CS599: Structure and Dynamics of Networked Information (Spring 2005)
03/23/2005: Small-World Networks
Scribes: Nupur Kothari and Viral Shah

The fact that social networks seem to be small worlds has long been the subject of anecdotal observation. The hypothesis that “any two persons in the world are connected by a small chain of intermediate social acquaintances” is known as the small world phenomenon. The hypothesis was first tested by Stanley Milgram in 1967 [2]. Milgram found that in the experiment he conducted, there were an average of six acquaintances between any two participants. A flurry of work followed the discovery. The phenomenon has been captured in several popular ways including various movies and plays featuring the concept, the famous phrase “Six Degrees of Separation”, the popular game “Six Degrees of Kevin Bacon” in which one attempts to find a shortest path from any actor to the actor Kevin Bacon (`http://www.louisville.com/loumag/mar/bacon.htm`). The mathematicians’ equivalent to the Bacon number is the *Erdős number*: if a scientist has published a paper with the famous Hungarian mathematician Paul Erdős, his number is 1, if someone has published with anyone who has published with Erdős, the number is 2, etc. So the Erdős number is the shortest path distance to Erdős in the co-authorship graph.

1 Milgram’s Experiment (1967)

In 1967, Stanley Milgram attempted to verify the small world phenomenon quantitatively. He selected random people from locations like Kansas or Nebraska, and had them start a chain of letter-forwarding. The targets of the letters were in Cambridge, MA and Boston. Each starter was to send a folder through the mail to the target person. But the game had rules. The starters could only mail the folder to someone they knew on a first-name basis. That person was to mail the folder on to another first-name acquaintance, etc. Returned tracer postcards tracked the progress of each chain.

His first target person was the wife of a divinity student living in Cambridge, and starters were located in Wichita, Kansas. Milgram found that the very first folder reached her in just four days and took only two intermediate acquaintances. In the second study, the starters were located in Nebraska, and the target was from Sharon, MA, working in Boston. Milgram reported that ”chains varied from two to ten intermediate acquaintances, with the median at five” [2, p. 65]. Any person appeared to be able to reach the target with an average of six jumps.

Graph-theoretically, Milgram was trying to find shortest paths in the given social network of people in the US. Instead of using the “correct” approach of BFS, he had to rely on DFS, as the typical branching factor of several hundred social acquaintances would have resulted both in much work for all participating individuals, and an illegal chain letter using the US Postal System. So what Milgram found was really just an upper bound on the shortest paths.

Out of 296 chains, only 217 chains started, and 64 completed. The number of intermediate nodes varied from 2–10, with a median of 5 and a mean of 6. The frequencies that Milgram measured were monotonically increasing to chains of length 4, then dropped slightly for chains of length 5, and increased to the highest frequency for length 6, dropping again afterwards until length 10.

Later analysis shows that the attrition rate (the probability of not forwarding the letter) was about 25% at each step. Thus, it seems likely that the observed distribution is skewed to shorter chains: if a chain was naturally longer, then it has a higher probability of not finishing at all, and thus not being counted towards the average. White [5] gave a simple calculation correcting for this attrition. The argument is that if we observe a fraction of $\frac{n_j}{n}$ chains of length j, then the true fraction of such chains would have been roughly proportional to $(4/3)^j \frac{n_j}{n}$, as each chain of length j does not complete with probability $(3/4)^j$. Correcting for attrition in this way, White suggests that Milgram's data is more supportive of a median distance of 7–8 between individuals.

Additional interesting details about the Milgram experiments were that many chains had the same person as a last step: out of all completed chains, 25% went through the same last person, and 50% through one of three persons. This may be taken as an indication that society relies on "hub persons" for connectivity.

Kleinfeld [1] revisits Milgram's experiment and finds that there were a number of loopholes in Milgram's experiment. Milgram's sampling technique to determine starting points for the chains was biased since his target was a stockbroker, and a number of his starters were stockholders, who would be more likely to know stockbrokers. He recruited people through advertisements for people who were well connected, and the experiment was generally more likely to attract economically better-off people, who tend to have more long-distance connections. In addition, out of the Kansas and Nebraska studies, only the Nebraska one, which was much more successful, was published.

2 Formalization, and Expander Graphs

If we want to reason about small-world networks analytically, we first need to come up with a good definition of the concept. A simple (and popular) definition is to consider a graph a small world if it has small (e.g., logarithmic) diameter. However, this ignores another characteristic of small worlds, namely that the graph exhibits a lot of clustering and regular structure, similar to lattices. Watts [4] proposes the notion of *clustering coefficient* C, defined as the number of triangles in the graph divided by the number of adjacent edge pairs. Thus, it can be interpreted as the probability that two nodes are connected, given that they share a mutual friend. For a complete graph, the clustering coefficient is 1, and for an empty graph, it is 0.

If we are interested merely in finding graphs with small diameter, our task is easy: the complete graph has diameter 1. The complete graph is certainly not a good model of social networks, as individuals tend to have a smaller number of contacts. So we are looking for graphs of small diameter and low degree.

One way in which we can achieve this goal is by virtue of *expanders*: A graph G is an α-expander (or α edge expander) if each vertex set $S \subseteq V$ with $|S| \leq n/2$ has at least $\alpha|S|$ edges leaving. Intuitively, this means that no two "large chunks" of the graph can be disconnected by removing few vertices or edges. Obviously, the larger α, the better an expander the graph is. Normally, people are particularly interested in families of expanders where α is constant, i.e., does not become smaller as the number of nodes n increases.

While the definition of expanders seems useful, we still have to find out whether they actually exist. It is pretty obvious that complete graphs are expanders (with $\alpha = n/2$), but again, we are looking for lower degrees. A binary tree is not a good expander, as the left half of the tree has $n/2$ nodes, but only one edge leaving. However, hypercubes are good expanders (with $\alpha = 1$), with degree $\log(n)$.

When we want to reduce the degree all the way down to constant, i.e., not increasing with n, the task becomes quite a bit more difficult. For a long time, no explicit constructions were known, and the first explicit constructions were based on Cayley graphs of rather difficult looking algebraic groups. More recently, Reingold et al. [3] presented a new way of constructing constant degree expanders via the zig-zag graph product. On the other hand, it is much easier to construct expanders randomly. For any $d \geq 3$, almost all d-regular graphs are actually expanders (notice that for $d = 2$, any d-regular graph is just a union of disjoint cycles, and thus certainly not an expander.)

It is not quite obvious how to generate a uniformly random d-regular graph. It is much easier to instead generate a d-regular *multigraph*, in which we also allow self-loops and parallel edges. For then, we can use the configuration model: each node has d edge ends sticking out, and we construct a uniformly random matching among those edge ends, by randomly picking two of the remaining edge ends and pairing them up until all edge ends are matched. For this model, whenever $d \geq 6$, a fairly straightforward analysis using standard tail bounds (Martingale or Chernoff bounds) shows that with high probability, the resulting graph is an expander. For $3 \leq d < 6$, the analysis gets a little messier, as tail bounds are not quite strong enough: one has to reason a bit more carefully about binomial coefficients and such.

2.1 Properties of expanders

Expander graphs, and the notion of expansion, are important for several reasons. From a network design perspective, expander graphs are resilient to node or edge failures, as few failures will not be able to disconnect large parts of the network.

Expansion also plays an important role in the analysis of random walks and their mixing time (and thus in the context of MCMC sampling). The reason is that a low expansion means that few edges connect two large components. Thus, a random walk starting in one of the components has small probability of crossing to the other component, and it will take a long time until the initial bias in the probability of being at any one node gets evened out.

For our purposes, the reason that expanders are interesting is that they have small diameter.

Lemma 1 *Expanders have diameter $O(\log n)$.*

Proof. Consider a BFS search of an expander starting from any node. For each layer S of the BFS, at least $\alpha|S|$ edges are leaving the set of all nodes already reached, hence at least $\frac{\alpha}{d}|S|$ (where d is the degree of G) new nodes are hit, so long as $|S| \leq n/2$. Thus, at each layer, the total number of nodes in the BFS tree grows by a factor of $1 + \alpha/d$, and $\frac{n}{2}$ nodes are reached in

$$\log_{1+\alpha/d}(n/2) \quad = \quad O(\tfrac{\log n}{\log(1+\alpha/d)})) \quad = \quad O(\tfrac{\log n}{\alpha/d}) \quad = \quad O(\tfrac{d}{\alpha}\log n)$$

steps. Carrying out BFS from both the source s and target t, they each reach $\frac{n}{2}$ nodes. After one more step of BFS, they must therefore intersect at some node v, and the concatenation of the s-v and t-v paths gives an s-t path of length $O(\frac{\alpha}{d}\log n)$. ■

References

[1] J. Kleinfeld. Could it be a big world after all? the 'six degrees of separation' myth. *Society*, 2002.

[2] S. Milgram. The small world problem. *Psychology Today*, 1:61–67, 1967.

[3] O. Reingold, S. Vadhan, and A. Wigderson. Entropy waves, the zig-zag graph product, and new constant-degree expanders. *Annals of Mathematics*, 155:157–187, 2002.

[4] D. Watts. Networks, dynamics and the small world phenomenon. *American Journal of Sociology*, 105:493–527, 1999.

[5] H.C. White. Search parameters for the small world problem. *Social Forces*, 49:259–264, 1970.