# Graph mining
# SD212
# Graph structure

Thomas Bonald

2018 – 2019

These lecture notes present 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 $p$ (so a proportion $p_k$ of nodes have degree $k$). An Erdös-Rényi graph, for instance, results in a Poisson degree distribution with parameter $\lambda$. Arbitrary degree distributions can be generated through the configuration model. We remove the isolated nodes so that $p_0 = 0$. It has been shown in [1] that the average path between two distinct nodes can be approximated by:

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

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

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

[2] Try `https://www.sixdegreesofwikipedia.com` !

For an Erdös-Rényi graph, conditionned on the fact that there are no isolated nodes, we get:

$$p_k = \frac{e^{-\lambda}}{1-e^{-\lambda}}\frac{\lambda^k}{k!},$$

so that:

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

Neglecting the term $\mathrm{E}(\ln X)$, 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 $\mathrm{E}(X^2)$ may now depend on $n$, depending on the parameter $\alpha$ of the power law. 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 $\alpha$ 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 the tendency to cluster: nodes having a common neighbor tend to be connected. This can be measured through the clustering coefficient, counting the fraction of closed triangles:

$$C = \frac{\sum_u \sum_{v<w} 1_{u\sim v} 1_{u\sim w} 1_{v\sim w}}{\sum_u \sum_{v<w} 1_{u\sim v} 1_{u\sim w}},$$

where we use the notation $u \sim v$ for a link between nodes $u, v$. Observe that each triangle is counted three times (i.e., there are 3 closed triangles).

The local clustering coefficient of each node $u$ counts the fraction of closed triangles involving $u$ and two of its neighbors:

$$C_u = \frac{\sum_{v<w} 1_{u\sim v} 1_{u\sim w} 1_{v\sim w}}{\sum_{v<w} 1_{u\sim v} 1_{u\sim w}}.$$

Denoting by $d_u$ the degree of node $u$ and assuming that there are no self-loops, we get:

$$\sum_{v<w} 1_{u\sim v} 1_{u\sim w} = \frac{1}{2}\sum_{v\neq w} 1_{u\sim v} 1_{u\sim w} = \frac{1}{2}(d_u - 1)\sum_v 1_{u\sim v} = \frac{1}{2} d_u(d_u-1),$$

so that

$$C_u = \frac{\sum_{v<w} 1_{u\sim v} 1_{u\sim w} 1_{v\sim w}}{\frac{1}{2} d_u(d_u-1)}. \tag{2}$$

Observe that the numerator is the number of pairs of neighbors of $u$ that are connected.

The average local clustering coefficient is not equal to the global clustering coefficient $C$ unless node $u$ is sampled with probability proportional to $d_u(d_u-1)$. There is a simple interpretation of the sampling

distribution proportional to $d_u(d_u - 1)$: this is the distribution induced by the common neighbors. More precisely, node $u$ is sampled with probability proportional to[3]:

$$\sum_{v<w} 1_{u\sim v} 1_{u\sim w}.$$

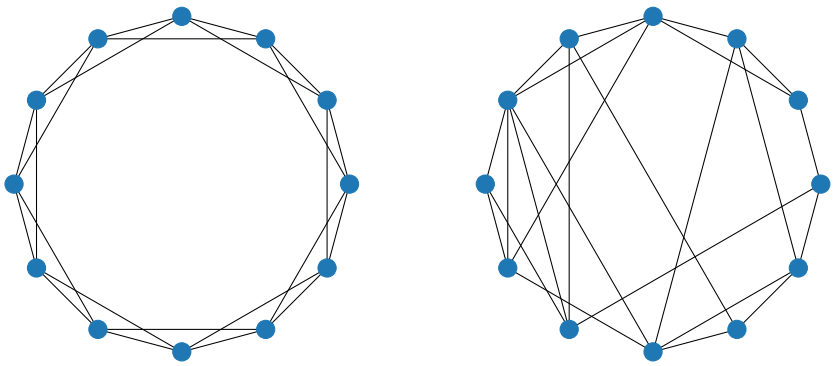
So the clustering coefficient $C$ is the probability that two nodes having a common neighbor are connected.

A natural question is whether clustering emerges from randomness, like the small-world property. The answer is no. Take an Erdös-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 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 $u, v$ of the initial graph, with $v \in \{u+1, \dots, u+\frac{d}{2}\} \mod n$, replace that edge by some random edge $u, v'$ with probability $p$, where $v'$ is chosen uniformly at random in the set $\mathcal{N} \setminus \{u\} \cup \{v\}$, where $\mathcal{N}$ is the set of nodes that are not neighbors of $u$ in the current graph. Note that we add $v$ so that the set $\mathcal{N} \setminus \{u\} \cup \{v\}$ 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 Erdös-Rényi graph[4]: it has the small-world property (average path length in $O(\ln n)$, for fixed $d$) but a low clustering coefficient. For some properly chosen value 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).



(a) Ring ($p = 0$)

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

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

[3] Note that the pair $v, w$ is sampled in proportion to its number of common neighbors.

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

# 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. We seek to compute the clustering coefficient $C$ of the graph, which is also the clustering coefficient of each node. We need to compute the number of pairs of neighbors of a given node $u$ that are connected.

Let $k = d/2$. Consider the $i$-th neighbor of $u$ on its right side, for $i = 1, \ldots, k$. This node is connected to $k$ nodes on its left (including $u$) and $k$ nodes on its right, among which $k - i$ are neighbors of $u$. Thus the total number of pairs of neighbors of $u$ that are connected is:

$$\frac{1}{2} \times 2 \left( k(k-1) + \frac{1}{2}k(k-1) \right) = \frac{3}{2}k(k-1),$$

where the factor 2 comes from the $k$ neighbors on the left side of $u$ and the factor $\frac{1}{2}$ from the fact that each pair is counted twice. We deduce from (2) that:

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

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.