It's a small world

but I wouldn't want to paint it

- Stephen Wright

The Small World Effect

Illustrations of the Small World

- Erdős numbers
  - http://www.ams.org/mathscinet/searchauthors.html
- Bacon numbers
  - http://oracleofbacon.org/
- LinkedIn
  - http://www.linkedin.com/
  - Privacy issues: the whole network is *not* visible to all
- Millgram's experiment

Sociometry, Vol. 32, No. 4. (Dec., 1969), pp. 425-443.

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

The City University of New York

Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target

64 of 296 chains succeed, avg chain length is 6.2

THE DOCUMENT. The 296 initial volunteers were sent a document which was the principal tool of the investigation.[3] The document contained:

a. a description of the study, a request that the recipient become a participant, and a set of rules for participation;
b. the name of the target person and selected information concerning him;
c. a roster, to which each participant was asked to affix his name;
d. a stack of fifteen business reply cards asking information about each participant.

Name, hometown, school, dates of military service, …

Rules for Participation. The document contained the following specific instructions to participants:

a. Add your name to the roster so that the next person who receives this folder will know whom it came from.
b. Detach one postcard from the bottom of this folder. Fill it out and return it to Harvard University. No stamp is needed. The postcard is very important. It allows us to keep track of the progress of the folder as it moves toward the target person.
c. If you know the target person on a personal basis, mail this folder directly to him (her). Do this only if you have previously met the target person and know each other on a first name basis.
d. If you do not know the target person on a personal basis, do not try to contact him directly. Instead, mail this folder to a personal acquaintance who is more likely than you to know the target person. You may send the booklet on to a friend, relative, or acquaintance, but it must be someone you know personally.

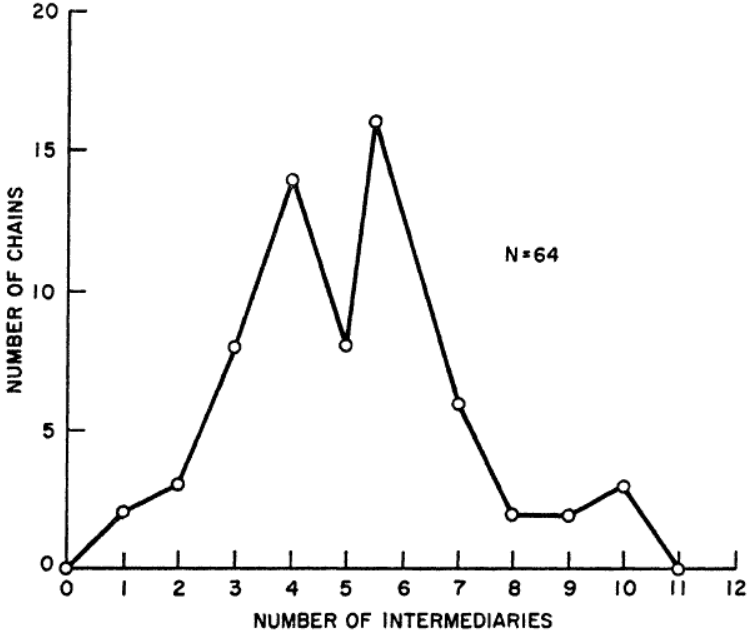


FIGURE 1

Lengths of Completed Chains

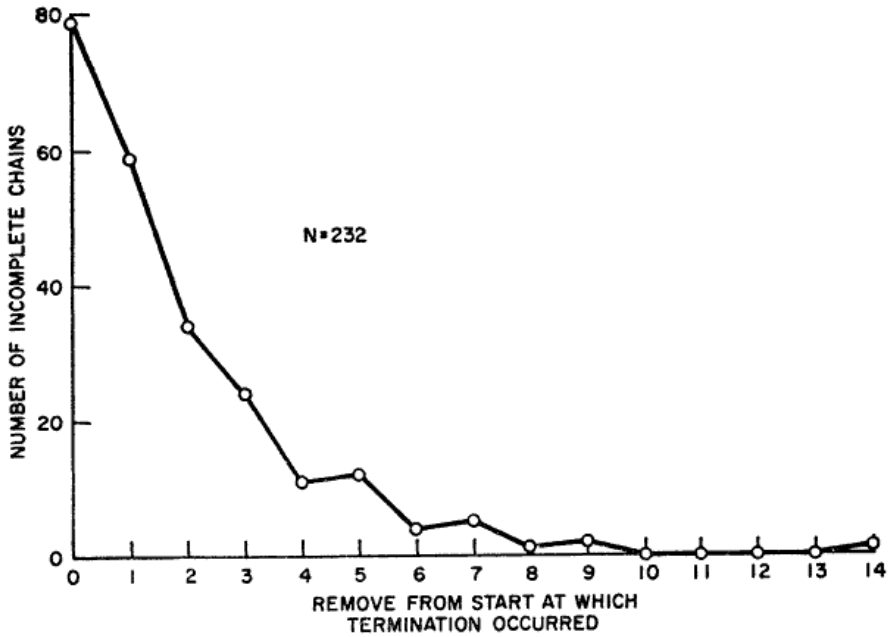


FIGURE 2

Lengths of Incomplete Chains

Bimodal?

Connections thru target's professional circle tended to be more direct; connections thru hometown take longer.

Means

Starting Population	Mean Chain Length
Nebraska Random	5.7
Nebraska Stockholders	5.4
All Nebraska	5.5
Boston Random	4.4
All	5.2

An observation and question

- It's easy to find a short path given the entire network
  - Breadth-first search
- The participants in Millgram's experiment did *not* see the whole network
  - Only their friends and (information about) the target node
- When can you navigate through a network using only *local* information?
  - LinkedIn
- More generally: is geography a bug or a feature?
  - Q1: what do social networks look like?
  - Q2: what *should* social networks look like?

The Small-World Phenomenon: An Algorithmic Perspective *

Jon Kleinberg †

A mathematical model

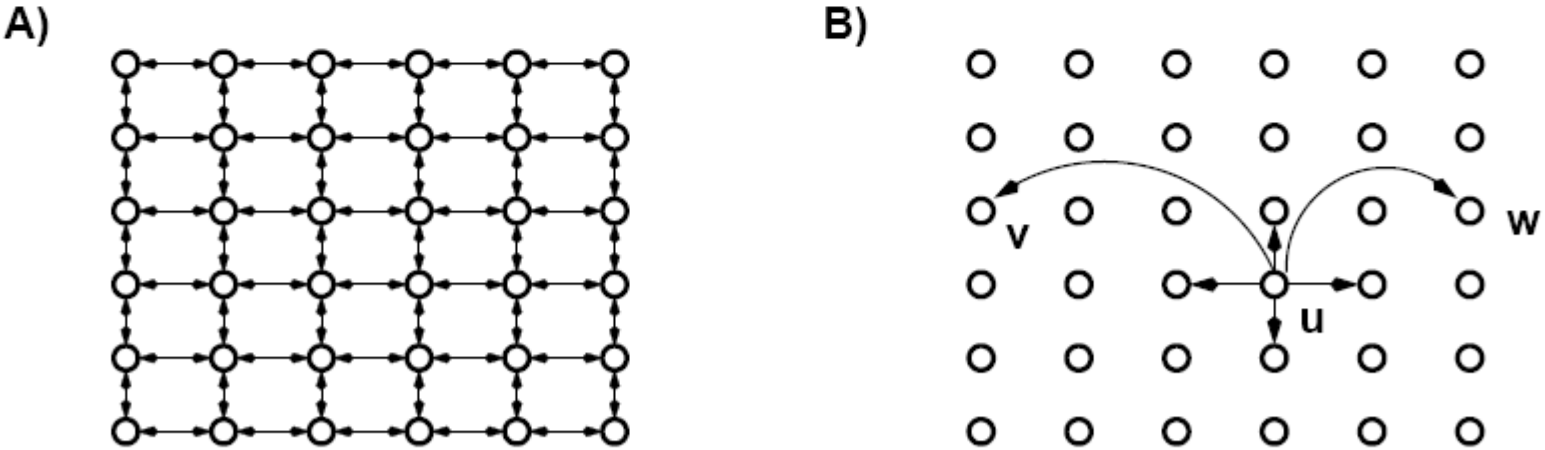


Figure 1: (A) A two-dimensional grid network with $n = 6$, $p = 1$, and $q = 0$. (B) The contacts of a node u with $p = 1$ and $q = 2$. v and w are the two long-range contacts.

- World is an n*n grid *plus* q "long-range" connections for each node
- Probability of a long-range link from u to v is

$$(1 / Z) * \mathrm{dist}(u,v)^{-r}$$

A mathematical model

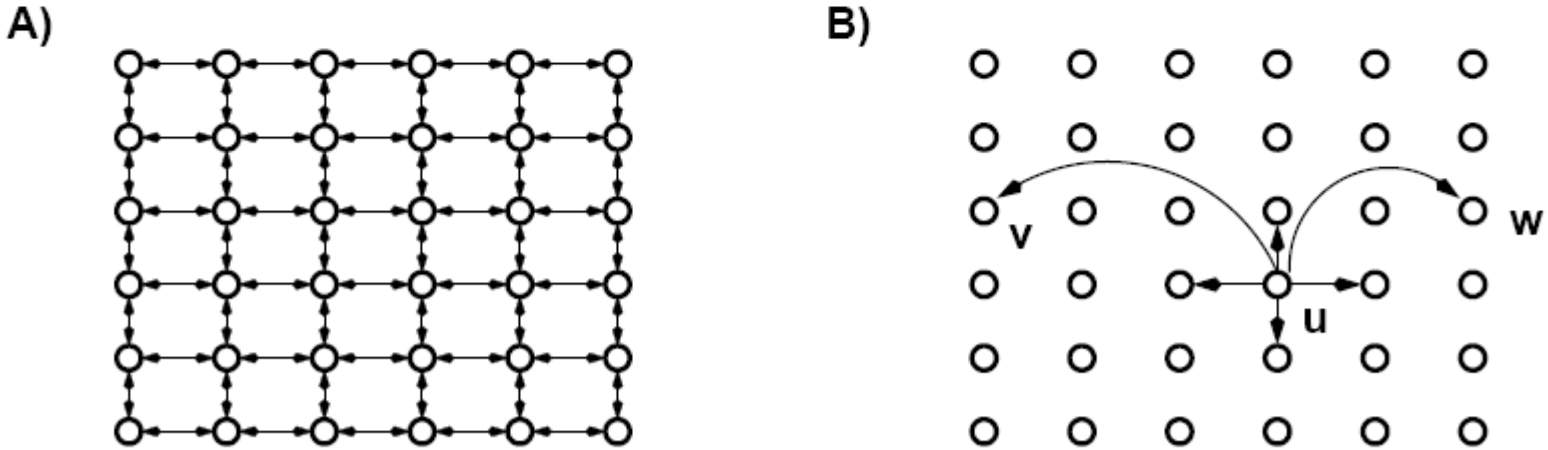


Figure 1: (A) A two-dimensional grid network with $n = 6$, $p = 1$, and $q = 0$. (B) The contacts of a node u with $p = 1$ and $q = 2$. v and w are the two long-range contacts.

- This is *very* similar to the 2-D version of Watt's "small world" model – a ring with fixed short-range connections and random "long-range" connections.
- Difference is that longer links are *progressively* less likely.

The task

- Simulate Milgrams's problem:
  - Packet starts at node *u* and is being sent to node *t*.
  - Each node in the chain knows
    - x,y coordinates of *t*
    - her own neighbors (and their x,y coordinates)
    - "history" of previous nodes that touched the packet
  - Each node must decide "locally" which neighbor to send it to
    - Greedy algorithm: send to neighbor closest to *t*
    - With no "long-distance" links, greedy takes time O(n)
    - When is it substantially *faster* than O(n)? i.e., when do the long-distance links really help?
      - Looking for polynomial in *logn vs* polynomial in *n*

The results

1. If r=0 (i.e., long-range contacts are uniformly distributed across the whole world) then expected delivery time is $\Omega(\alpha_{p,q} n^{2/3})$
2. If r=2 (i.e., long-range contacts follow an inverse square law) then expected delivery time is $O(\alpha_{p,q} \log^2(n))$
3. Asymptotically *only* r=2 leads to logarithmic delivery time (r=1.9 or r=2.01 are not good).

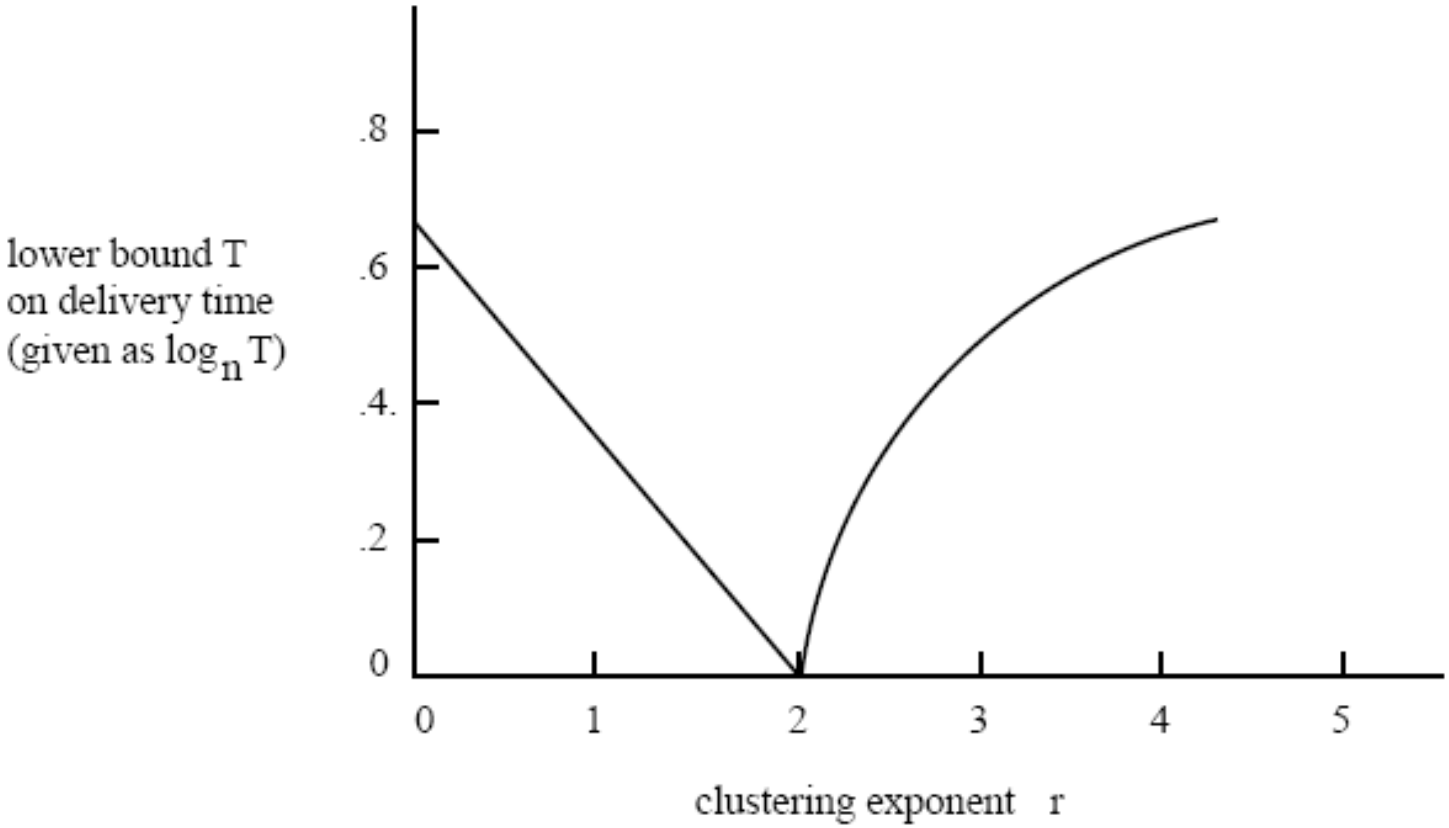
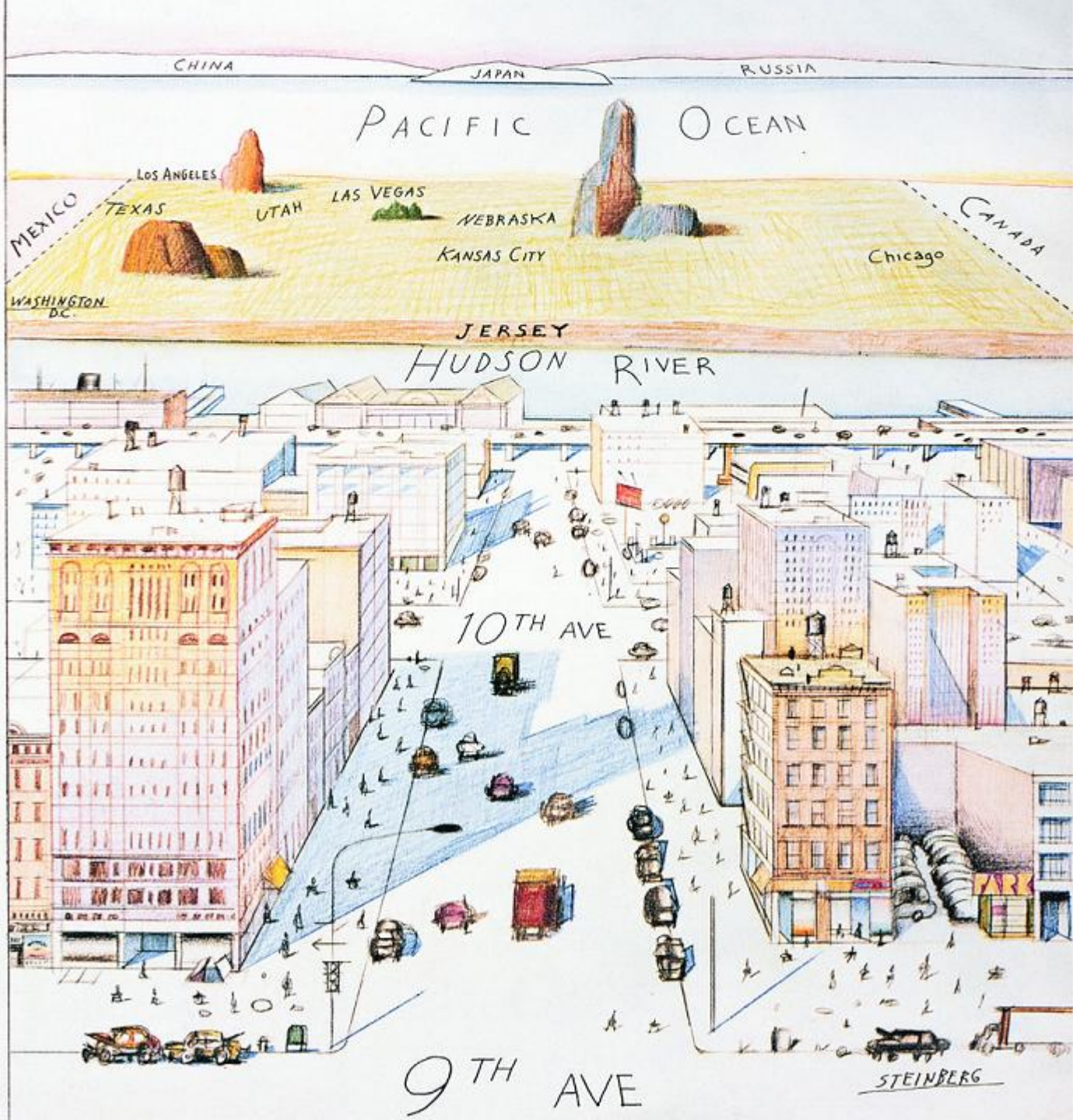
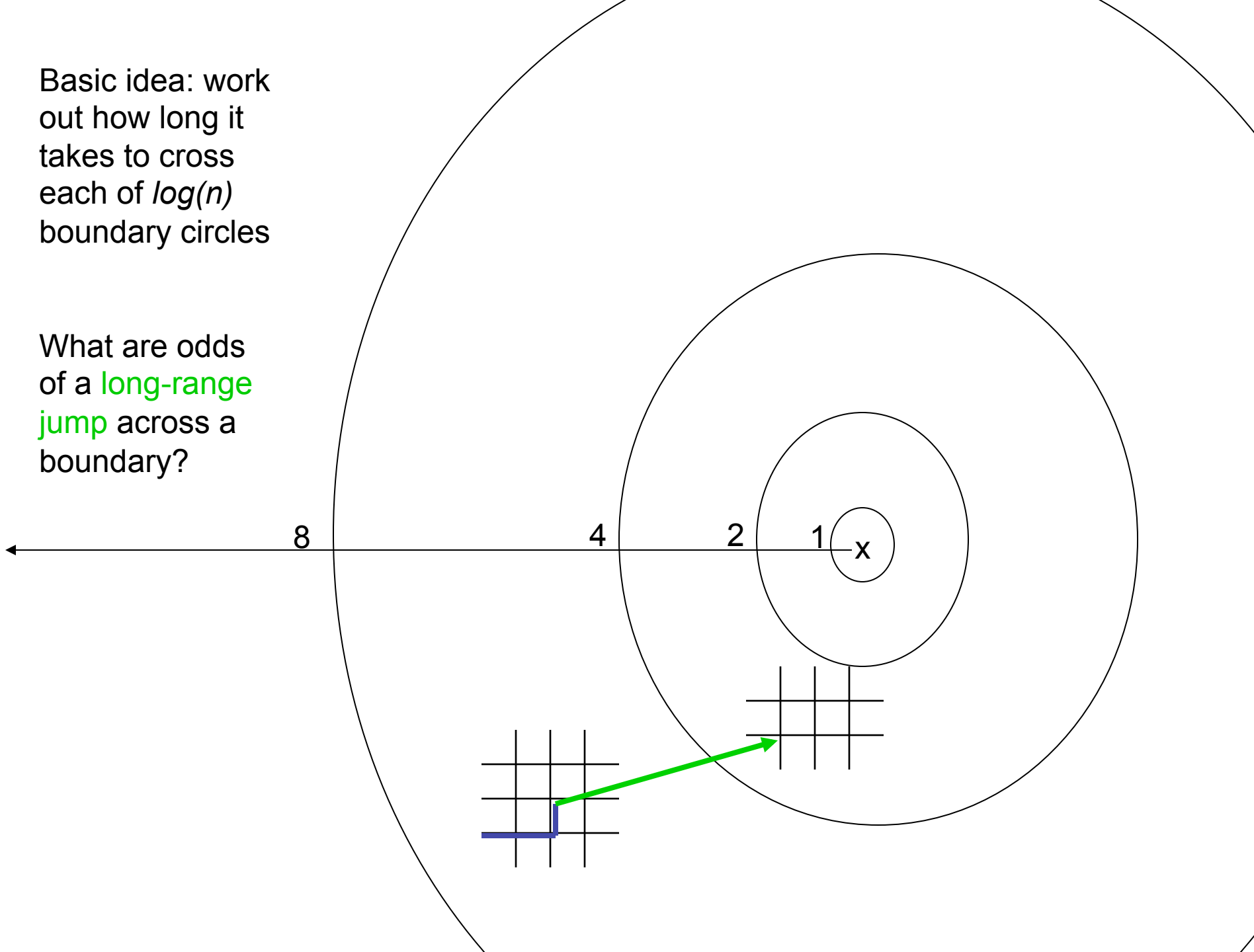


Figure 2: The lower bound implied by Theorem 3. The x-axis is the value of r; the y-axis is the resulting exponent on n.

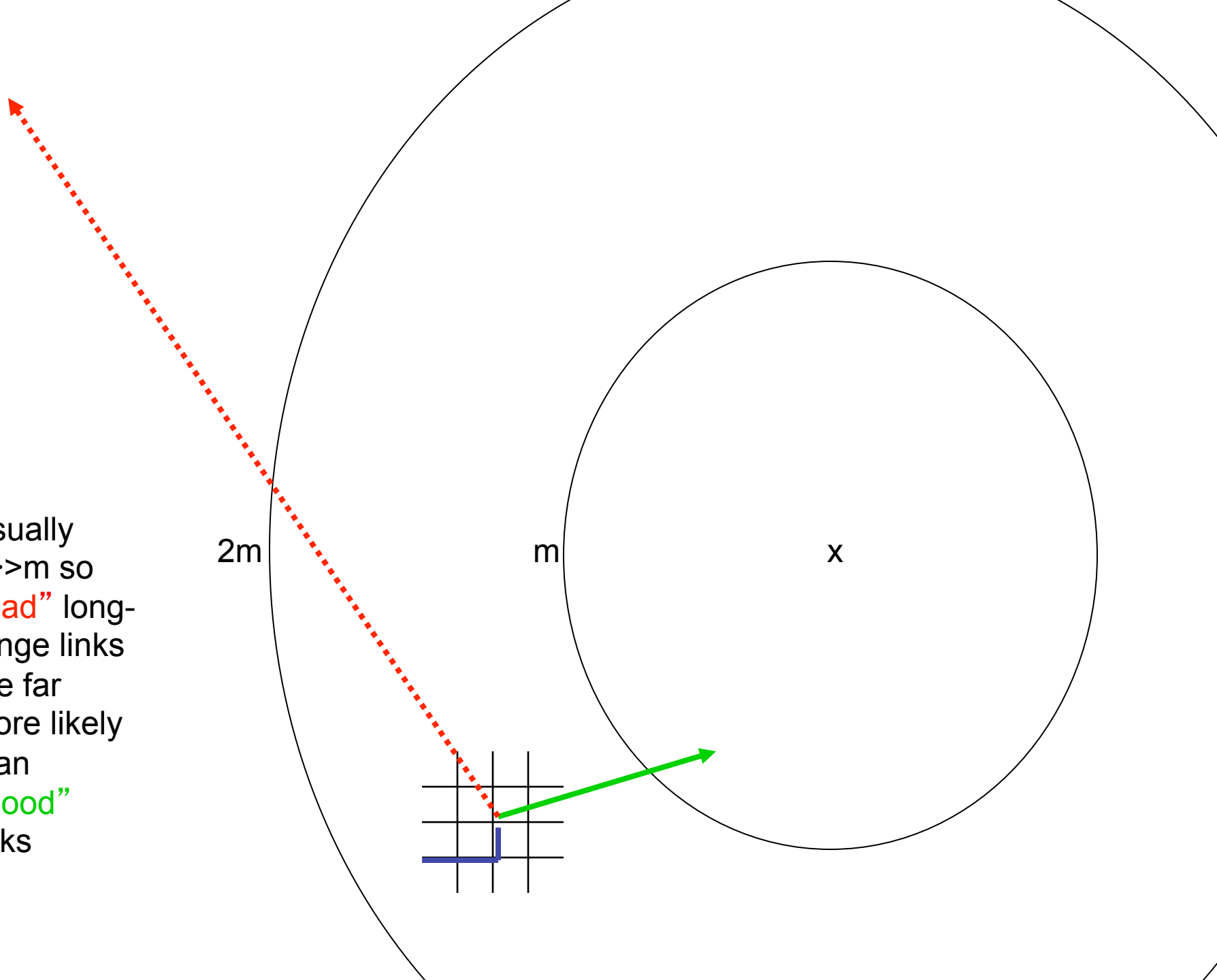

CHINA
JAPAN
RUSSIA
PACIFIC OCEAN
LOS ANGELES
MEXICO
TEXAS
UTAH
LAS VEGAS
NEBRASKA
KANSAS CITY
CANADA
Chicago
WASHINGTON D.C.
JERSEY
HUDSON RIVER
10TH AVE
9TH AVE
STEINBERG

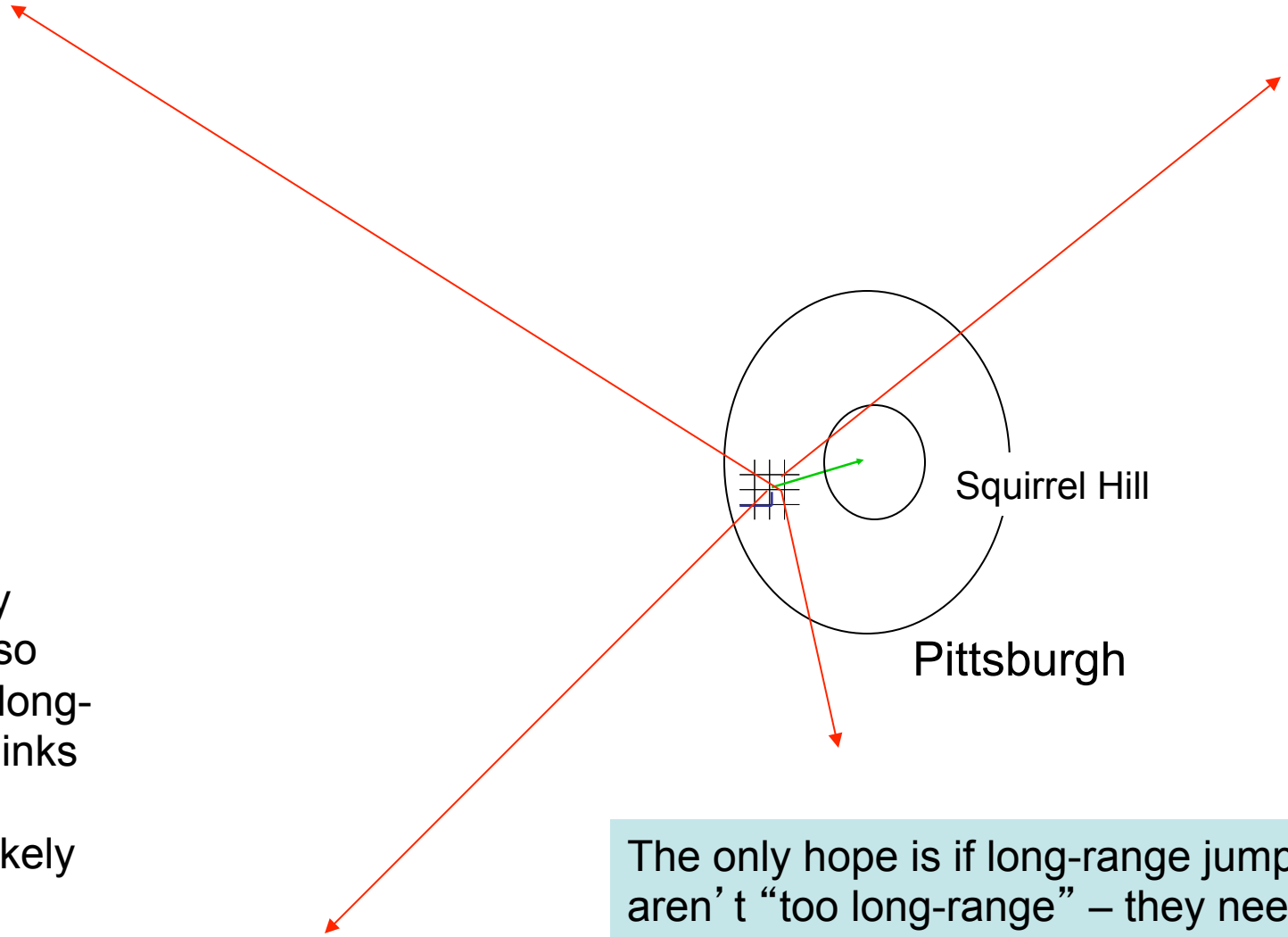
Basic idea: work out how long it takes to cross each of *log(n)* boundary circles

What are odds of a long-range jump across a boundary?



Usually n>>m so "bad" long-range links are far more likely than "good" links





Usually n>>m so "bad" long-range links are far more likely than "good" links

The only hope is if long-range jumps aren't "too long-range" – they need to have a pretty good shoot at being near but not too near

North America

Claim 1:

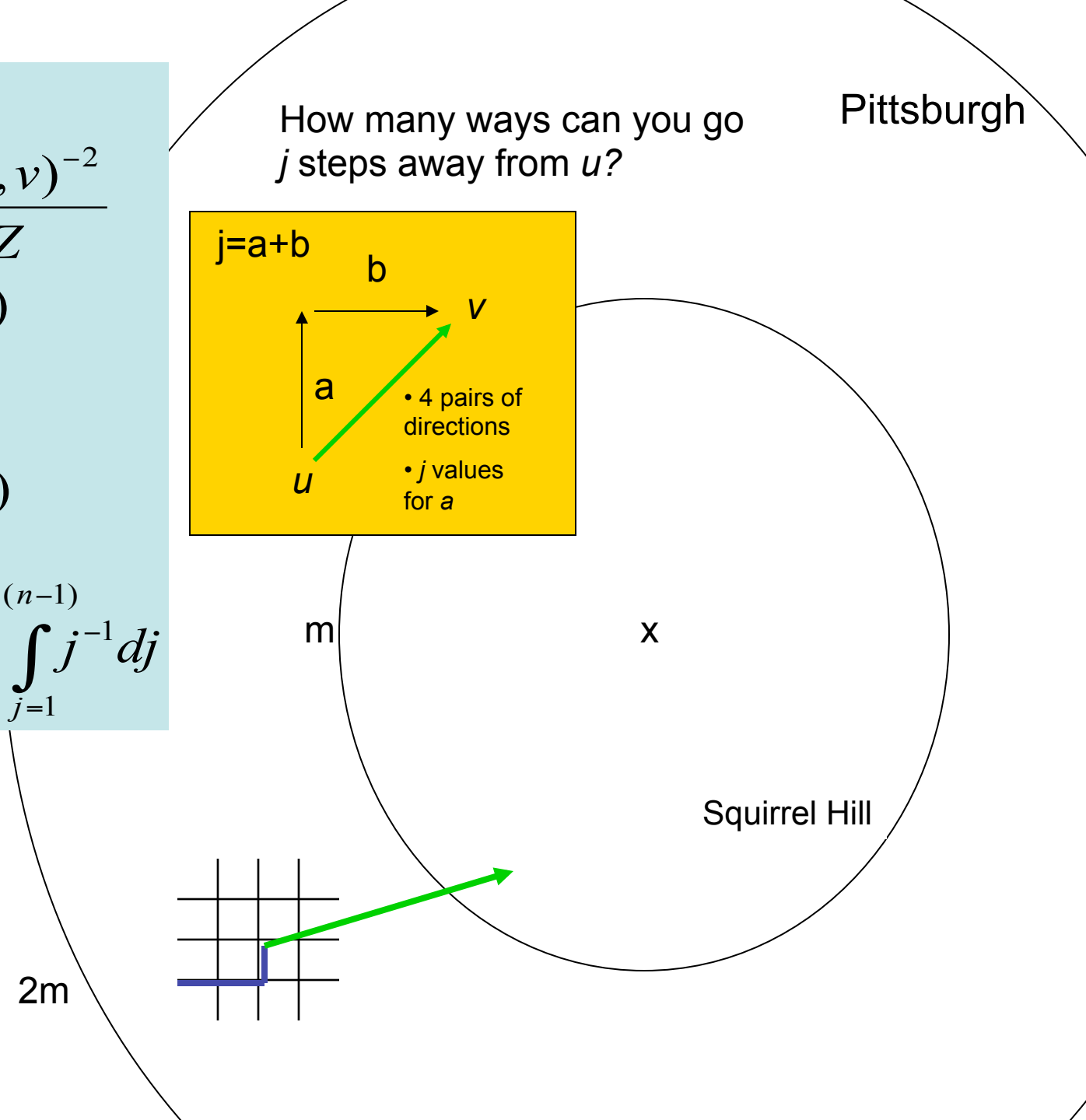
$$\Pr(u \to v) = \frac{d(u,v)^{-2}}{Z}$$

$$\Rightarrow Z = \ldots \log(\ldots n)$$

Proof:

$$Z \le \sum_{j=1}^{2(n-1)} (4j)(j^{-2})$$

$$Z \le 4 \sum_{j=1}^{2(n-1)} j^{-1} \approx 4 \int_{j=1}^{2(n-1)} j^{-1} dj$$



Claim 2 :

$$\Pr(u \to v \mid u \in \mathrm{Pgh}, v \in \mathrm{SqH}) \geq \frac{(3m)^{-2}}{Z}$$

max dist from u→v is 3m

$$\Pr(u \to \mathrm{SqH} \mid u \in \mathrm{Pgh}) \geq (\pi m^2)\frac{9m^{-2}}{Z} = \ldots \frac{1}{\log(\ldots n)}$$

Pittsburgh

So … if you "wait" (walk locally) for about *logn* transfers, you should get lucky and get passed to someone with a friend from Squirrel Hill.

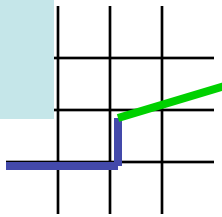
This holds for each of the *logn* concentric circles that we've imagined…

m

x

Squirrel Hill

2m



So …we should expect about O(logn * logn) transfers before reaching the target

Geographic routing in social networks

David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins

Extensions to Kleinberg's result

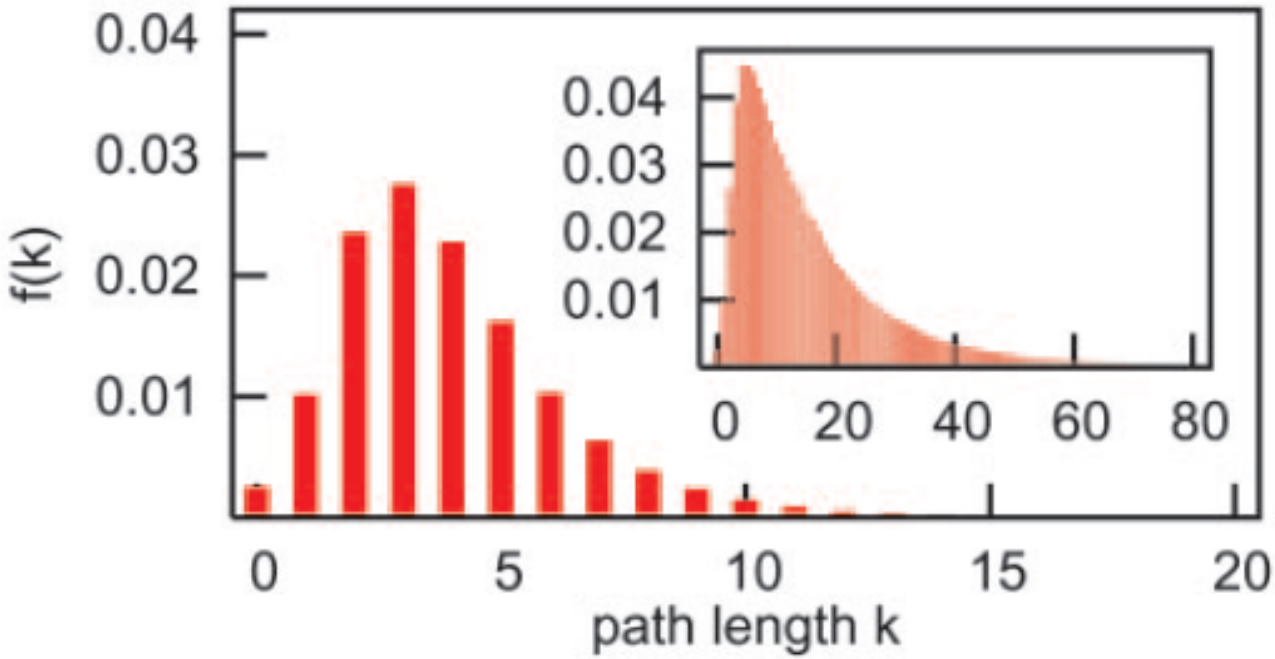
- "Geographic routing in social networks" – Liben-Nowell,Novak,Kumar,Raghavan,Tomkins, PNAS 102(33) pp 11623-11628
  - Model: Pr($u \rightarrow v$) = $1/Z * (\text{number of closer people})^{-1}$
  - About 2/3 of relationships fit this model in data mined from LiveJournal

Liben-Nowell *et al* experiment

- LiveJournal site, *c.* 2004:
  - 1.3M bloggers, who can list
    - Friends (other LJ bloggers)
    - Location
    - Interests, …
  - 500k LJ bloggers list home town and state that can be geomapped (to lat & long)
    - Only approximate (to within the city)
  - About 4M "friendship" links between these bloggers
    - mostly reciprocal links
    - 385k bloggers are in one connected component
    - In-degree/out-degree plots look roughly lognormal

Idea: simulate the Millgram experiment

- Pick random start node *u* and target *t*
- Repeat until message is at *u*'s hometown:
  - If *u* is closer to *t* than any of *t's* friends:
    - Give up (failing)
  - Else:
    - Pass the message to the friend of *u* closest to *t,* geographically



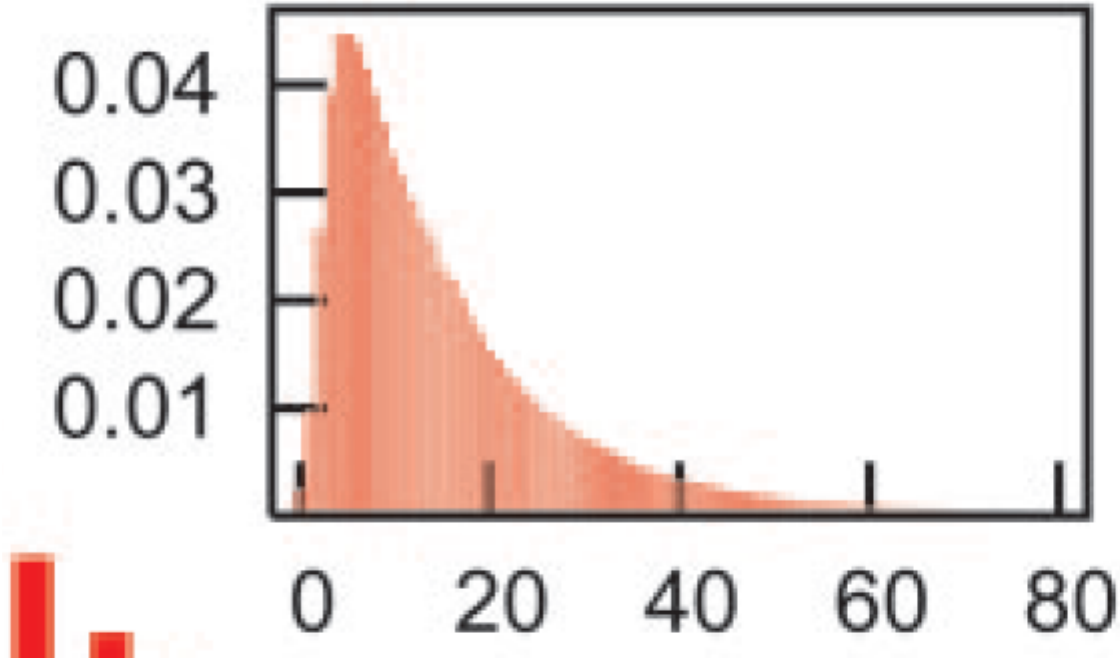
vs Millgram:

- 13% completed vs 18% (21%?)
- mean chain length 4 *vs* 6 (but they don't reach *t* just his hometown)

Fig. 2. Results of GEOGREEDY on LiveJournal. In each of 500,000 trials, a source s and target t are chosen randomly; at each step, the message is forwarded from the current message-holder u to the friend v of u geographically closest to t. If $d(v, t) > d(u, t)$, then the chain is considered to have failed. The fraction $f(k)$ of pairs in which the chain reaches t's city in exactly k steps is shown (12.78% chains completed; median 4, $\mu = 4.12$, $\sigma = 2.54$ for completed chains). (*Inset*) For 80.16% completed, median 12, $\mu = 16.74$, $\sigma = 17.84$; if $d(v, t) > d(u, t)$ then u picks a random person in the same city as u to pass the message to, and the chain fails only if there is no such person available.

Idea: simulate the Millgram experiment

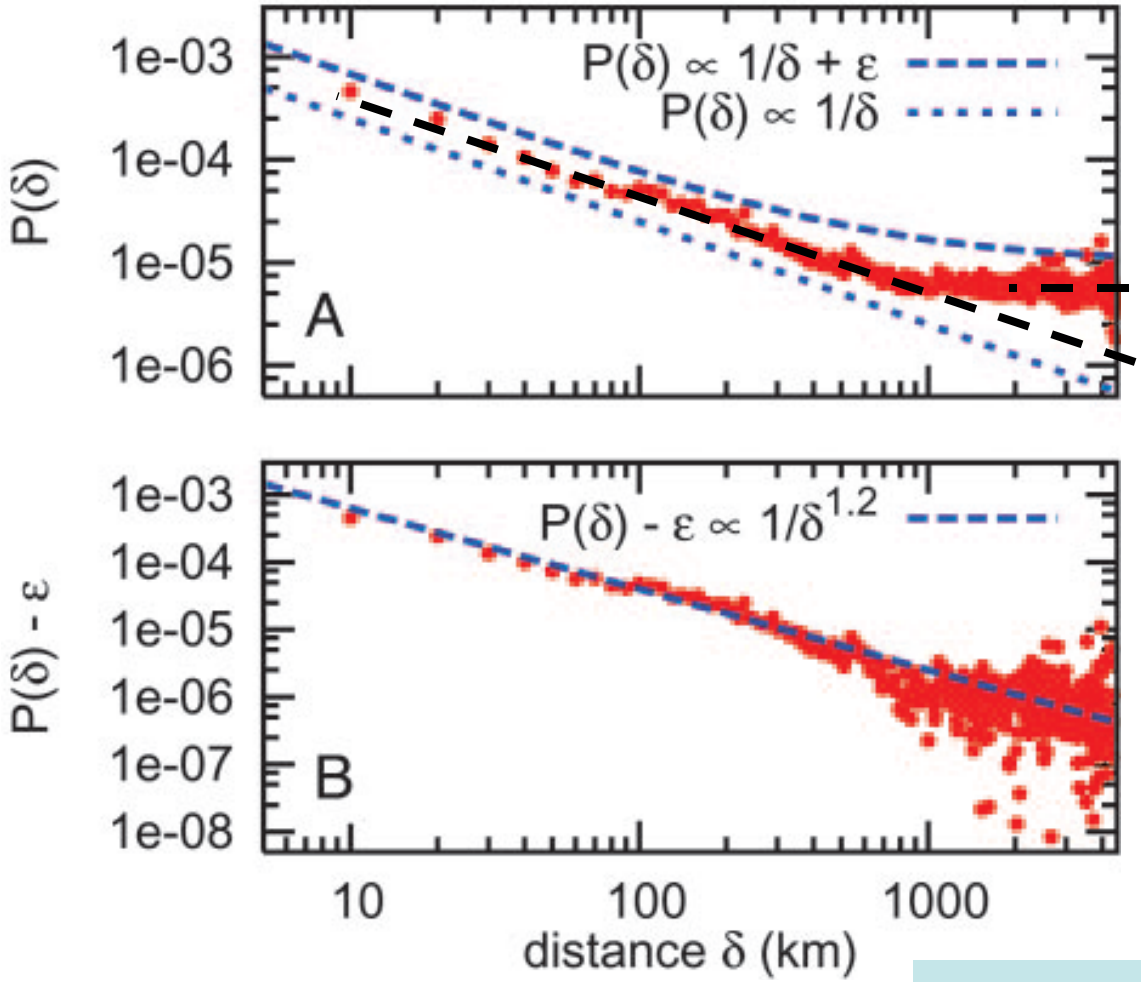
- Pick random start node *u* and target *t*
- Repeat until message is at *u*'s hometown:
  - If *u* is closer to *t* than any of *t's* friends:
    - ~~Give up (failing)~~ **Forward to a random person from *u*'s hometown**
  - Else:
    - Pass the message to the friend of *u* closest to *t,* geographically



vs Millgram:

- 80% completed vs 18%
- mean chain length 16 *vs* 6 (but they don't reach *t* just his hometown)

Fig. 2. Results of GEOGREEDY on LiveJournal. In each of 500,000 trials, a source s and target t are chosen randomly; at each step, the message is forwarded from the current message-holder u to the friend v of u geographically closest to t. If $d(v, t) > d(u, t)$, then the chain is considered to have failed. The fraction $f(k)$ of pairs in which the chain reaches t's city in exactly k steps is shown (12.78% chains completed; median 4, $\mu = 4.12$, $\sigma = 2.54$ for completed chains). (*Inset*) For 80.16% completed, median 12, $\mu = 16.74$, $\sigma = 17.84$; if $d(v, t) > d(u, t)$ then u picks a random person in the same city as u to pass the message to, and the chain fails only if there is no such person available.

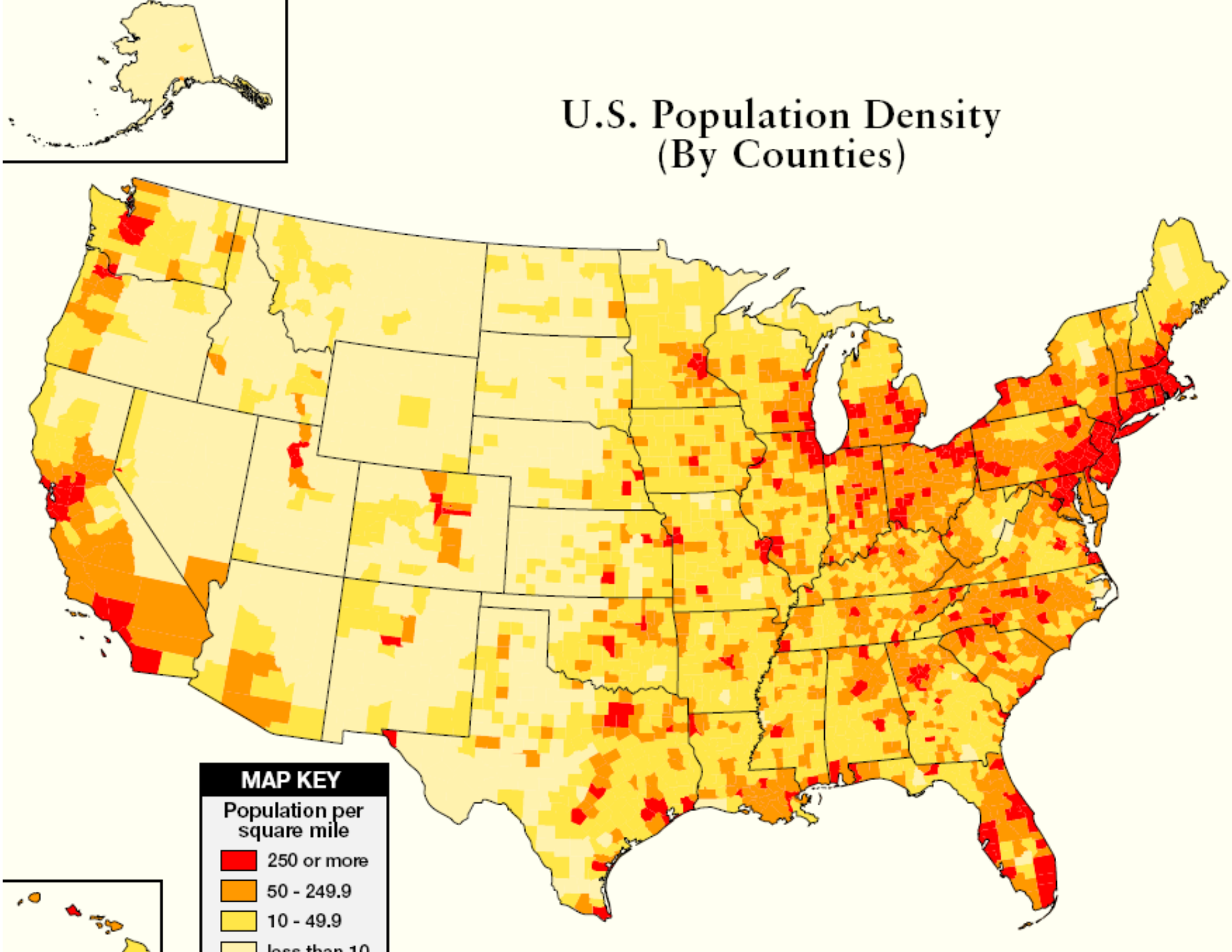


Mixture of power law (local connections) and uniformly-distributed long-range links?

Fitting the mixture model (?) people average 5.5 local and 2.5 long-range links.

Problem: Kleinberg's paper predicts that short paths are *not* locally findable with

$\text{Prob}(u \rightarrow v) = 1/Z\ d(u,v)^{-1.2}$


U.S. Population Density
(By Counties)
MAP KEY
Population per square mile
250 or more
50 - 249.9
10 - 49.9
less than 10

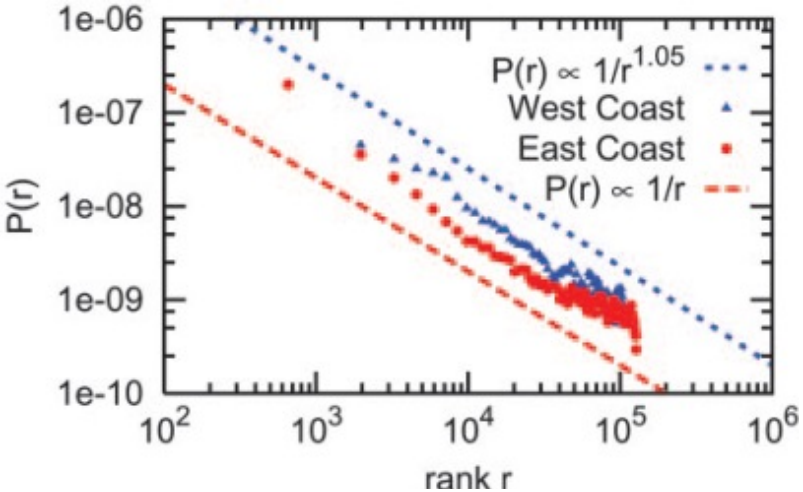
Resolution of the issue?

Claim: The same positive result can be worked through in a new model:

* $Pr(u \rightarrow v) = 1/Z \, rank(u,v)^{-1}$

where rank(u,v)=number of people closer to u than v.

(No proof in paper: but notice that (*) holds for inverse-square links also).



Results on LJ data (smoothed, split into East coast/West coast, and correcting for the "background probability" of friendship)

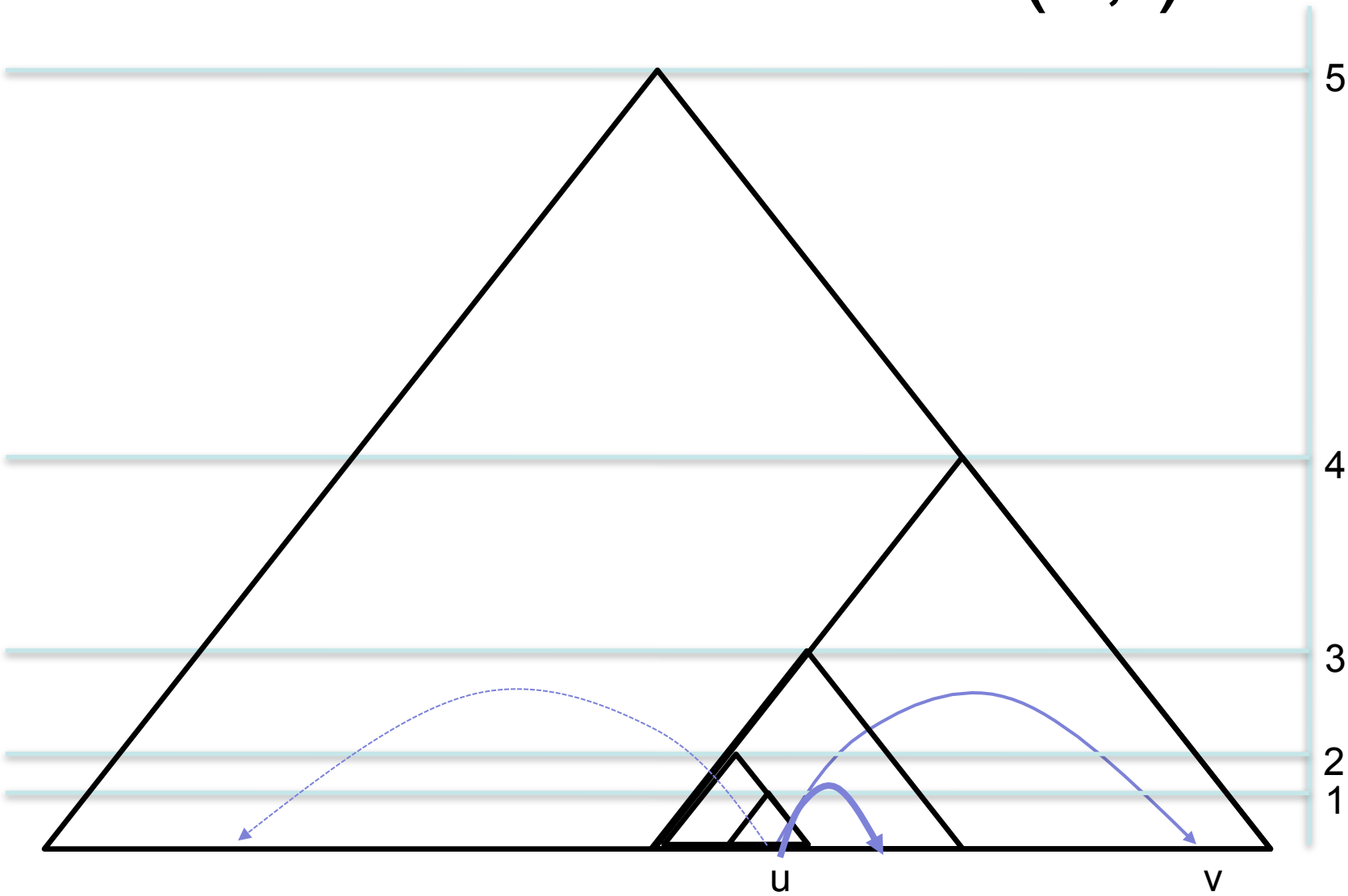
Small-World Phenomena and the Dynamics of Information (NIPS 2001)

Jon Kleinberg

Kleinberg's Social Distance Model

- The social tree:
  - A completely balanced b-ary tree
  - Internal nodes are groups, leaves are people
  - Social distance:
    - Based on h(u,v)=height(least common anc. u,v)
    - SocialDist(u,v)=f(h(u,v))=O($b^{-\alpha h(u,v)}$)
- The graph
  - Each node has
    - c*logn friends
    - chosen as Pr(u→v) = (1/Z) $b^{-\alpha h(u,v)}$

Social distance and h(u,t)



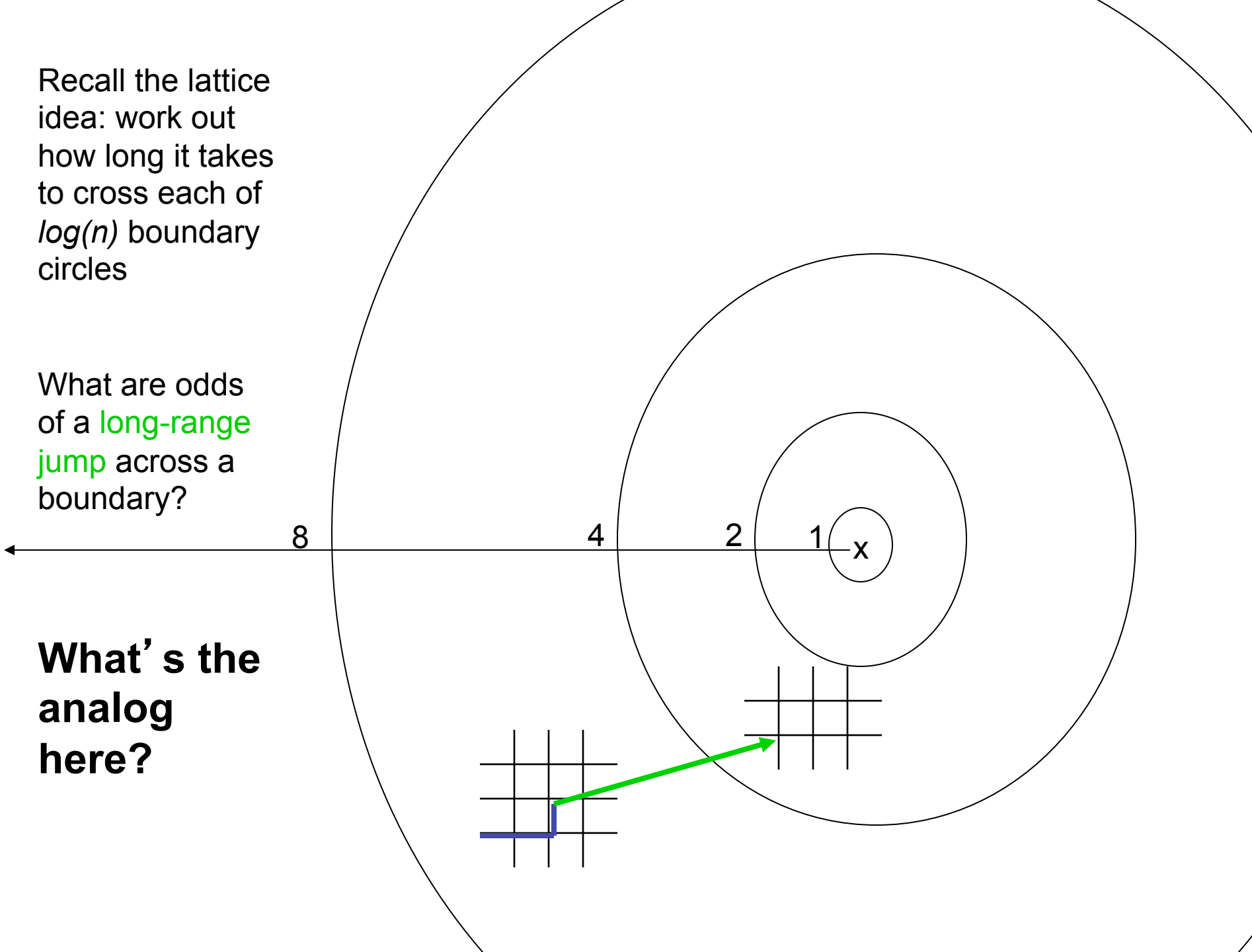
Kleinberg's Social Distance Model

- The social tree:
  - A completely balanced b-ary tree
  - Internal nodes are groups, leaves are people
  - Social distance:
    - Based on h(u,v)=height(least common anc. u,v)
    - SocialDist(u,v)=f(h(u,v))=O($b^{-\alpha h(u,v)}$)
- The graph
  - Each node has c*logn links chosen as
    - Pr(u→v) = (1/Z) $b^{-\alpha h(u,v)}$
- The navigation problem:
  - Greedy, using *social distance* to target *t* but nothing else
  - Result: another sweet spot when α=1
  - I'll give the positive result

Recall the lattice idea: work out how long it takes to cross each of *log(n)* boundary circles

What are odds of a long-range jump across a boundary?

What's the analog here?

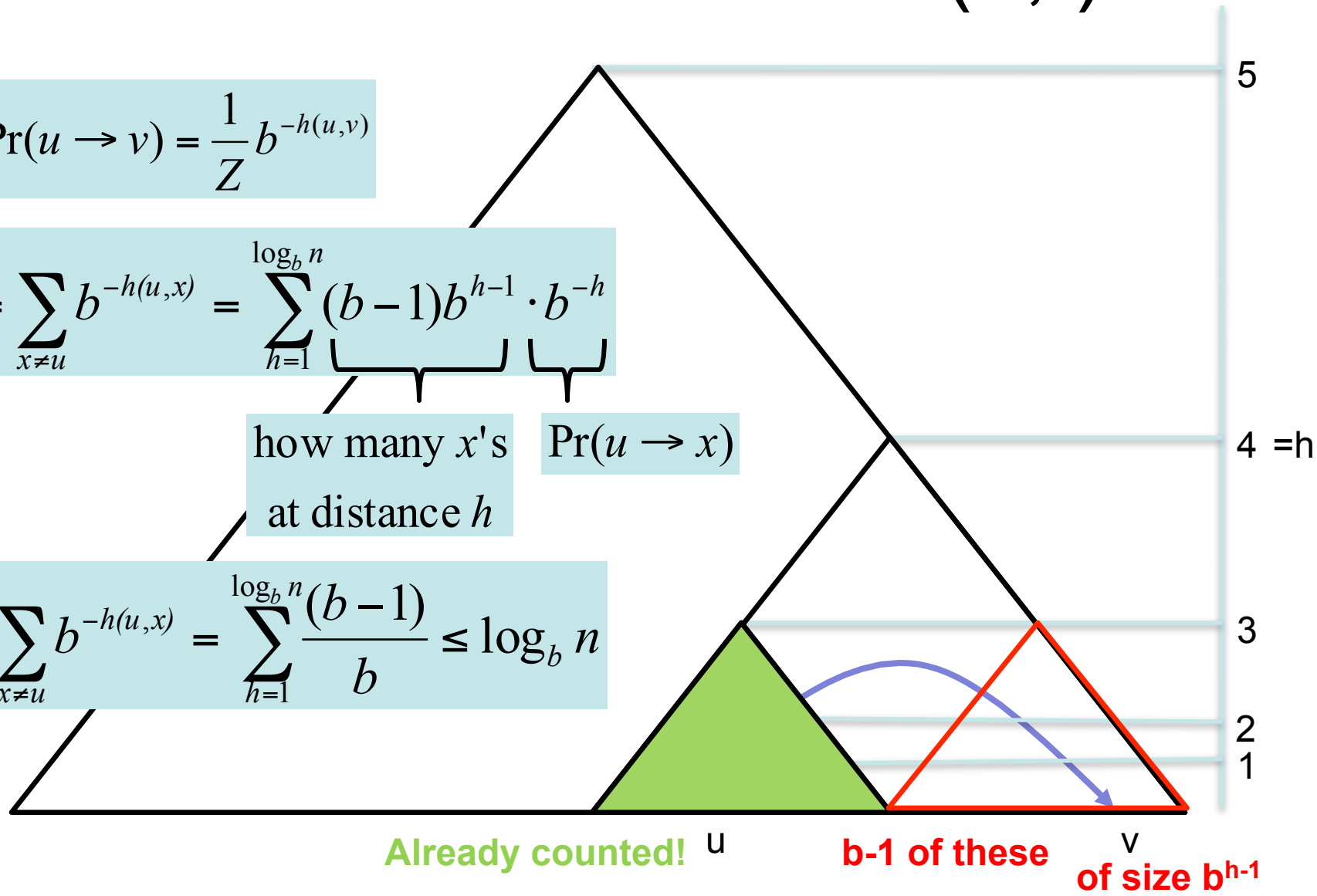


Social distance and h(u,t)

$$\Pr(u \to v) = \frac{1}{Z} b^{-h(u,v)}$$

$$Z = \sum_{x \neq u} b^{-h(u,x)} = \sum_{h=1}^{\log_b n} \underbrace{(b-1)b^{h-1}}_{\text{how many } x\text{'s at distance } h} \cdot \underbrace{b^{-h}}_{\Pr(u \to x)}$$

$$Z = \sum_{x \neq u} b^{-h(u,x)} = \sum_{h=1}^{\log_b n} \frac{(b-1)}{b} \leq \log_b n$$



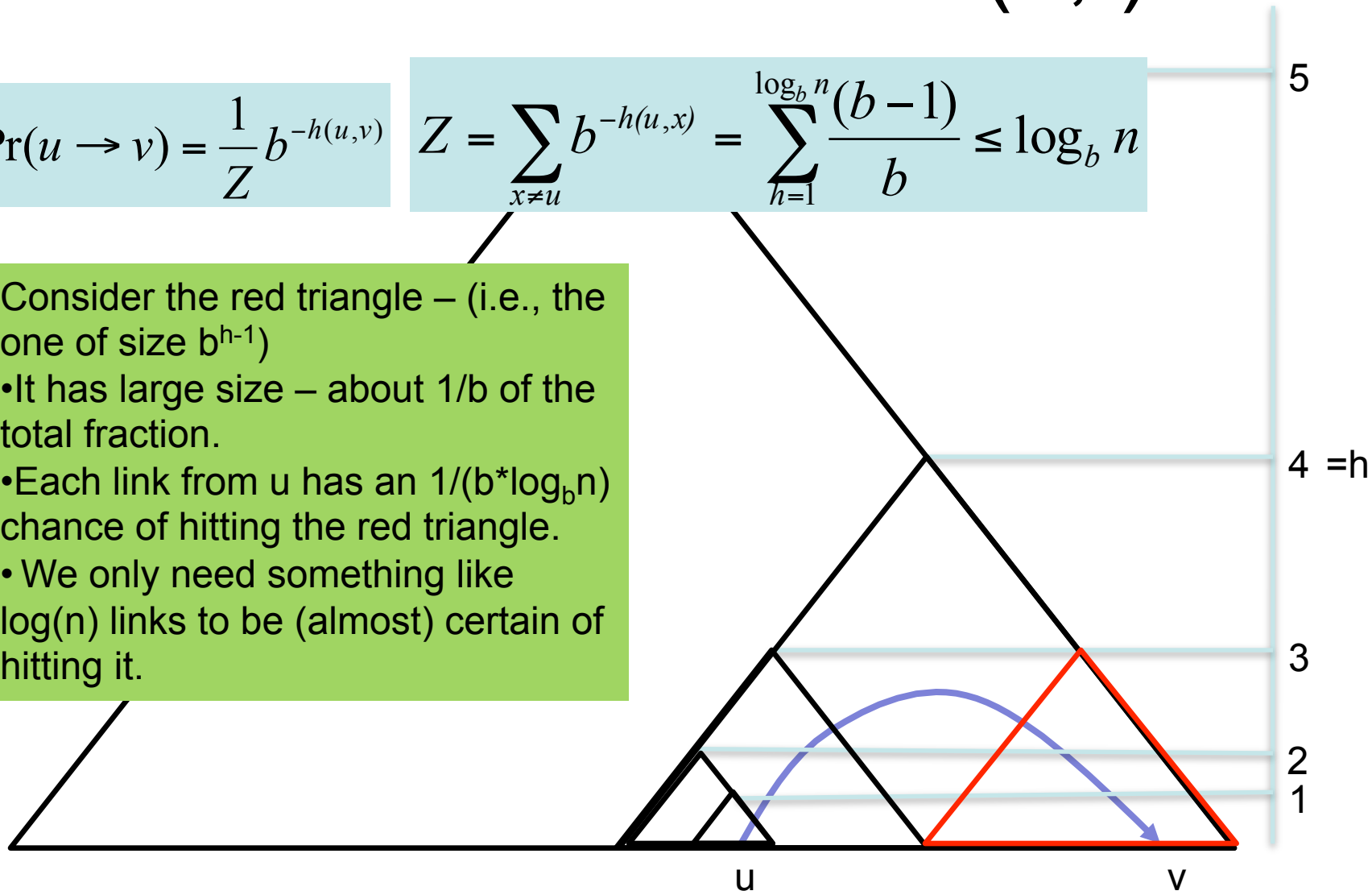
Social distance and h(u,t)

$$\Pr(u \to v) = \frac{1}{Z} b^{-h(u,v)}$$

$$Z = \sum_{x \neq u} b^{-h(u,x)} = \sum_{h=1}^{\log_b n} \frac{(b-1)}{b} \leq \log_b n$$

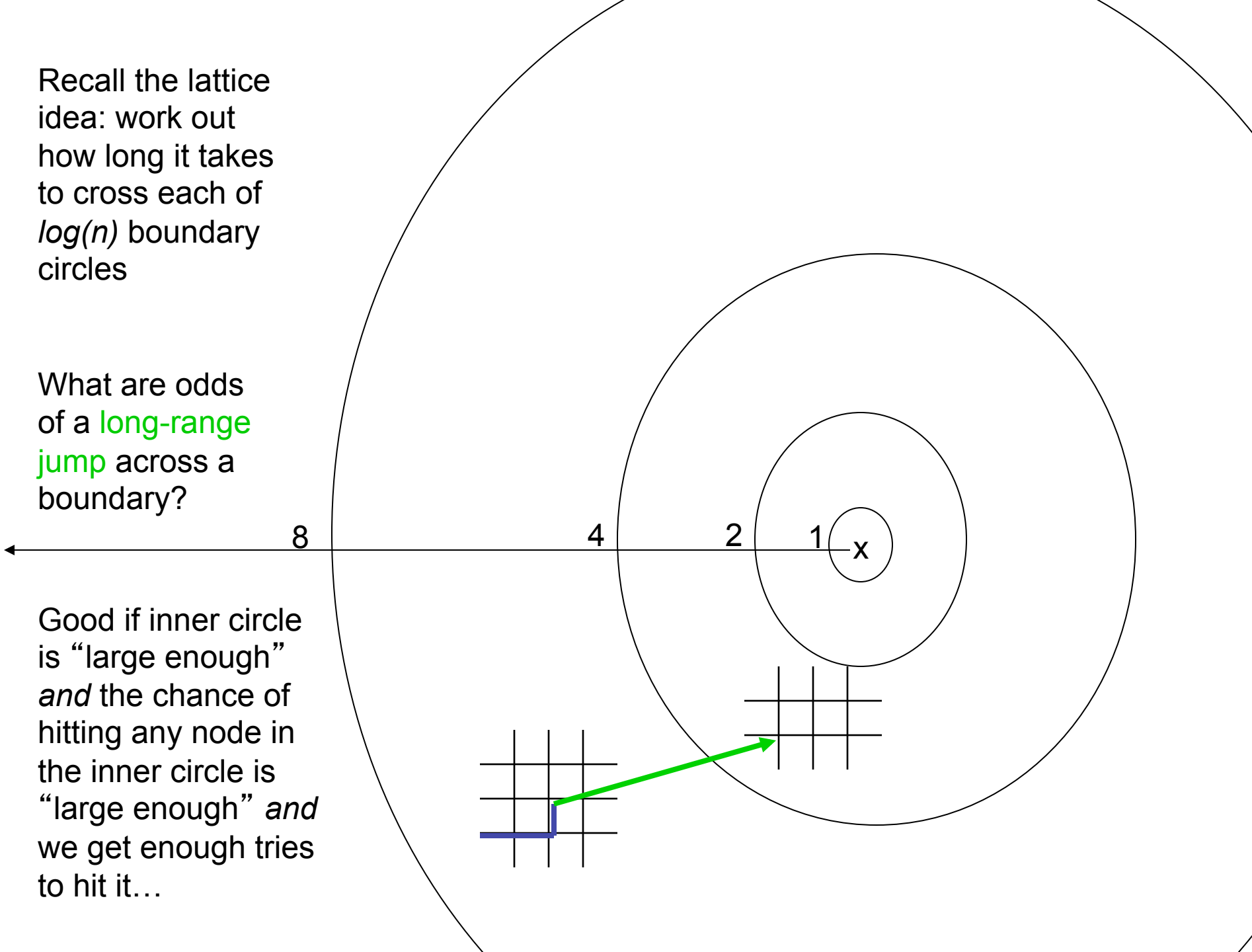
Consider the red triangle – (i.e., the one of size b^{h-1})

- It has large size – about 1/b of the total fraction.
- Each link from u has an $1/(b*\log_b n)$ chance of hitting the red triangle.
- We only need something like log(n) links to be (almost) certain of hitting it.



Recall the lattice idea: work out how long it takes to cross each of *log(n)* boundary circles

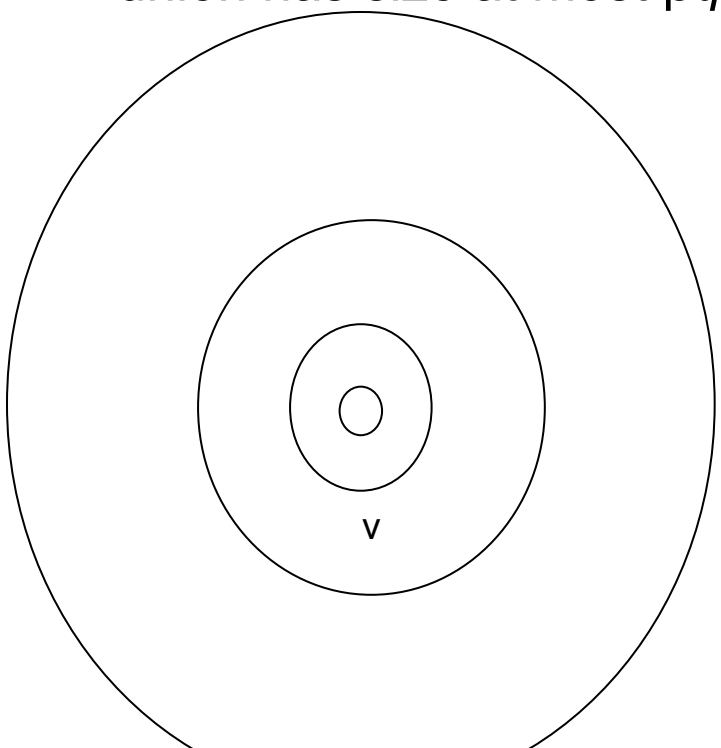
What are odds of a long-range jump across a boundary?

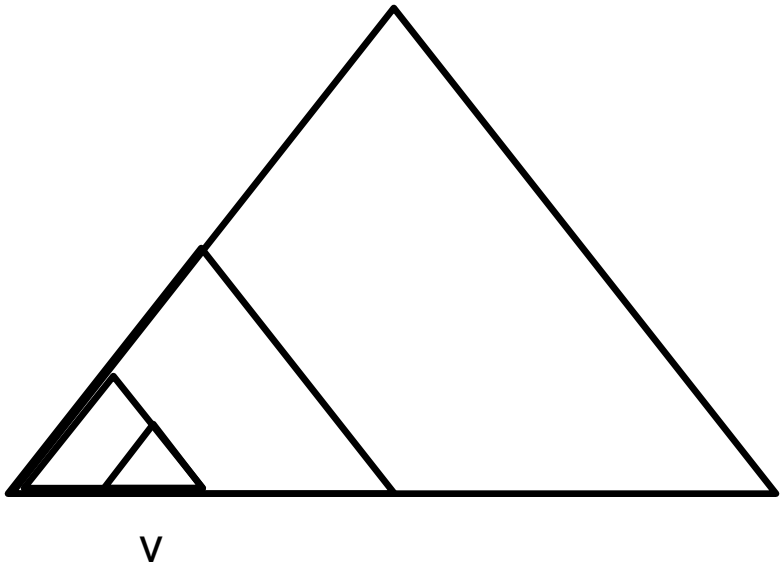


Good if inner circle is “large enough” *and* the chance of hitting any node in the inner circle is “large enough” *and* we get enough tries to hit it…

Kleinberg's Group Structure Model

- Group structure (λ, β):
  - If R has size |R|>1 containing *v* then there is a smaller group R' containing v and size of R' is at least λ*|R|
  - If R1,R2,… all have size at most *q* and all contain *v,* then their union has size at most β*q*





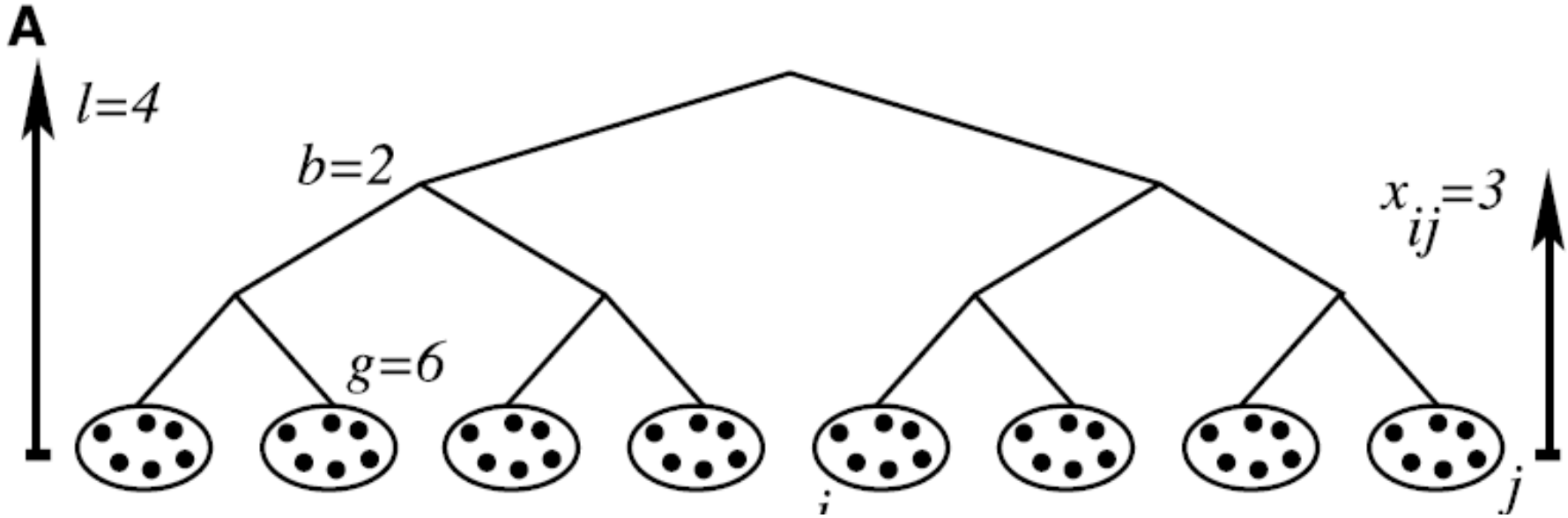
Kleinberg's Group Structure Model

- Group structure – for constants (λ, β):
  - If R has size |R|>1 containing *v* then there is a smaller group R' containing v and size of R' is at least λ*|R|
  - If R1,R2,… all have size at most *q* and all contain *v,* then their union has size at most β*q*
- Graph structure: given the groups $\{R_i\}$:
  - Define *group distance* q(u,v) as *size of smallest group* containing both u and v
  - For each u, create polylogarithmically many links from u→v with Prob=(1/Z)*f(q(u,v))
  - Call this a *group-induced graph with exponent α* if $f(q)=O(q^{-\alpha})$
- Theorem: *α=1 leads to efficient decentralized search and α<1 does not.*

Identity and Search in Social Networks

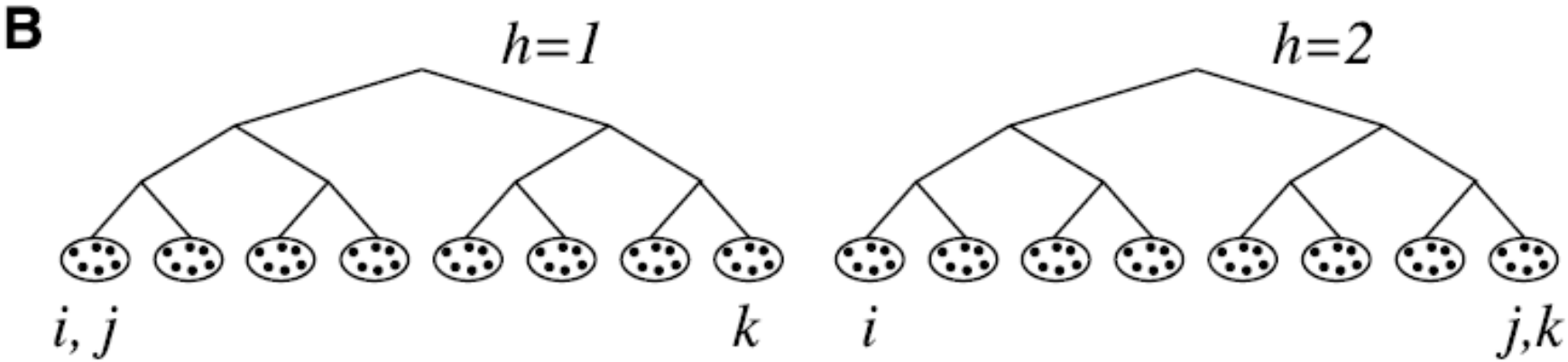
Duncan Watts, Peter Sheridan Dodds, Mark Newman

Science 2002



Assume a heirarchy of groups: LTI, CSD, CMU, …

Social distance is distance to nearest common group (e.g., 1 for two LTI members, 2 for two CSD members, 3 for two CMU faculty, ….)



Assume *k* heirarchies of groups:

- LTI, CSD, CMU, …
- Squirrel Hill, Fox Chapel, …
- ….

Social distance is *minimum* distance in any heirarchy

- Pass a message to someone you're connected to that is *socially closest* to the target

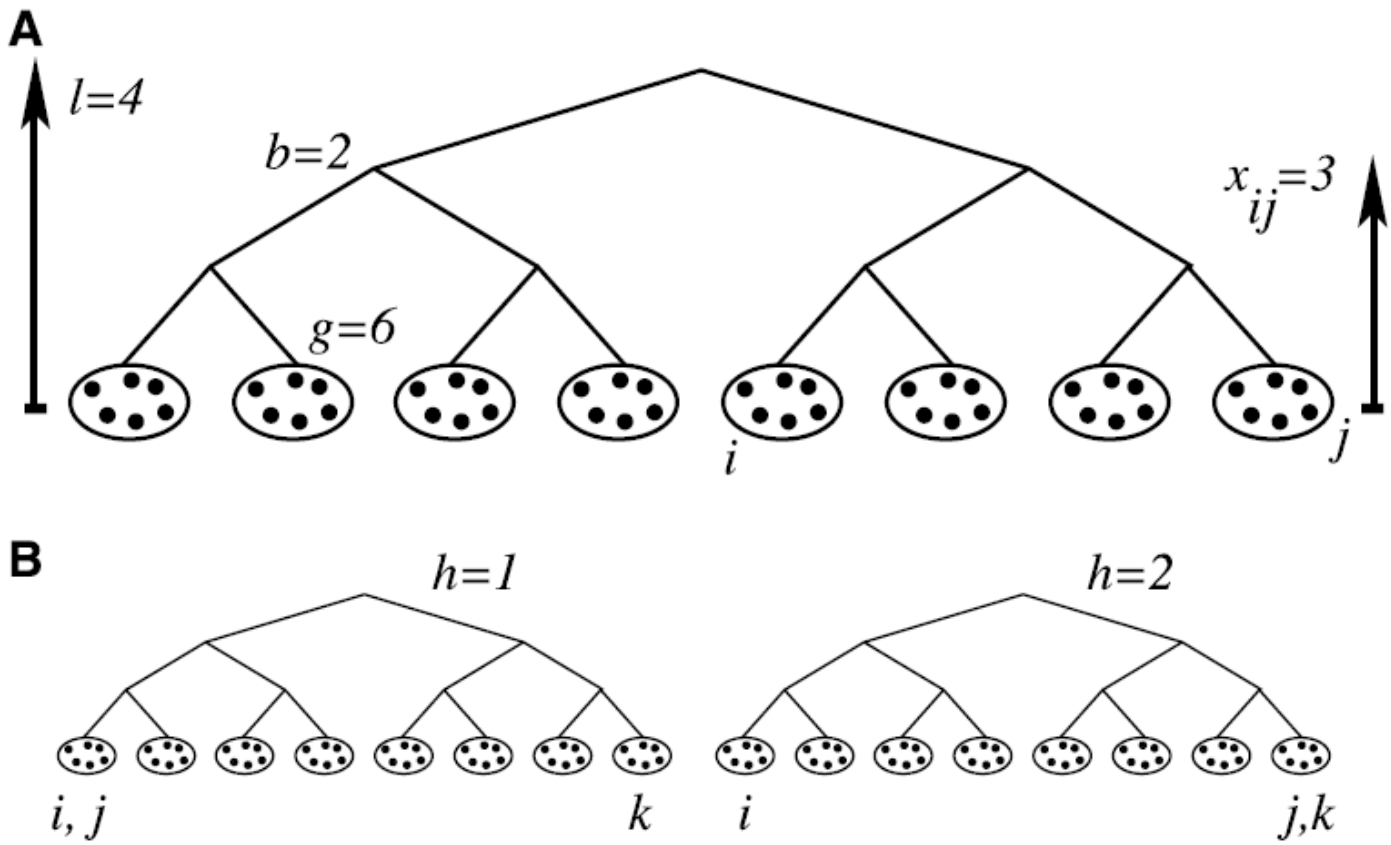
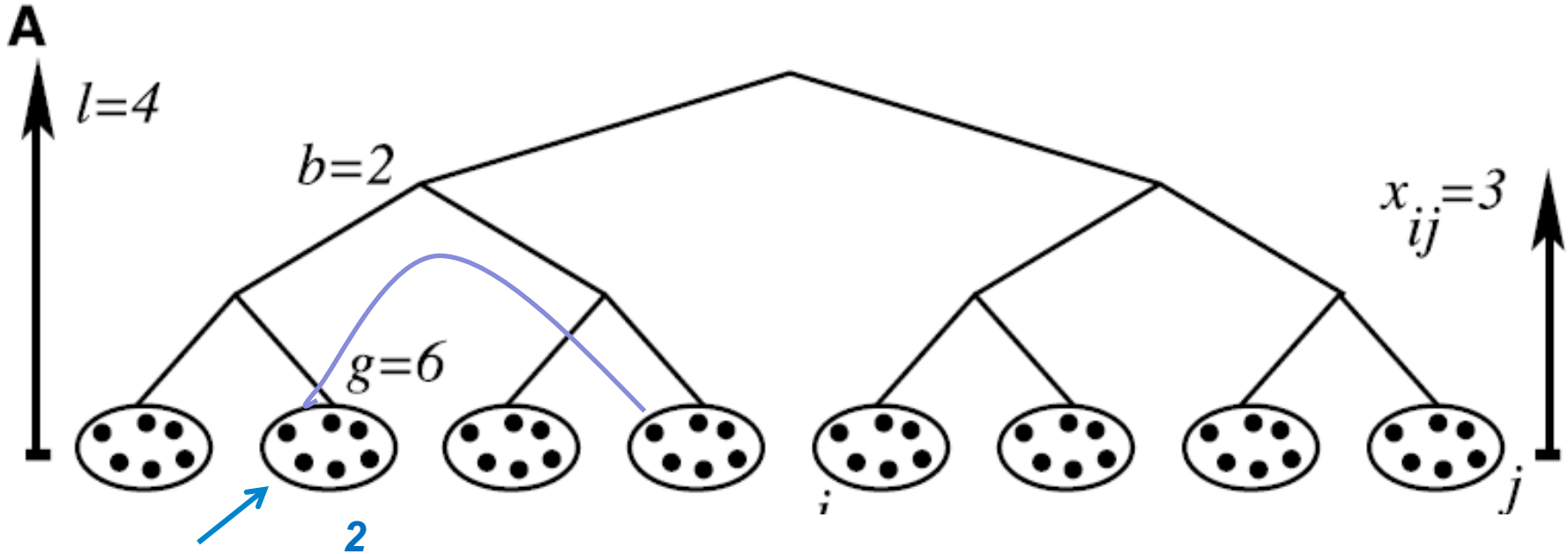
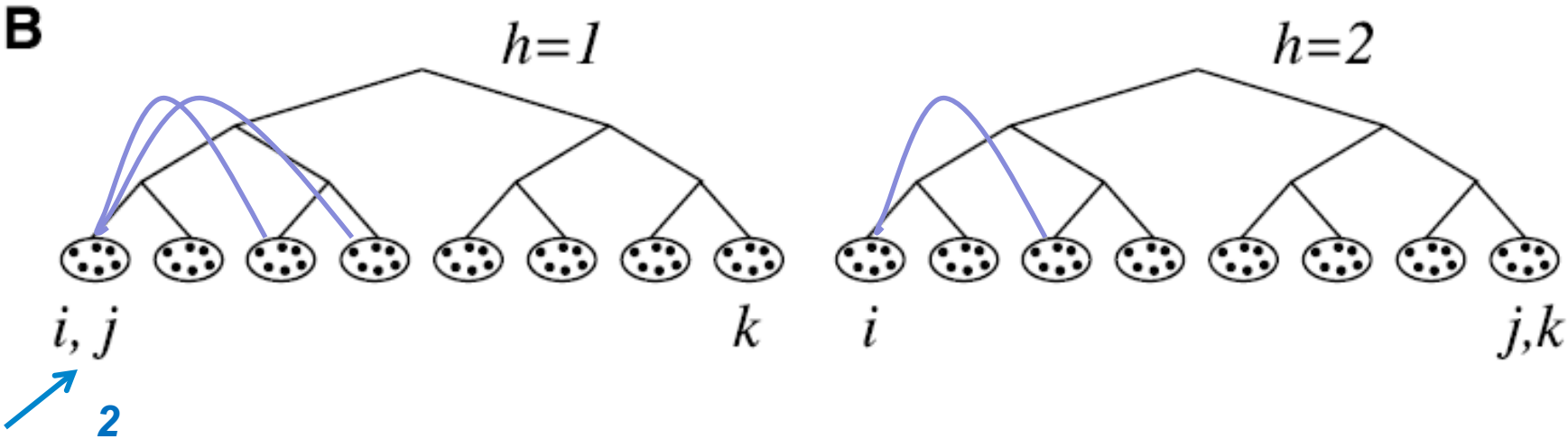


Fig. 1. (**A**) Individuals (dots) belong to groups (ellipses) that in turn belong to groups of groups, and so on, giving rise to a hierarchical categorization scheme. In this example, groups are composed of $g = 6$ individuals and the hierarchy has $l = 4$ levels with a branching ratio of $b = 2$. Individuals in the same group are considered to be a distance $x = 1$ apart, and the maximum separation of two individuals is $x = l$. The individuals i and j belong to a category two levels above that of their respective groups, and the distance between them is $x_{ij} = 3$. Individuals each have z friends in the model and are more likely to be connected with each other the closer their groups are. (**B**) The complete model has many hierarchies indexed by $h = 1 \ldots H$, and the combined social distance y_{ij} between nodes i and j is taken to be the minimum ultrametric distance over all hierarchies $y_{ij} = \min_h x_{ij}^h$. The simple example shown here for $H = 2$ demonstrates that social distance can violate the triangle inequality: $y_{ij} = 1$ because i and j belong to the same group under the first hierarchy and similarly $y_{jk} = 1$ but i and k remain distant in both hierarchies, giving $y_{ik} = 4 > y_{ij} + y_{jk} = 2$.



Network model:

- Fix average #links *z* and "homophily" parameter α
- Repeat until enough links:
  - pick source node *i*
  - pick distance *x* ~ Pr(x) = (1/Z) * $e^{-\alpha x}$
  - pick destination *j* uniformly at random among all nodes distance *x* from *i*

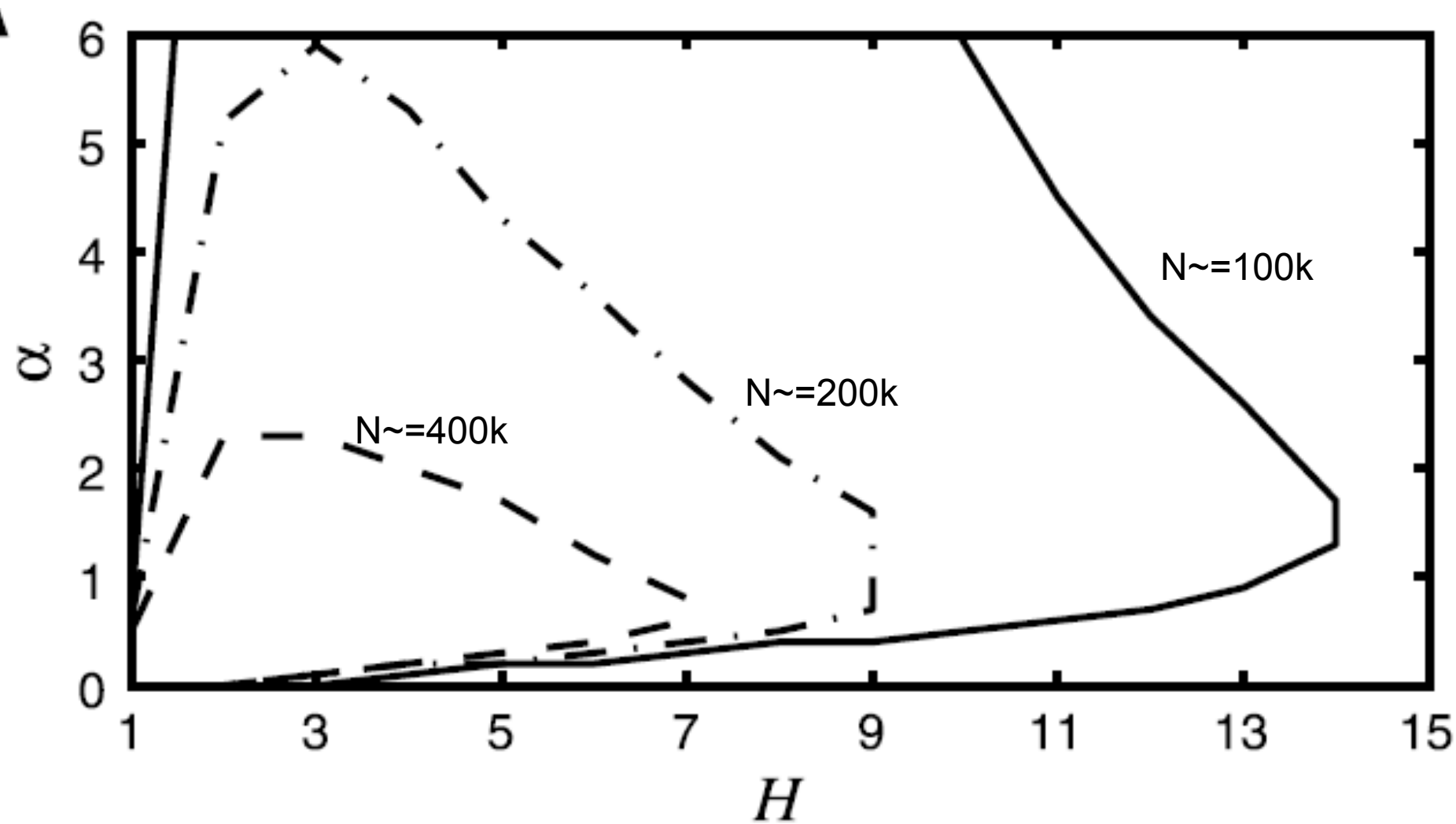


Network model:

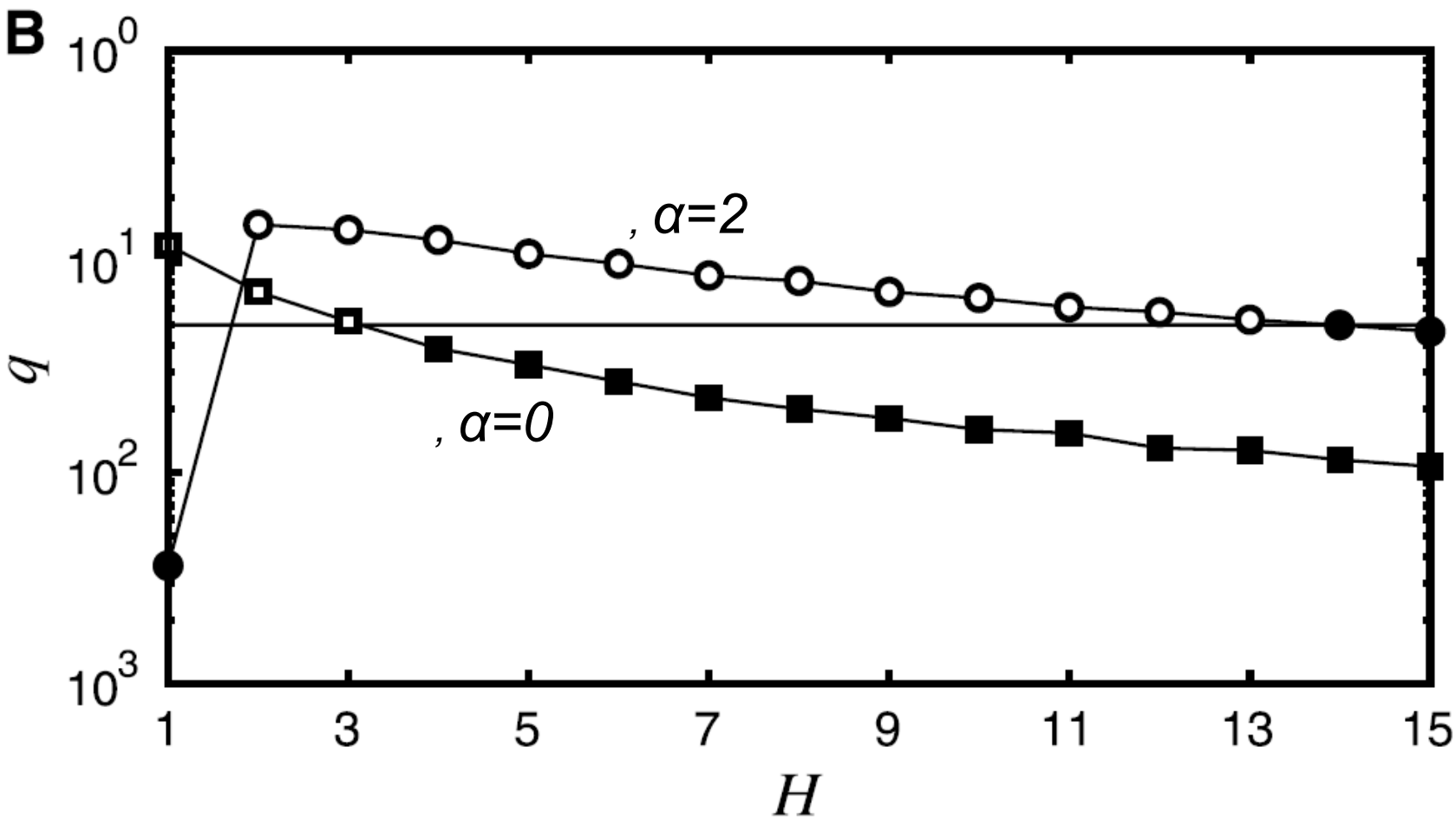
- Fix average #links *z* and "homophily" parameter α
- Repeat until enough links:
  - pick source node *i*
  - pick distance *x* ~ Pr(x) = (1/Z) * $e^{-\alpha x}$
  - pick destination *j* uniformly at random among *all* nodes distance *x* from *i* (in any heirarchy)

Recap

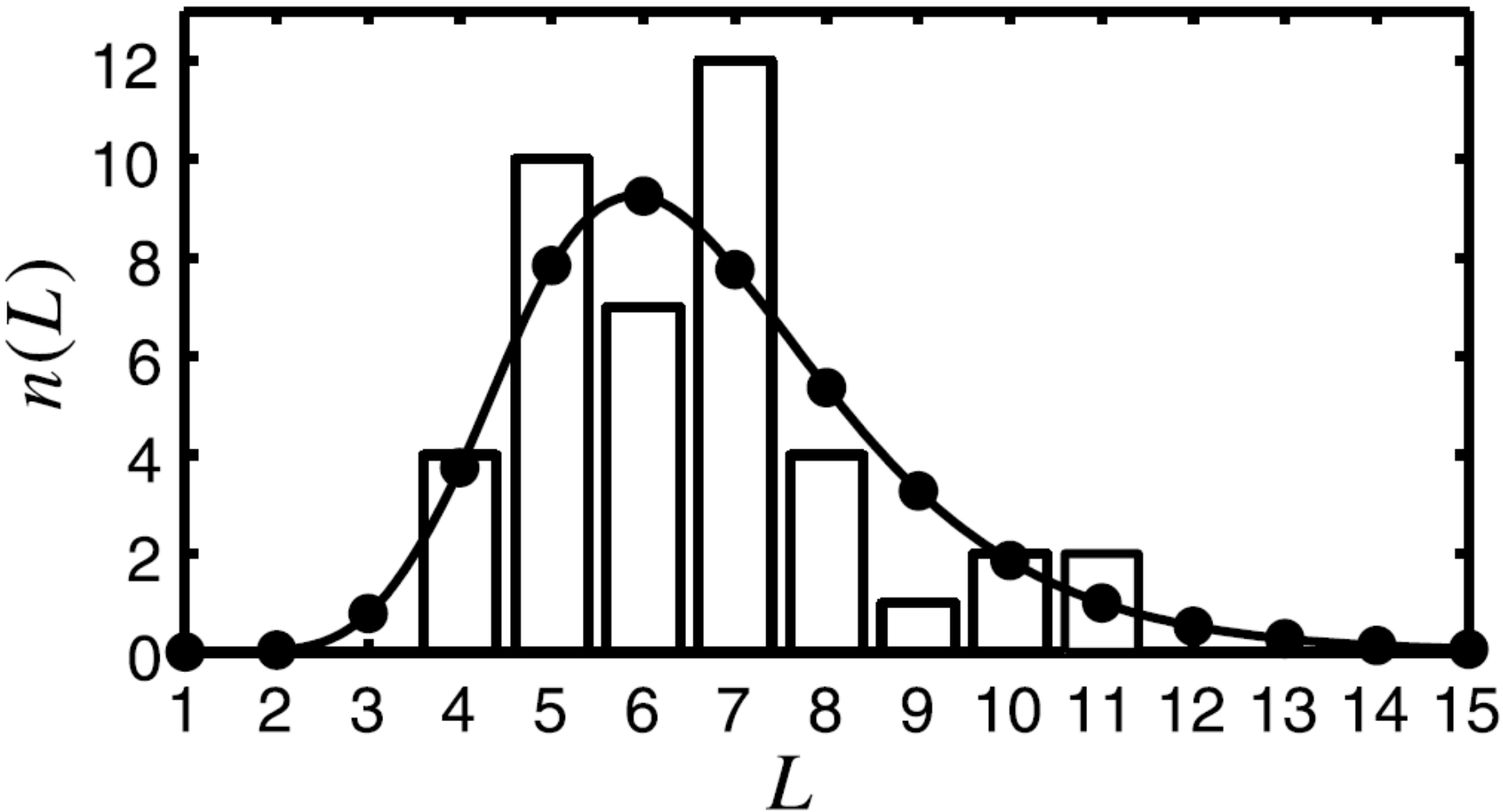
- We now have:
  - A family of networks
  - A distance metric (social distance) for nodes
  - A greedy message-passing algorithm
  - A problem (local navigation)
- Does the [greedy] algorithm work [for local navigation on this class of networks]?
  - Specifically:
    - Fix: *network size N, group size g, mean neighbors z=99, H, α*
    - Let <L> be the average *navigated* path in a random network between randomly-chosen *i* and *j*
    - *Question:* is *q=Pr(*message from *i→j* completes) >= 0.05?
    - Letting p=Pr(message terminates at some stage)=0.25 then this implies <L> small (<= 10.4).



Regions where network is “searchable”



Probability of completion, *q*, for *N=100k*



$N=10^8$, H=2, α=1, g=100, z=300

n(L) based on attrition=0.25 and 1M chain sample