Graph Learning
SD212
1. Graph Structure

Thomas Bonald
Institut Polytechnique de Paris

2023 – 2024

These lecture notes focus on some key properties of graphs: the friendship paradox (your friends tend to have more friends than you), the small-world property (you can reach anyone on earth through a small chain of friends) and the tendency to cluster (my friends tend to be friends).

1 The friendship paradox

Consider a graph of n nodes and m edges. The graph is assumed to be undirected and without self-loops. We denote by A its adjacency matrix and by $d = A1$ the vector of node degrees, which we assume positive. Observe that:

$$\sum_{i=1}^{n} d_i = 1^T A 1 = 2m.$$

Let $X \in \{1, \ldots, n\}$ be a random node and D its degree.

Node sampling. If the node is sampled uniformly at random, we have:

$$\forall i = 1, \ldots, n, \quad P(X = i) = \frac{1}{n}.$$

The corresponding degree distribution is given by:

$$\forall k \geq 0, \quad P(D = k) = \sum_{i=1}^{n} P(X = i) 1_{\{d_i = k\}} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{d_i = k\}}.$$

This is the empirical degree distribution, which we denote by p. The expected degree is:

$$\mathrm{E}(D) = \sum_{k \geq 0} k p_k = \frac{1}{n} \sum_{i=1}^{n} d_i = \frac{2m}{n}.$$

Edge sampling. Now choose an edge uniformly at random and one of the two ends of this edge uniformly at random. Since the graph is undirected, the sampling distribution is now:

$$\forall i = 1, \ldots, n, \quad P'(X = i) = \frac{1}{2m} \sum_{i=1}^{n} A_{ij} = \frac{d_i}{2m}.$$

Observe that each node is sampled with a probability proportional to its degree. The corresponding degree distribution is given by:

$$\forall k \geq 0, \quad P'(D = k) = \sum_{i=1}^{n} P'(X = i) 1_{\{d_i = k\}} = \frac{1}{2m} \sum_{j=1}^{n} k 1_{\{d_i = k\}}.$$

This is the *size-biased* empirical degree distribution p', given by:

$$\forall k \geq 0, \quad p'_k \propto k p_k = \frac{k p_k}{\mathrm{E}(D)}$$

Observe that:

$$\mathrm{E}'(D) = \frac{\mathrm{E}(D^2)}{\mathrm{E}(D)} \geq \mathrm{E}(D),$$

with equality if and only if $\mathrm{var}(D) = 0$, that is, the graph is regular (all nodes have the same degree).

Neighbor sampling. Finally, we sample a node uniformly at random and one of its neighbors uniformly at random. The sampling distribution for this node is:

$$\forall i = 1, \ldots, n, \quad P''(X = i) = \frac{1}{n} \sum_{j=1}^{n} P_{ji},$$

where $P_{ji} = A_{ji}/d_j$ is the probability of choosing neighbor i from node j. The corresponding degree distribution is given by:

$$\forall k \geq 0, \quad P''(D = k) = \sum_{i=1}^{n} P''(X = i) 1_{\{d_i = k\}} = \frac{1}{n} \sum_{i,j=1}^{n} 1_{\{d_i = k\}} P_{ji}.$$

The following result shows the *friendship paradox*: your friends have more friends than you on average.

The friendship paradox

The random neighbor of a a random node has a higher degree than that of a random node on average.

Proposition 1 *We have* $\mathrm{E}''(D) \geq \mathrm{E}(D)$ *with equality if and only if each connected component of the graph is regular.*

Proof. We have:

$$\mathrm{E}''(D) = \sum_{k \geq 0} k P''(D = k) = \frac{1}{n} \sum_{i,j=1}^{n} d_i P_{ji} = \frac{1}{n} \sum_{i,j=1}^{n} \frac{d_i}{d_j} A_{ji}.$$

By symmetry,

$$\mathrm{E}''(D) = \frac{1}{2n} \sum_{i,j=1}^{n} \left(\frac{d_i}{d_j} + \frac{d_j}{d_i} \right) A_{ij}.$$

Using the fact that $x + 1/x \geq 2$ for all $x > 0$ with equality if and only if $x = 1$, we get

$$\mathrm{E}''(D) \geq \frac{2m}{n} = \mathrm{E}(D)$$

with equality if and only if $d_i = d_j$ for all edges i, j (all pairs i, j such that $A_{ij} = 1$), that is, if and only if each connected component of the graph is regular. □

2 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].

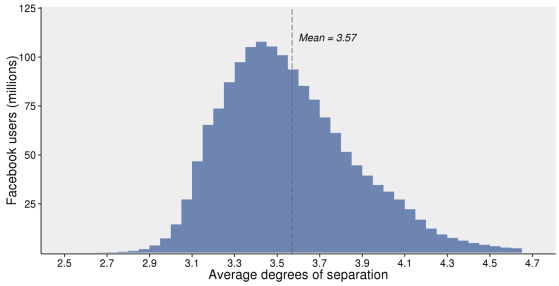

Figure 1: The small-world of Facebook.

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.

Small world

A graph of n nodes has the small-world property if the length of the shortest path between any two nodes of the graph is small compared to n. This length is typically (at most) logarithmic in n.

Consider a large graph where nodes are connected independently at random. Let's remove the isolated nodes so that each node has at least degree 1. It has been shown in [1] that the average length of a 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 the degree of a random node. Observe that this expression is well-defined whenever $\mathrm{E}(D) > 2$, using the fact that $\mathrm{E}(D^2) \geq \mathrm{E}(D)^2 > \mathrm{E}(D)$.

For a large Erdős-Rényi graph, the degree distribution is approximately Poisson with parameter $\lambda \approx np$ where p is the probability of connection between any two nodes. This distribution is independent of n so that the average path length is logarithmic in n. For a power-law degree distribution, the computation is more involved as the second moment $\mathrm{E}(D^2)$ typically depends on n. It can be proved that, for an exponent $\alpha < 3$, the average path length remains finite when n grows to infinity [1].

3 Clustering property

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:

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

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

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

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:

$$C = \frac{3\sum_{i<j<k} A_{ij}A_{ik}A_{jk}}{\sum_i \binom{d_i}{2}}. \tag{2}$$

The numerator is 3 times the number of closed triangles; the denominator is the number of candidate triangles (open or closed).

Global clustering coefficient

The clustering coefficient C of a graph is the fraction of closed triangles in the graph:

$$C = \frac{3 \times \#\text{triangles}}{\sum_i \binom{d_i}{2}}.$$

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

$$C_i = \frac{\sum_{j<k} A_{ij}A_{ik}A_{jk}}{\sum_{j<k} A_{ij}A_{ik}} = \frac{\sum_{j<k} A_{ij}A_{ik}A_{jk}}{\binom{d_i}{2}}. \tag{3}$$

Local clustering coefficient

The clustering coefficient C_i of node i is the fraction of closed triangles containing i:

$$C_i = \frac{\#\text{triangles containing } i}{\binom{d_i}{2}}.$$

Note that 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 $\binom{d_i}{2}$:

$$C = \frac{\sum_i \binom{d_i}{2} C_i}{\sum_i \binom{d_i}{2}}.$$

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.

References

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