CSC200: Lecture 36

Allan Borodin

Announcements and today lecture

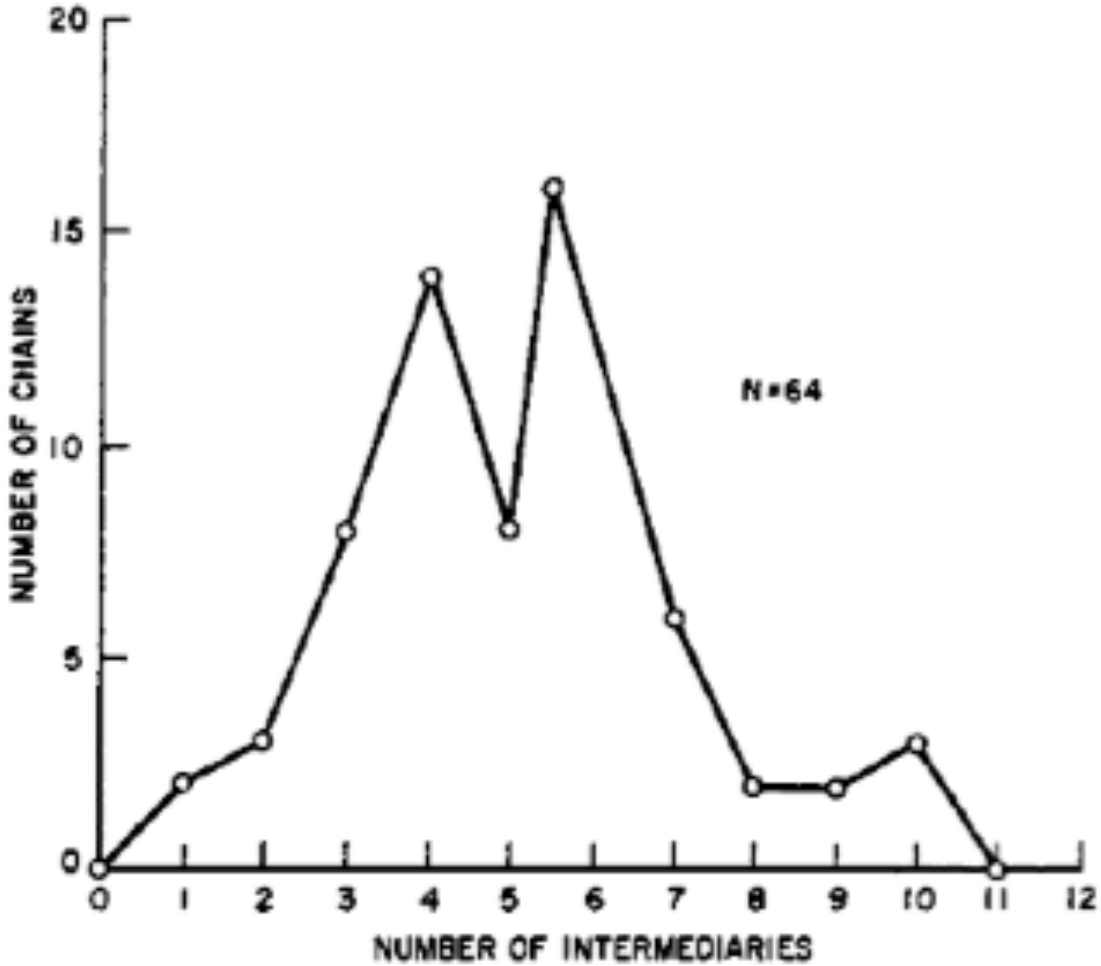
- Announcements
  1. Quiz 7 this Friday, March 11. Quiz will cover influence spread in a social network.
  2. I plan to post the completed Assignment 4 today or tomorrow. There is now a question on influence spread in a social network, one on the small worlds phenomena (the topic for this week) and one on voting rules (which we will not discuss for a week or two).

- Lecture outline
  1. Chapter 20: small world phenomena.
     Note: It seems to me that Chapter 21 should come next but I will follow text sequence this week.
  2. Watts-Strogatz model
  3. Kleinberg's explanation of navigation in small worlds
  4. Liben-Nowell study
  5. Backstrom et al study
  6. Social distance
  7. Adamic and Adar study

The Small World Phenomenon (Chapter 20)

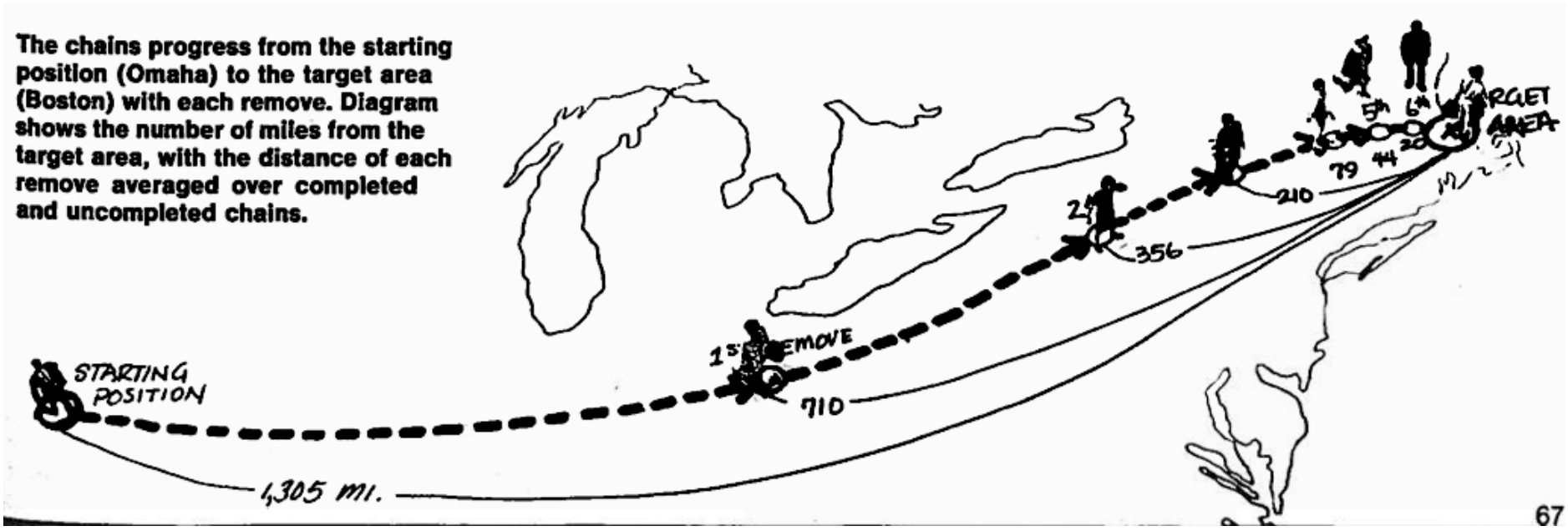
- We now move from a study of wide range spreading of technology, contacts and influence to the issue of focused or targeted search.
- Popularized in the famous concept of "six degrees of separation".
- At the start of this course, we briefly discussed the original 1960s Milgram experiment as it was 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]



- An image from Milgram's original article in *Psychology Today*, 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.

Two remarkable aspects of experiment

- There are short paths (of friendship) between people even though they are seemingly very unrelated.
  - We have also seen this phenomena when we spoke of one's Erdos number (amongst mathematicians or all scientists) and Bacon number (amongst actors).

- But the even more striking fact is that the Milgram letter chain succeeded without individuals knowing anything globally about the network structure.

- That is, without any centralized coordination, individuals were reasonably successful in reaching the target. (They did have 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

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 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]

Two settings for finding someone

- 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?

Two settings for finding someone

- 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?

- Suppose on the other hand that we want to reach someone and it either costs real money/effort to pass a message (e.g. postal mail) or perhaps I would prefer to not let everyone know that I am looking for person X.

Two settings for finding someone

- 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?
- Suppose on the other hand that we want to reach someone and it either costs real money/effort to pass a message (e.g. postal mail) or perhaps I would prefer to not let everyone know that I am looking for person X.
- 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 there is no triadic closure (i.e. your friends are not mutual friends, etc.), it is easy to see why every 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.

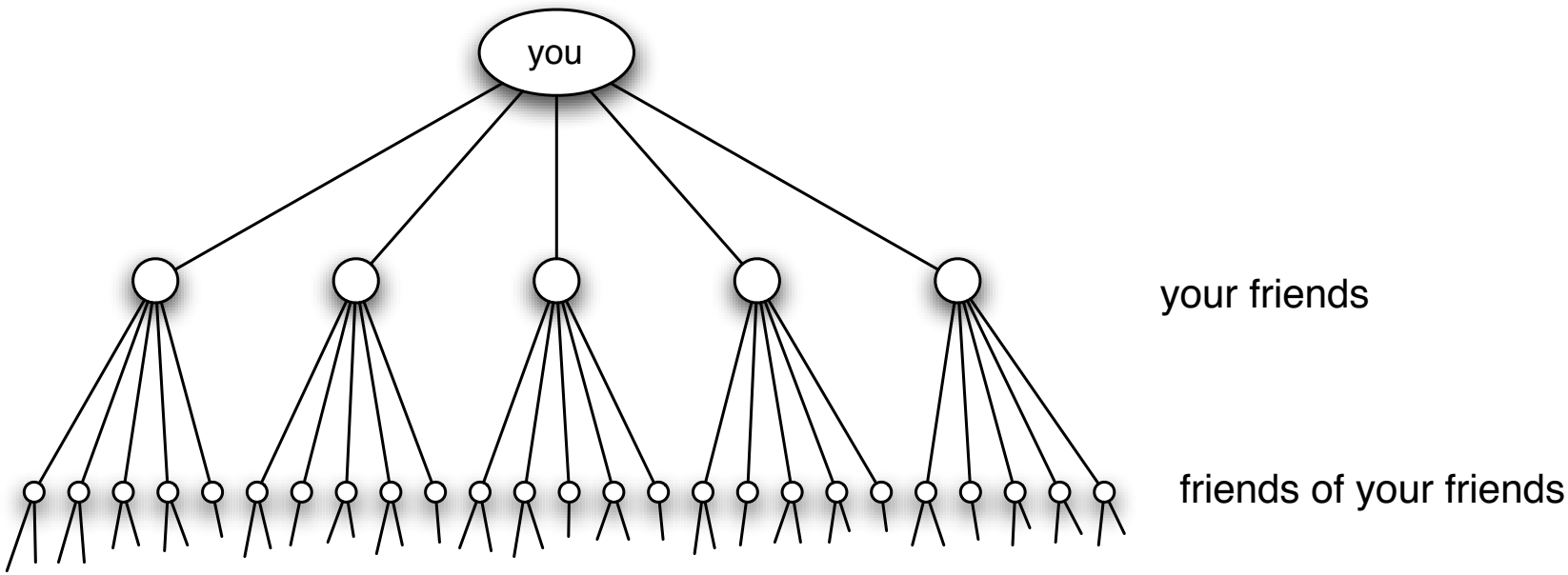


Figure : Pure exponential growth produces a small world [Fig 20.1 (a), E&K]

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.

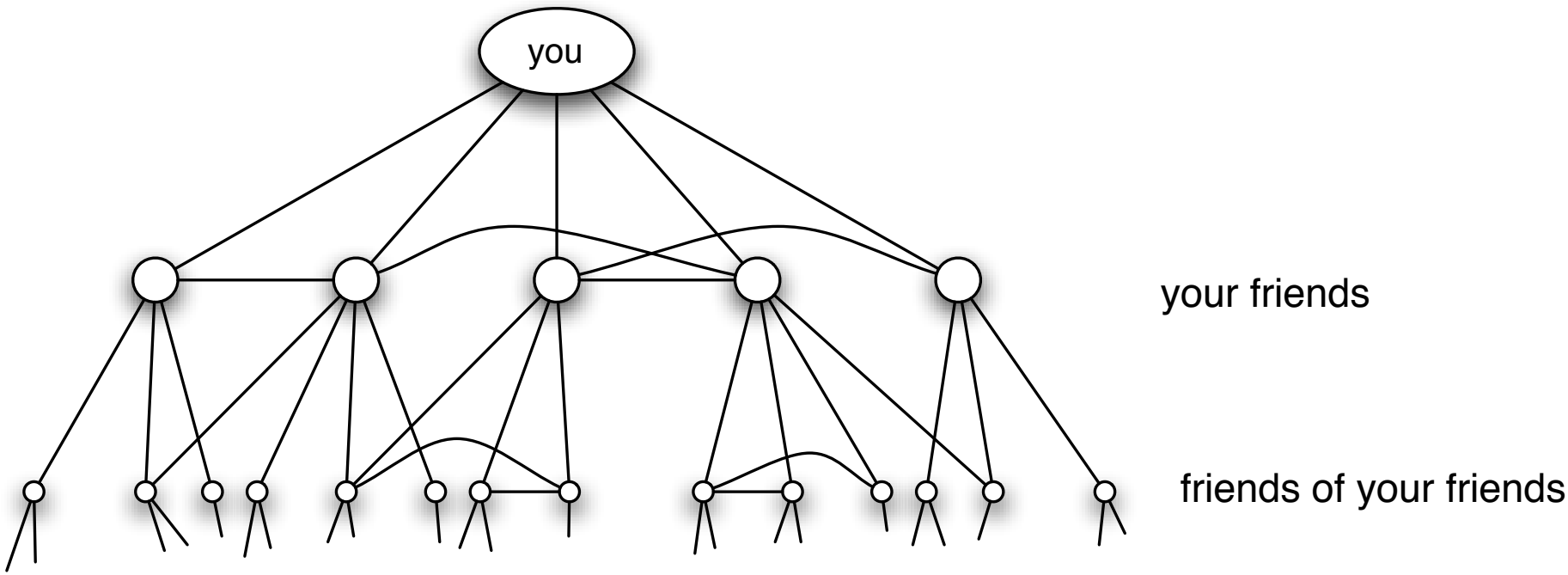


Figure : Triadic closure reduces the growth rate [Fig 20.1 (b), E&K]

The Watts-Strogatz model

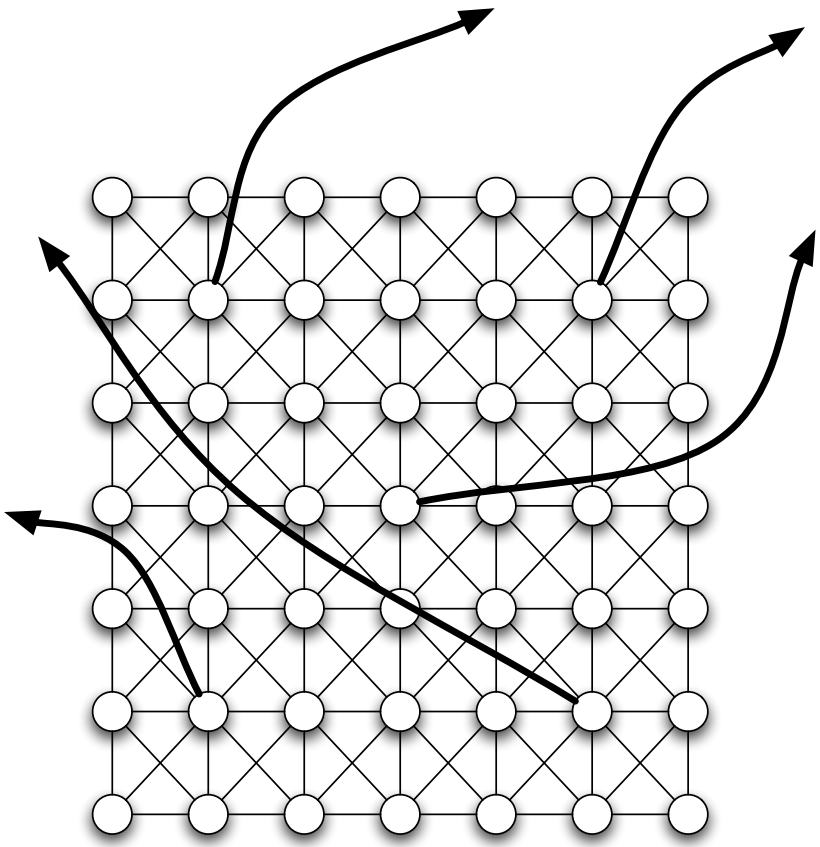
- Is it possible to have extensive triadic closure and still have short paths?

- Homophily is consistent with triadic closure especially for strong ties whereas weak ties can connect different communities and thereby provide the kind of branching that yields short paths to many nodes.

- One stylized model to demonstrate the effect of these different kinds of ties is the Watts-Strogatz model, which considers nodes lying in a two dimensional grid and then having two types of edges:
  - Short-range edges to all nodes within some small distance r. This captures an idealized sense of homophily
  - A small number of random longer-distance edges to other nodes in the network; in fact, one needs very few such random edges to achieve the effect of short paths.

Very few random edges are needed

- A k by k "town" with probability $1/k$ that a person has a random weak tie.
- This would be sufficient to establish short paths.



[Fig 20.3, E&K]

But how 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.
  - This is analogous to the Milgram experiment where individuals seem to use geographic information to guide the search.
  - However, completely random edges does no reflect real social networks

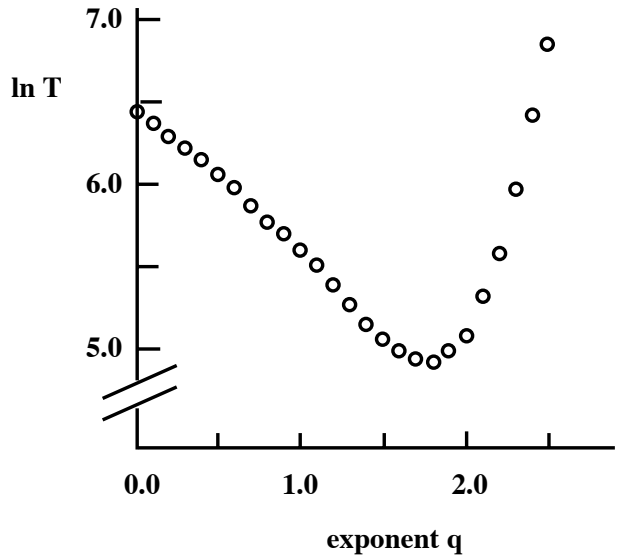
- Furthermore, having uniformly random edges will not work in general as:
  - Completely random edges (i.e. going to a random node anywhere in the network) are too random.
  - A random edge in an $n \times n$ grid is likely to have grid distance $\Theta(n)$.
  - Without some central guidance, such random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

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 r 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 $n^{1/3} \approx 27$.