Graph mining
SD212
2. Graph structure

Thomas Bonald
Institut Polytechnique de Paris

2019 – 2020

These lecture notes focus on two key properties of real graphs: the small-world property and the clustering structure (my friends tend to be friends). A graph model having both properties is then described.

1 Small-world property

The small-world property refers to the fact that any pair of nodes is connected by some short path compared to the size of the graph. In social networks, this is the well-known *six degrees of separation* principle stating that all people are at most six links from each other. This somewhat surprising result was originally imagined by Karinthy as early as 1929, well before the advent of online social networks:

> A fascinating game grew out of this discussion. One of us suggested performing the following experiment to prove that the population of the Earth is closer together now than they have ever been before. We should select any person from the 1.5 billion inhabitants of the Earth - anyone, anywhere at all. He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances. For example, "Look, you know Mr. X.Y., please ask him to contact his friend Mr. Q.Z., whom he knows, and so forth."

This idea was verified experimentally by Milgram in 1967. Recent experiments on Facebook have shown typical degrees of separation of 3 or 4 [1] Similar results have been shown for other graphs, like Wikipedia [2].

Where does this property come from? Can we formalize it? Clearly, a graph structured as a grid (like the streets of a city) have shortest paths of order $O(\sqrt{n})$, about 100 hops for 10,000 nodes. There is no small-world phenomenon there. What about random graphs? We shall see that shortest paths are of order $O(\ln n)$, that is closer to what is observed in real graphs.

Consider a large graph where nodes are connected independently at random, with some degree distribution μ. A large Erdős-Rényi graph, for instance, results in a Poisson degree distribution with parameter $\lambda \approx np$ where n is the number of nodes and p the probability of connection between any two nodes. We remove the isolated nodes so that the support of μ is on $\{1, 2, \ldots, n\}$. It has been shown in [1] that the average path between two distinct nodes can be approximated by:

$$\frac{\ln n - \gamma + \ln(\mathrm{E}(D^2) - \mathrm{E}(D)) - 2\mathrm{E}(\ln D)}{\ln(\mathrm{E}(D^2) - \mathrm{E}(D)) - \ln \mathrm{E}(D)} + \frac{1}{2}, \tag{1}$$

[1] See `https://research.fb.com/three-and-a-half-degrees-of-separation/`.
[2] Try `https://www.sixdegreesofwikipedia.com` !

where n is the number of nodes, $\gamma \approx 0.58$ is Euler's constant and D is a random variable with distribution μ (the empirical degree distribution on non-isolated nodes). Observe that this expression is well-defined whenever $\mathrm{E}(D) > 2$, using the fact that $\mathrm{E}(D^2) \geq \mathrm{E}(D)^2$.

For an Erdős-Rényi graph, given that we have removed isolated nodes, we get:

$$\mu(k) = \frac{e^{-\lambda}}{1 - e^{-\lambda}} \frac{\lambda^k}{k!},$$

so that:

$$\mathrm{E}(D) = \frac{\lambda}{1 - e^{-\lambda}}, \quad \mathrm{E}(D^2) = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}}.$$

Neglecting the term $\mathrm{E}(\ln D)$, we get the following conservative estimate of the average path length:

$$\frac{\ln n - \gamma + \ln \lambda}{\ln \lambda - \ln(1 - e^{-\lambda})} + \frac{3}{2}.$$

The term $\ln(1 - e^{-\lambda})$ being negligible whenever $\lambda > 4$, we obtain the simple approximation:

$$\frac{\ln n - \gamma}{\ln \lambda} + \frac{5}{2}.$$

For $n = 10,000$ nodes, this means an average path length of 7.9 for $\lambda = 5$ and 6.2 for $\lambda = 10$.

For power-law graphs, the computation is more involved as the second moment $\mathrm{E}(D^2)$ may now depend on n. We obtain the following average path length [1]:

$$\begin{cases} \frac{2}{3-\alpha} + \frac{1}{2} & \text{for } \alpha \in (2,3), \\ \frac{\ln n}{\ln \ln n} + \frac{3}{2} & \text{for } \alpha = 3. \end{cases}$$

Note that the average path length becomes *independent* of n for $\alpha < 3$ (and the approximation becomes bad for α close to 3). For $n = 10,000$ nodes, we get an average path length of 4.5 for $\alpha = 2.5$ and 5.6 for $\alpha = 3$.

2 Clustering coefficient

Another key property of real graphs is their tendency to cluster: nodes having a common neighbor tend to be connected (my friends tend to be friends). This can be measured through the clustering coefficient, counting the fraction of closed triangles:

$$C = \frac{\sum_{i,j<k} A_{ij} A_{ik} A_{jk}}{\sum_{i,j<k} A_{ij} A_{ik}},$$

where A is the adjacency matrix of the graph, assumed undirected, unweighted and without self-loops. Observe that each triangle is counted 3 times in the numerator (one for each vertex of the triangle). Now

$$\sum_{i,j<k} A_{ij} A_{ik} A_{jk} = \frac{1}{2} \sum_{i,j,k} A_{ij} A_{ik} A_{jk}$$

and, denoting by d_i the degree of node i,

$$\sum_{j<k} A_{ij} A_{ik} = \frac{1}{2} \sum_{j \neq k} A_{ij} A_{ik} = \frac{1}{2} \sum_{k} (d_i - 1) A_{ik} = \frac{1}{2} d_i (d_i - 1).$$

We deduce that:

$$C = \frac{\sum_{i,j,k} A_{ij} A_{ik} A_{jk}}{\sum_i d_i (d_i - 1)}. \tag{2}$$

Each triangle is now counted 6 times in the numerator (one per permutation of the vertices of the triangle).

The clustering coefficient of a node i counts the fraction of closed triangles involving i and two of its neighbors. We get similarly:

$$C_i = \frac{\sum_{j<k} A_{ij}A_{ik}A_{jk}}{\sum_{j<k} A_{ij}A_{ik}} = \frac{\sum_{j,k} A_{ij}A_{ik}A_{jk}}{d_i(d_i-1)}. \tag{3}$$

The average clustering coefficient of all nodes is not equal to the clustering coefficient of the graph C unless node i is sampled with probability proportional to $d_i(d_i-1)$:

$$C = \frac{\sum_i d_i(d_i-1)C_i}{\sum_i d_i(d_i-1)}.$$

There is a simple interpretation of the sampling distribution proportional to $d_i(d_i-1)$: this is the distribution induced by common neighbors. Recall that $d_i(d_i-1) = \sum_{j\neq k} A_{ij}A_{ik}$. Sampling in proportion to $d_i(d_i-1)$ reduces to first sample two nodes j,k uniformly at random and then to select one of their common neighbors uniformly at random, if any (otherwise, resample j and k). So the clustering coefficient C is simply the probability that two nodes having a common neighbor are connected.

A natural question is whether clustering emerges from randomness, like the small-world property. The answer is no. Take an Erdős-Rényi graph for instance, with n nodes and probability of connection p. Then the probability that two nodes are connected is p, independently of whether they have a common neighbor or not. So the expected clustering coefficient is p, which is equal to the density of the graph and is typically very low (e.g., a graph of $n = 10,000$ with average degree $d = 10$ means a density $p = d/(n-1) \approx 10^{-3}$). The clustering coefficient of real graphs like social or information networks is typically much larger.

3 Watts-Strogatz graphs

The Watts-Strogatz model [2] is a random graph built as follows. We start with a ring of n nodes, with each node connected to its d closest neighbors on the ring, for some $d < n$ (see Figure 1). The parameter d is assumed to be even. Then each edge is rewired at random with probability p, while avoiding multi-edges and self-loops. Specifically, for each edge i,j of the initial graph, with $j \in \{i+1,\dots,i+\frac{d}{2}\} \mod n$, replace that edge by some random edge i,j' with probability p, where j' is chosen uniformly at random in the set $\mathcal{N} \setminus \{i\} \cup \{j\}$, where $\mathcal{N}$ is the set of nodes that are not neighbors of i in the current graph. Note that we add j so that the set $\mathcal{N} \setminus \{i\} \cup \{j\}$ is not empty.

For $p = 0$, the graph has a high clutering coefficient (see the appendix for the exact computation) but no small-world property (average path length in $O(n)$, for fixed d). For $p = 1$, the graph is close to an Erdős-Rényi graph[3]: it has the small-world property (average path length in $O(\ln n)$, for fixed d) but a low clustering coefficient. For suitable values of p, the graph has both a high clustering coefficient (due to the initial ring topology) and the small-world property (due to the random edges). This is the simplest model capturing these two key properties of real graphs.

Appendix

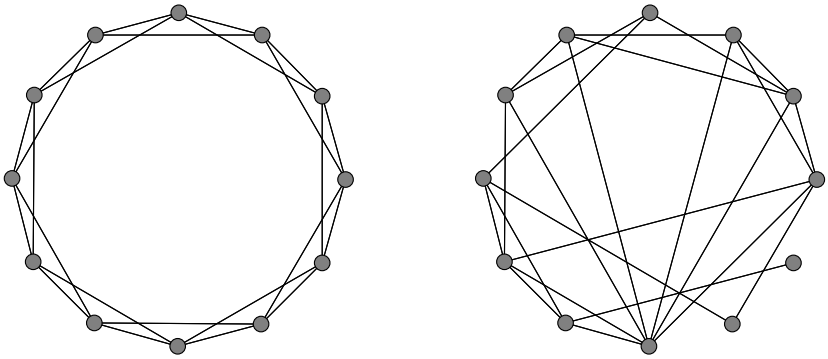
Consider the Watts-Strogatz model with $p = 0$. We assume that $d < n/2$. Each node is connected to its d closest neighbors on the ring. Let $K = d/2$. The total number of triangles is:

$$n\binom{K}{2}.$$

It then follows from (2) that:

$$C = \frac{3K(K-1)}{d(d-1)} = \frac{3}{4}\frac{d-2}{d-1}.$$

[3]It is not an Erdős-Rényi graph since only one of the edge ends is modified; in particular, each node has degree at least $d/2$.



(a) Ring ($p = 0$)

(b) Random rewiring ($p = 0.4$)

Figure 1: Watts-Strogatz graphs with $n = 12$ nodes and average degree $d = 4$.

Thus the clustering coefficient is close to $3/4$ for large values of d.

References

[1] A. Fronczak, P. Fronczak, and J. A. Hołyst. Average path length in random networks. *Physical Review E*, 2004.

[2] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 1998.