CSC200: Lecture 37

Allan Borodin

Announcements and today lecture

- Announcements
  1. Quiz 7 this Friday. Scope: influence spread in a social network.
  2. The completed Assignment 4 has been posted. There is a question on voting rules (which I plan to start discussing next week).
  3. For next few weeks, office hours by appointment.
  4. I will deal with the following two issues at the end of a class if you bring the quizes with you:
     - ★ If you lost some points in quiz 6 for not explaining your answer (assuming the answer was correct).
     - ★ If you lost a point in assignment 3 for question 3 because the grader assumed it was an undirected network.

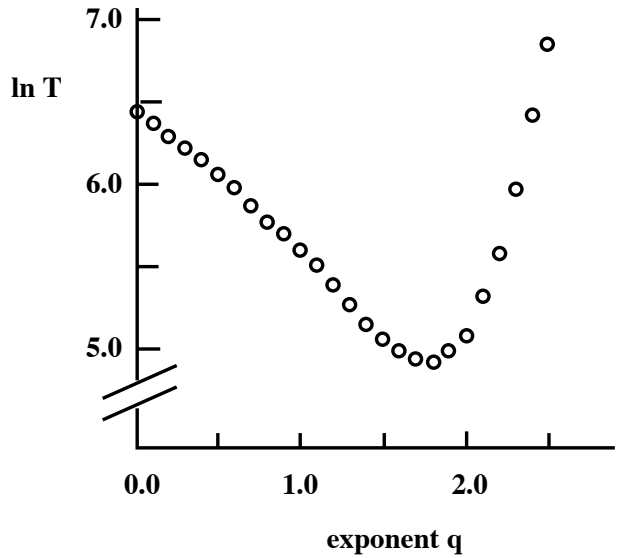
- Lecture outline
  1. Review Watts-Strogatz-Kleinberg model and Kleinberg's analysis of navigation in this small world model.
  2. Liben-Nowell and Backstrom et al studies
  3. Social distance
  4. Adamic and Adar study

A modification of the model

- Random edges outside of ones "close community" are still more likely to reflect some relation to closeness.
- So assume as in the Watts-Strogatz model, from every node v we have edges to all nodes x within some grid distance k from v.
- And now in addition random edges are generated as follows: we (independently) create an edge from v to w with probability proportional to $d(v,w)^{-q}$ where $d(v,w)$ is the grid distance from v to w and $q \geq 0$ is called the clustering exponent.
- The smaller $q \geq 0$ is, the more completely random is the edge whereas large $q \geq 0$ leads to edges which are not sufficiently random and basically keeps edges within or very close to ones community.
- What is the best choice of $q \geq 0$?

So what is a good or the best choice of the clustering coefficient q?

- It turns out that in this 2-dimensional grid model decentralized search works best when $q = 2$. (This is a result that holds and can be proven for the limiting behaviour, in the limit as the network size increases.)



[Fig 20.6, E&K]

  - Simulation of decentralized search in the grid-based model with clustering exponent q.
  - Each point is the average of 1000 runs on (a slight variant of) a grid with 400 million nodes.
  - The delivery time is best in the vicinity of exponent $q = 2$, as expected.
  - But even with this number of nodes, the delivery time is comparable over the range between 1.5 and 2.

More precise statements of Kleinberg's results on navigation in small worlds

The Milgram-like experiment

- Consider a grid network and construct (local contact) directed edges from each node u to all nodes v within grid distance $d(u, v) = k > 1$.
- Also probabilistically construct m (long distance) directed edges where each such edge is chosen with probability proportional to $d(v, w)^{-q}$ for $q \geq 0$.
- We think of k and m as constants and consider the impact of the clustering coefficient q as the network size n increases.
- At every node, we assume we know the directed edges and the location of a target node t.
- The Milgram-like experiment is that at each node we try to move from a node u to a node v that is closest to t (in grid distance).

Navigation in small worlds results

Theorem

(a) *For $0 \leq q < 2$, the (expected) delivery time T of any "decentralized algorithm" in the $n \times n$ grid-based model is $\Omega\left(n^{\frac{2-q}{3}}\right)$.*

(b) *For $q = 2$, there is a decentralized algorithm with delivery time $O(\log n)$.*

(c) *For $q > 2$, the delivery time of any decentralized algorithm in the grid-based model is $\Omega\left(n^{\frac{q-2}{q-1}}\right)$.*

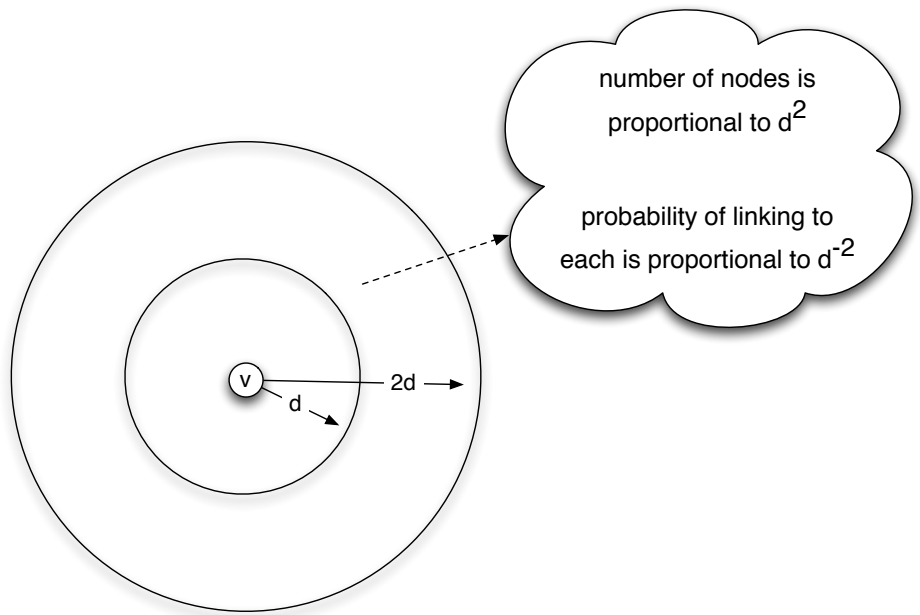
(The lower bounds in (a) and (c) hold even if each node has an arbitrary constant number of long-range contacts, rather than just one.)

Notes

"Big O" and "big Omega" mean asymptotic behaviour as a function of n.
Note: In Figure 20.6, $n = 20,000$ so that $n^{1/3} \approx 27$.

Intuition as to why $q = 2$ is best for grid

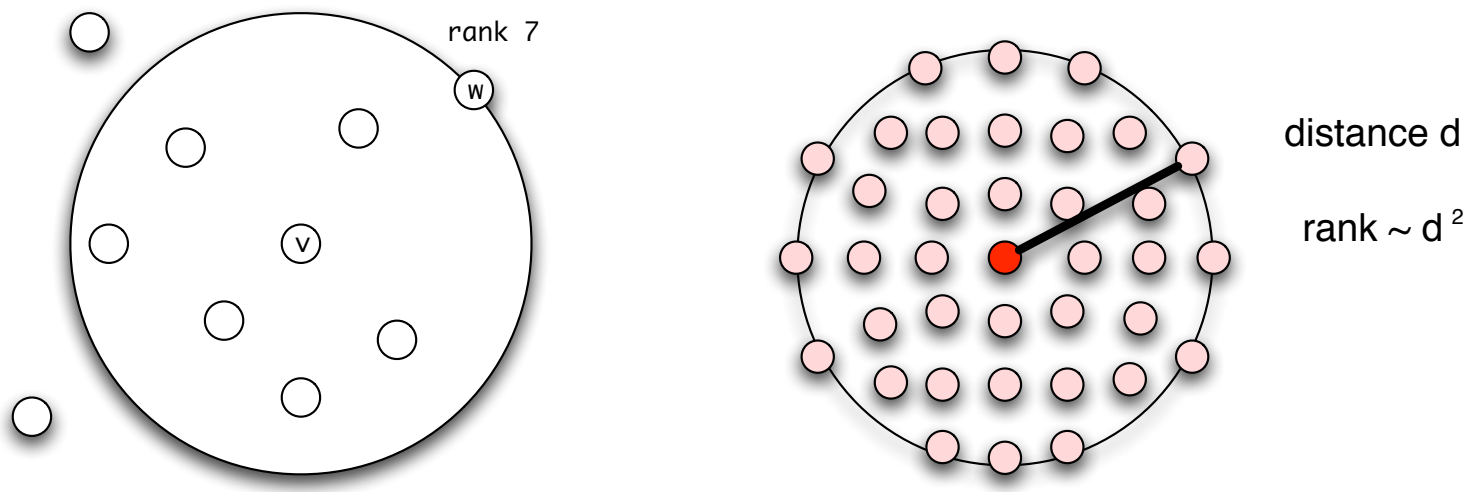
- It is instructive to see why this choice of q provides links at the different "scales of resolution" seen in the Milgram experiment.
- That is, if D is the maximum distance to be travelled, then we would like links with distances between d and $2d$ for all d
- Given that we have a 2-dimensional grid, the number of points with distance say d from a given node v will be $\approx d^2$.
- We are choosing such a node with probability proportional to $1/d^2$ and hence we expect to have a link to some node whose distance from v is between d and $2d$ for all d.



[Fig 20.7, E&K]

More realistic (nonuniformly spread population) data

- In the grid model, the population density is completely uniform which is not what one would expect in real data.
- How can this $1/d^2$ (inverse-square) distribution be modified to account for population densities that are very non-uniform?
- The idea is to replace distance $d(v, w)$ from v to w by the rank of w relative to v.
  - For a fixed v, define the $rank(w)$ to be the number of nodes closer to v than w.
  - In the 2D grid case, when $d(v, w) \sim d$, then $rank(w) \sim d^2$.

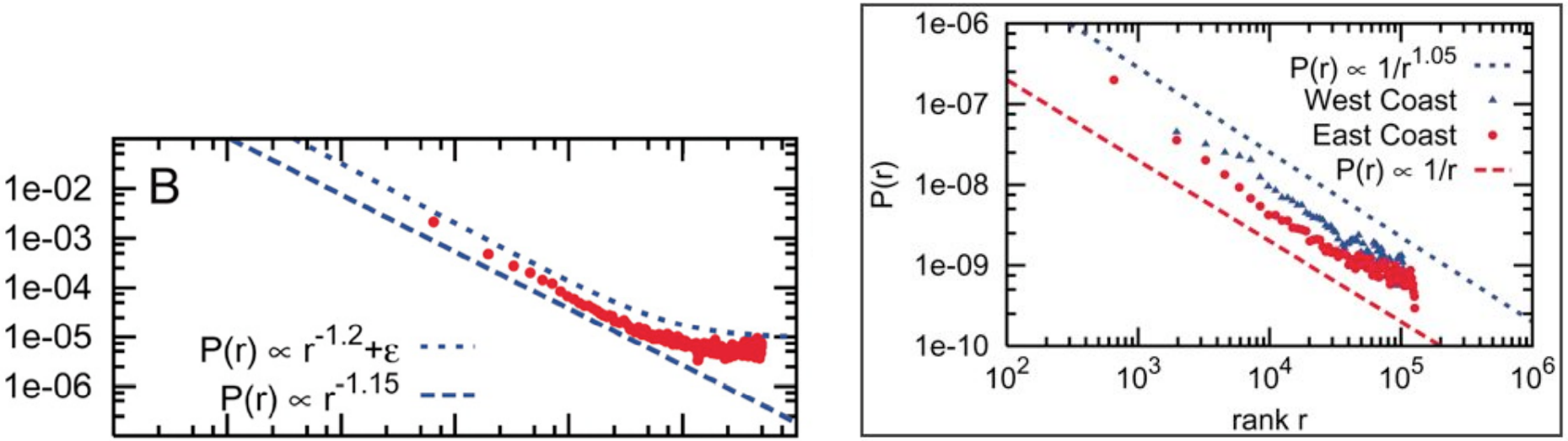


[Fig 20.9, E&K]

More realistic data continued

- We can then restate the inverse-square distribution by saying that the probability that v links to w is proportional to $1/rank(w)$.
- Using zip code information, for every pair of nodes (500,000 users on Live-Journal) one can assign ranks.
- Liben-Nowell et al did such an study in 2005, and then for different rank values examined the fraction of edges that are actually friends.
- The theory tells us that this fraction f should be a decreasing function proportional to $1/rank$.
- That is, $f \sim rank^{-1}$. Taking logarithms, $\log f \sim (-1) \log rank$.

More realistic (LiveJournal) friendship data



(a) *Rank-based friendship on LiveJournal*

(b) *Rank-based friendship: East and West coasts*

[Fig 20.10, E&K]

- In Figure 20.10 (a), the Lower (upper) line is exponent $= -1.15$ (resp. -1.12).
- In Figure 20.10 (b), the Lower (upper) line is exponent $= -1.05$ (resp. -1). The red data is East Coast data and the blue data is West Coast data.

The Backstrom et al study

- Backstrom et al study US Facebook user data (from around 2010
- The Facebook data:
  1. Roughly 100 million users
  2. About 6% of which enter home address info and of that population about 60% can be parsed into longitude and lattiude information.
  3. This gave a set of 3.5 million users (of which 2.9 million had at least one friend with a well specified address and each of these 2.9 million users had an average of 10 friends with specified addresses resulting in 30.6 million edges.
  4. Although a small part of facebook, this 2.9 million person "geolocated data set" is sufficently large and representative for experimental study.

The Backstrom et al study

- Backstrom et al study US Facebook user data (from around 2010
- The Facebook data:
  1. Roughly 100 million users
  2. About 6% of which enter home address info and of that population about 60% can be parsed into longitude and lattiude information.
  3. This gave a set of 3.5 million users (of which 2.9 million had at least one friend with a well specified address and each of these 2.9 million users had an average of 10 friends with specified addresses resulting in 30.6 million edges.
  4. Although a small part of facebook, this 2.9 million person "geolocated data set" is sufficently large and representative for experimental study.
- They study probability of friendships vs distance and rank and how those probabilities depend on population densities for where people live. As previously mentioned this study provides more evidence as to the power law relation between distance/rank and probability of friendship.

The Backstrom et al study

- Backstrom et al study US Facebook user data (from around 2010
- The Facebook data:
  1. Roughly 100 million users
  2. About 6% of which enter home address info and of that population about 60% can be parsed into longitude and lattiude information.
  3. This gave a set of 3.5 million users (of which 2.9 million had at least one friend with a well specified address and each of these 2.9 million users had an average of 10 friends with specified addresses resulting in 30.6 million edges.
  4. Although a small part of facebook, this 2.9 million person "geolocated data set" is sufficently large and representative for experimental study.
- They study probability of friendships vs distance and rank and how those probabilities depend on population densities for where people live. As previously mentioned this study provides more evidence as to the power law relation between distance/rank and probability of friendship.
- Furthermore, using the distribution of ones friends, they can better predict (than say using IP information) geographic locations!

Who provides address information on Facebook

Table 1: Demographic Statistics of Geolocated Users

	Located	All US Users
% Male	57.51%	44.81%
% Female	42.49%	55.19%
Age, Median	30	30
Age, Mean	33.89	33.44
Account Age (days), Median	413	325
Account Age (days), Mean	558.9	453
Friend Count, Median	105	47
Friend Count, Mean	189.4	129.5

[Table 1 from Backstrom et al]

- What is noticeable about this data?

Probability of friendship wrt. distance

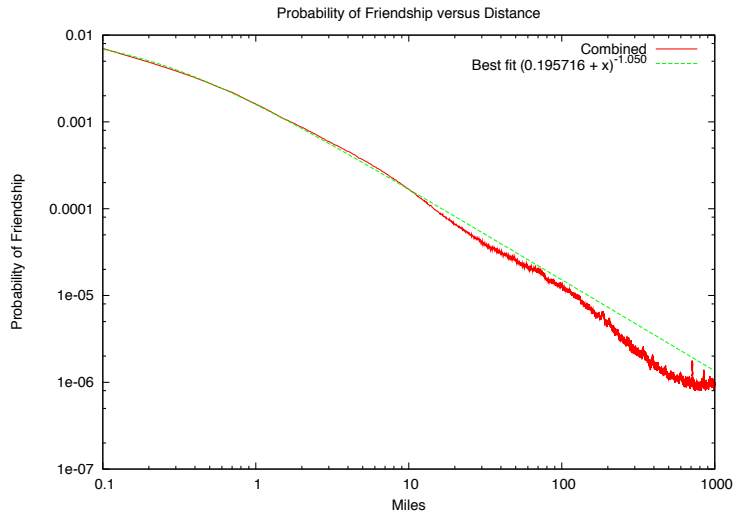


Figure 7: Probability of friendship as a function of distance. By computing the number of pairs of individuals at varying distances, along with the number of friends at those distances, we are able to compute the probability of two people at distance d knowing each other. We see here that it is a reasonably good fit to a power-law with exponent near -1.

[Figure 7 from Backstrom et al]

Probability of friendship wrt. distance relative to population density

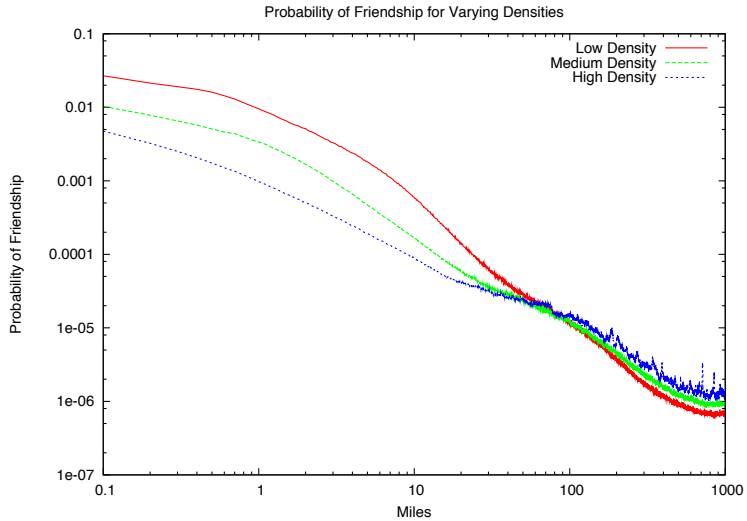
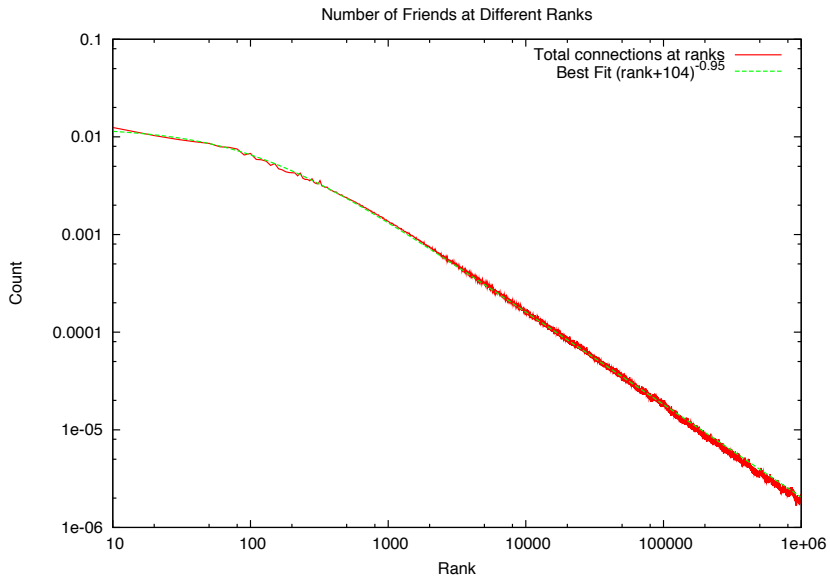


Figure 8: Looking at the people living in low, medium and high density regions separately, we see that if you live in a high density region (a city), you are less likely to know a nearby individual, since there are so many of them. However, you are more likely to have contact with someone far away.

[Figure 8 from Backstrom et al]

Number of friends wrt. rank



[Figure 9 from Backstrom et al]

From geographic distance to social distance

- What if there is no (reliable) distance information in a social network?

From geographic distance to social distance

- What if there is no (reliable) distance information in a social network?

- It is, of course, natural that we tend to have more common interests with people who live closer to us (e.g. based on ethnicity, economic status, etc), but clearly there are other notions of social distance that should be considered.

From geographic distance to social distance

- What if there is no (reliable) distance information in a social network?
- It is, of course, natural that we tend to have more common interests with people who live closer to us (e.g. based on ethnicity, economic status, etc), but clearly there are other notions of social distance that should be considered.
- Early in the course we considered social foci (clubs, shared interests, language, etc.) and we all tend to share a number of focal interests with the same person.
- But, of course, belonging to a small group of people in a course, is different than attending the same University, and speaking Mandarin is different than being interested in Esperanto.

From geographic distance to social distance

- What if there is no (reliable) distance information in a social network?
- It is, of course, natural that we tend to have more common interests with people who live closer to us (e.g. based on ethnicity, economic status, etc), but clearly there are other notions of social distance that should be considered.
- Early in the course we considered social foci (clubs, shared interests, language, etc.) and we all tend to share a number of focal interests with the same person.
- But, of course, belonging to a small group of people in a course, is different than attending the same University, and speaking Mandarin is different than being interested in Esperanto.
- So the suggestion is made that we can define social distance $s(v, w)$ between two individuals v, w to be the smallest size foci they share.

Smallest size shared foci as distance

- Kleinberg gives theoretical results indicating that when friendships follow a distribution proportional to $1/s(v, w)$ then the resulting social network will support efficient decentralized search.

- This is somewhat verified in a study (by Adamic and Adar) of 'who talks to whom' friendship data (based on frequency of email exchanges) amongst a small group of HP employees.

- The focal groups are defined by the organizational hierarchy of the company.

- The Adamic and Adar study shows that the distribution for this friendship relationship is proportional to the inverse of $s(v, w)^{-3/4}$ so that it doesn't match as closely with the previous geographical rank based results but still observes a power law relation governing how social ties decrease with "distance".

Probability of email exchanges vs distance in the organization hierarchy

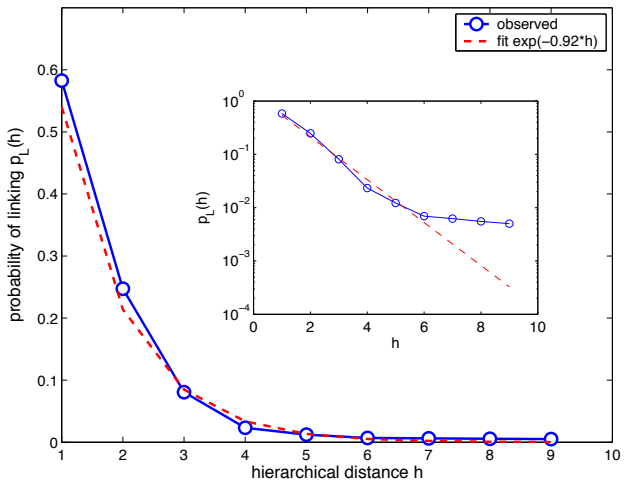


Figure 4: Probability of linking as a function of the separation in the organizational hierarchy. The exponential parameter $\alpha = 0.92$, is in the searchable range of the Watts model [13]

[Figure 4 from Adamic and Adar]

Probability of email exchanges vs size of smallest common organizational unit

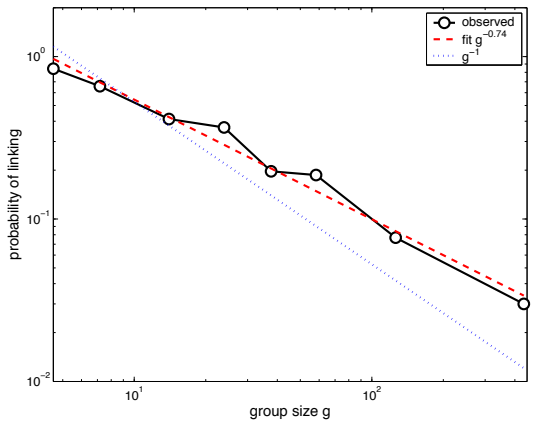


Figure 5: Probability of two individuals corresponding by email as a function of the size of the smallest organizational unit they both belong to. The optimum relationship derived in [7]is $p \sim g^{-1}$, g being the group size. The observed relationship is $p \sim g^{-3/4}$.

[Figure 5 from Adamic and Adar]

Final observations in chapter

- The text suggests viewing the Milgram experiment as an example of decentralized problem solving (in this case solving a shortest path problem). Advertisement for distributed systems course.

- The text asks what other problem solving tasks might be amenable to such decentralized problem solving and how to analyze what can be done especially in large online networks.

- Finally the text briefly suggests the role of social status in determining the effectiveness of reaching a given target.
  - An email forwarding Milgram type study by Dodds et al shows that completion rates to all targets were low but were highest for "high status" targets and particularly small for "low status" targets.

- In section 12.6, the text speculates on structural reasons for the impact of status. This discussion leaves me with the sense that we are far from having any comprehensive understanding of such phenomena.

Redux: The punch line of the chapter, text, course

The plots in Figure 20.10, and their follow-ups, are thus the conclusion of a sequence of steps in which we start from an experiment (Milgrams), build mathematical models based on this experiment (combining local and long-range links), make a prediction based on the models (the value of the exponent controlling the long-rang links), and then validate this prediction on real data (from LiveJo urnal and Facebook, after generalizing the model to use rank-based friendship). This is very much how one would hope for such an interplay of experiments, theories, and measurements to play out. But it is also a bit striking to see the close alignment of theory and measurement in this particular case, since the predict predictions come from a highly simplified model of the underlying social network, yet these predictions are approximately borne out on data arising from real social networks.

[From E&K Ch.20, p.549]