CS-E5740 Complex Networks, Answers to exercise set 8

Marco Di Francesco, Student number: 100632815

November 15, 2022

Problem 1

a) Look at table 1.

samp.	triangles	two-stars	transit.	triang.frac.	two-st.frac.
node	345	472	0.7309	0.0135	0.0128
edge	258	1802	0.1432	0.0101	0.0491
star	3174	7653	0.4147	0.1244	0.2083
orig.	25524	36732	0.6949	1.0000	1.0000

Figure 1: Table exercise 1 section a

How do sampling schemes compare in the fraction of triangles/two-stars they preseve? Do sampling schemes affect two-stars/triangles in the same way?

Sampling shemes given that they work in different ways have very different results for the number of two-starts and triangles. Explanation for each sampling sheme:

- Node sampling: we have very little number of two-stars and triangles (less than 2%) because in general we have very little edges. This is given because we have an edge only in the case we select both nodes it connectes, thus the probability an edge exists is very small.
- Edge sampling: we have an higher number of edges compared to node sampling, thus we have a higher number of two-stars fraction, but because the selecting the edges is random (not like in star sampling) the number of triangles is still very small.
- Star sampling: we have way more edges compared to the other sampling techniqes, and they are not selected randomly but each node selectes also its neighbors, thus we have not only a high number of two-stars included, but also a high number of triangles.

Transitivity via node sampling should be similar to the real value, what could be the reason for this?

As explained above, the edges are sampled randomly in all cases, and it's way more likely to get two-stars than triangles. This is because getting 3 nodes that constitute a triangle is way more unlikely than getting 2 nodes than consitute a two-stars.

b)
- Node sampling two-stars: p^3. This is because to get an edge all 3 nodes must be selected.
- Node sampling triangles: p^3. As for two-stars, because to get an edge all 3 nodes must be selected.
- Star sampling two-star: p. This is beause for each node we get we also have the neighbors. (Assumption: we start from 55 cliques of 12 nodes and rewiring only 10% of the nodes we are not considering the degree of each node)
- Star sampling triangles: $\binom{3}{2}p^2(1-p)+p^3$. This is because we need a binomial to select 2 out of 3 neighbor nodes to create a triangle, or we are just choosing all 3 nodes (p^3).

c) *Why the HT estimator corrects for sampling bias?*

The HT estimator corrects for sampling bias because it normalizes the count of two-stars or three stars by the probability of observing them after sampling. This probability is distoreted and will not always be p.

HT estimators for the number of two-stars, triangles and transitivity for the three sampling schemes we explored

Two-stars

- Nodes $\hat{\tau}_{\angle}^{HT} = \hat{\tau}_{\angle}^{n}/p^3$
- Edges $\hat{\tau}_{\angle}^{HT} = \hat{\tau}_{\angle}^{e}/p^2$
- Stars $\hat{\tau}_{\angle}^{HT} = \hat{\tau}_{\angle}^{s}/p$

Triangles

- Nodes $\hat{\tau}_{\triangle}^{HT} = \hat{\tau}_{\triangle}^{n}/p^3$
- Edges $\hat{\tau}_{\triangle}^{HT} = \hat{\tau}_{\triangle}^{e}/p^3$
- Stars $\hat{\tau}_{\triangle}^{HT} = \hat{\tau}_{\triangle}^{s}/(\hat{\tau}_{\angle}^{s}\binom{3}{2}p^2(1-p)+p^3)$

Transitivity

- Nodes $\hat{\tau}_{C}^{HT} = \hat{\tau}_{\triangle}^{n}/\hat{\tau}_{\angle}^{n}$
- Edges $\hat{\tau}_{C}^{HT} = \hat{\tau}_{\triangle}^{e}/(p\hat{\tau}_{\angle}^{e})$
- Stars $\hat{\tau}_{C}^{HT} = \hat{\tau}_{\triangle}^{s}/(p\hat{\tau}_{\angle}^{s}\binom{3}{2}(1-p)+p)$

d) Looks plots in figure 2.

What is the effect of the sampling probability p on your estimators?

Higher probability raises variance in the prediction. We can see that the distributions with $p = 0.35$ have higher variance than with $p = 0.5$.

How the HT distribuitions differ between sampling schemes? In all 3 cases the HT distributions were able to allighn the to the mean value the estimation, the difference lays on the variance these distribution have. Having lower probability of having more two-stars and triangles (lower bias) led to having less variance.

In the last plot, the empirical estimators (without HT correction) should be centered around the true value, why could this be?

This happens because there is no bias for $p = 0.5$, thus the distributions are centered around the true value.

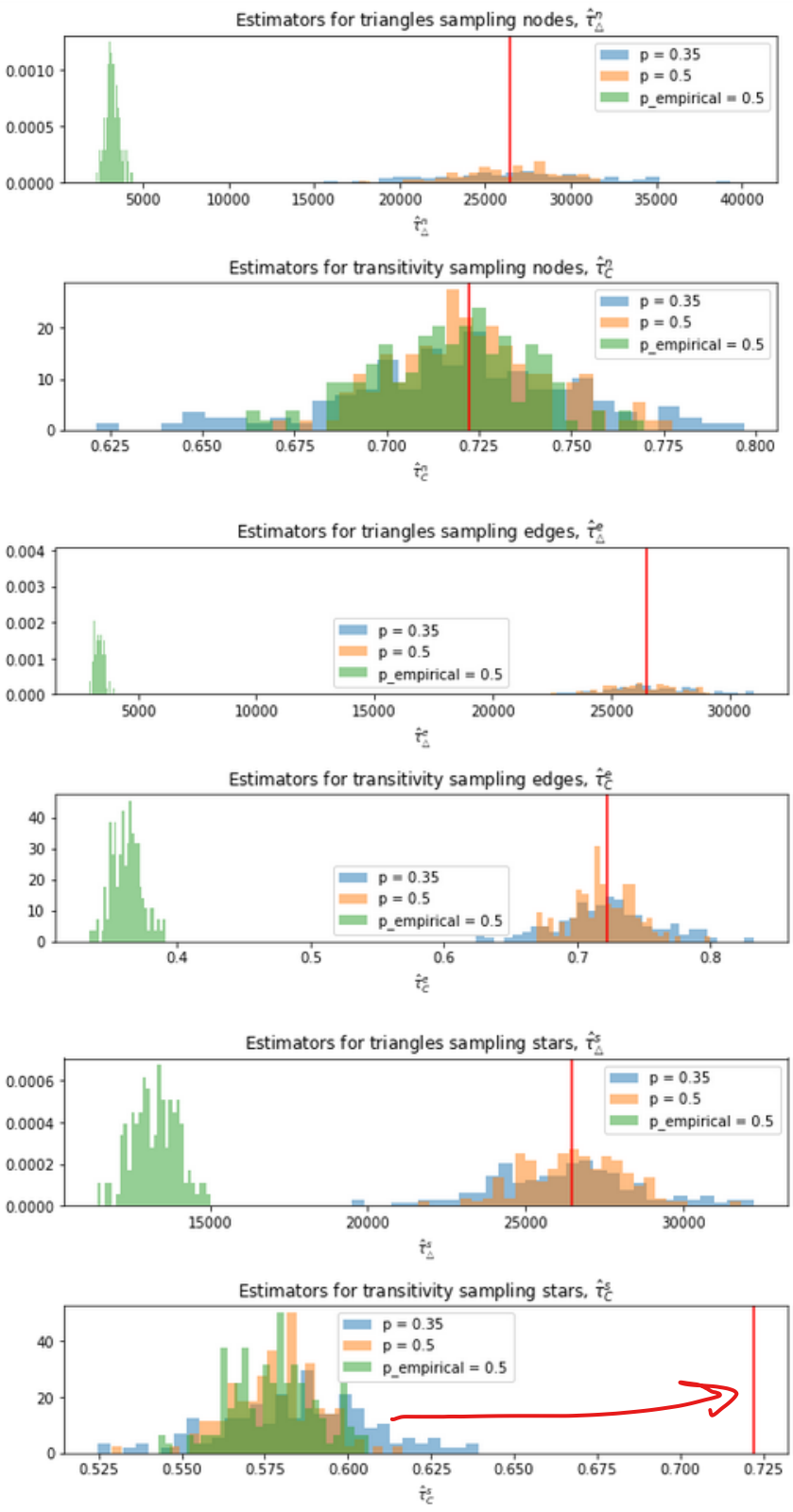

Figure 2: Exercise 1 part d

Problem 2

a) Look at image 3.

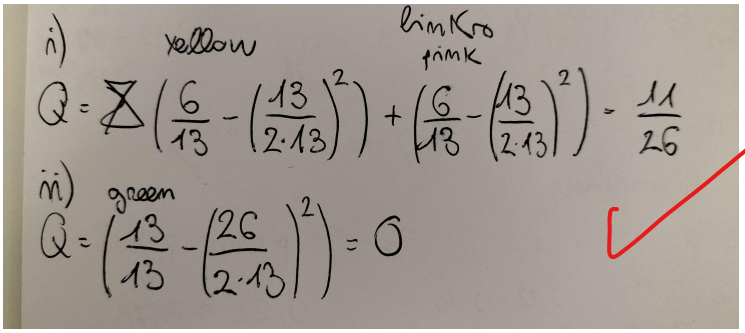

Figure 3: Exercise 2 part a

b) Look at images 4 and 5.

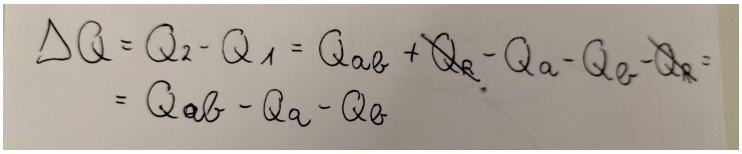

Figure 4: Exercise 2 part b image 1

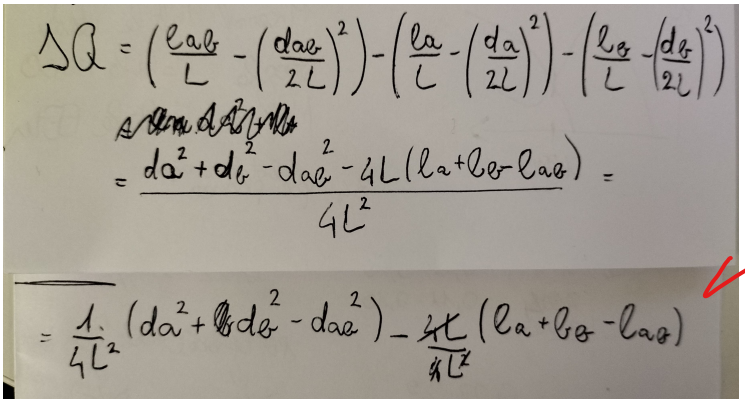

Figure 5: Exercise 2 part b image 2

c) Look at image 6.

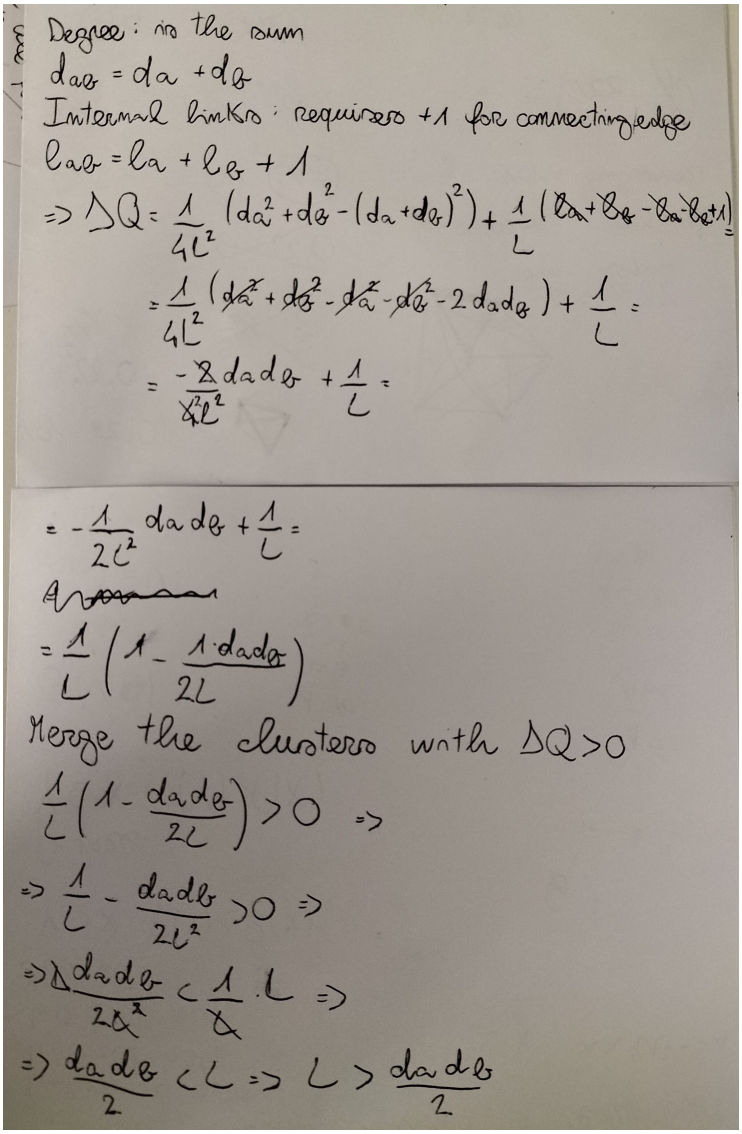

Figure 6: Exercise 2 part c

Explain why this result indicates that modularity optimization must have a resolution limit, or a minimum size of community that can be found that depends on the size of the network

Increasing the number of links L increases we need to have a resolution limit because otherwise the number of edges required to connect 2 communities drops below 1, thus connecting clusters of nodes that are in reality weakly connected.

d) Look at image 7

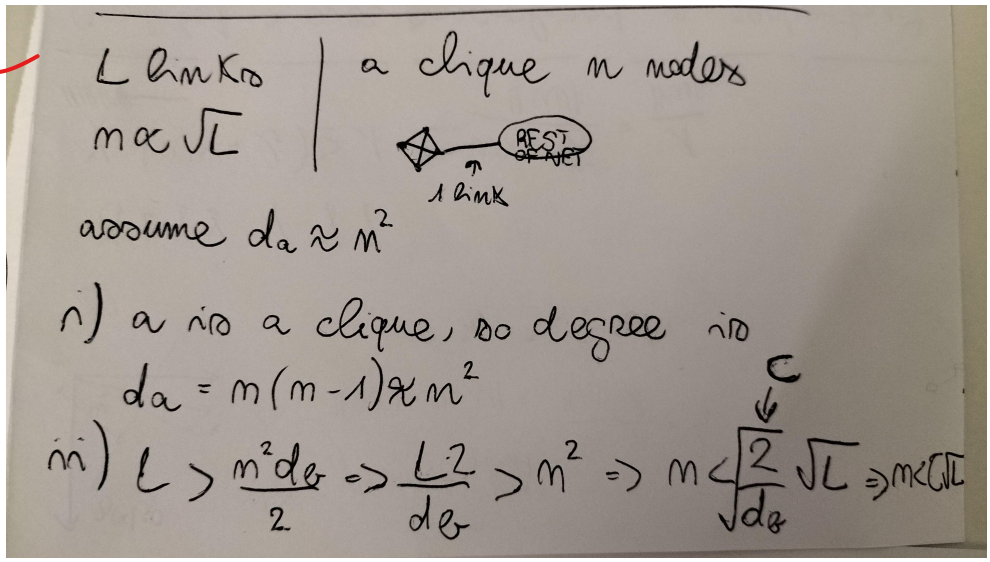

Figure 7: Exercise 2 part d