Social and Information Networks

University of Toronto CSC303
Winter/Spring 2024

Week 5: Feb 5-9

This week's high-level learning goals

- Recall the Milgram Small World Experiment, and be able to explain both the term six degrees of separation, and their connection to small world phenomena
- Define the Watts-Strogatz model, and the changes required to make it a small world
- Perform decentralized search and centralized search in a given graph
- Explain the differences between decentralized and centralized search, including how the pertain to small worlds
- Recall conditions that provably allow for efficient decentralized search
  - in a grid
  - under non-uniform population density
- Explain why the results of the empiric studies on distributions of friendship seen are consistent with theory, and why their methodology is applicable
- At a very high level, explain one practical application of friendship distributions

The Small World Phenomenon (Chapter 20)

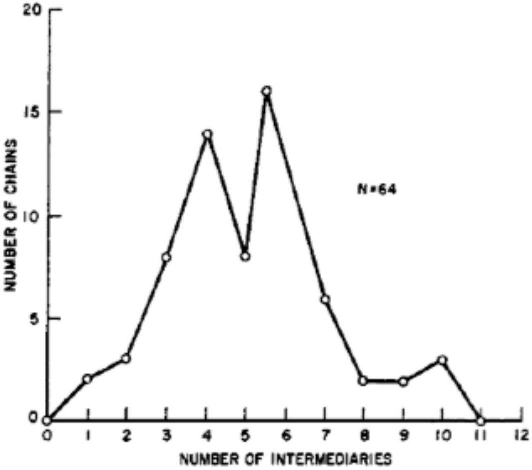
- We now move from a study of selection, influence, and balance in networks, to the issue of *focused or targeted search*.
  - Popularized in the famous concept of "six degrees of separation".

- At the start of this course, we briefly discussed the original 1960s Milgram experiment as it was introduced in Chapter 2 of the text.

- Milgram asked 296 randomly chosen people in Omaha to forward a letter to a target person (a stockbroker) living in a Boston suburb.

- Of the 64 chains that succeeded the median length of the letter chain was 6, the motivation for the play and movie that came to popularize the phenomena.

Lengths of the successful letter chains



- From Milgram (1967), "The Small World Problem," Psychology Today [297]

[Fig 20.4, E&K]

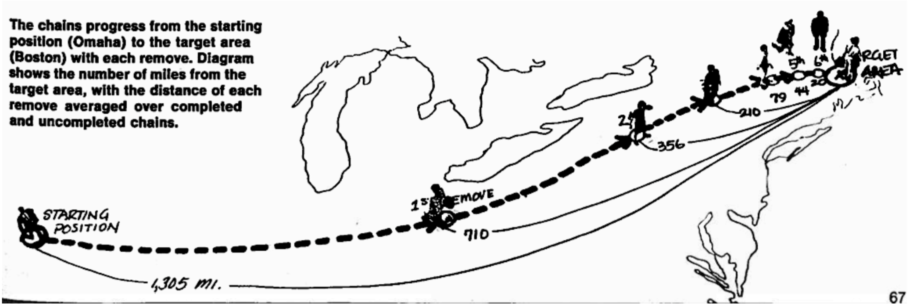


Image from Milgram (1967)

- Milgram's diagram showing a "composite" of the successful paths converging on the target person
  - Each intermediate step is positioned at the average distance of all chains that completed that number of steps
  - Anything interesting about the spacing?

Two remarkable aspects of experiment

1. There are short paths (of friendship) between seemingly very unrelated people
   - We've seen this phenomena with the Erdos number (amongst mathematicians or all scientists) and Bacon number (amongst actors)
   - e.g., Week 1: the Oracle of Bacon uses a centralized search – i.e., BFS on the graph of the social network to find a shortest path
     - ★ Were Milgram's human participants doing this?
2. The Milgram letter chain succeeded without individuals knowing anything globally about the network structure
   - i.e., without any centralized coordination, individuals were reasonably successful in reaching the target using only geographic and occupational information

- Chapter 20 studies how we can better understand this interesting phenomena.

Looking ahead: The punch line of the chapter, text, course

... we start from an experiment (Milgram's), 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 LiveJournal 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 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]

Trying to find a path of friendships to someone

- Given the full social network, we can use a centralized search (e.g., run BFS on the graph). Problems?

- We could ask all of our friends to tell all of their friends to tell all of their friends. . . (i.e. a traditional chain letter) that I am looking for person X.

- Now say assuming your online social network has a "broadcast to all" feature, this can be done easily but it has its drawbacks. Drawbacks?
  - It either costs real money/effort to pass a message (e.g. postal mail)
  - I would prefer to not let everyone know that I am looking for person X
  - Possible "social cost" in terms of annoyance to people in the network receiving multiple requests to pass on a message.

- Clearly if everyone cooperates, the broadcast method ensures the shortest path to the intended target X in the leveled tree/graph of reachable nodes.

Reachable nodes without triadic closure

- If we assume the social network has a tree structure (therefore, no triadic closure), then it follows that every simple path is a shortest path to everyone in the network.
- Consider the number of people that you could reach by a path of length at most t if every person has say at least 5 friends.

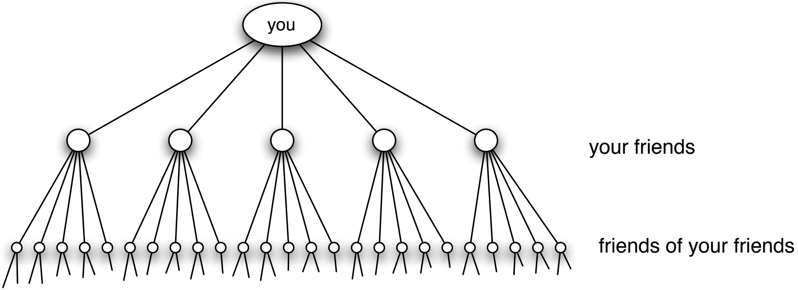


Figure: Pure exponential growth produces a small world [Fig 20.1 (a), E&K]

Reachable nodes with triadic closure

- Given that our friends tend to be mostly contained within a few small communities, the number of people reachable will be much smaller.

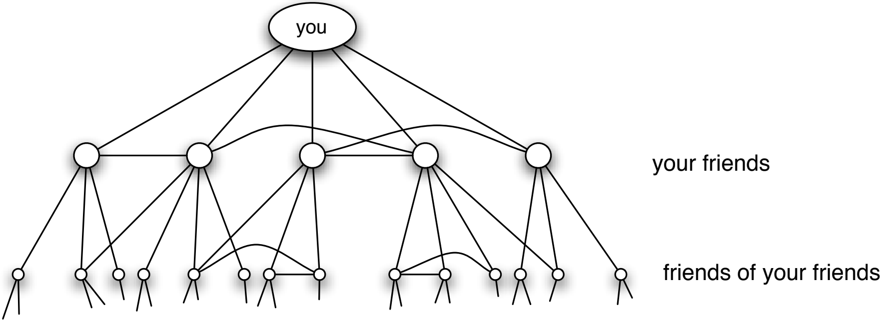


Figure: Triadic closure reduces the growth rate [Fig 20.1 (b), E&K]

The Watts-Strogatz model

- Is it possible to have extensive triadic closure and still have short paths?

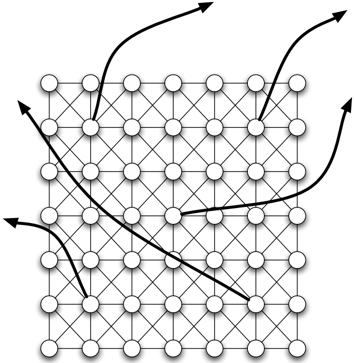
- Consider dense communities of strong ties consistent with triadic closure (possibly homophily?), and different communities attached by weak ties
  - Weak ties provide the kind of branching that yields short paths to many nodes

- Watts-Strogatz model: A stylized model with two types of ties:
  - Nodes lie in a two dimensional grid
  - Short-range edges connect all nodes within some small distance r
    - ★ Why? Short-range edges capture an idealized sense of homophily
  - A small number of random longer-distance edges to other nodes in the network
    - ★ Very few random edges are needed to achieve the effect of short paths

- Aside: This is actually a variant of the Watts-Strogatz model, but the core idea is the same

Very few random edges are needed

- A $1/k$ probability that a person has a random weak tie is sufficient to establish short paths, for a sufficiently large grid



[Fig 20.3, E&K]

- Image subdividing the grid into k by k "towns", each with $k^2 \times \frac{1}{k} = k$ long distance edges (on average)

Question: Are short paths enough to explain the Miligram experiment?

Does this explain the ability to find people in a decentralized manner?

- In the Watts-Strogatz type of model, we can use the random edges (in addition to the short grid edges) and the geometric location of nodes to keep trying to reduce the grid distance to a target node
  - Analogous to the Milgram experiment where individuals seem to use geographic information to guide the search
  - However, completely random edges does not reflect real social networks!

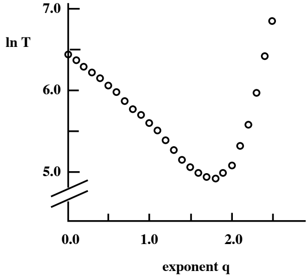
- Having uniformly random edges will not work in general as:
  - Completely random edges (i.e. going to a random node anywhere in the network) are too random.
  - A random edge in an $n \times n$ grid is likely to have grid distance $\Theta(n)$.
  - Without some central guidance, such random edges will essentially just have us bounce around the network causing a substantially longer path to the target than the shortest path.

A modification of the model

- Random edges outside of ones "close community" represent weak ties, thus it seems like they should reflect some relation to closeness.
- As in the Watts-Strogatz model, from every node v we have edges to all nodes x within some grid distance r from v.
- However, the random edges are instead generated as follows:
  - We independently create an edge from v to w with probability proportional to $d(v, w)^{-q}$
  - $d(v, w)$ is the grid distance from v to w
  - $q \geq 0$ 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 exponent q?

- It turns out that in this 2-dimensional grid model decentralized search works best when $q = 2$
  - Provably optimal, in the limit as the network size increases
  - What about in practice where the network is finite?



[Fig 20.6, E&K]

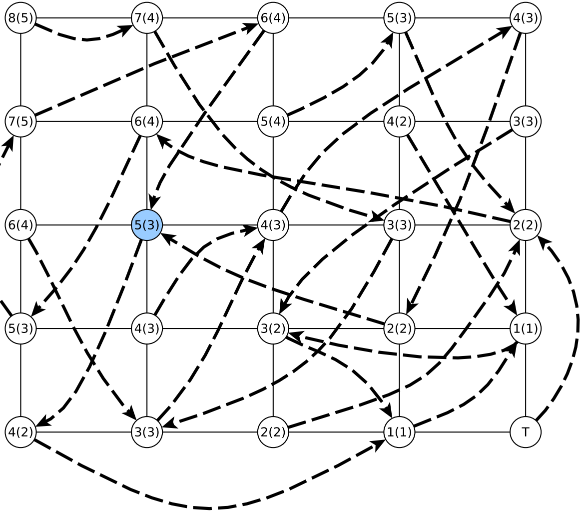
  - Simulation of decentralized search in the grid-based model with clustering exponent q
  - Each point is the average of 1000 runs on (a slight variant of) a grid with 400 million nodes
  - The delivery time is best in the vicinity of exponent $q = 2$, as expected
  - But even with this number of nodes,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

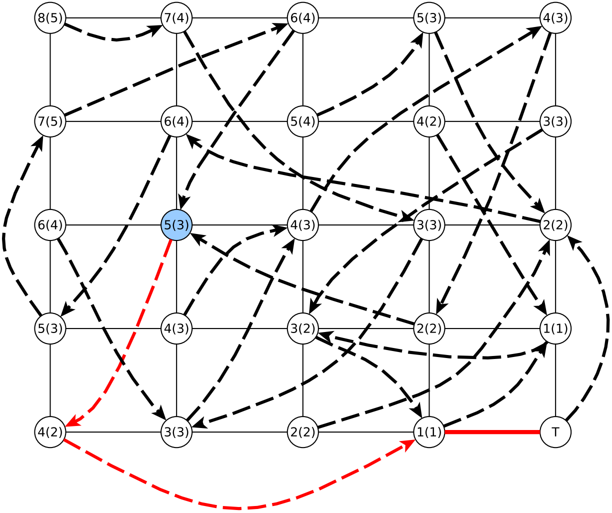
- Consider a grid network and construct (local contact) directed edges from each node u to all nodes v within grid distance $d(u, v) = k > 1$.
- For each node u we also probabilistically construct $m > 0$ (long distance) directed edges where each such edge is chosen with probability proportional to $d(u, w)^{-q}$ for $q \geq 0$.
- We think of k and m as constants and consider the impact of the clustering exponent q as the network size n increases.
- We assume that each node knows its location and the location of its adjacent edges and its distance to the location of a target node t.
- The Milgram-like experiment is to produce a path to node t from node s, where each node on the path is followed by it's neighbouring node v that is closest to t (in *grid* distance).

Example for $m = 1$



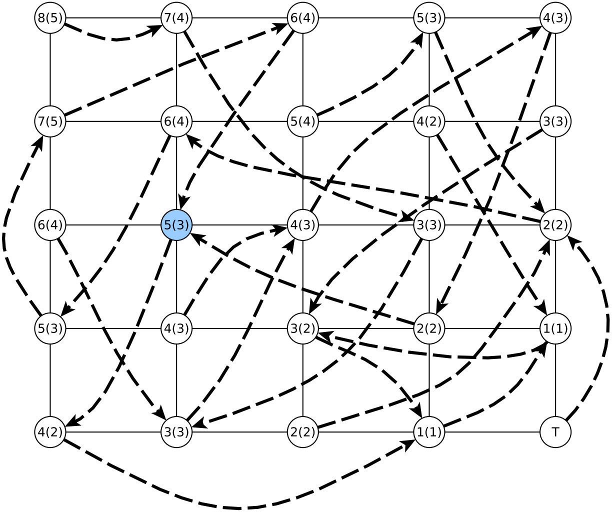
Note: Numbers are numbered with *grid distance (optimal distance)*
Although normally not allowed, for legibility $k = 1$

Example: Shortest Path



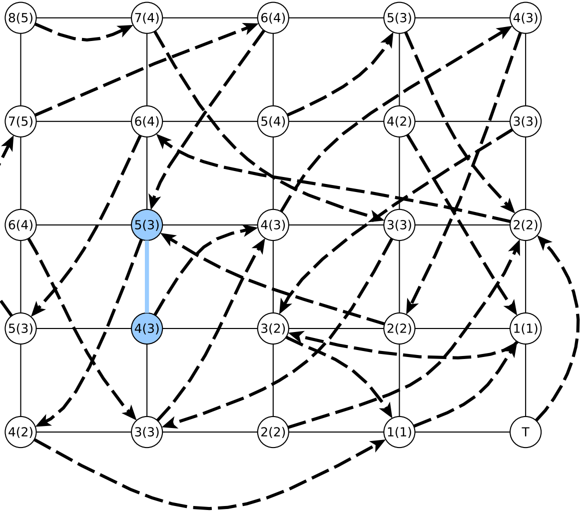
Note: Numbers are numbered with *grid distance (optimal distance)*
Although normally not allowed, for legibility $k = 1$

Example: Decentralized Search



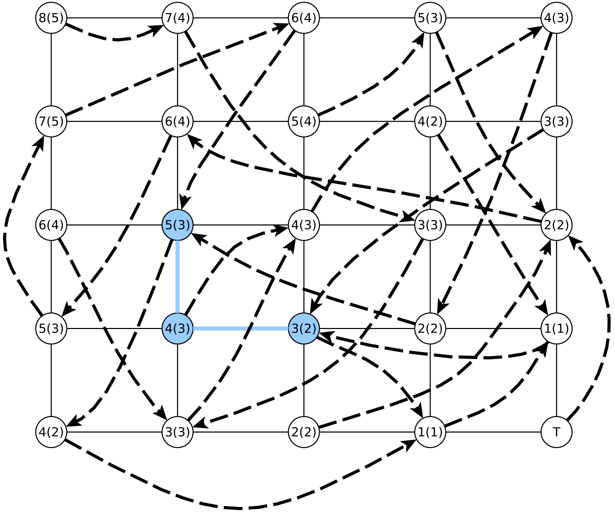
Note: When multiple neighbours have the same grid distance, tie breaking is random

Example: Decentralized Search



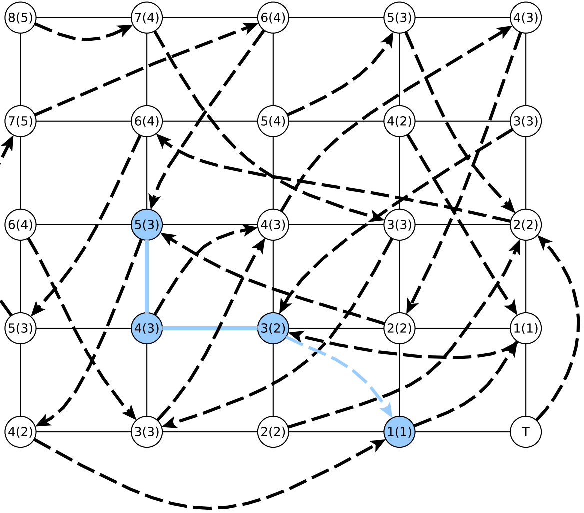
Note: When multiple neighbours have the same grid distance, tie breaking is random

Example: Decentralized Search



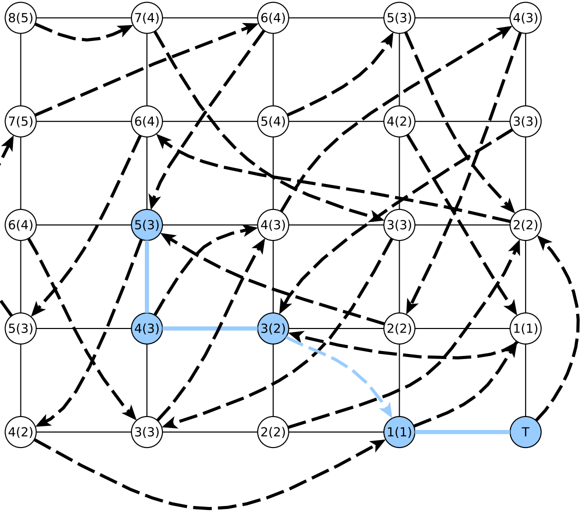
Note: When multiple neighbours have the same grid distance, tie breaking is random

Example: Decentralized Search



Note: When multiple neighbours have the same grid distance, tie breaking is random

Example: Decentralized Search



Note: When multiple neighbours have the same grid distance, tie breaking is random

Reflection on the Kleinberg-Milgram model

As we said at the start of this topic, the real surprise is that a “short” (but not shortest) path is (probably w.r.t. to the randomly generated network) being found by a decentralized search.

It is true that each node will pursue a “greedy strategy” but this is different than say Dijkstra’s least cost/distance algorithm which entails a centralized search.

Navigation in small worlds results

Theorem

(J. Kleinberg 2000)

(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.)

Aside: Clustering coefficient

NOTE: It is no accident that the exponent in the Strogatz-Kleinberg model is called the "clustering exponent".

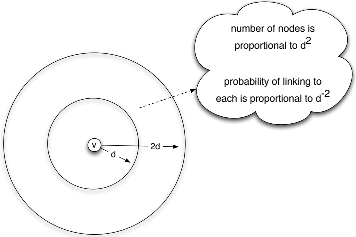
Recall (from chapter 2) the definition of the *clustering coefficient* of a node which is the ratio:

$$\frac{\left|\left\{(B,C) \in E : (B,A) \in E \text{ and } (C,A) \in E\right\}\right|}{\left|\left\{\{B,C\} : (B,A) \in E \text{ and } (C,A) \in E\right\}\right|}$$

- As the the clustering exponent increases, ones friends are all close and therefore (in this geometric model), ones friends are mutual friends which means the clustering coefficient goes to 1
- And when the clustering exponent goes to 0, friends are randomly scattered and unlikely to be mutual friends so that the clustering coefficient goes to zero

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

- It is instructive to see why this choice of q provides links at the different "scales of resolution" seen in the Milgram experiment.
- If D is the maximum distance to be travelled, then we would like links with distances between d and $2d$ for all $d < \log D$
  - You'll be seeing why in tutorial
- Given that we have a 2-dimensional grid, the number of points with distance say d from a given node v will be $\propto d^2$.
- 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.
  - Area of a circle: πr^2
  - $\pi(2d)^2 - \pi d^2 = 3\pi d^2 \propto d^2$



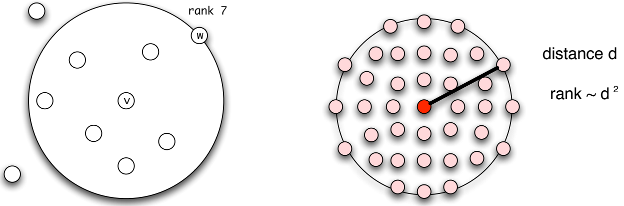
[Fig 20.7, E&K]

Recap

- Milgram Small World Experiment
- Watts-Strogatz model
- Efficient decentralized search
  - in a grid

More realistic population (non-uniform density)

- 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 is to v.
  - In the 2D grid case, when $d(v, w) \sim d$, then $rank(w) \sim d^2$.

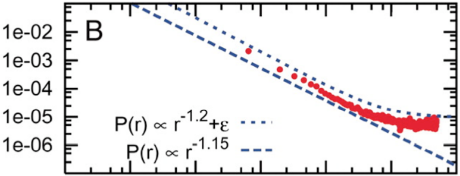


[Fig 20.9, E&K]

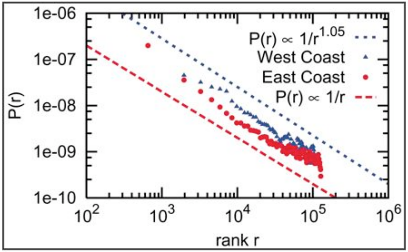
More realistic geographic data continued

- 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 the blogging site LiveJournal) one can assign ranks.
- Liben-Nowell et al did such a study in 2005, and then for different rank values examined the fraction f 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.

Liben-Nowell: practice closely matches theory

Liben-Nowell prove that for "essentially" any population density (i.e. no matter where people are located) if links are randomly constructed so that the probability of a friendship is proportional to $rank^{-1}$, then the resulting network is one that can be efficiently searched in a decentralized manner.

That is, Kleinberg's result for the grid generalizes. This is a rather exceptional result in that the abstraction from d^{-2} to $rank^{-1}$ is not at all an obvious generalization.

How surprised should we be that natural populations locate themselves in this probabilistic manner since there is no centralized organizing mechanism that is causing this phenomena?

The text refers to a 2008 article by Oscar Sandberg who analyzes a network model where decentralized search takes place which in turn causes links to "re-wire" so as to fascilitate more efficient decentralized search.

It remains an intriguing question as to the extent this does happen in social networks and the implicit mechanisms that would cause networks to evolve this way.

The punch line (again) of 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 (Milgram's), 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 LiveJournal 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 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.

- And not clear why real friendships are so arranged.

The Backstrom et al rank-based study

- Backstrom et al study US Facebook 2010 geographic user 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 latitude information.
  3. This gave a set of 3.5 million users and 30.6 million edges
     - ⋆ 2.9 million had at least one friend with a well specified address (these averaged 10 friends with a known address)
  4. Although a small part of Facebook, this 2.9 million person "geolocated data set" is sufficiently large and representative for experimental study.
- Studied probability of friendships vs distance and rank and how those probabilities depend on population densities for where people live
  - This study provides more evidence as to the approximately inverse relation between distance/rank and probability of friendship ($\approx rank^{-.95}$)
  - This relation is known as a power law

  Question: What can we do with this knowledge?
- They utilize this relationship between friends and distance to create an algorithm that will predict the location of an individual from a small set of users with known locations. They claim their algorithm can predict geographic locations better than using IP information!

Some statistics for geolocated data

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

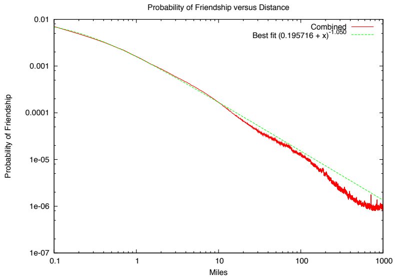


Figure 7: Probability of friendship as a function of distance. By computing the number of pairs of individuals at varying distances, along with the number of friends at those distances, we are able to compute the probability of two people at distance d knowing each other. We see here that it is a reasonably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

[Figure 7 from Backstrom et al]

- Interestingly, w.r.t. distance we still get a power law relation!
  - The exponent is -1 instead of -2, but this is not surprising given the non-uniform population

Probability of friendship wrt. distance relative to population density

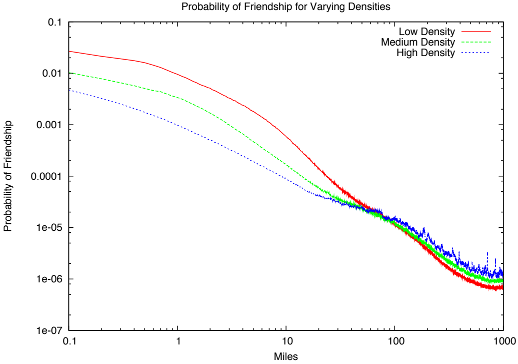
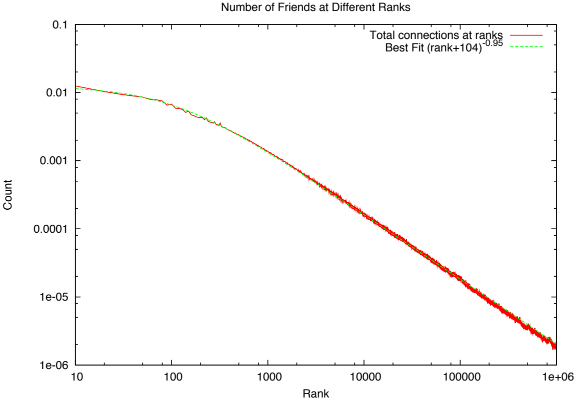


Figure 8: Looking at the people living in low, medium and high density regions separately, we see that if you live in a high density region (a city), you are less likely to know a nearby individual, since there are so many of them. However, you are more likely to have contact with someone far away.

Number of friends wrt. rank



[Figure 9 from Backstrom et al]

- With respect to rank, the exponent very close to the optimal -1 predicted by Liben-Nowell

Predicting locations

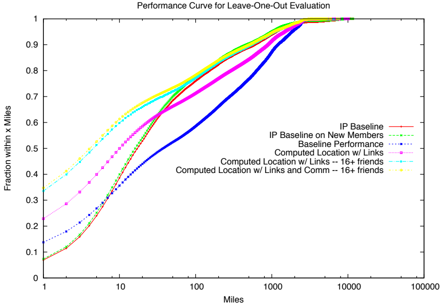


Figure 11: Location Prediction Performance. This figure compares external predictions from an IP geolocation service, the same service constrained to users who have recently updated their address, a baseline of randomly choosing the location of a friend, along with three predictions: our algorithm with all links, for users with 16+ friends, and finally for users with 16+ friends constraining to only those with whom they have communicated recently.

[Figure 11 from Backstrom et al]

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.) we 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 define social distance $s(v, w)$ between individuals v, w to be the minimum size of a common focus.

Smallest size shared focus as a distance measure

- Kleinberg (2001) 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 2005 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 social distance

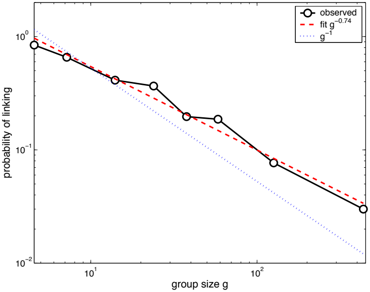


Figure 5: Probability of two individuals corresponding by email as a function of the size of the smallest organizational unit they both belong to. The optimum relationship derived in [7]is $p \sim g^{-1}$, g being the group size. The observed relationship is $p \sim g^{-3/4}$.

[Figure 5 from Adamic and Adar]

Aside: Probability of email exchanges vs distance in the organizational hierarchy

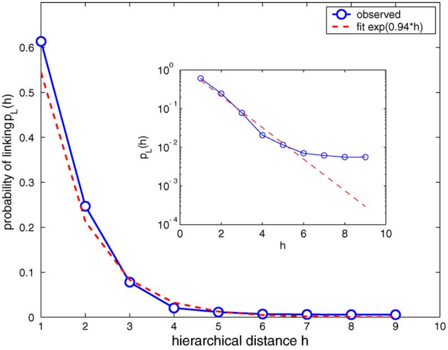


Fig. 4. Probability of linking as a function of the separation in the organizational hierarchy. The exponential parameter $\alpha = 0.94$, is in the searchable range of the Watts model (Watts et al., 2002).

[Figure 4 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).

- 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 2003 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, however, we are far from having a 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 (Milgram's), 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 LiveJournal 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]

Recap

With practice & review, you'll be able to:

- Recall the Milgram Small World Experiment, and be able to explain both the term six degrees of separation, and their connection to small world phenomena
- Define the Watts-Strogatz model, and the changes required to make it a small world
- Perform decentralized search and centralized search in a given graph
- Explain the differences between decentralized and centralized search, including how the pertain to small worlds
- Recall conditions that provably allow for efficient decentralized search
  - in a grid
  - under non-uniform population density
- Explain why the results of the empiric studies on distributions of friendship seen are consistent with theory, and why their methodology is applicable
- At a very high level, explain one practical application of friendship distributions