Social network Analysis: small world phenomenon and decentralized search

Donglei Du

Faculty of Business Administration, University of New Brunswick, NB Canada Fredericton E3B 9Y2 (ddu@unb.ca)

Table of contents

The small-world phenomenon (a.k.a, six degrees of separation) I

- The materials is adopted form Chapter 10 of (Easley and Kleinberg, 2010).
- Social networks are so rich in short paths, known as the *small-world phenomenon*, or the "six degrees of separation"; and it has long been the subject of both anecdotal and scientific fascination.
- Mathematically, **small world** networks of size n have an average distance $O(\log n)$, meaning that between any two random nodes, the expected distance is $O(\log n)$.

$$\langle L \rangle \propto \log n$$

The small-world phenomenon (a.k.a, six degrees of separation) II

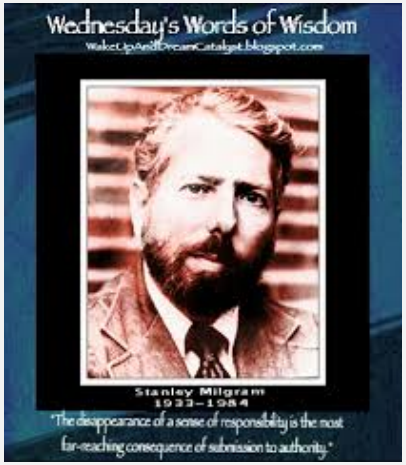
- Compare to the **ultra-small world**, where the average distance become significantly smaller and scale as

$$\langle L \rangle \propto \log \log n$$

Small-world networks are abundant in real life

- Small-world properties are found in many real-world phenomena:
  - Transportation networks in ground, air or sea;
  - Biology network such as food webs, gene network, protein network, neuron network, metabolism network, immune network;
  - Technology network like the Internet, electric power grids, wireless network, cable network, telephone call graphs;
  - Various social networks.

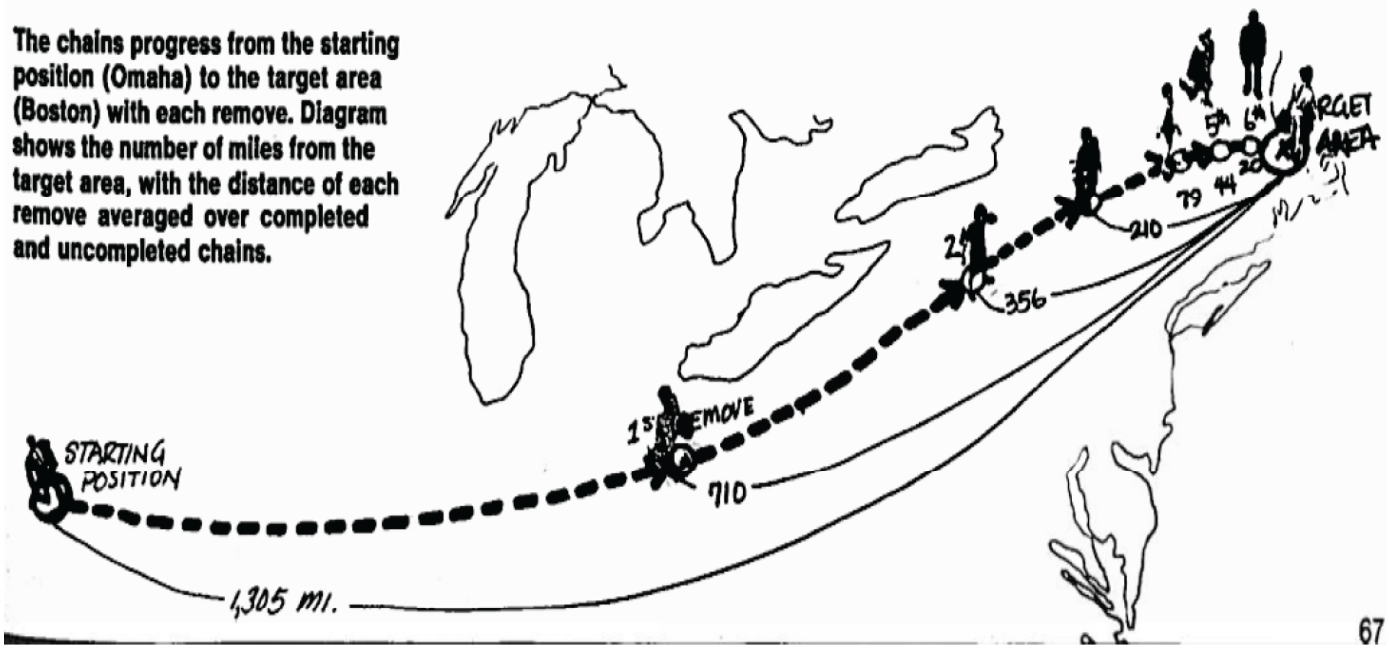
Milgram's experiment (Travers and Milgram, 1969)



http://stanleymilgram.com/milgram.php

- The first significant empirical study of the small-world phenomenon was undertaken by the social psychologist Stanley Milgram on the global friendship network as follows.
  - Randomly chosen "starter" individuals each tries forwarding a letter to a designated "target" person living in the town of Sharon, MA, a suburb of Boston.
  - The target's name, address, occupation, and some personal information are provided,
  - The participants could not mail the letter directly to the target; rather, each participant could only advance the letter by forwarding it to a **single** acquaintance that he or she knew on a first-name basis, with the goal of reaching the target as rapidly as possible.

Result from Milgram's experiment



- 20% of initiated chains reached target
- average chain length = 6.5
- median =6
- Hence the famous "Six degrees of separation"

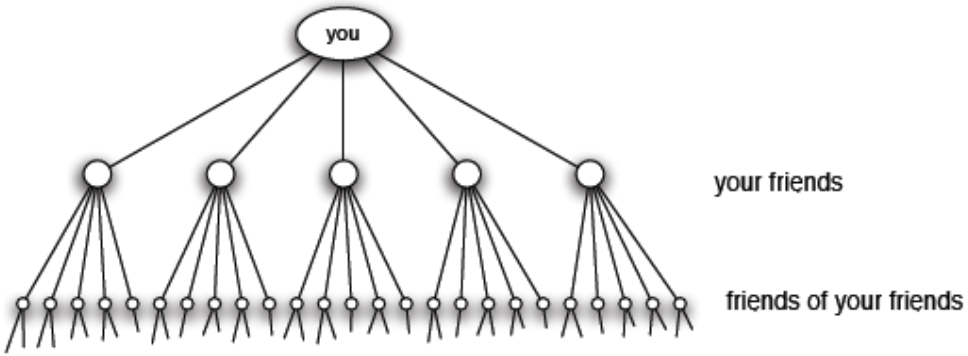
Milgram's experiment repeated (Dodds et al., 2003)

- 60,000+ participants
- 24,163 message chains
- 384 reached their targets
- average path length 4.0
- Median 5-7

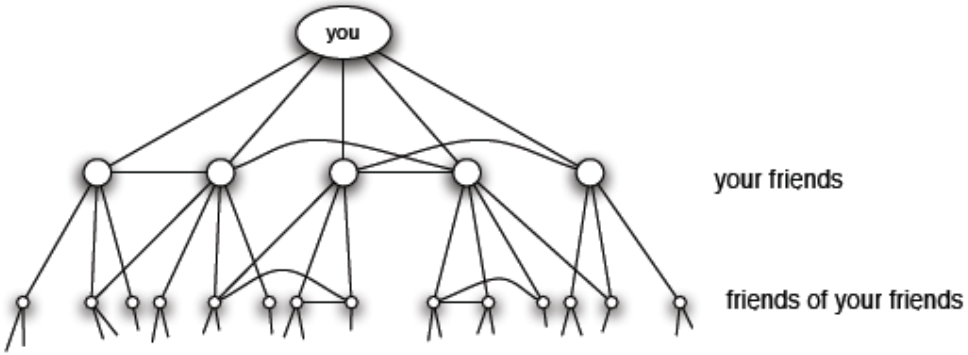
What can we learn from Milgram's experiment?

- Milgram's experiment really demonstrated two striking facts about large social networks:
  - First, that short paths are abundant;
  - Second, that people, acting without any sort of global "map" of the network, are effective at collectively finding these short path.

Question 1: The existence of short paths



(a) *Pure exponential growth produces a small world*



(b) *Triadic closure reduces the growth rate*

- Network grows exponentially, leading to the the existence of short paths!
  - The average person has between 500 and 1500 acquaintances, leading to $500^2 = 25K$ in one step, $500^3 = 125M$ in two steps, $500^4 = 62.5B$ in four (Figure (a)).
- However, the effect of *triadic closure* works to limit the number of people you can reach by following short paths (Figure (b)).
  - Triadic closure: If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.
- Question: Can we make up a simple model that exhibits both of the features: many closed triads (high clustering), but also very short path (small-world)?

The Watts-Strogatz small-world network (Watts and Strogatz, 1998) I

- Small-world network satisfies two properties according to Watts and Strogatz:
  - small average shortest path (global)
  - high clustering coefficient (local)
- Such a model follows naturally from a combination of two basic social-network ideas:
  - Homophily: the principle that we connect to others who are like ourselves, and hence creates many triangles.
  - Weak ties: the links to acquaintances that connect us to parts of the network that would otherwise be far away, and hence the kind of widely branching structure that reaches many nodes in a few steps.

The Watts-Strogatz small-world network (Watts and Strogatz, 1998) II

- The crux of the Watts-Strogatz model: introducing a tiny amount of randomness–in the form of long-range weak ties–is enough to make the world "small" with short paths between every pair of nodes.
- Bollobás and Chung (1988) shows mathematically that with high probability that the diameter is no more than $O(\log n)$, and hence the small world phenomenon.

How to generate the Watts-Strogatz small-world network

- Step 1. Start with a lattice of n nodes, and join each vertex to r of its neighbors to each side.
- Step 2. (Rewiring) For each edge, one end of this edge is rewired to another vertex independently and with probability p to a new vertex chosen randomly.
- Step 2'. (Adding) Alternatively, add a small number of new edges to randomly selected pairs of vertices.

Small world arises as randomness increases

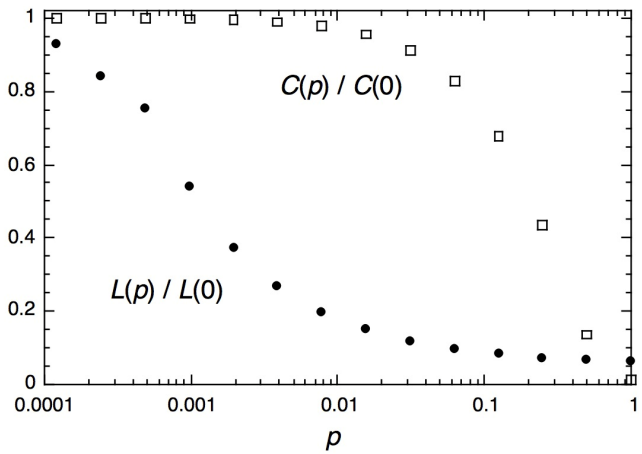


Figure: Characteristic path length $L(p)$ and clustering coefficient $C(p)$ for the family of randomly rewired graphs described (Watts and Strogatz, 1998)

- The data shown in the figure are averages over 20 random realizations of the rewiring process, and have been normalized by the values $L(0)$, $C(0)$ for a regular lattice.
- All the graphs have $n = 1,000$ vertices and an average degree of $k = 10$ edges per vertex.
- A logarithmic horizontal scale has been used to resolve the rapid drop in $L(p)$, corresponding to the onset of the small-world phenomenon.
- During this drop, $C(p)$ remains almost constant at its value for the regular lattice, indicating that the transition to a small world is almost undetectable at the local level.

The Watts-Strogatz random grid model: Illustration in **Netlogo**

- http://ccl.northwestern.edu/netlogo/
- Go to File/Model Library/Networks/Small Worlds

The Watts-Strogatz random grid model: Illustration in R packge: igraph

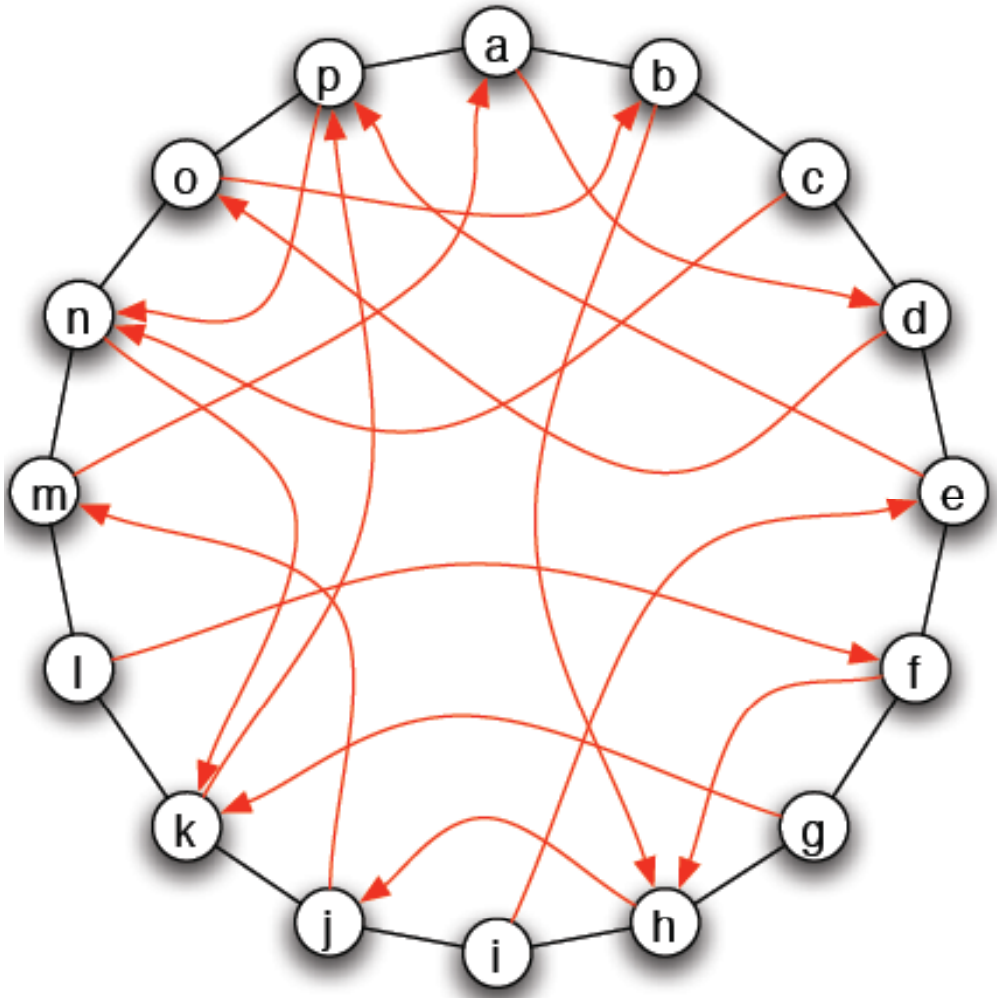
```
# R script: small_world.R
library(igraph)
g <- watts.strogatz.game(1, 100, 5, 0.05)
plot(g,layout=layout.circle)
average.path.length(g)
transitivity(g, type="average")
```

Watts-Strogatz small-world network vs Random network

	Random (chaos)	Watts-Strogatz (???)	regular (order)
average path	$\Omega\left(\frac{\log n}{\log\langle K\rangle}\right)$	$\Omega(\log n)$	$\Omega\left(\frac{n}{2r}\right)$
average clustering	$\Omega(\langle K\rangle/n)$	$\approx \frac{3r-3}{4r-2}(1-p)^3$ (Barrat and Weigt, 2000)	$\approx \frac{3r-3}{4r-2}$

- Random network is a small world, but not navigable.
  - In a random graph, although a short path exists, a local algorithm must be lucky to find it as it can do little better than a random walk on the network.
- Random network has much smaller average clustering coefficient, compared to that of the Watts-Strogatz small-world network.
- It is still an open question on the exact quantity of these measures for Watts-Strogatz small-world network.

The one-dimensional case: the people who lives on a ring society



Question 2: The Kleinberg's decentralized search model based on geographical distance (Kleinberg, 2000) I

- Nodes on a q-dimensional grid as before, and each node still has edges to each other node within r grid steps.
- But in generating a random edge out of v, we have this edge link to w with probability proportional to $d(v,w)^{-q}$ where $q \geq 0$. More formally, for each node v, connect to w with probability

$$\mathbb{P}(v \text{ linked to } w) = \frac{d(v,w)^{-q}}{\sum_{w} d(v,w)^{-q}}.$$

- The Watts-Strogatz model therefore corresponds to the special case where $q = 0$.

Question 2: The Kleinberg's decentralized search model based on geographical distance (Kleinberg, 2000) II

- We will need the following bound when $q = 1$ later for the proof of efficiency of decentralized search:

$$\mathbb{P}(v \text{ linked to } w) \geq \frac{d(v,w)^{-1}}{2 \log n}. \tag{1}$$

  - There are two nodes at distance 1 from v, two at distance 2, and more generally two at each distance d up to $n/2$ (assuming n is even):

$$\sum_{w} \frac{1}{d(v,w)} = 2 \sum_{k=1}^{n/2} \frac{1}{k} \leq 2 \left(1 + \ln \frac{n}{2}\right) \leq 2 \log_2 n.$$

The Kleinberg model: Illustration in **Netlogo**

- http://ccl.northwestern.edu/netlogo/
- Go to File/Model Library/Networks/Small Worlds

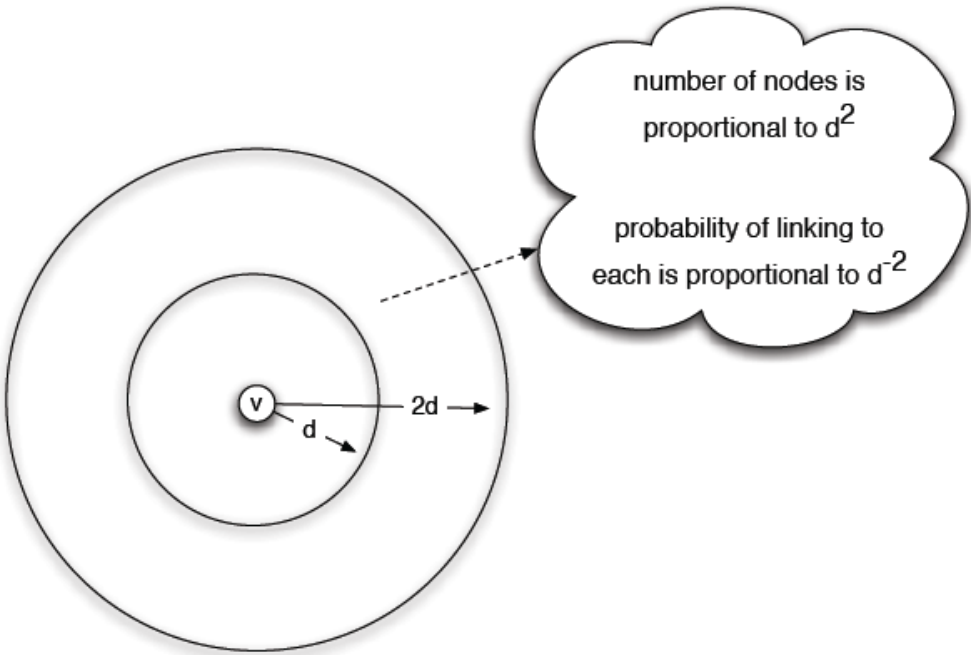
The Kleinberg model: Illustration in R packge: igraph

```
# R script: small_world.R
library(igraph)
g <- watts.strogatz.game(1, 100, 5, 0.05)
plot(g,layout=layout.circle)
average.path.length(g)
transitivity(g, type="average")
```

The main result for Kleinberg's model:

- In the limit of large network size,
  - When $q = 0$: long range contacts are chosen uniformly, resulting in the random network which has short paths between every pair of vertices, but no decentralized algorithm capable of finding these paths.
  - If $q < n$: more likely to choose distant-friends where decentralized algorithm quickly approaches the neighborhood of the target, but then slows down till finally reaches target itself.
  - If $q > n$: more likely choose close-friends where decentralized algorithm quickly finds target in its neighborhood, but reaches the target slowly.
  - If $q = n$, decentralized search is most efficient: next slide $\Rightarrow \ldots$

Why $q = n$?



- We look at the two-dimensional case where $n = 2$.
- Taking a node v in the network, and a fixed distance d, and considering the ring area where group of nodes lying at distances between d and $2d$ from v:
- What is the probability that v forms a link to some node inside this group?
  - Area $\propto d^2 \implies$ number of nodes therein $\propto d^2$.
  - Probability that v links to any one node therein $\propto d^{-2}$ according to the model.
  - These two terms approximately cancel out.
  - Therefore the probability that a random edge links from v into some node therein is approximately independent of the value of d.
  - Consequently, when $q = n = 2$: long-range weak ties are being formed in a way that is spread roughly uniformly over all different scales of resolution.

From geographic data on friendship to rank-based friendship (Liben-Nowell et al., 2005)

- Question to answer: how friendship links scale with distance and look for evidence of the exponent $q = 2$, that is, the probability of a random chosen link $p(v, w) \propto d(v, w)^{-2}$!
- Data: Blogging site LiveJournal
  - Roughly 500,000 users who provided a U.S. ZIP code for their home address, as well as links to their friends on the system.
  - We now have a friendship network with location as one of the node attributions

Method I

- Consider pairs of nodes which are d distance away from each pother, and calculate what fraction f of these pairs are actually friends, as a function of d.
- One difficulty: the inverse-square distribution is useful for finding targets when nodes are uniformly spaced in two dimensions
- But the population density of the users within any country is extremely non-uniform...
- What's a reasonable generalization to the case in which they can be spread very non-uniformly?
- One approach that works well is to determine link probabilities not by physical distance, but by rank.

Method II

- Suppose that as a node v looks out at all other nodes, it ranks them by proximity: the rank of a node w, denoted $rank(w)$, is equal to the number of other nodes that are closer to v than w is.
- if a node w in a uniformly-spaced grid is at distance d from v, then it lies on the circumference of a disc of radius d, which contains about d^2 closer nodes - so its rank is approximately d^2.
- Thus, linking to w with probability proportional to d^{-2} is approximately the same as linking with probability $rank(w)^{-1}$, so this suggests that exponent $p = 1$ is the right generalization of the inverse-square distribution.

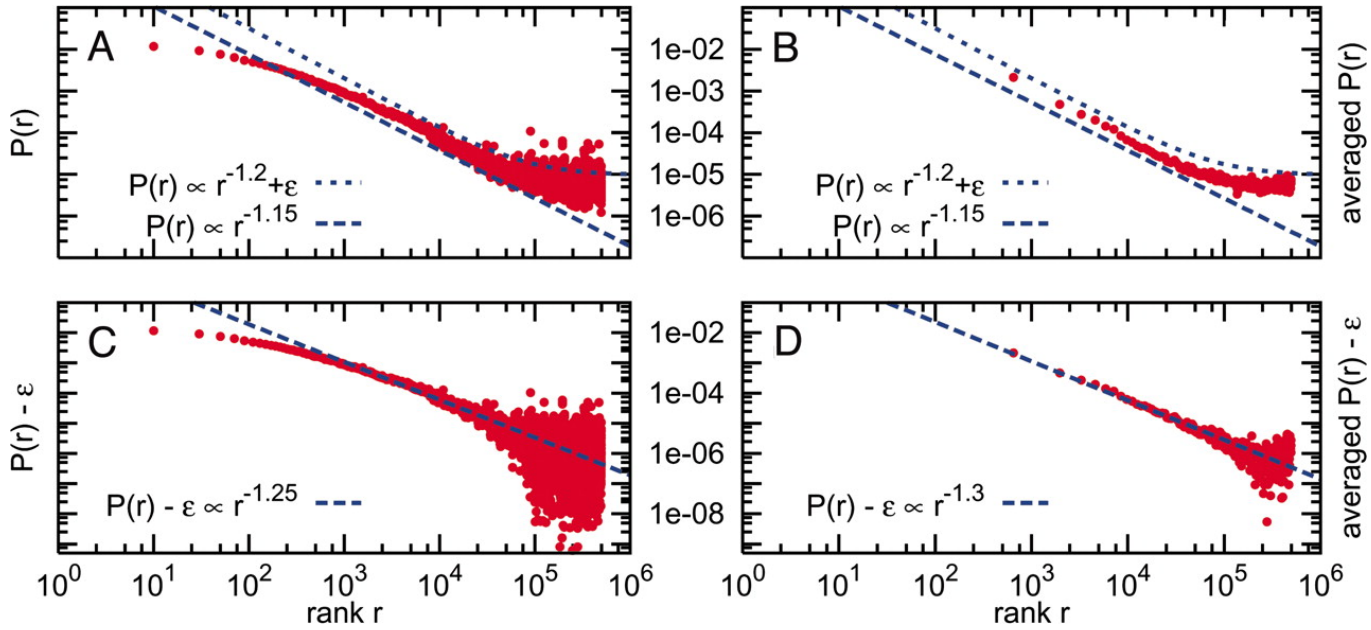
- Liben-Nowell et al. were able to prove that for essentially any population density, if random links are constructed using rank-based friendship with exponent 1, the resulting network allows for efficient decentralized search with high probability.

Method III

- In addition to generalizing the inverse-square result for the grid, this result has a nice qualitative summary: to construct a network that is efficiently searchable, create a link to each node with probability that is inversely proportional to the number of closer nodes.

Results

The relationship between friendship probability and rank.



Liben-Nowell D et al. PNAS 2005;102:11623-11628

PNAS

Back into the future I

- In this case study, one follows a sequence of steps in which
  1. start from an experiment (Milgram's),
  2. build mathematical models based on this experiment (combining local and long-range links),
  3. make a prediction based on the models (the value of the exponent controlling the long-range links), and then
  4. 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.

Back into the future II

- 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.

Social Distance

- The social distance between two people is the size of the smallest focus that includes both of them.

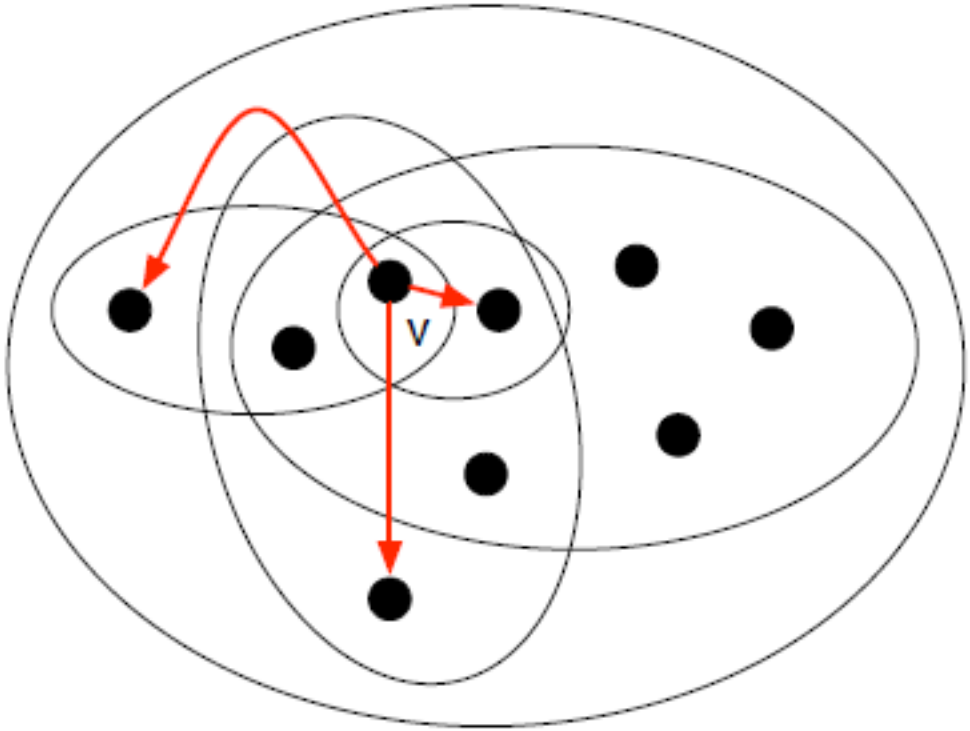


Figure: the node labeled v belongs to five foci of sizes 2, 3, 5, 7, and 9 (with the largest focus containing all the nodes shown).

Network model based on social distance (Kleinberg, 2001; Adamic and Adar, 2005) I

- Following the style of earlier models, construct a link between each pair of nodes v and w with probability proportional to $dist(v, w)^{-p}$.
- One can show, subject to some technical assumptions on the structure of the foci, that when links are generated this way with exponent $p = 1$, the resulting network supports efficient decentralized search with high probability.
- Two conclusions:
  - As with rank-based friendship, there is a simple description of the underlying principle: when nodes link to each other with probability inversely proportional to their social distance, the resulting network is efficiently searchable.

Network model based on social distance (Kleinberg, 2001; Adamic and Adar, 2005) II

- Moreover, the exponent $p = 1$ is again the natural generalization of the inverse-square law for the simple grid model.

Decentralized Problem-Solving

- The notion that social networks can be effective at this type of decentralized problem solving is an intriguing and general premise that applies more broadly than just to the problem of path-finding that Milgram considered.
- There are many possible problems that people interacting in a network could try solving, and it is natural to suppose that their effectiveness will depend both on the difficulty of the problem being solved and on the network that connects them

Ways to generate small-world networks

- As the output of an optimization problem (Mathias and Gopal, 2001; Gastner and Newman, 2006).
- As the output of a growth process: add links with probability depending on property of existing nodes, edges (preferential attachment, link copying).
- As the equilibrium of a game: simulate nodes as agents deciding whether to rewire or add links.

The mathematics behind: myoptic search is efficient in expectation I

- Choose a random start node s and a random target node t on the random ring network equipped with Kleinberg's inverse power distribution.
- The goal is to forward a message from s to t, with each intermediate node on the way only knowing the locations of its own neighbors, and the location of t, but nothing else about the full network.
- **Myoptic search**: when a node v is holding the message, it passes it to the contact that lies as close to t on the ring as possible

The mathematics behind: myoptic search is efficient in expectation II

- We will show the myoptic search constructs a path that is exponentially smaller: proportional to $\log^2 n$, although myoptic usually cannot give us the shortest path.
- Namely, we will show that $\mathbb{E}[X] \leq O(\log^2 n)$, where, X is a random variable indicating the number of steps required by myopic search.

Idea of the proof I

- Given s and t, as the message moves from s to t, it is in phase j of the search if its distance from the target is between 2^j and 2^{j+1}.
  - There are at most $\log_2 n$ different phases.

Idea of the proof II

- Let X_j $(j = 1, \ldots, \log_2 n)$ be the number of steps taken in Phase j. Then

$$
\begin{aligned}
X &= \sum_{j=1}^{\log_2 n} X_j \\
&\Downarrow \\
\mathbb{E}[X] &= \sum_{j=1}^{\log_2 n} \mathbb{E}[X_j] \\
&= \sum_{j=1}^{\log_2 n} \sum_{k=1}^{\infty} \mathbb{P}(X_j \geq k)
\end{aligned}
$$

Idea of the proof III

- So it suffices to bound from above each $\mathbb{P}(X_j \geq k)$, which is the probability that Phase j runs for at least k steps, implying that phase j failed to terminate $k-1$ steps in a row.
  - We shall show that

$$\begin{aligned}
\mathbb{P}(X_j \geq k) &\leq \left(1 - \frac{1}{3\log n}\right)^{k-1} \qquad (2)\\
&\Downarrow \\
\mathbb{E}[X_j] &= \sum_{k=1}^{\infty} \mathbb{P}(X_j \geq k) \\
&\leq \sum_{k=1}^{\infty} \left(1 - \frac{1}{3\log_2 n}\right)^{k-1} = 3\log_2 n \\
&\Downarrow (1) \\
\mathbb{E}[X_j] &\leq O(\log_2^2 n).
\end{aligned}$$

Idea of the proof IV

- To show (2), it suffices to show that (due to the independence of the steps within a phase)

$$\mathbb{P}(\text{Phase } j \text{ terminates after one step}) \geq \frac{1}{3 \log n.} \tag{3}$$

  - Suppose the message is at a node v whose distance to the target t is some number $d \in [2^j, 2^{j+1}]$.
  - Phase j terminates after one step only if the next connected node w is at most $d(v,t)/2$ distance away from t.
  - Let S be the set of nodes at distance $d(v,t)/2$ from t, namely $S = \{w : d(w,t) \leq d(v,t)/2\}$.

Idea of the proof V

- For each $w \in S$, we have

$$d(v,w) \le d(v,t) + d(w,t) \le 3d(v,t)/2$$

implying that

$$\mathbb{P}(v \text{ linked to } w) \underbrace{\geq}_{(1)} \frac{d(v,w)^{-1}}{2\log n}$$

$$\geq \frac{1}{2\log n}\frac{1}{3d(v,t)/2} = \frac{1}{3d(v,t)\log n}.$$

- Since $|S| = d(v,t)+1$, there are more than $d(v,t)$ nodes in S. Therefore the probability that one of them is linked to v is at least

$$d(v,t)\frac{1}{3d(v,t)\log n} = \frac{1}{3\log n}$$

and this proves (3).

Fragility and caveats of the small-world phenomenon

- Myth or fact? (Kleinfeld, 2002; Marvel et al., 2013)

References I

Adamic, L. and Adar, E. (2005). How to search a social network. *Social Networks*, 27(3):187–203.

Barrat, A. and Weigt, M. (2000). On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560.

Bollobás, B. and Chung, F. R. K. (1988). The diameter of a cycle plus a random matching. *SIAM Journal on discrete mathematics*, 1(3):328–333.

Dodds, P. S., Muhamad, R., and Watts, D. J. (2003). An experimental study of search in global social networks. *science*, 301(5634):827–829.

Easley, D. and Kleinberg, J. (2010). Networks, crowds, and markets. *Cambridge Univ Press*, 6(1):6–1.

References II

Gastner, M. T. and Newman, M. E. (2006). The spatial structure of networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 49(2):247–252.

Kleinberg, J. (2001). Small-world phenomena and the dynamics of information. In *NIPS*, pages 431–438.

Kleinberg, J. M. (2000). Navigation in a small world. *Nature*, 406(6798):845–845.

Kleinfeld, J. (2002). Could it be a big world after all? the six degrees of separation myth. *Society, April*, 12:5–2.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628.

References III

Marvel, S. A., Martin, T., Doering, C. R., Lusseau, D., and Newman, M. (2013). The small-world effect is a modern phenomenon. *arXiv preprint arXiv:1310.2636*.

Mathias, N. and Gopal, V. (2001). Small worlds: How and why. *Physical Review E*, 63(2):021117.

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32(4):425–443.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442.