6

Social Networks

6.1 Introduction

A social network has people as its vertices, and social connections representing, e.g., friendships, acquaintances, or business relationships, as its edges. In this chapter we explore two key properties exhibited by many social networks: the *small-world* phenomenon, which says that any two individuals are connected via a short path, and *high clustering*, i.e., If two people in a social network have a friend in common, then they are more likely to have an edge between them (as compared with two randomly selected nodes).

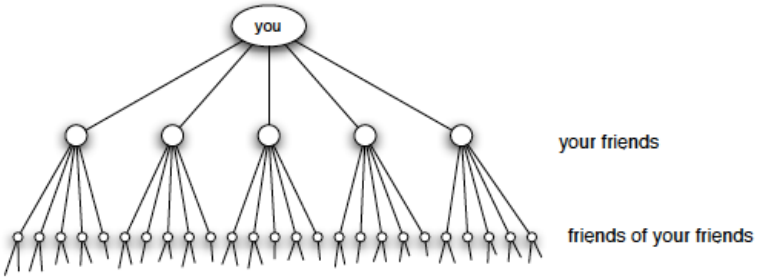

Figure 6.1: An exponential growth in the k-hop neighborhood. Image from [1].

6.2 Small-World Phenomena

The term *small world network* was coined by the sociologist Stanley Milgram in the 1960s in the context of his experiments on the structure of social networks. Milgram was interested in determining the distance in hops separating two arbitrary persons. In an ingenious experiment, Milgram mailed letters to randomly chosen individuals in Nebraska. The task of these individuals was to send a letter to a target individual living near Boston, but the letter could be sent only through chains of social acquaintances.

The outcome of these experiments was surprising: a relatively large fraction of these letters did indeed arrive at their target; furthermore, they did so after traversing only a small number of social edges; in fact, the median number of edges traversed was 6, the finding now canonised as "six degrees of separation". It is rather surprising to think that in a country with a population on the order of 10^8 people, two completely unrelated, arbitrarily chosen individuals, who might lead very different lives, belong to different social classes, and live thousands of kilometers apart, would nevertheless be so close in the social network.

However, on second thought, the intuition behind this finding is not so surprising. Let us assume all people in the network have degree d. Then, you will have d friends, and if you extrapolate out, you see that you would have approximately d^2 friends of friends. This is your "two-hop" neighborhood. Similarly, you would expect your "k-hop" neighborhood to have approximately d^k friends. Hence, the farthest person in the network would be approximately $\log_d(n) \approx \log(n)/\log(d)$ steps away. In fact, as we will discuss in Section 6.5, for random graphs such as $G(n,p)$ and $G(n,r)$ this intuition is largely correct.

More formally, consider a graph G of maximum degree Δ and of diameter D. We can establish the following inequality bounding the order of G. A vertex v can have at most Δ neighbors. Each of these neighbors in turn can have at most $\Delta - 1$ new neighbors, and so forth. Therefore

$$n \leq 1 + \Delta \left(\sum_{i=1}^{D} (\Delta - 1)^{(i-1)} \right) = 1 + \Delta \frac{(\Delta - 1)^D - 1}{\Delta - 2}. \tag{6.1}$$

Graphs that have equality in (6.1) are called Moore graphs. Moore graphs exist only for particular sets of values of Δ and D.[1]

Informally, a small world graph (or, a graph with "small diameter") has diameter close to the above bound. Sometimes, the small world property is also defined in terms of the average distance, rather than the diameter (i.e., the maximum distance). The key is that to achieve small diameter, it is necessary for the size of the k-hop neighborhood to grow exponentially, as in (6.1) and the intuition provided in Figure 6.1. This is in contrast to, e.g., regular lattices $\mathbb{L}^d$ of dimension d, whose k-hop

[1] Note that trees are not Moore graphs, as their diameter is twice their height.

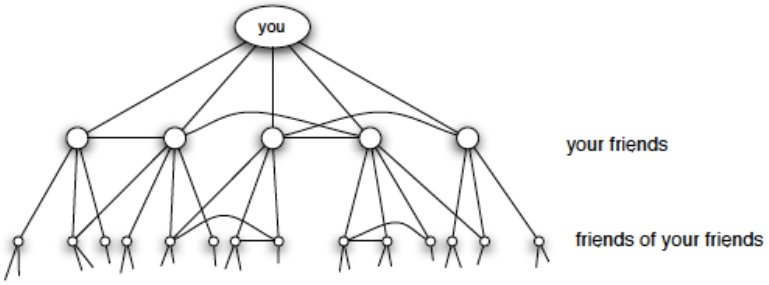

Figure 6.2: Social networks have high clustering – your friends are likely to also know each other. Image from [1].

neighborhood only grows polynomially in k^d, and whose diameter is therefore on the order of $n^{1/d}$. Do social networks have small diameter? Before we get back to this question, we first consider an important property of social networks that could be a barrier to small diameters.

6.3 Clustering

A feature that many social networks have is high "clustering", i.e., the likelihood that two vertices have an edge between them is higher if they have a neighbor in common (see Figure 6.2). A natural example is an square lattice of size n in which a vertex is connected to the eight vertices that surround it. Clearly, two random nodes have probability 0 a.a.s. of having an edge between them as n goes to infinity, but if two random vertices have a neighbor in common then there is a constant probability that they have an edge between them regardless of n. Intuitively, this seems at odds with the small world property – indeed, in the example above, one can easily see that average distances between nodes are $\Omega(\sqrt{n})$ as opposed to $O(\log(n))$. Hence, one would imagine that high clustering could also slow down the k-hop growth in social networks, however, that is not the case – networks can have high clustering as still exhibit the small world property! Before we explain one such network model, let us formalize our definition of clustering.

We first define the clustering coefficient C_v of a vertex:

$$\begin{aligned} C_v &= \frac{\text{number of edges between neighbors of } v}{\text{number of possible edges between neighbors of } v} \\ &= \frac{|\{(u,w) \in E(G) : (u,v) \in E(G), (w,v) \in E(G)\}|}{\binom{d(v)}{2}} \end{aligned} \tag{6.2}$$

Using this, two distinct definitions of the clustering coefficient of a graph are common in the literature. The first, also known as the *local* clustering coefficient is simply an average clustering coefficient of all vertices in the graph:

$$C_{local}(G) = \frac{1}{n} \sum_{v \in V(G)} C_v(G).$$

The second, known as the *global* clustering coefficient, considers what fraction of "potential triangles" (i.e., paths of length two) in a graph actually form a triangle.

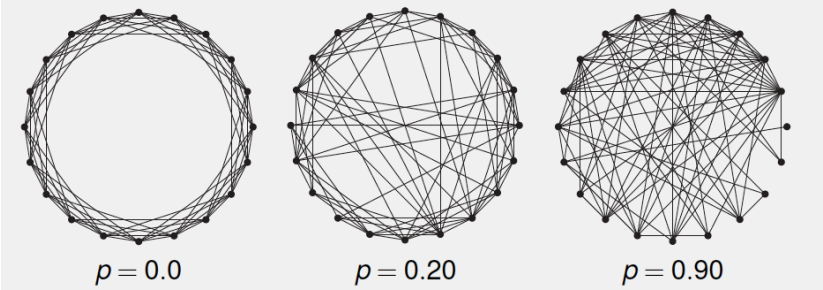

Figure 6.3: Examples of Watts-Strogatz networks for $n = 20$ and $k = 4$.

$$C_{global}(G) = \frac{3 \times \text{number of triangles}}{\text{number of pairs of adjacent edges}} \tag{6.3}$$

$$= \frac{\sum_{v \in V(G)} \binom{d(v)}{2} C_v(G)}{\sum_{v \in V(G)} \binom{d(v)}{2}} \tag{6.4}$$

As shown, this definition can also be re-written in terms of the vertex clustering coefficient.

Note that for the random graph $G(n, p)$, every possible edge exists independently of everything else with probability p. Therefore

$$\mathbb{E}\left[C_v(G(n,p))\right] = p. \tag{6.5}$$

Also, note that we had seen that for $G(n, r)$, the number Z_3 of triangles is asymptotically Poisson with mean independent of n, which shows that the clustering coefficient would decrease at a rate of $1/n$.

Note that the clustering coefficient is limited in that it only captures local connectivity within one hop. It is easy to construct graphs that have rich local connectivity over more than one hop, but that have $C(G) = 0$.[2]

6.4 Watts-Strogatz Networks

Watts-Strogatz networks [**?**] can simultaneously have high clustering and the small diameter. The basic idea in this model is to start with a regular lattice, and then select each edge independently with probability p to "rewire" it, i.e., change one of the (or in some variations, both) endpoints to a vertex selected uniformly at random.[3] This model can thus be viewed as an interpolation between the lattice (when $p = 0$), and a random graph (when $p = 1$); see Figure 6.3.

More formally, we construct a Watts-Strogatz network $WS(n, k, p)$ as follows:

- Arrange the n vertices in a cycle.

[2] As an example, consider transforming a graph by breaking every edge in half and "inserting" an additional vertex. The transformed graph has $C = 0$, even though it has "almost" the same structure as the initial graph.

[3] In yet another variation which we will later use for ease of analysis, edges are not rewired, instead new random edges are added to the existing lattice.

- Form an edge between every vertex and the k vertices to it's left and the k vertices to it's right; i.e., every vertex will have $2k$ edges.
- Rewire one endpoint of each edge to another vertex selected uniformly at random with probability p.

We can now consider the clustering coefficient of such a network. First, note that before rewiring, a neighbor at distance i of v, for all $1 \leq i \leq k$ has $2k-1-i$ edges to other neighbors of v. Therefore, the number of edges between neighbors of v is $2\sum_{i=1}^{k}(2k-1-i) = 3k(k-1)$. Hence, for $p=0$, the clustering coefficient is $C(WS(n,k,0)) = C_v = \frac{3(k-1)}{2(2k-1)}$ For $p>0$, we can approximate the clustering coefficient by noting that a triangle survives rewiring with probability $(1-p)^3$, and

$$C(WS(n,k,p)) \approx C(WS(n,k,0))(1-p)^3. \tag{6.6}$$

This confirms our intuition that if the fraction of rewired edges is low, the impact on the clustering coefficient is quite small.

On the other hand, even for small p, the fact that the subgraph of rewired edges resembles a random graph will allow the average distance and the diameter to drop very quickly even for small p. We will give a formal proof for a variation of this model below, but to get an intuition as to how the diameter could be small, let us first consider some $p \gg \frac{1}{\sqrt{n}}$. First, let us split the cycle into "towns" of size $\sqrt{n}$ so there are $\sqrt{n}$ towns in total. Before rewiring, there are approximately n edges in a town, each of which is rewired with probability p to another town. Hence, in expectation, approximately 1 edge gets rewired from every town to every other town. A path can then be found by walking within a town to the cross-town edge, taking it, and then walking across that town to the desired node. Within a town, as long as p is not too large, a path is of length at most $O(\sqrt{n})$ Hence, the entire path is similarly of length $O(\sqrt{n})$; this is not quite the desired $O(\log n)$ (which will require much more mathematical technology to prove), but still a vast improvement over the diameter $\frac{n}{k}$ of the graph before rewiring!

Hence, intermediate values of p model the two features of real networks: small worlds (similar to random graphs) and large clustering coefficients (similar to lattices).

6.5 Random Graphs have Small Diameter

The diameter of many randomly generated graphs, such as $G(n,p)$ and $G(n,r)$, are surprisingly small. At the risk of making an overly sweeping statement, we can say that randomness produces rapidly expanding, and hence compact, networks. In this section, we study a slightly different model from the random graphs considered so far, to avoid certain technical difficulties. Specifically, we study a random network obtained by adding a random matching to an n-cycle. This also similar to the Watts-Strogatz network considered above with $k=1$ where the rewiring is performed in a very controlled manner.

Theorem 6.1 (Cycle + Random Matching has Small Diameter [**?**])**.** *Let G be an undirected graph formed by adding a random matching to an n-cycle. Then G has diameter $diam(G)$ satisfying* a.a.s.

$$\log_2 n - c \leq diam(G) \leq \log_2 n + \log_2 \log n + c, \tag{6.7}$$

for constant $c \leq 10$.

Proof:

The idea of the proof is to show that most chords, i.e., edges in the random matching, lead to new vertices that are sufficiently far away from any previously visited vertices when we explore the

graph starting from a fixed vertex v. In the proof, we have to proceed in two phases. In the first phase, we consider distances that are relatively short with respect to the diameter, and when most vertices have not been visited yet; in the second, we consider distances above that threshold l.

Let C denote the n-cycle, and M the random matching, so that $G = C \cup M$. Also, let $d_C(u, v)$ denote the distance between u and v in C; note that $d_C(u, v) \leq n/2$. We start at a vertex v, and define circles and balls around v as follows.

$$S_i = \{u : d(u, v) = i\}, B_i = \cup_{j \leq i} S_j = \{u : d(u, v) \leq i\} \tag{6.8}$$

Short distances $i \leq l_1(n) = (1/5) \log_2 n$. Consider a chord (u, v) where $u \in S_i$ and $v \in S_{i+1}$. We call such a chord *local* if v is close on the cycle to at least one other vertex in B_{i+1}, i.e., if $d_C(v, v') \leq 2 \log_2 n$ for any other $v' \in B_{i+1}$.

Note that $|B_i| \leq 3 \cdot 2^{i-1}$, because after the initial 3 neighbors of v, each node in the previous stage gives rise to at most 2 children.

The probability that a new chord after step i is local is at most

$$\frac{2|B_{i+1}| 2 \log_2 n}{n} \leq \frac{4 \cdot 3 \cdot 2^i \log_2 n}{n}, \tag{6.9}$$

because in the worst case, there are $2 \log_2 n$ forbidden nodes on both sides for each node in B_{i+1}.

We now want to compute the probability that local chords are rare while $i \leq l_1$. Specifically, consider all the chords traversed in the first l_1 steps, of which there are $|B_{l_1}|$. Therefore, the probability that there are two or more chords in this set is

$$\mathbb{P}\{\text{at least two local chords}\} \leq \left(\begin{array}{c} 3 \cdot 2^{l_1} \\ 2 \end{array} \right) \left(\frac{12 \cdot 2^{l_1} \log_2 n}{n} \right)^2 = O(n^{-6/5} (\log_2 n)^2) = o(n^{-1}). \tag{6.10}$$

A union bound over all n starting vertices v then shows that *a.a.s.* for every starting vertex v, there is at most one local chord up to step l_1. From now on, we condition on the event E_1 that this is true.

Fresh neighbors on one or two sides. We have shown that most chords lead to vertices that are at least $2 \log_2 n$ from other already discovered vertices. We now use this property to define two sets of vertices at step i. The set C_i contains vertices that have at least $2 \log_2 n$ untouched vertices on one side (on the cycle), and there is a unique edge connecting each such vertex to B_{i-1} (which guarantees two new edges in the next step); the set D_i contains vertices that have at least $2 \log_2 n$ vertices on both sides, and a single connecting edge as well.

Consider a vertex $v \in C_i$. A neighbor u of v on the circle becomes an element of C_{i+1}, unless a local chord falls into the interval of free vertices next to u at step $i + 1$ (or hits u itself). Also, because $v \in C_i$, there is a fresh chord (v, u) to another $u \in D_{i+1}$ unless (v, u) is a local chord. Similarly, for $v \in D_i$, both neighbors of v become elements of C_{i+1}, unless a local chord hits on either side of v. Thus, in the absence of any local chords, the two sets C_i and D_i satisfy

$$\begin{aligned} |C_{i+1}| &= |C_i| + 2|D_i| \\ |D_{i+1}| &= |C_i|, \end{aligned} \tag{6.11}$$

because each element of D_i has two unexplored neighbors, giving rise to two elements in C_{i+1}; each element of C_i has one neighbor and one chord emanating from it, giving rise to one element in C_{i+1} and one in D_{i+1}.

If there are local chords, clearly $C_i \cup D_i \subset S_i$. Conditional on E_1, the worst case (giving smallest C_i and D_i) is when the unique local chord goes to one of the neighbors of v. In that case,

$|C_1| = 2, |D_1| = 0$, $|C_2| = 2$, $|D_2| = 2$, and generally

$$\begin{aligned} |C_i| &\geq 2^{i-2} \\ |D_i| &\geq 2^{i-3}. \end{aligned} \tag{6.12}$$

Long distances $l_1(n) < i \leq l_2(n) = (3/5)\log_2 n$. The probability that a chord is local is

$$p = \mathbb{P}\{\text{chord local}\} \leq \frac{12 \cdot 2^i \log_2 n}{n} < n^{-1/6}. \tag{6.13}$$

There are at most 2^i chords leaving the set S_i. The probability that there are at least $2^i n^{-1/10}$ local chords leaving S_i is at most

$$\left(\begin{array}{c} 2^i \\ 2^i n^{-1/10} \end{array} \right) p^{2^i n^{-1/10}} \leq n^{-5}. \tag{6.14}$$

A union bound over all n starting vertices and all $(2/5)\log_2 n$ time steps i shows that *a.a.s.* , at most $2^i n^{-1/10}$ chords leave every S_i. Call this event E_2.

We now lower-bound the sizes of C_i and D_i, conditional on the event $E_1 \cap E_2$, i.e., of having at most one local chord in the first phase, and at most $2^i n^{-1/10}$ local chords in each step of the second phase.

A neighbor u of $v \in C_i$ is sure to be an element of C_{i+1}, except if a local chord falls into the interval of free vertices next to u at step $i+1$. The difference equations can be bounded as follows:

$$\begin{aligned} |C_{i+1}| &\geq |C_i| + 2|D_i| - 2^{i+1} n^{-1/10} \\ |D_{i+1}| &\geq |C_i| - 2^{i+1} n^{-1/10} \end{aligned} \tag{6.15}$$

Therefore, for n large enough, over the entire range of $i \leq l_2(n)$,

$$\begin{aligned} |C_i| &\geq 2^{i-3} \\ |D_i| &\geq 2^{i-4} \end{aligned} \tag{6.16}$$

All vertices are close. Set $i^\star = (1/2)(\log_2 n + \log_2 \log n + c) \leq (3/5)\log_2 n$. Suppose we go through the discovery process described above for two different starting vertices v' and v'', to generate two sets $C_{i^\star}(v')$ and $C_{i^\star}(v'')$. Assuming that the two balls around v and v'' have not touched yet, we compute the probability that the set $C_{i^\star}(v')$ and $C_{i^\star}(v'')$ will be connected by one of the edges generated in the next step. Specifically, we can conservatively focus only on the chords generated in the next step by vertices in $C_{i^\star}(v')$, and ask whether they will hit any vertex in $C_{i^\star}(v'')$. The probability that none hits is upper-bounded by

$$\begin{aligned} \mathbb{P}\{d(v', v'') > 2i^\star + 1 | E_1 \cap E_2\} &\leq \left(1 - \frac{2^{i^\star - 3}}{n}\right)^{2^{i^\star - 3}} \\ &\leq \exp\left(-2^{2i^\star - 7}/n\right) \qquad \text{using } (p \leq -\log(1-p)) \\ &\leq \exp\left(-2^{c-7}\log n\right) \\ &\leq n^{-4} = o(n^{-2}), \end{aligned} \tag{6.17}$$

for $c \geq 9$.

We can now bound the diameter of the entire graph.

$$\mathbb{P}\{\text{diam}(G) > 2i^\star + 1\} \leq \mathbb{P}\left\{\bar{E_1}\right\} + \mathbb{P}\left\{\bar{E_2}\right\} + \sum_{v', v''} \mathbb{P}\{d(v', v'') > 2i^\star + 1 | E_1 \cap E_2\} = o(1). \tag{6.18}$$

Therefore, G has diameter $2i^\star + 1 = \log_2 n + \log_2 \log n + 10$ *a.a.s.*

■

Therefore, the addition of the random matching has decreased the diameter significantly, from $n/2$ to approximately $\log_2 n$. Similar results are known for other classes of random graphs; while the techniques are similar, the proofs are typically (even) more involved than the one above. We give the statement of two such results, for which the diameter $\mathrm{diam}(G)$ refers to the diameter of the largest connected component of the graph.

Theorem 6.2 (Diameter of $G(n,p)$ [**?**])**.** *The diameter of the giant component of $G(n,p)$ for $\log n > np \to \infty$ satisfies*

$$diam(G(n,p)) = (1+o(1))\frac{\log n}{\log np} \text{ a.a.s. .} \tag{6.19}$$

Theorem 6.3 (Diameter of $G(n,r)$ [**?**])**.** *Let $r \geq 3$ and $\epsilon > 0$ be fixed. Then the diameter of the largest connected component of $G(n,r)$*

$$\frac{diam(G(n,r))}{\log n/\log(r-1)} \in (1-\epsilon, 1+\epsilon) \text{ a.a.s. .} \tag{6.20}$$

Theorem 6.4 (Average Shortest Path Length of $WS(n,k,p)$ [**?**])**.** *Then the average shortest path length of $WS(n,k,p)$*

$$\frac{\log(n/k)}{\log k}. \tag{6.21}$$

Thus, random graphs do possess the small world property, in that their diameter behaves like $\log n$. In a sense, the absence of structure in random graphs, i.e., that edges are independent, ensure that the k-hop neighborhood of a vertex v grows quickly with k, a fact we had explicitly used in the proof of the emergence of the giant component for $G(n, c/n)$.

Bibliography

[1] B. Bollobas and F. R. K. Chung. The Diameter of a Cycle plus a Random Matching. *SIAM J. Disc. Math.*, 1(3), August 1988.

[2] B. Bollobas and F. de la Vega. The Diameter of Random Regular Graphs. *Combinatorica*, 2:125–134, 1982.

[3] F. Chung and L. Lu. The Diameter of Sparse Random Graphs. *Advances in Applied Mathematics*, 26:257–279, 2001.

[4] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[5] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 363:202–204, June 1998.