CS-E5740 Complex Networks, Answers to exercise set 8

Tommaso Brumani, Student number: 100481325

November 15, 2022

1. Network Sampling

a) The three sampling schemes (Bernoulli node, edge, and star sampling) were implemented.
Empirical estimates of the number of triangles, two-stars, and transitivity were reported for each of the sampling schemes and collected in the table in Figure 1.

samp.	triangles	two-stars	transit.	triang.frac.	two-st.frac.
node	297	388	0.7655	0.0111	0.0106
edge	306	1884	0.1624	0.0114	0.0514
star	3922	8698	0.4510	0.1464	0.2371
orig.	26792	36686	0.7303	1.0000	1.0000

Figure 1: Empirical estimates of the properties.

The node and edge sampling schemes seem to preserve only about 1% of the triangles in the original network, with the star sampling method performing significantly better by preserving about 15%.
In terms of two-stars it also appears to present a much-improved performance, maintaining 23% of the structures compared to the 5% and 1% of edge and node sampling, respectively.
Notably, node sampling appears to preserve the same percentage of triangles and two-stars with respect to the original, meaning that transitivity, defined as the ratio between the two, would remain unchanged even after sampling: the other two methods would instead alter the relationship between the two.

b) The derivations and explanations of the probability of sampling two-stars/triangles for each sampling scheme are reported in the scanned sheets attached at the end of this file.

c) The HT estimator for the count of structures in sampled networks is defined as:

$$\hat{\tau}^{HT} = \frac{1}{p_\tau}\hat{\tau}$$

Where $\hat{\tau}$ is the empirical count of a structure found in a sampled network, and p_τ is the probability of observing these structures.
The HT estimator corrects for sampling bias because it normalizes the count of the structure by the probability of observing it during the sampling, thus taking into account the implicit distortion caused by the sampling method.
This can be observed by interpreting τ as:

$$\tau = p_\tau \cdot \tau_{original}$$

So that the sampling probability is simplified through the HT estimator.
The calculated HT estimators for each sampling scheme and structure are reported in the scanned sheets attached at the end of this file.

d) The HT estimator for the three sampling schemes was implemented, and the network sampled 150 times for each sampling scheme and for two probabilities $p = 0.35$ and $p = 0.5$.
The number of triangles and transitivity was calculated for each, and plotted against the empirical estimator for $p = 0.5$ and a red line representing the value of the original network.
The results are reported in Figures 2, 3, and 4.

The sampling probability p causes different effects based on the sampling method as explained in the previous sections, as it factors in differently on the number of triangles/two-stars counted and can thus shift the estimation by a significant amount.
The HT distributions succeed in aligning the mean value of the estimation to that of the original network for all sampling schemes and appear to all present largely the same shape and values, although the distance from the empirical estimation varies based on the latter's bias with respect to the original network depending on the sampling method and the quantity being plotted.
In the last plot, the empirical distribution is centered around the original value because for $p = 0.5$ the bias in observing two-stars and triangles when star sampling is the same, and thus, when calculating the transitiviy, the two biases cancel each other out.

1 2. Modularity

a) The modularity for the two partitions was calculated and reported in the scanned sheets attached at the end of this file.

b) The derivation of ΔQ as a function of $L, d_a, d_b, d_{ab}, l_a, l_b$, and l_{ab} was reported in the scanned sheets attached at the end of this file.

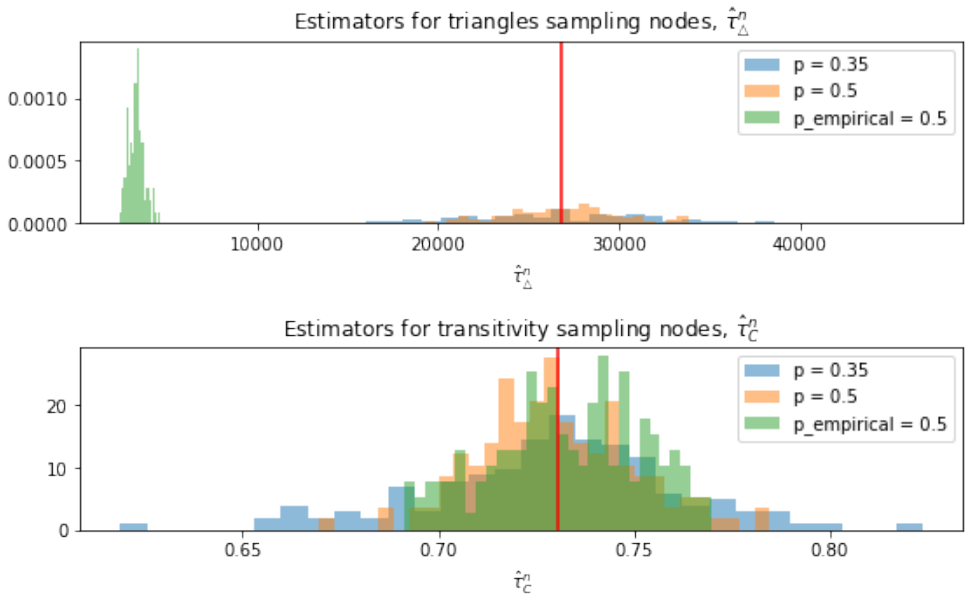


Figure 2: Triangles and transitivity for node sampling.

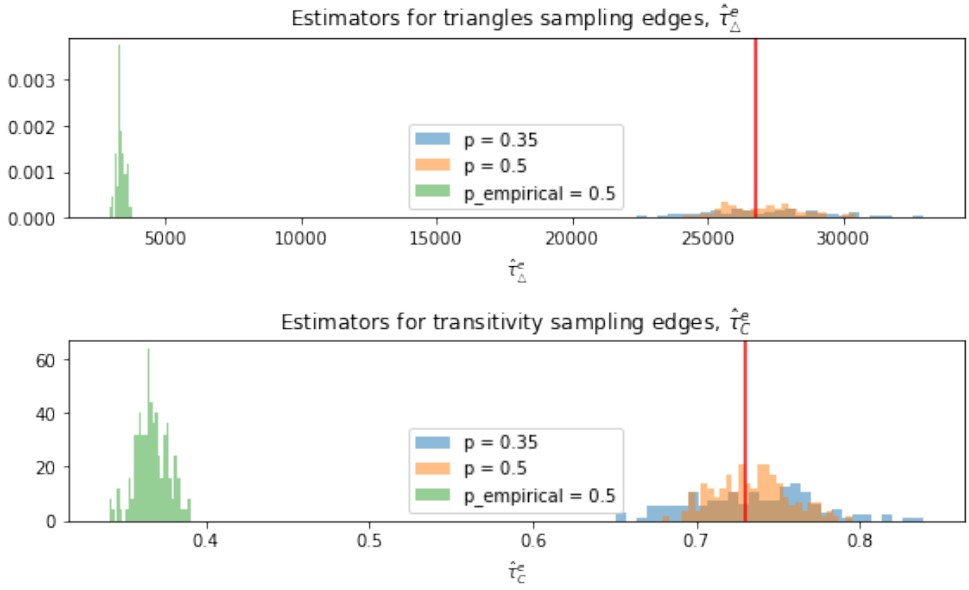


Figure 3: Triangles and transitivity for edge sampling.

c) The required calculations are reported in the scanned sheets attached at the end of this file.
The result indicates that modularity optimization has a resolution limit because, if

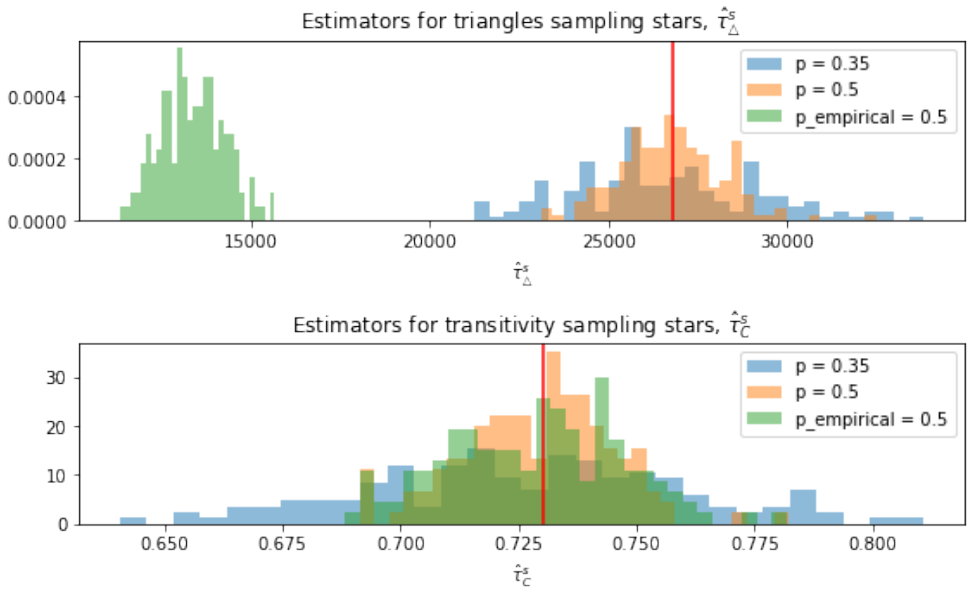


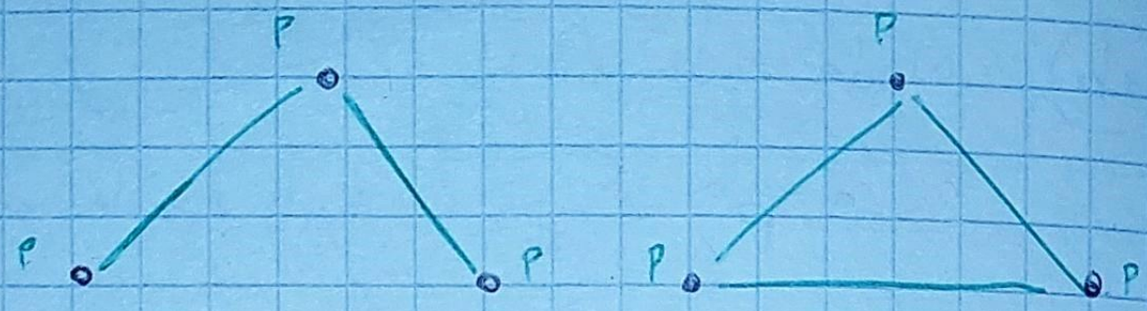
Figure 4: Triangles and transitivity for star sampling.

L increases while d_a and d_b remain approximately constant, it is easy to observe that for every cluster a and b there will be a value of L sufficiently large to render $L > \frac{d_a d_b}{2}$, and thus to cause the merging of the two clusters.

d) The requested arguments are presented in the scanned sheets attached at the end of this file.

Complex Networks Ex 8:

(1b)



- Node sampling:

  since in order to select an edge both its ends must be observed, then in both cases all three nodes must be observed to sample the two-star and triangle. Thus:

$$P_{\angle}^{n} = P_{\Delta}^{n} = p^3$$

- Star sampling:

  in order to select a two-star, only the node at its core must be selected (this will include the edges to its neighbors), thus:

$$P_{\angle}^{s} = p$$

to observe a triangle, however, at least two of the three nodes must be observed (edges between neighbors are not observed when observing the two-star's core node). The probability of choosing two out of the three nodes is $Bi(2,3,p)$, whereas that of choosing all three is p^3. Thus:

$$p_\Delta^S = Bi(2,3,p) + p^3$$

$$= \binom{3}{2} p^2(1-p) + p^3 = p^2\left(\binom{3}{2}(1-p)+p\right)$$

1c) • Node sampling:

$$\hat{\tau}_\wedge^{HT} = \frac{\hat{\tau}_\wedge^n}{p^3} \qquad \hat{\tau}_\Delta^{HT} = \frac{\hat{\tau}_\Delta^n}{p^3} \qquad \hat{\tau}_{tr}^{HT} = \frac{\hat{\tau}_\Delta^n}{\cancel{p^3}} \frac{\cancel{p^3}}{\hat{\tau}_\wedge^n} = \frac{\hat{\tau}_\Delta^n}{\hat{\tau}_\wedge^n}$$

• Edge sampling:

$$\hat{\tau}_\wedge^{HT} = \frac{\hat{\tau}_\wedge^e}{p^2} \qquad \hat{\tau}_\Delta^{HT} = \frac{\hat{\tau}_\Delta^e}{p^3} \qquad \hat{\tau}_{tr}^{HT} = \frac{\hat{\tau}_\Delta^e}{p^3}\frac{p^2}{\hat{\tau}_\wedge^e} = \frac{\hat{\tau}_\Delta^e}{p\,\hat{\tau}_\wedge^e}$$

• Star sampling:

$$\hat{\tau}_\wedge^{HT} = \frac{\hat{\tau}_\wedge^S}{p} \qquad \hat{\tau}_\Delta^{HT} = \frac{\hat{\tau}_\Delta^S}{Bi(2,3,p)+p^3} \qquad \hat{\tau}_{tr}^{HT} = \frac{\hat{\tau}_\Delta^S}{p\hat{\tau}_\wedge^S\left(\binom{3}{2}(1-p)+p\right)}$$

2a) Modularity:

$$Q = \sum_{c \in P} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right]$$

- Partition 1:

$$Q = 2 \cdot \left[\frac{6}{13} - \left(\frac{3 \cdot 3 + 1 \cdot 4}{2 \cdot 13} \right)^2 \right]$$

$$= 0.423$$

- Partition 2:

$$Q = \left[\frac{13}{13} - \frac{6 \cdot 3 + 2 \cdot 4}{2 \cdot 13} \right] = 0$$

2b) Difference in modularity:

$$\Delta Q = Q_2 - Q_1$$

1) Two real modules are separate:

$$Q_1 = \left[\frac{l_a}{L} - \left(\frac{d_a}{2L} \right)^2 \right] + \left[\frac{l_b}{L} - \left(\frac{d_b}{2L} \right)^2 \right] + Q_R$$

with: $Q_R = \sum_{c \in P'} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right]$

2) Two real modules are merged:

$$Q_2 = \left[\frac{l_{ab}}{L} - \left(\frac{d_{ab}}{2L} \right)^2 \right] + Q_R$$

Then:

$$\Delta Q = \left[\frac{l_{ab}}{L} - \left(\frac{d_{ab}}{2L} \right)^2 \right] + Q_R - \left[\frac{l_a}{L} - \left(\frac{d_a}{2L} \right)^2 \right]$$

$$- \left[\frac{l_b}{L} - \left(\frac{d_b}{2L} \right)^2 \right] - Q_R$$

$$= \frac{1}{L} \left(l_{ab} - l_a - l_b \right) - \frac{1}{4L^2} \left(d_{ab}^2 - d_a^2 - d_b^2 \right)$$

2c) $l_{ab} = l_a + l_b + 1$

$d_{ab} = d_a + d_b$ — link connecting the 2 clusters

Then:

$$\Delta Q = \frac{1}{L}(1) - \frac{1}{4L^2}(2 d_a d_b) = \frac{1}{L} - \frac{2 d_a d_b}{4L^2}$$

When $\Delta Q > 0$ we merge the clusters, so:

$$\Delta Q = \frac{1}{L}\left(1 - \frac{d_a d_b}{2L}\right) > 0 \quad \text{with } L > 0$$

$$\rightsquigarrow \frac{d_a d_b}{2L} < 1 \rightsquigarrow L > \frac{d_a d_b}{2}$$

2d) I) A clique is a complete subgraph, and thus if a has n nodes, each of them will have degree $n-1$. Thus:

$$d_a = n \cdot (n-1) \approx n^2$$

II) Assuming another community b, a and b are merged if:

$$L > \frac{d_a d_b}{2} = \frac{n^2 d_b}{2}$$

meaning when:

$$n^2 < \frac{2L}{d_b} \quad \text{where } n, \frac{2L}{d_b} > 0$$

thus:

$$h < \sqrt{\frac{2L}{d_b}} = \underbrace{\sqrt{\frac{2}{d_b}}}_{c} \cdot \sqrt{L} = c \cdot \sqrt{L}$$