Graph mining
SD212
1. Sampling

Thomas Bonald
Institut Polytechnique de Paris

April 2020

These lecture notes present some properties related to node sampling in graphs. In particular, we show the so-called *friendship paradox*: in a social network, your friends have more friends than you on average.

1 Graphs

We first consider a graph of n nodes and m edges. The graph is assumed to be undirected, 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:

$$\forall j = 1, \ldots, n, \quad \mathrm{P}_0(X = j) = \frac{1}{n}.$$

The corresponding degree distribution is given by:

$$\forall k \geq 0, \quad \mathrm{P}_0(D = k) = \sum_{j=1}^n \mathrm{P}_0(X = j) 1_{\{d_j = k\}} = \frac{1}{n} \sum_{j=1}^n 1_{\{d_j = k\}}.$$

This is the empirical degree distribution, which we denote by μ_0. The expected degree is:

$$\mathrm{E}_0(D) = \sum_{k \geq 0} k \mu_0(k) = \frac{1}{n} \sum_{j=1}^n d_j = \frac{2m}{n}.$$

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[1]:

$$\forall j = 1, \ldots, n, \quad \mathrm{P}_\infty(X = j) = \frac{1}{2m} \sum_{i=1}^n A_{ij} = \frac{d_j}{2m}.$$

[1]The notation P_∞ is related to the interpretation in terms of random walk, see Proposition 2 below.

The corresponding degree distribution is given by:

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

This is the *size-biased* empirical degree distribution μ_∞, given by:

$$\forall k \geq 0, \quad \mu_\infty(k) \propto k\mu_0(k) = \frac{k\mu_0(k)}{\mathrm{E}_0(D)}$$

Observe that:

$$\mathrm{E}_\infty(D) = \frac{\mathrm{E}_0(D^2)}{\mathrm{E}_0(D)} \geq \mathrm{E}_0(D),$$

with equality if and only if $\mathrm{var}_0(D) = 0$, that is, the graph is regular (all nodes have the same degree).

Neighbor sampling. Now choose a node uniformly at random and one of its neighbors uniformly at random. The sampling distribution is:

$$\forall j = 1, \ldots, n, \quad \mathrm{P}_1(X = j) = \frac{1}{n} \sum_{i=1}^n P_{ij},$$

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

$$\forall k \geq 0, \quad \mathrm{P}_1(D = k) = \sum_{j=1}^n \mathrm{P}_1(X = j) 1_{\{d_j = k\}} = \frac{1}{n} \sum_{i,j=1}^n k 1_{\{d_j = k\}} P_{ij}.$$

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

Proposition 1 *We have* $\mathrm{E}_1(X) \geq \mathrm{E}_0(X)$ *with equality if and only if each connected component of the graph is regular.*

Proof. We have:

$$\mathrm{E}_1(D) = \sum_{k \geq 0} k \mathrm{P}_1(D = k) = \frac{1}{n} \sum_{i,j=1}^n d_j P_{ij} = \frac{1}{n} \sum_{i,j=1}^n \frac{d_j}{d_i} A_{ij}.$$

By symmetry,

$$\mathrm{E}_1(D) = \frac{1}{2n} \sum_{i,j=1}^n \left(\frac{d_i}{d_j} + \frac{d_j}{d_i} \right) A_{ij}.$$

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

$$\mathrm{E}_1(D) \geq \frac{2m}{n} = \mathrm{E}_0(D)$$

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. □

Random walk. One may consider the random neighbor of a random neighbor. More generally, consider the sampling distribution obtained after t hops of the random walk in the graph, starting from the uniform distribution. Let π_t be the corresponding distribution, expressed as a row vector. Since the random walk defines a Markov chain with transition matrix P, we have:

$$\pi_{t+1} = \pi_t P,$$

with

$$\pi_0 = \frac{1}{n}(1, \dots, 1).$$

In particular,

$$\pi_t = \pi_0 P^t.$$

Proposition 2 *If the graph is connected and not bipartite, the sampling distribution after t hops of the random walk converges to the edge sampling distribution,*

$$\lim_{t \to +\infty} \pi_t = \frac{d^T}{2m}.$$

Proof. Since the graph is connected and not bipartite, the Markov chain is irreducible and aperiodic. In particular, the distribution π_t has a limit π, which is the unique solution to the balance equations:

$$\pi = \pi P.$$

Now

$$d^T P = 1^T A = d^T,$$

which shows that $\pi \propto d^T$. □

Self-loops. The results apply in presence of self-loops. We still define the degree vector by $d = A1$ (so that a self-loop is counted once in the degree). Denoting by ℓ the number of self-loops and by m the number of other edges, we get:

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

Edge sampling consists in selecting a positive entry of the adjacency matrix A uniformly at random. Observe that regular edges are sampled twice as often as self-loops. In terms of random walk, this reflects the fact that, unlike self-loops, a regular edge i, j with $i \neq j$ can be visited in each direction, $i \to j$ and $j \to i$.

Weighted graphs. The results also apply to weighted graphs, with $d = A1$ the vector of node weights. We are interested in the sampled node X and its weight W:

- **Node sampling:** X has the uniform distribution over $\{1, \dots, n\}$ and W has the empirical weight distribution.
- **Edge sampling:** Edges are sampled in proportion to their weights; W has the sized-biased empirical weight distribution, which is higher in expectation (unless all nodes have the same weight). This is the limit of the sampling distribution obtained after t hops of the random walk, when t tends to $+\infty$.
- **Neighbor sampling:** Neighbors are sampled in proportion to the edge weights; W is higher in expectation (unless node weights are the same in each connected component of the graph). This is the sampling distribution obtained after 1 hop of the random walk.

2 Directed graphs

We now consider a directed graph of n nodes and m edges, without self-loops. We denote by A the adjacency matrix A and by $d^+ = A1$ and $d^- = A^T 1$ the vectors of out-degrees and in-degrees. We have:

$$\sum_{i=1}^n d_i^+ = \sum_{i=1}^n d_i^- = m.$$

Let $X \in \{1, \dots, n\}$ be a random node, D^+ and D^- its out-degree and in-degree.

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

$$\forall j = 1, \dots, n, \quad \mathrm{P}_0(X = j) = \frac{1}{n}.$$

The corresponding degree distributions are given by:

$$\forall k \geq 0, \quad \mathrm{P}_0(D^+ = k) = \frac{1}{n} \sum_{j=1}^n 1_{\{d_j^+ = k\}}, \quad \mathrm{P}_0(D^- = k) = \frac{1}{n} \sum_{j=1}^n 1_{\{d_j^- = k\}}.$$

This are the empirical degree distributions, which we denote by μ_0^+ and μ_0^-. The expected degrees are:

$$\mathrm{E}_0(D^+) = \mathrm{E}_0(D^-) = \frac{m}{n}.$$

Edge sampling. Now choose an edge uniformly at random. Let D^+ be the out-degree of the origin of this edge and D^- the in-degree of the end of this edge. We have:

$$\forall k \geq 0, \quad \mathrm{P}_\infty(D^+ = k) = \frac{1}{m} \sum_{j=1}^n k 1_{\{d_j^+ = k\}}, \quad \mathrm{P}_\infty(D^- = k) = \frac{1}{m} \sum_{j=1}^n k 1_{\{d_j^- = k\}}.$$

These are the *size-biased* empirical degree distributions:

$$\forall k \geq 0, \quad \mu_\infty^+(k) \propto k \mu_0^+(k), \quad \mu_\infty^-(k) \propto k \mu_0^-(k).$$

Observe that:

$$\mathrm{E}_\infty(D^+) \geq \mathrm{E}_0(D^+) \quad \text{and} \quad \mathrm{E}_\infty(D^-) \geq \mathrm{E}_0(D^-),$$

with equality if and only if $\mathrm{var}_0(D^+) = 0$ and $\mathrm{var}_0(D^-) = 0$, respectively.

Random successor, random predecessor. Now choose a node uniformly at random among nodes of positive out-degrees (thus excluding sinks). Denote by D^- the in-degree of one of its successors, chosen uniformly at random. We have:

$$\forall k \geq 0, \quad \mathrm{P}_1(D^- = k) = \frac{1}{n^+} \sum_{i,j=1}^n 1_{\{d_i^+ \geq 1\}} 1_{\{d_j^- = k\}} P_{ij},$$

where n^+ is the number of nodes of positive out-degrees and $P_{ij} = A_{ij}/d_i^+$ is the probability of choosing successor j from node i. Thus

$$\mathrm{E}_1(D^-) = \sum_{k \geq 0} k \mathrm{P}_1(D^- = k) = \frac{1}{n^+} \sum_{i,j=1}^n 1_{\{d_i^+ \geq 1\}} \frac{d_j^-}{d_i^+} A_{ij}.$$

There is no obvious relationship with $\mathrm{E}_0(D^-)$. The same conclusion holds for a random predecessor. The friendship paradox does not apply to directed graphs.