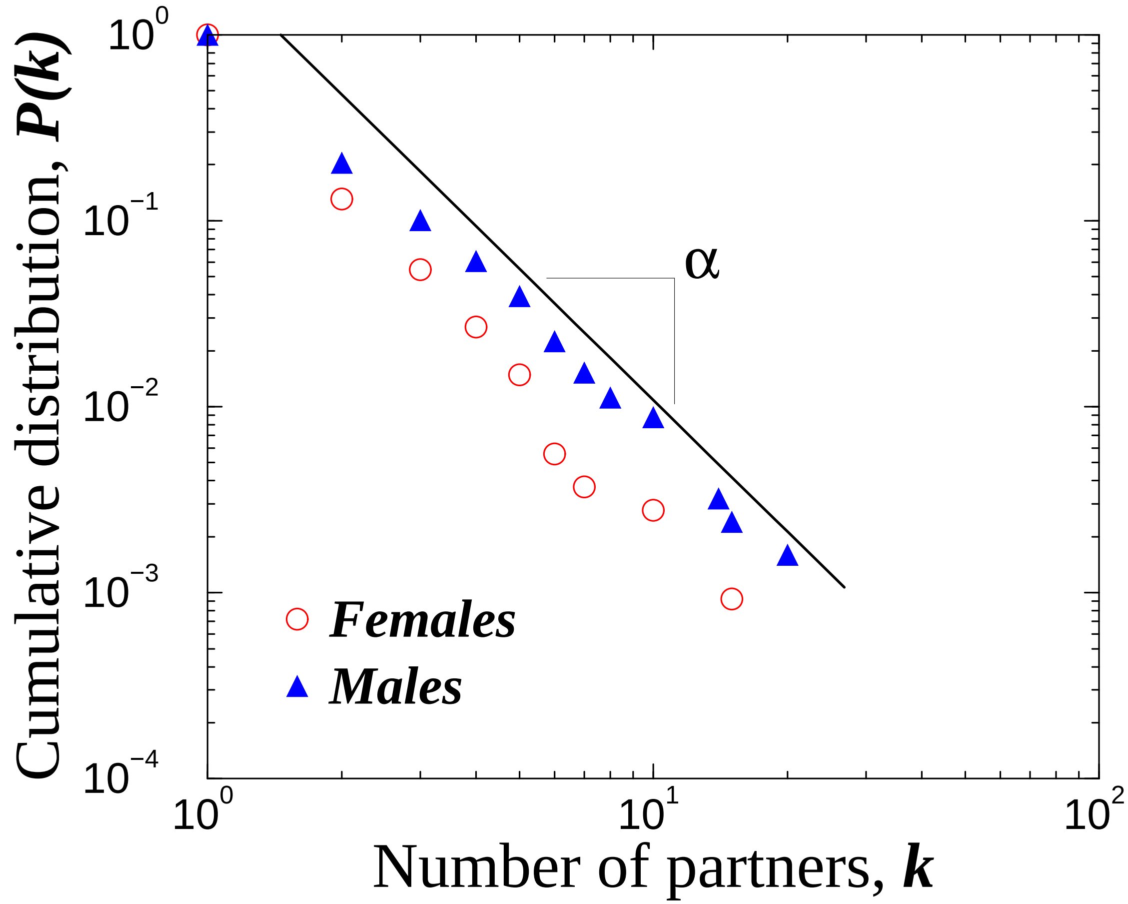
Evolution of Social Networks

Implicitly, this is a model of the evolution of social networks.
I decide to make a new friend u. *How do I pick u?*

- I choose u according to (rank-based) geography.
  With probability $1/3$, I choose u uniformly at random.
  With probability $2/3$, I choose u according to (rank-based) geography.
- OR: I choose u according to occupational proximity.
- OR: I choose a u with whom I share a mutual friend.
- OR:

Networks of Swedish Sexual Contacts

[Liljeros et al 2001]



Number of sexual contacts within last year (Swedish survey).

$\alpha \approx 2.3$ to 2.5

Preferential Attachment

➡ Power-law degree distribution.

- proportion of people with $\geq k$ friends proportional to $k^{-\alpha}$.
- $\alpha \in [2, 2.5]$ is a reasonably good model for social networks.
- (Or is it? [Mitzenmacher 2001] esp. Mandelbrot vs. Simon)

Preferential Attachment

Power-law degree distribution.

- proportion of people with $\geq k$ friends proportional to $k^{-\alpha}$.
- $\alpha \in [2, 2.5]$ is a reasonably good model for social networks.
- (Or is it? [Mitzenmacher 2001] esp. Mandelbrot vs. Simon)

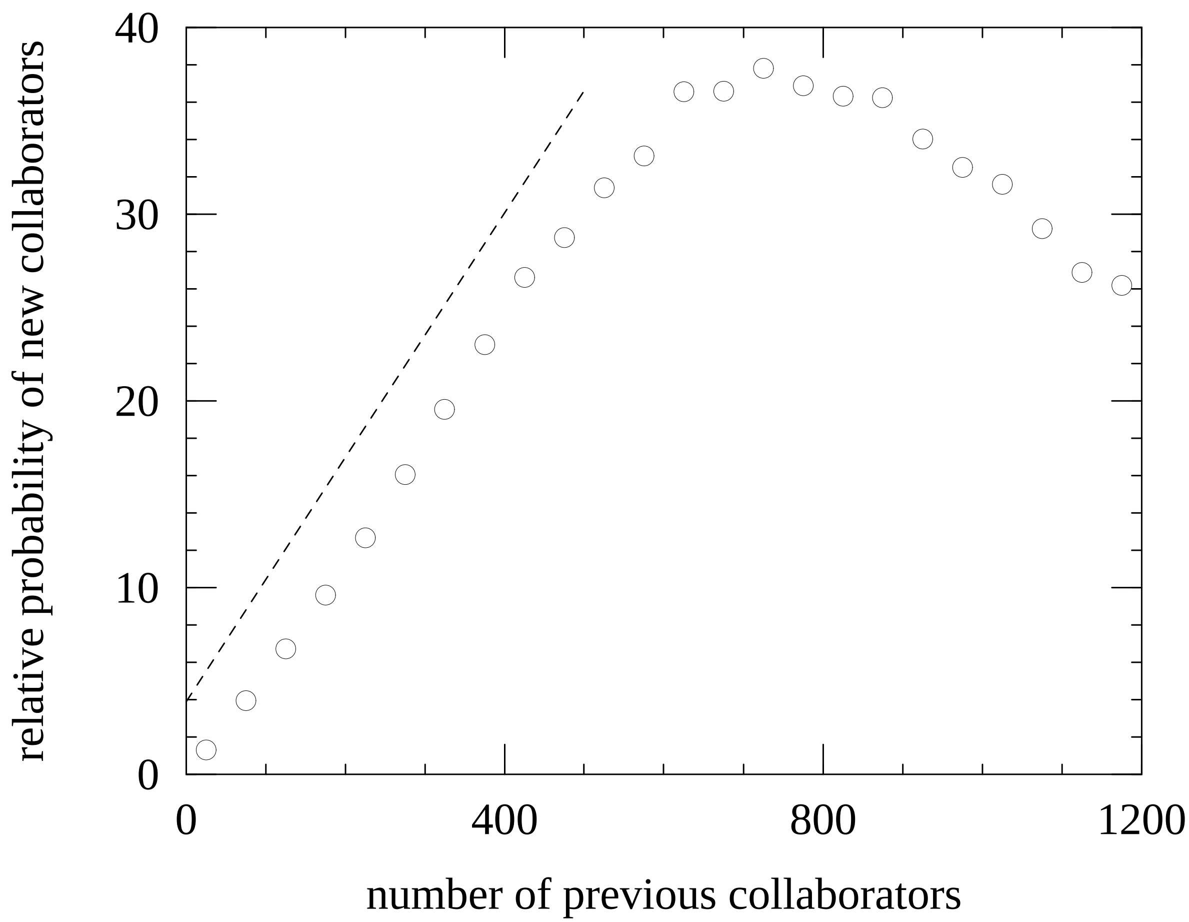
A model generating power-law networks:

$$\texttt{Pr}[u \text{ befriends } x] \propto (\text{number of friends that } x \text{ already has})$$

"preferential attachment" [Barabasi Albert 1999]

Statistics on Preferential Attachment

Do the rich really get richer?
Collaboration network in biology/medicine. [Newman 2001]



Evolution of Social Networks

Implicitly, this is a model of the evolution of social networks.
I decide to make a new friend u. *How do I pick u?*

- I choose u according to (rank-based) geography.
  With probability $1/3$, I choose u uniformly at random.
  With probability $2/3$, I choose u according to (rank-based) geography.

- OR: I choose u according to occupational proximity.

- OR: I choose a u with whom I share a mutual friend.

- OR: I choose u by preferential attachment ($\Pr[u] \propto \mathrm{degree}(u)$).

Evolution of Social Networks

Implicitly, this is a model of the evolution of social networks.
I decide to make a new friend u. *How do I pick u?*

- I choose u according to (rank-based) geography.
  With probability $1/3$, I choose u uniformly at random.
  With probability $2/3$, I choose u according to (rank-based) geography.

- OR: I choose u according to occupational proximity.

- OR: I choose a u with whom I share a mutual friend.

- OR: I choose u by preferential attachment ($\Pr[u] \propto \text{degree}(u)$).

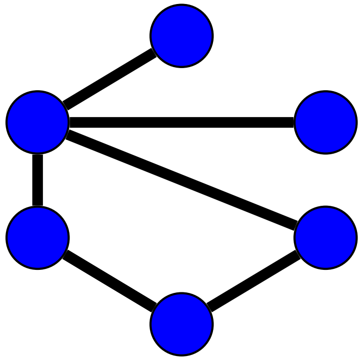
What's a principled way to evaluate these models?

The Link-Prediction Problem

[DLN Kleinberg 2003]

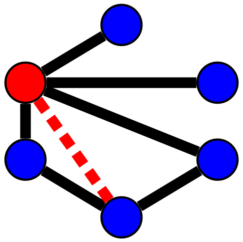
The link-prediction problem:

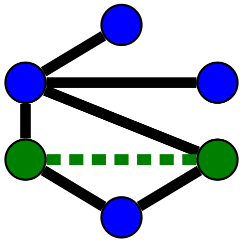
predict the next edges that will appear in the social network.



You are given a social network.

What's the next friendship that will occur?





The Link-Prediction Problem Formalized

The Link-Prediction Problem:

Given: all social network edges during training interval.

Output: predictions of all **new** interacting pairs during subsequent test interval.

How much information does the network contain about its own future?

Link Prediction and Network Growth

➡ A principled way to evaluate network growth models.

- Many models of the growth of networks.
  (preferential attachment, ...)

- Typical evaluation:
  "does this model produce the correct
  power-law degree distribution?"
  (or diameter, or clustering coefficient, or ...)

- Link prediction:
  **A network-growth model is good
  if it accurately predicts network growth.**

Related Link-Prediction Formulations

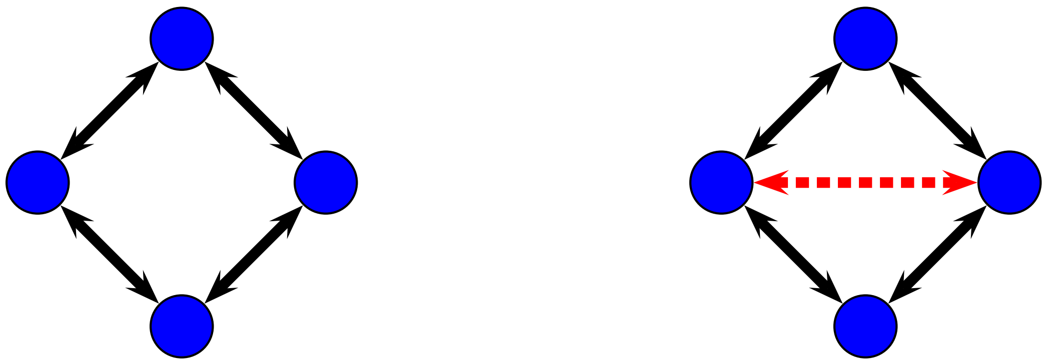
- Our approach: how does the social network evolve?
  (As opposed to inferring hidden links in a static network.)

- Our approach: predict *new* interactions only.
  (What new information can we extract?)

- Our approach: purely graph-based.
  (Why? Interested in information content of network itself!)

Common-Neighbors Predictor

The more common friends we have,
then the more likely we are to become friends ...



Common Neighbors predictor:
predict pairs with largest number of common neighbors.

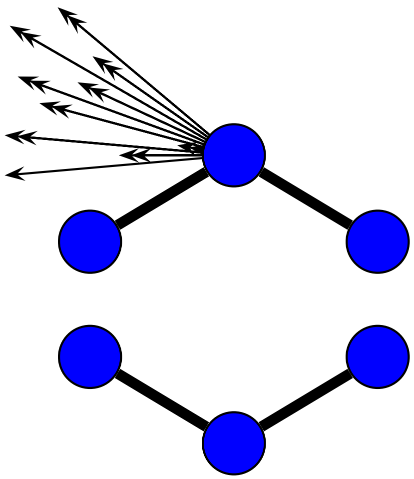
Preferential-Attachment Predictor

The more friends I have,
then the more likely I am to make new friends ...

Preferential Attachment predictor:
predict pairs x and y with largest $\mathsf{degree}(x) \cdot \mathsf{degree}(y)$.

Adamic/Adar Predictor

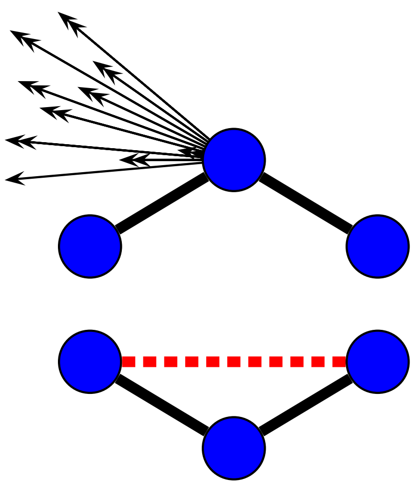
The more (un?)popular common friends we have, then the more likely we are to become friends ...



Adamic/Adar Predictor

The more unpopular common friends we have, then the more likely we are to become friends ...

"rare features are more telling"



$$\text{score}_{x,y} := \sum_{z \in \text{CN}(x,y)} \frac{1}{\log \text{degree}(z)}$$

$\text{CN}(x,y) =$ common neighbors of x, y.

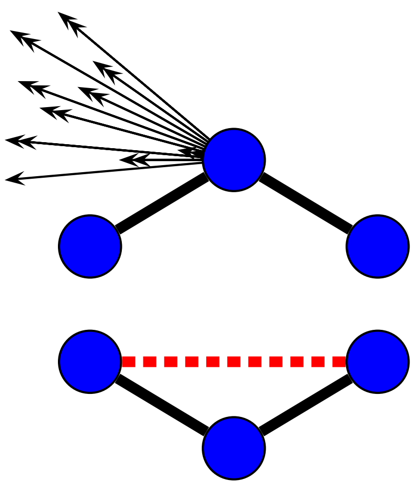
[Adamic Adar 2003]

for measuring "relatedness"
of personal home pages.

Adamic/Adar Predictor

The more unpopular common friends we have,
then the more likely we are to become friends ...

"rare features are more telling"



$$\mathrm{score}_{x,y} := \sum_{z \in \mathrm{CN}(x,y)} \frac{1}{\log \mathrm{degree}(z)}$$

$\mathrm{CN}(x,y) =$ common neighbors of x, y.

[Adamic Adar 2003]

for measuring "relatedness"
of personal home pages.

Adamic/Adar (frequency-weighted common neighbors):
predict pairs with largest score.

Katz Predictor

> If there are many short chains of friends connecting us, then we are more likely to become friends.

$$\boxed{\mathsf{score}_{x,y} := \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot (\text{number of } x \leftrightarrow y \text{ paths of length } \ell)}$$

(for a parameter $\beta > 0$.)

Katz predictor:

predict new edges between nodes with largest score.

Katz Predictor

> If there are many short chains of friends connecting us, then we are more likely to become friends.

$$\boxed{\mathsf{score}_{x,y} := \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot (\text{number of } x \leftrightarrow y \text{ paths of length } \ell)}$$

(for a parameter $\beta > 0$.)

Originally measured "social standing" in a network [Katz 1953]:

$$\mathsf{standing}(x) = \sum_{y} \mathsf{score}_{x,y}.$$

Katz predictor:

predict new edges between nodes with largest score.

Link-Prediction Experiments

- For our experiments, we use collaboration networks:
  - moderately large-scale, time-resolved, $\approx$social network.
  - x, y linked $\iff$ x, y coauthor an academic paper
  - 5 subfields of physics, from `www.arxiv.org`.

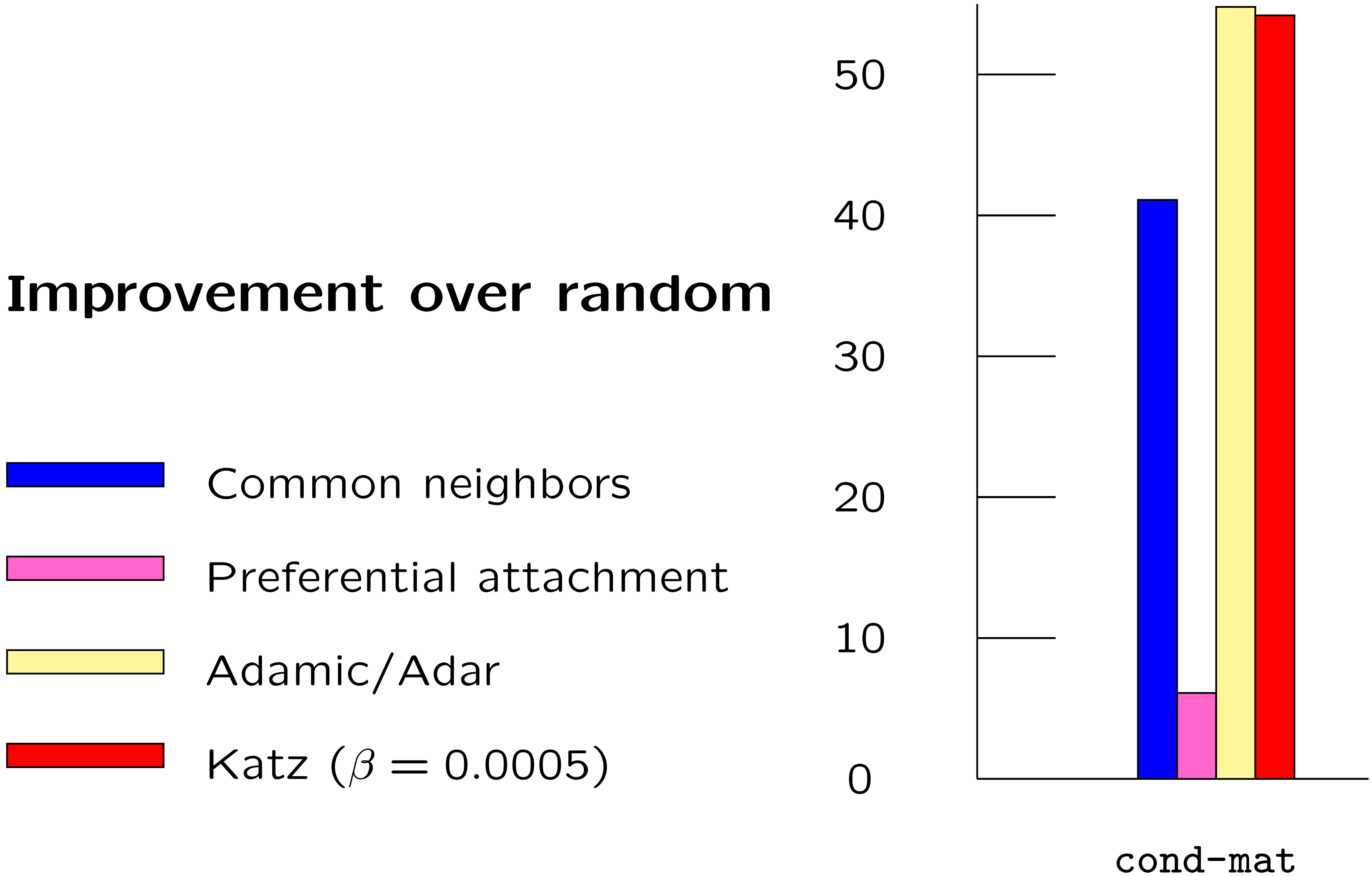
Link-Prediction Experiments

- For our experiments, we use collaboration networks:
  - moderately large-scale, time-resolved, ≈social network.
  - x, y linked $\iff$ x, y coauthor an academic paper
  - 5 subfields of physics, from `www.arxiv.org`.

- Measure each predictor using performance versus random:
  - "How much better is this predictor than randomly guessing new edges?"
  - "How much information did we extract from the network?"

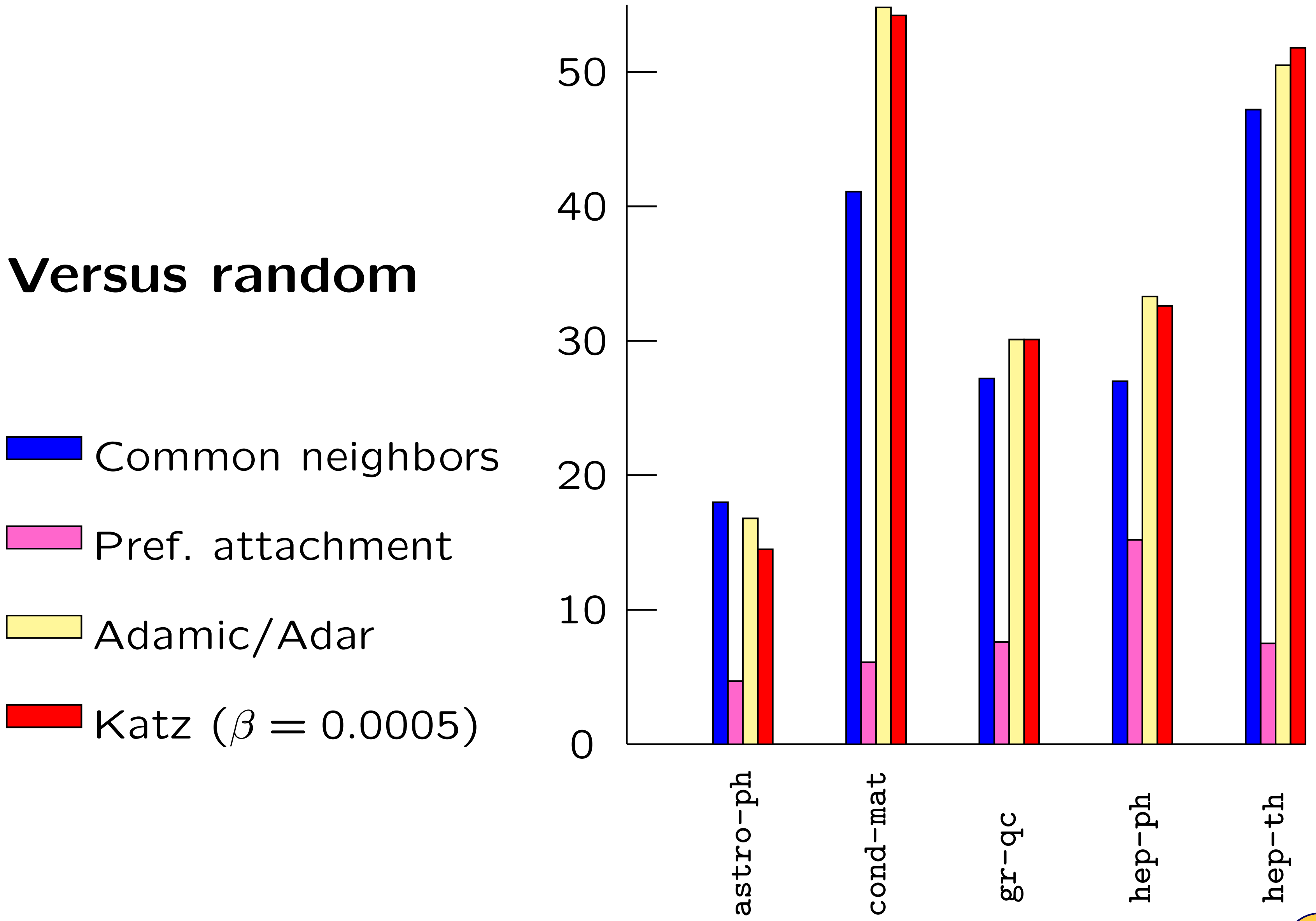
Results on `cond-mat`

Improvement over random



Results on the arXiv

Versus random



Small-World Problem

- Social networks form a "small world."
  - normally seen as crucial for scientific progress.

    cross-disciplinary research.
    quick dissemination of new ideas.

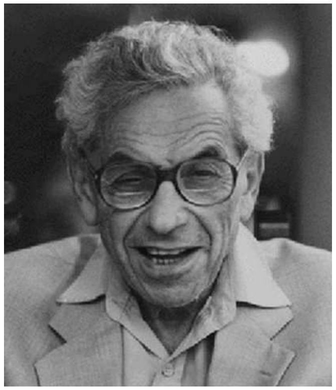
  - here: a major problem!

    short (but tenuous) paths between
    people who are "really" far apart.

graph-distance predictor beats random ...
... but it is not especially good!

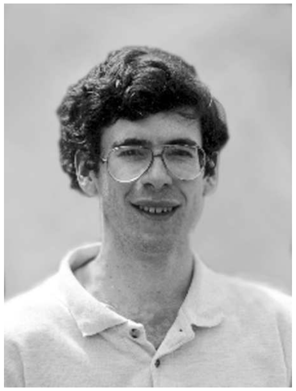
Small-World Problem: Erdős Numbers

- Your Erdős number: graph distance from you to Paul Erdős.



Paul Erdős

Erdős Number 0

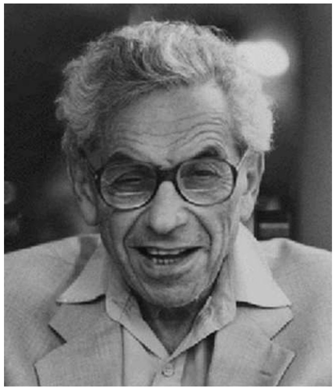


JMK

Erdős Number 3

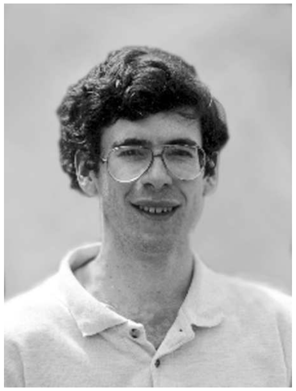
Small-World Problem: Erdős Numbers

- Your Erdős number: graph distance from you to Paul Erdős.



Paul Erdős

Erdős Number 0



JMK

Erdős Number 3

Jean Piaget

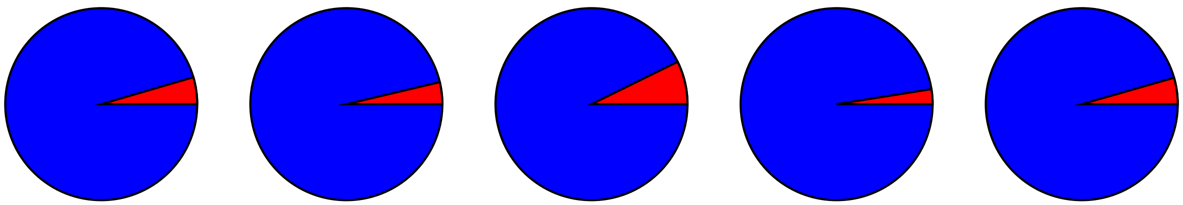
Erdős Number 3

- good predictors:

proximity measures robust to spurious connections.

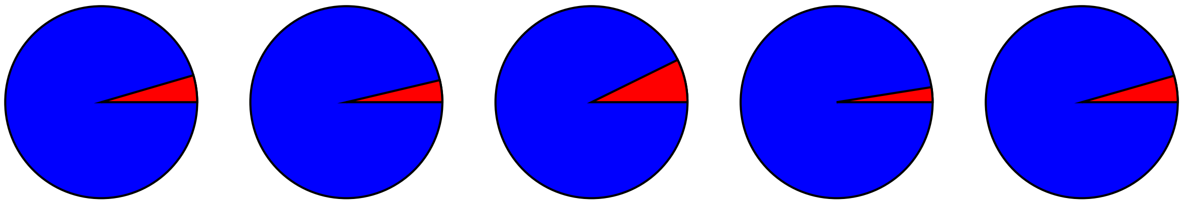
New Edges Forming at Distance 2

Proportion of distance-2 pairs that form an edge:

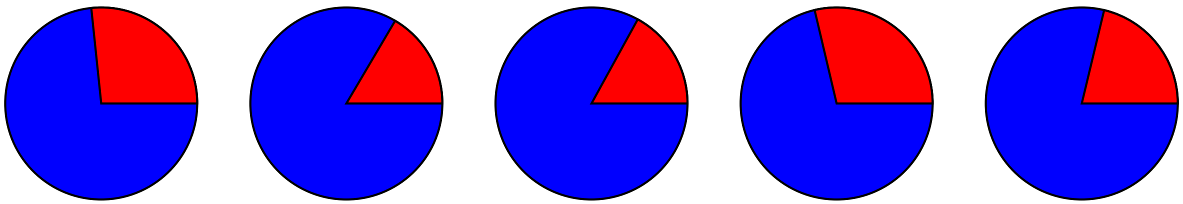


New Edges Forming at Distance 2

Proportion of distance-2 pairs that form an edge:



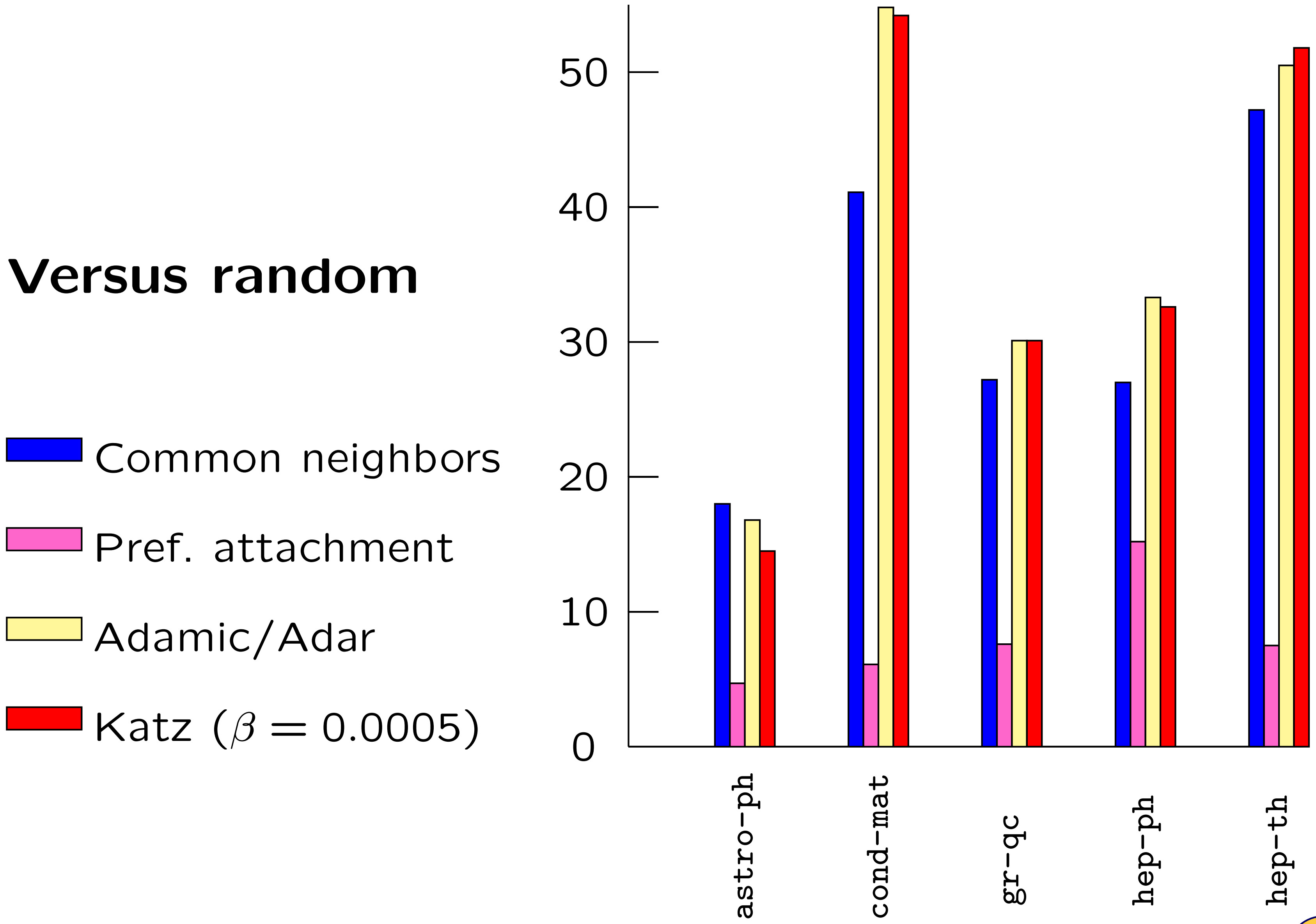
Proportion of new edges between distance-2 pairs:



Most new edges are at distance 3 or more!

Results on the arXiv

Versus random



Social Networks versus Social Networks?

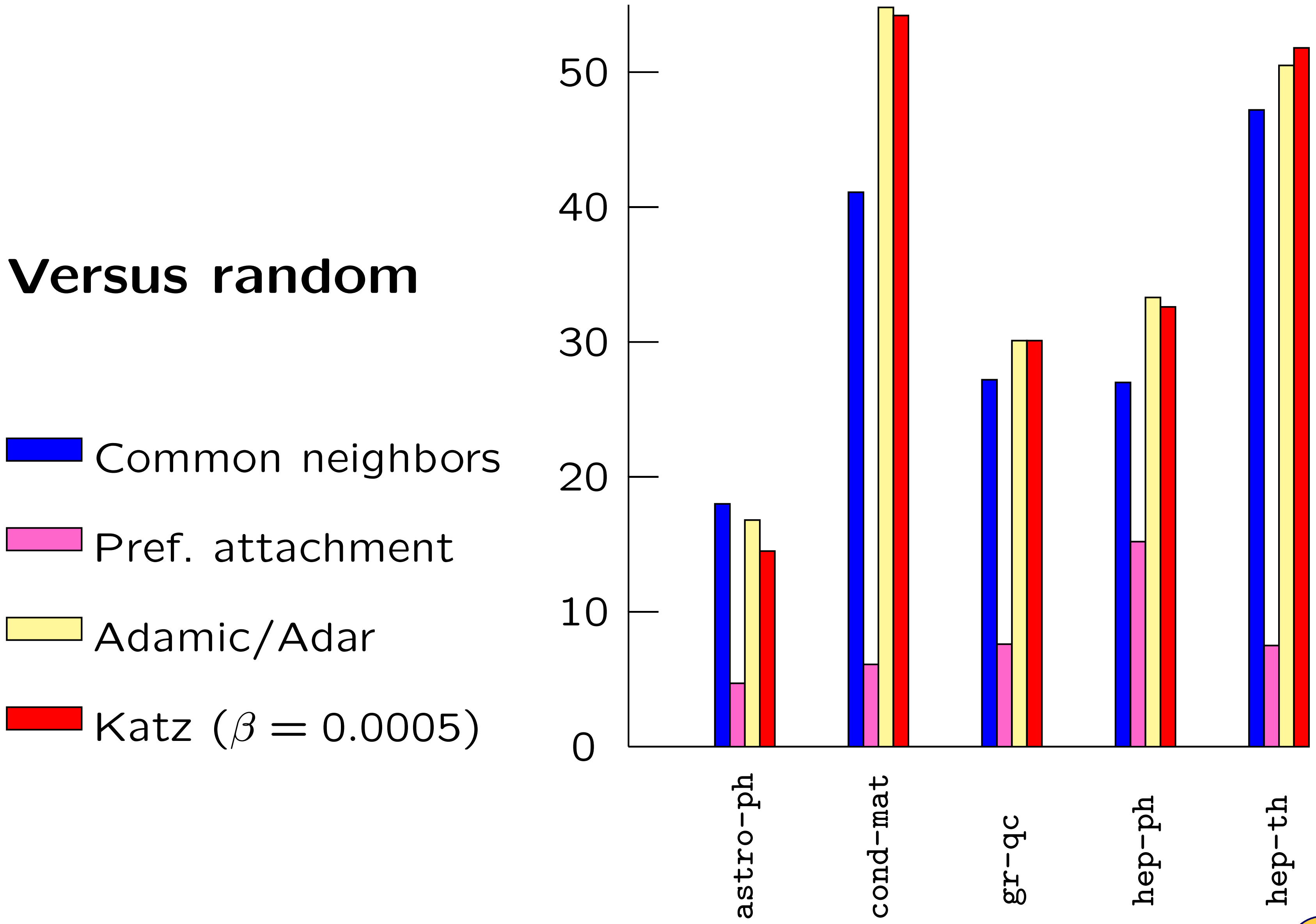
Social networks are not all equal!

- `astro-ph` is 'unsystematic.'

  all predictors do worse than on other datasets.

Results on the arXiv

Versus random



Low-Rank Approximations

Observation: the graph is noisy.

Want to remove "tenuous" edges.

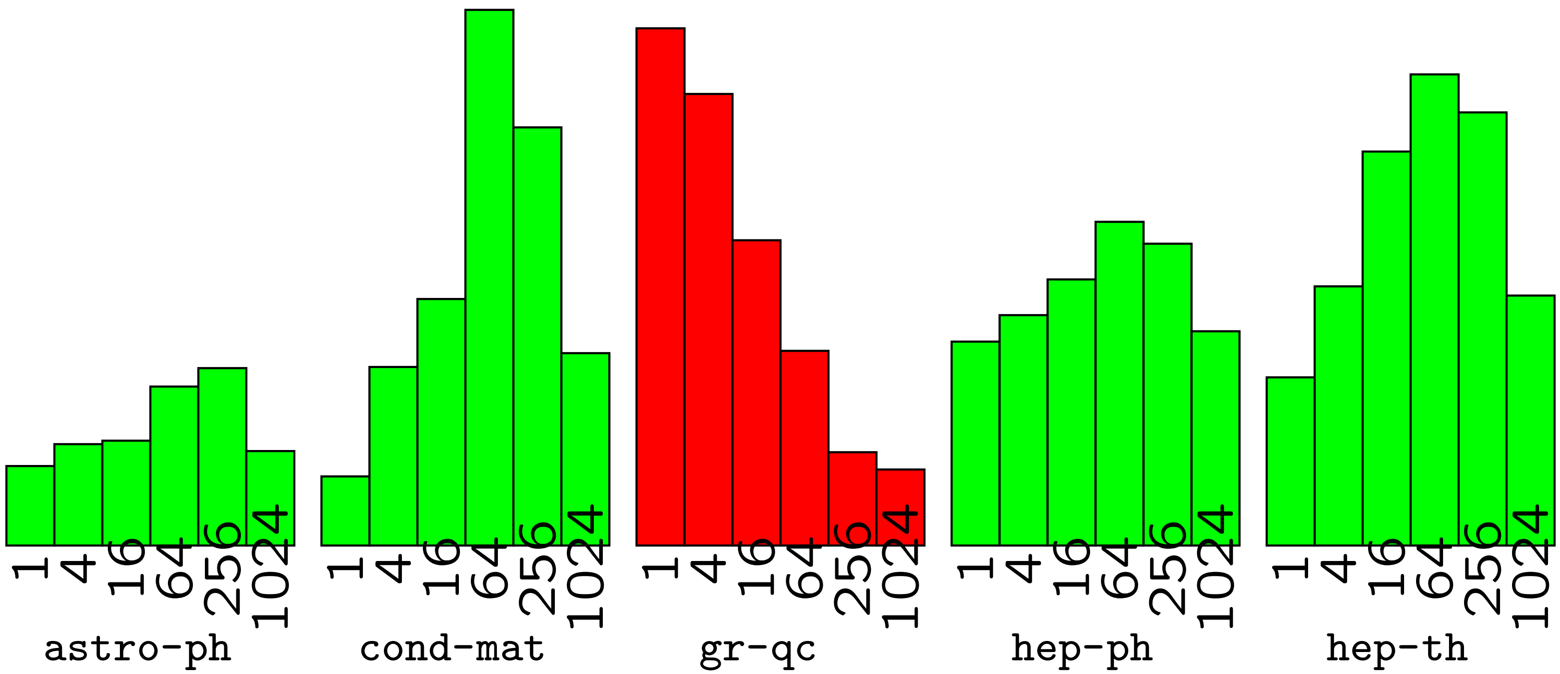
Look at the adjacency matrix M for the training interval.

One approach: approximate M by a simpler matrix.

- Compute a low-rank approximation M^k to M.
  (using Singular Value Decomposition)
- Run some predictor on M^k instead.
- E.g.: $M_{i,j}$ big if i, j collaborated $\Rightarrow$ predict i, j if $M^k_{i,j}$ is big.

Implications of Link Prediction?

Low-rank approximation: Matrix Entry Predictor



- all best at intermediate ranks—except gr-qc!
- do collaborations in general relativity and quantum cosmology have a simpler structure?

Social Networks versus Social Networks?

Social networks are not all equal!

- `astro-ph` is 'unsystematic.'

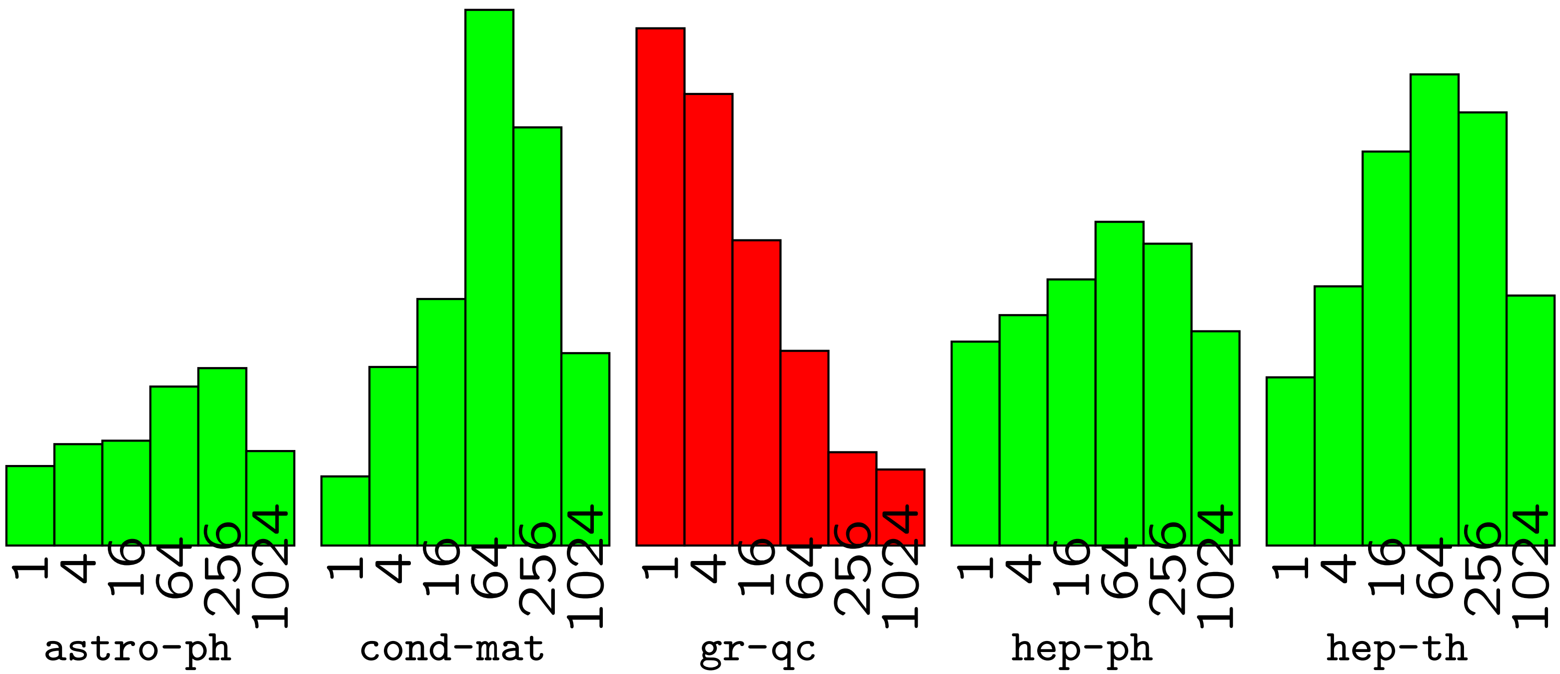
  all predictors do worse than on other datasets.

- `gr-qc` is 'simplistic.'

  low-rank predictors are best at rank$= 1$.

Implications of Link Prediction?

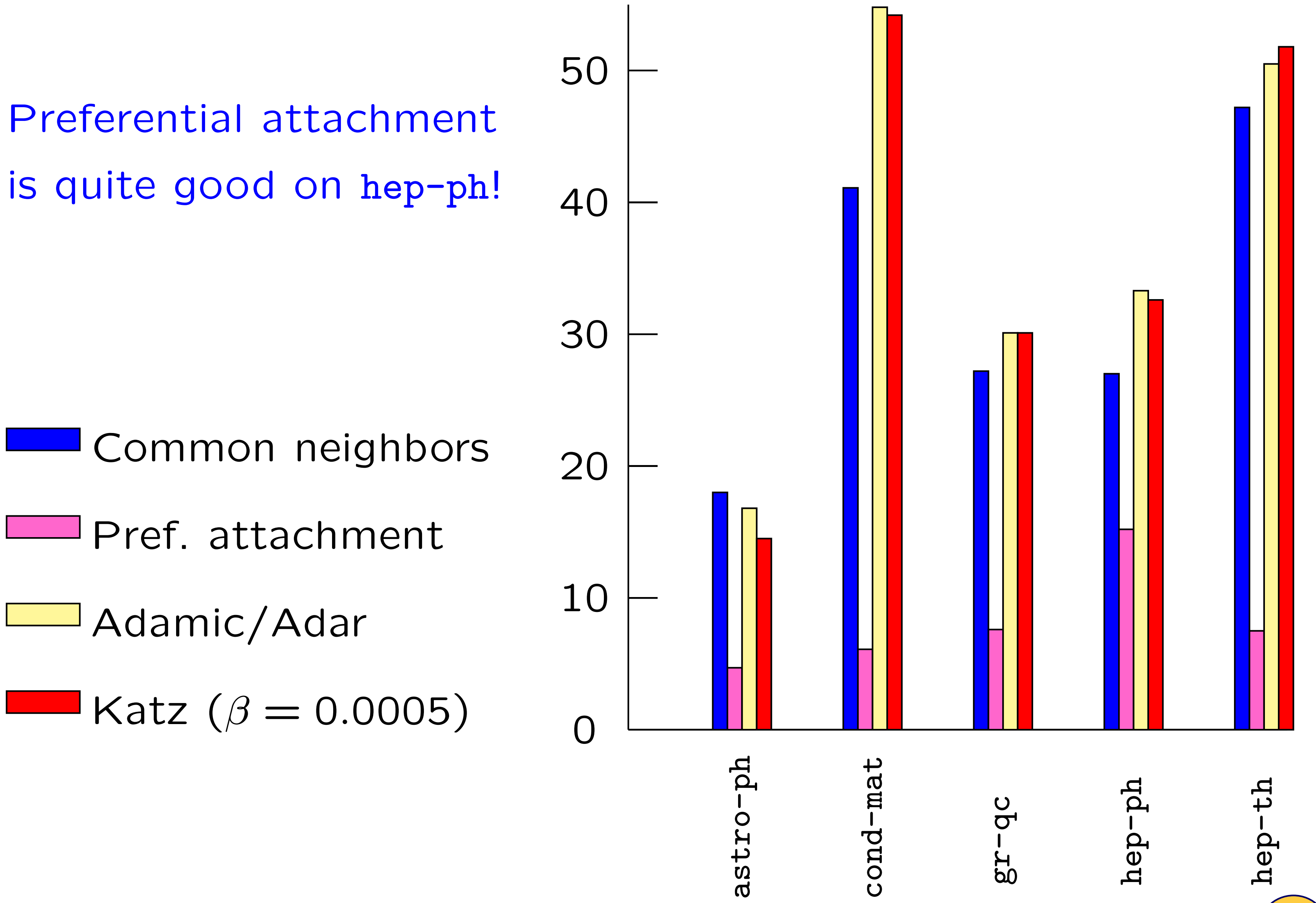
Low-rank approximation: Matrix Entry Predictor



- all best at intermediate ranks—except gr-qc!
- do collaborations in general relativity and quantum cosmology have a simpler structure?

Results on the arXiv

Preferential attachment is quite good on `hep-ph`!



Social Networks versus Social Networks?

Social networks are not all equal!

- `astro-ph` is ‘unsystematic.’

  all predictors do worse than on other datasets.

- `gr-qc` is ‘simplistic.’

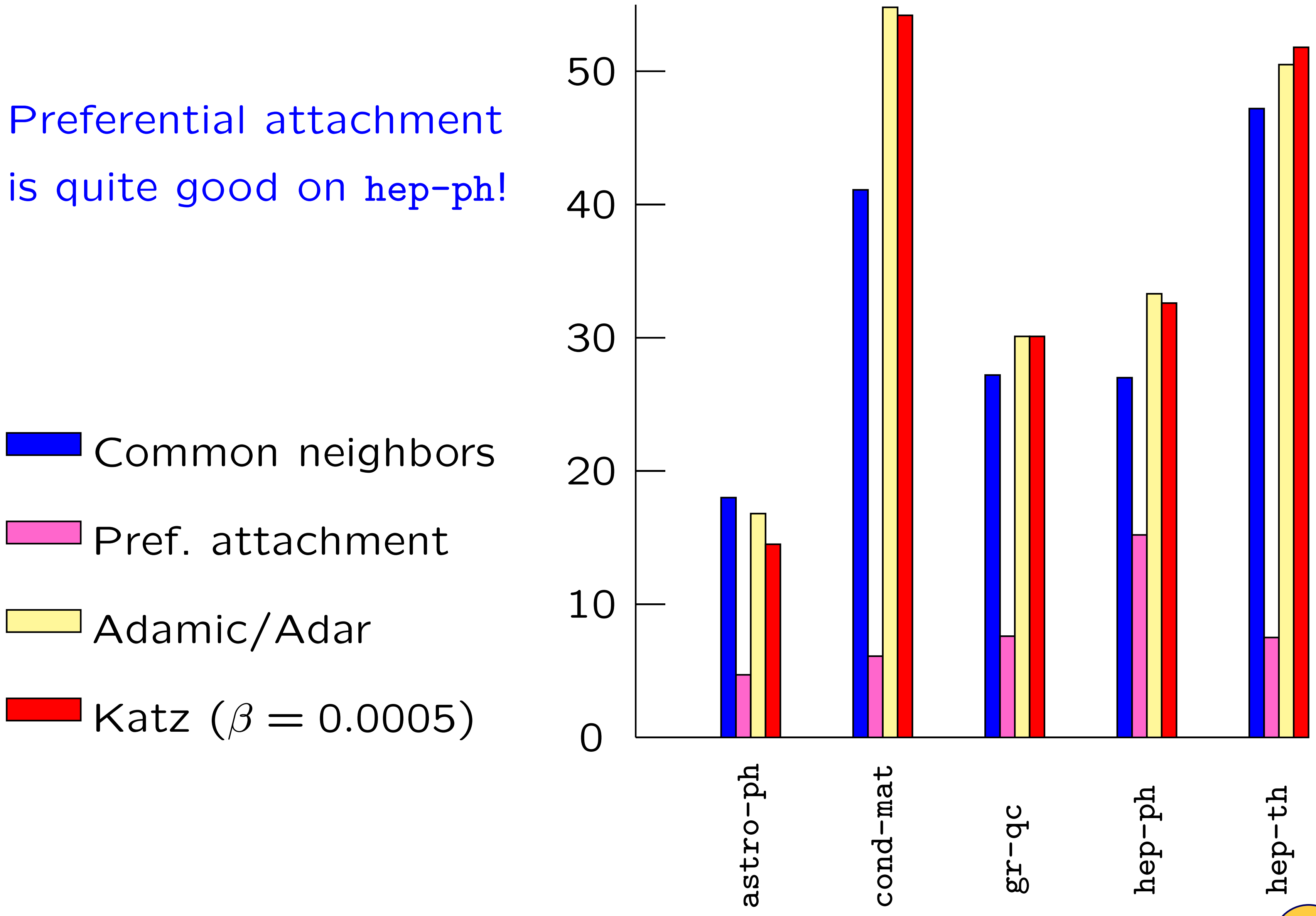
  low-rank predictors are best at rank$=1$.

- `hep-ph` is a ‘popularity contest.’

  preferential attachment is uncharacteristically good.

Results on the arXiv

Preferential attachment is quite good on `hep-ph`!



Conclusions from Link Prediction

... so what does all of this say?

- There a lot of information in social networks!
  - We can predict new links 25–50$\times$ better than random!
  - Friendship probability decays smoothly as similarity drops.

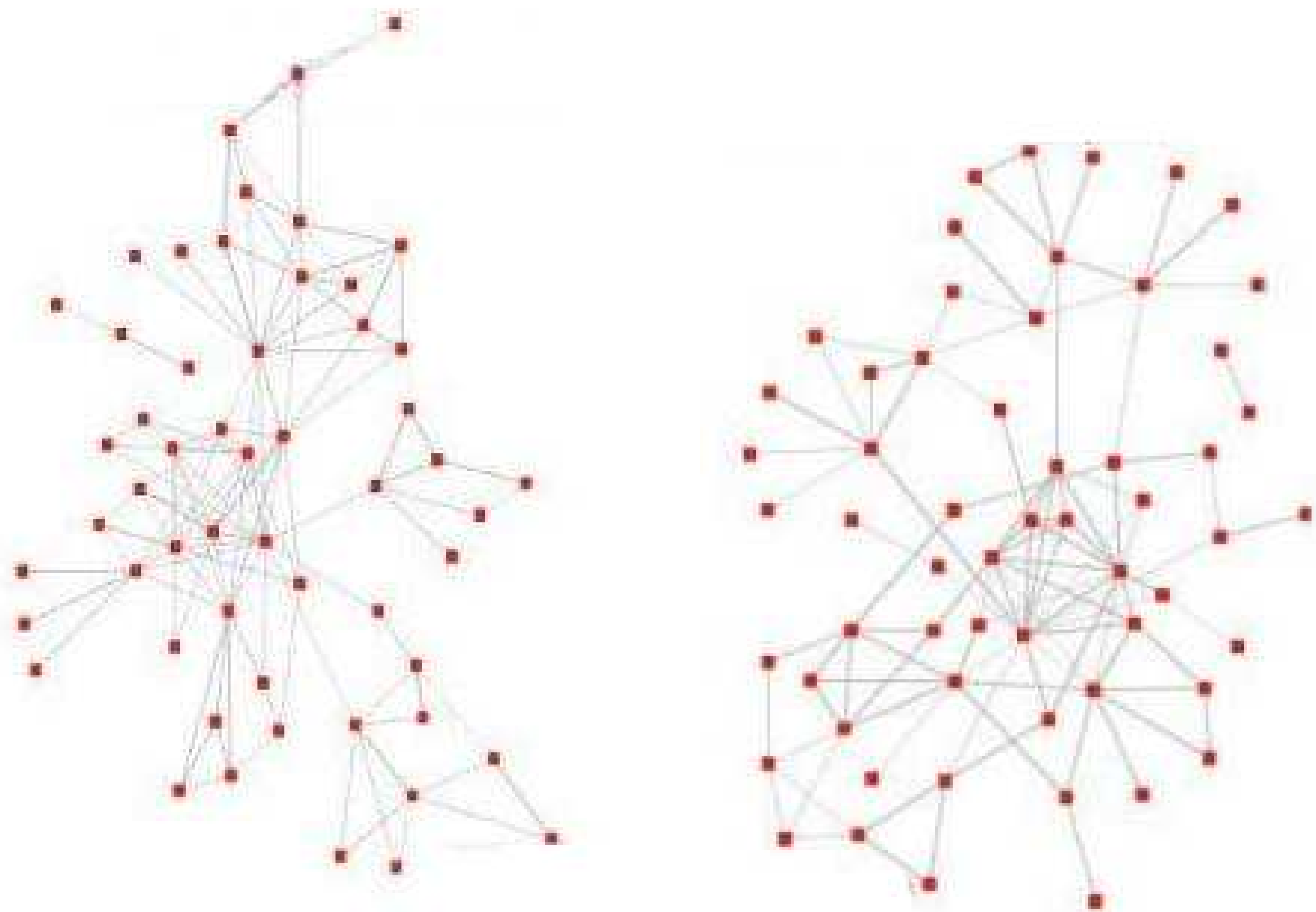
Conclusions from Link Prediction

... so what does all of this say?

- There a lot of information in social networks!
  - We can predict new links 25–50$\times$ better than random!
  - Friendship probability decays smoothly as similarity drops.

- There's not a lot of information in social networks!
  - Best predictors get $>80\%$ of predictions wrong.
  - Only $\sim$5–10% of even the most similar people are friends!

Two Social Networks



One Fortune 500 executives; one terrorist network.
[Krebs 2002]

Social Networks versus Social Networks?

Social networks are not all equal!

- `astro-ph` is ‘unsystematic.’

  all predictors do worse than on other datasets.

- `gr-qc` is ‘simplistic.’

  low-rank predictors are best at rank$=1$.

- `hep-ph` is a ‘popularity contest.’

  preferential attachment is uncharacteristically good.

- Terrorists versus Fortune 500 executives?

  online votes: about 50%/50%.

Conclusions from Link Prediction

... so what does all of this say?

- There a lot of information in social networks!
  - We can predict new links 25–50$\times$ better than random!
  - Friendship probability decays smoothly as similarity drops.
- There's not a lot of information in social networks!
  - Best predictors get $> 80\%$ of predictions wrong.
  - Only $\sim$5–10% of even the most similar people are friends!
- One of the most interesting challenges:

What do the differences between these networks really mean?

Thank you!

David Liben-Nowell

dlibenno@carleton.edu

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

Some references

Milgram 1967
"The Small World Problem." *Psychology Today*.

Kleinberg 2000
"The Small-World Phenomenon: An Algorithmic Perspective." STOC'00.

Dodds Muhamad Watts 2003
"An Experimental Study of Search in Global Social Networks." *Science*.

Liben-Nowell Novak Kumar Raghavan Tomkins 2005
"Geographic Routing in Social Networks." *Proc. Natl. Acad. Sci.*

Kumar Liben-Nowell Tomkins 2006
"Navigating Low-Dimensional and Hierarchical Population Networks." ESA'06.

Adamic Adar 2005
"How to search a social network." *Social Networks*.

Liben-Nowell Kleinberg 2003
"The Link-Prediction Problem for Social Networks." CIKM'03.