Graph mining
SD212
Sampling nodes and edges

Thomas Bonald

2017 – 2018

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

1 Undirected graphs

We first consider an undirected graph $G = (V, E)$ of n nodes and m edges. We denote by d_u the degree of node $u \in V$. We assume that $d_u \geq 1$ for all $u \in V$. We have:

$$\sum_{u \in V} d_u = 2m.$$

Random node. The empirical degree distribution is given by

$$\forall k \geq 0, \quad p_k = \frac{1}{n} \sum_{u \in V} 1_{\{d_u = k\}}.$$

This is the degree distribution of a node chosen uniformly at random in V. Let X be a random variable having this distribution. The average degree is

$$\mathrm{E}(X) = \sum_{k \geq 0} k p_k = \frac{1}{n} \sum_{u \in V} d_u = \frac{2m}{n}.$$

Random edge. Now choose an edge uniformly at random and one of the two ends of this edge uniformly at random. Denote by $\hat{X}$ the degree of this node.

Proposition 1 *The distribution of the random variable $\hat{X}$ is the size-biased distribution:*

$$\forall k \geq 0, \quad \mathrm{P}(\hat{X} = k) = \frac{k p_k}{\mathrm{E}(X)}.$$

Proof. By definition,

$$\forall k \geq 0, \quad \mathrm{P}(\hat{X} = k) = \frac{1}{2m} \sum_{u,v \in V} 1_{\{u,v\} \in E} \left(\frac{1}{2} 1_{\{d_u = k\}} + \frac{1}{2} 1_{\{d_v = k\}} \right),$$
$$= \frac{1}{2m} \sum_{u,v \in V} 1_{\{u,v\} \in E} 1_{\{d_u = k\}},$$
$$= \frac{1}{2m} \sum_{u \in V} k 1_{\{d_u = k\}},$$
$$= \frac{n}{2m} k p_k = \frac{k p_k}{\mathrm{E}(X)}.$$

□

In particular, we have

$$\mathrm{E}(\hat{X}) = \frac{\mathrm{E}(X^2)}{\mathrm{E}(X)} \geq \mathrm{E}(X),$$

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

Random neighbor. Now choose a node uniformly at random and one of its neighbors uniformly at random. Denote by Y the degree of this node.

Proposition 2 *We have* $\mathrm{E}(Y) \geq \mathrm{E}(X)$ *with equality if and only if each connected component of the graph is regular.*

Proof. By definition,

$$\forall k \geq 0, \quad \mathrm{P}(Y = k) = \frac{1}{n} \sum_{u,v \in V} 1_{\{u,v\} \in E} \frac{1_{\{d_v = k\}}}{d_u}.$$

In particular,

$$\mathrm{E}(Y) = \sum_{k \geq 0} k \mathrm{P}(Y = k) = \frac{1}{n} \sum_{u,v \in V} 1_{\{u,v\} \in E} \frac{d_v}{d_u} = \frac{1}{n} \sum_{\{u,v\} \in E} \left(\frac{d_v}{d_u} + \frac{d_u}{d_v} \right).$$

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

$$\mathrm{E}(Y) \geq \frac{2m}{n} = \mathrm{E}(X)$$

with equality if and only if $d_u = d_v$ for all edges $\{u,v\}$, i.e., each connected component of the graph is regular. □

Observe that the random edge considered above is *not* chosen uniformly at random. Specifically, the probability of choosing edge $\{u,v\}$ is

$$\frac{1}{n} \left(\frac{1}{d_u} + \frac{1}{d_v} \right).$$

Thus edges whose ends have lower degrees are chosen more frequently. To get a uniform distribution over the edges, the first node needs to be drawn from the size-biased distribution (that is, with a probability proportional to its degree). The probability of choosing edge $\{u,v\}$ then becomes

$$\frac{1}{2m} \left(\frac{d_u}{d_u} + \frac{d_v}{d_v} \right) = \frac{1}{m}.$$

Moreover, both ends of this edge have the same (size-biased) degree distribution. Denoting by $\hat{Y}$ the degree of the second node, we have

$$P(\hat{Y}=k) = \frac{1}{2m}\sum_{u,v\in V} 1_{\{u,v\}\in E}\frac{d_u}{d_u}1_{\{d_v=k\}} = \frac{k}{2m}\sum_{v\in V} 1_{\{d_v=k\}} = \frac{nkp_k}{2m} = P(\hat{X}=k).$$

Degree correlation. Let $\hat{X}$ and $\hat{Y}$ be the degrees of the ends of a random edge. The covariance of these random variables (or its normalized version, the coefficient of correlation) quantifies the *assortativity* of the graph, that is the tendency of nodes to be connected to nodes of similar degrees. If $\mathrm{cov}(\hat{X},\hat{Y})>0$ then high-degree nodes tend to be connected to high-degree nodes; otherwise, high-degree nodes tend to be connected to low-degree nodes, like in a star graph. A similar metric consists in considering the degree of of a random node, X, and that of one of its neighbors, Y.

2 Directed graphs

We now consider a directed graph $G=(V,E)$ of n nodes and m edges. We denote by d_u^+ and d_u^- the out-degree and the in-degree of node $u\in V$. We have:

$$\sum_{u\in V} d_u^+ = \sum_{u\in V} d_u^- = m.$$

Random node. The empirical out-degree and in-degree distributions are given by

$$\forall k\geq 0,\quad p_k^+ = \frac{1}{n}\sum_{u\in V} 1_{\{d_u^+=k\}},\quad p_k^- = \frac{1}{n}\sum_{u\in V} 1_{\{d_u^-=k\}}.$$

These are the out-degree and in-degree distributions of a node chosen uniformly at random in V. Let X^+ and X^- be random variables having these distributions. The average out-degree is

$$E(X^+) = \sum_{k\geq 0} kp_k^+ = \frac{1}{n}\sum_{i\in V} d_u^+ = \frac{m}{n}.$$

It is equal to the average in-degree.

Random edge. Choose an edge uniformly at random. Denote by $\hat{X}^+$ the out-degree of the origin of this edge and by $\hat{X}^-$ the in-degree of the end of this edge.

Proposition 3 *The distributions of the random variables $\hat{X}^+$ and $\hat{X}^-$ are the size-biased distributions:*

$$\forall k\geq 0,\quad P(\hat{X}^+=k) = \frac{kp_k^+}{E(X^+)},\quad P(\hat{X}^-=k) = \frac{kp_k^-}{E(X^-)}.$$

Proof. By definition,

$$\begin{aligned}\forall k\geq 0,\quad P(\hat{X}^+=k) &= \frac{1}{m}\sum_{(u,v)\in E} 1_{\{d_u^+=k\}},\\ &= \frac{1}{m}\sum_{u\in V} k1_{\{d_u^+=k\}},\\ &= \frac{n}{m}kp_k^+ = \frac{kp_k^+}{E(X^+)}.\end{aligned}$$

The proof for $\hat{X}^-$ is similar. □

We have $E(\hat{X}^+)\geq E(X^+)$ and $E(\hat{X}^-)\geq E(X^-)$ with equality if and only if the graph is regular (all nodes have the same in-degree and the same out-degree).

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

$$\forall k \geq 0, \quad \mathrm{P}(Y^- = k) = \frac{1}{n^+} \sum_{(u,v)\in E} 1_{\{d_u^+ \geq 1\}} \frac{1_{\{d_v^- = k\}}}{d_u^+},$$

where n^+ is the number of nodes of positive out-degrees. Thus

$$\mathrm{E}(Y^-) = \sum_{k\geq 0} k\mathrm{P}(Y^- = k) = \frac{1}{n^+} \sum_{(u,v)\in E} 1_{\{d_u^+ \geq 1\}} \frac{d_v^-}{d_u^+}.$$

There is no obvious relationship with $\mathrm{E}(X^-)$.

Observe that the probability of choosing edge (u, v) is

$$\frac{1}{n^+ d_u^+}.$$

Thus edges with lower out-degree origins are chosen more frequently. To get a uniform distribution over the edges, the origin needs to be drawn from the size-biased distribution (that is, with a probability proportional to its out-degree). The probability of choosing edge (u, v) then becomes

$$\frac{1}{m} \frac{d_u^+}{d_u^+} = \frac{1}{m}.$$

Moreover, both ends of this edge have the size-biased distribution. Denoting by $\hat{Y}^-$ the in-degree of the end of the edge, we have

$$\mathrm{P}(\hat{Y}^- = k) = \frac{1}{m} \sum_{(u,v)\in E} \frac{d_u^+}{d_u^+} 1_{\{d_v^- = k\}} = \frac{k}{m} \sum_{v\in V} 1_{\{d_v^- = k\}} = \frac{nkp_k^-}{m} = \mathrm{P}(\hat{X}^- = k).$$

The results are similar when we choose the end of the edge first.

Degree correlation. Let $\hat{X}^+$ and $\hat{X}^-$ be the out-degree of the origin and the in-degree of the end of a random edge. Again, the covariance of these random variables quantifies the *assortativity* of the graph. If $\mathrm{cov}(\hat{X}^+, \hat{X}^-) > 0$ then high out-degree nodes tend to be connected to high in-degree nodes. Similar metrics consist in considering the out-degree of of a random node, X^+ and the in-degree of one of its successors, Y^-, or the in-degree of of a random node, X^- and the out-degree of one of its predecessors, Y^+.