# CS-E5740 Complex Networks, Answers to exercise set 8

Sara Cabodi, Student number: 784287

November 18, 2019

Compile with `pdflatex ex_template.tex`

## Problem 1

a) On the case of Bernoulli sampling of nodes, where each node is sampled with probability $p$, I calculate the following probabilities:

(i)
$$\Pr(\textit{observing a two star}) = p \cdot p \cdot p = p^3$$

(ii)
$$\Pr(\textit{observing a triangle}) = p \cdot p \cdot p = p^3$$

In both cases I apply the conjunction rule of probability of three independent events

$$\Pr(A \cap B \cap C) = \Pr(A) \cdot \Pr(B|A) \cdot \Pr(C|A \cap B) =$$

$$= \Pr(A) \cdot \Pr(B) \cdot \Pr(C) = p \cdot p \cdot p = p^3$$

where

- A is the event of sampling node i
- $\Pr(B|A)$ is the probability to sample node j, given that I already sampled i
- $Pr(C|A \cap B)$ is the probability to sample node z, given that I already sampled i and j

Now, I show two HT estimators for the sampling of nodes:

- HT estimator for the total number of two-stars

$$\hat{\tau}^n_{\angle} = \sum_{i \in S} \frac{y_i}{\pi_i} = |S| \cdot \frac{1}{p^3} \cdot \frac{|S_{\angle}|}{|S|} = \frac{|S_{\angle}|}{p^3}$$

  where $y_i = \frac{|S_{\angle}|}{|S|}$ is the number of two-star of node $i$, $\pi_i$ is the probability computed above $(i)$, $S$ is the set of sampled nodes and $S_{\angle}$ is the set of sampled two-stars.

- HT estimator for the total number of triangles

$$\hat{\tau}^n_{\triangle} = \sum_{i \in S} \frac{y_i}{\pi_i} = |S| \cdot \frac{1}{p^3} \cdot \frac{|S_{\triangle}|}{|S|} = \frac{|S_{\triangle}|}{p^3}$$

  where $y_i = \frac{|S_{\triangle}|}{|S|}$ is the number of triangles centered in the node $i$, $\pi_i$ is the probability computed above $(i)$, $S$ is the set of sampled nodes and $S_{\triangle}$ is the set of sampled triangles.

b) On the case of Bernoulli sampling of edges, where each edge is sampled with probability $p$, I calculate the following probabilities:

(i)
$$\Pr(\textit{observing a two star}) = p \cdot p = p^2$$

(ii)
$$\Pr(\textit{observing a triangle}) = p \cdot p \cdot p = p^3$$

As shown in point a), I applied the conjunction rule. In the first case, it's enough to sample two edges in order to form a two star, while, in the case of a triangle, I need to sample three.

Now, I show two HT estimators for the sampling of edges:

- HT estimator for the total number of two-stars

$$\hat{\tau}^e_{\angle} = \sum_{i \in S} \frac{y_i}{\pi_i} = |S| \cdot \frac{1}{p^2} \cdot \frac{|S_{\angle}|}{|S|} = \frac{|S_{\angle}|}{p^2}$$

  where $y_i = \frac{|S_{\angle}|}{|S|}$ is the average number of two-star of the edge $i$, $\pi_i$ is the probability computed above $(i)$, $S$ is the set of sampled edges and $S_{\angle}$ is the set of sampled two-stars.

– HT estimator for the total number of triangles

$$\hat{\tau}^e_{\triangle} = \frac{y_i}{\pi_i} = |S| \cdot \frac{1}{p^3} \cdot \frac{|S_{\triangle}|}{|S|} = \frac{|S_{\triangle}|}{p^3}$$

where $y_i = \frac{|S_{\triangle}|}{|S|}$ is the average number of triangles for the edge $i$, $\pi_i$ is the probability computed above $(i)$, $S$ is the set of sampled edges and $S_{\triangle}$ is the set of sampled triangles.

c) On the case of star sampling, where each node is sampled with probability $p$ and then we observe all of its edges, I calculate the following probabilities:

(i)

$$\Pr(\textit{observing a two star}) = p$$

(ii)

$$\Pr(\textit{observing a triangle}) = 2 \cdot p \cdot p = p^2$$

In the first case it is enough to compute the probability of sampling the central node of the two and the two-star will be automatically be there by the way the sampling is done. The second case, computes the probability that the first one is picked and that at least one of the other two is too. The case of the three nodes picked is included in the latter.

Now, I show two HT estimators for the star sampling:

– HT estimator for the total number of two-stars

$$\hat{\tau}^s_{\angle} = \sum_{i \in S} \frac{y_i}{\pi_i} = |S| \cdot \frac{1}{p} \cdot \frac{|S_{\angle}|}{|S|} = \frac{|S_{\angle}|}{p}$$

where $y_i = \frac{|S_{\angle}|}{|S|}$ is the number of two-star of node $i$, $\pi_i$ is the probability computed above $(i)$, $S$ is the set of sampled nodes and $S_{\angle}$ is the set of sampled two-stars.

– HT estimator for the total number of triangles

$$\hat{\tau}^s_{\triangle} = \frac{y_i}{\pi_i} = |S| \cdot \frac{1}{2 \cdot p^2} \cdot \frac{|S_{\triangle}|}{|S|} = \frac{|S_{\triangle}|}{2 \cdot p^2}$$

where $y_i = \frac{|S_{\triangle}|}{|S|}$ is the number of triangles centered in the node $i$, $\pi_i$ is the probability computed above $(i)$, $S$ is the set of sampled nodes and $S_{\triangle}$ is the set of sampled triangles.

d) Now, I construct our HT estimators for network transitivity $C$ under the sampling schemes analysed above:

– sampling of nodes:

$$\hat{\tau}_C^n = \frac{\hat{\tau}_\triangle^n}{\hat{\tau}_\angle^n} = \frac{|S_\triangle|}{|S_\angle|}$$

– sampling of edges:

$$\hat{\tau}_C^e = \frac{\hat{\tau}_\triangle^e}{\hat{\tau}_\angle^e} = \frac{p^2 \cdot |S_\triangle|}{p^3 \cdot |S_\angle|} = \frac{|S_\triangle|}{p \cdot |S_\angle|}$$

– star sampling:

$$\hat{\tau}_C^e = \frac{\hat{\tau}_\triangle^e}{\hat{\tau}_\angle^e} = \frac{p \cdot |S_\triangle|}{2 \cdot p^2 \cdot |S_\angle|} = \frac{|S_\triangle|}{2 \cdot p \cdot |S_\angle|}$$

Where $S$ is the set of random samples.

e) Here are my results on three tables, one for each sampling scheme, where the rows represent the different sampling probabilities, and the columns represent the empirical number of triangles, two-stars, and transitivity.

--------- Sampling nodes ---------

| p | triangles | two-stars | transitivity |
|---|---|---|---|
| 0.10 | 39 | 30 | 1.3000 |
| 0.30 | 1575 | 1918 | 0.8212 |
| 0.50 | 6456 | 9013 | 0.7163 |
| 1.00 | 52374 | 72940 | 0.7180 |

.....

--------- Sampling edges ----------

| p | triangles | two-stars | transitivity |
|---|---|---|---|
| 0.10 | 54 | 373 | 0.1448 |
| 0.30 | 1383 | 6017 | 0.2298 |
| 0.50 | 6195 | 17648 | 0.3510 |
| 1.00 | 52374 | 72940 | 0.7180 |

.....

--------- Star sampling ----------

| p | triangles | two-stars | transitivity |
|---|---|---|---|
| 0.10 | 1557 | 7620 | 0.2043 |
| 0.30 | 11754 | 22853 | 0.5143 |
| 0.50 | 26607 | 37027 | 0.7186 |
| 1.00 | 52374 | 72940 | 0.7180 |

Figure 1: Results for each sampling scheme

Overall, data confirm that

- the theoretical computations of probabilities for triangles and two-stars, under the three sampling schemes, are correct (data are clearly more reliable at higher probabilities, whereas data lower probabilities are more subject to variance around expected values
- the estimation of transitivity needs the HT formula, based on probabilities, because the empirical computation is more and more unreliable at decreasing sampling probabilities.

$p = 1.0$ this is the set of reference data, providing the number of triangles, two-stars and the resulting transitivity, in the real (non sampled) network

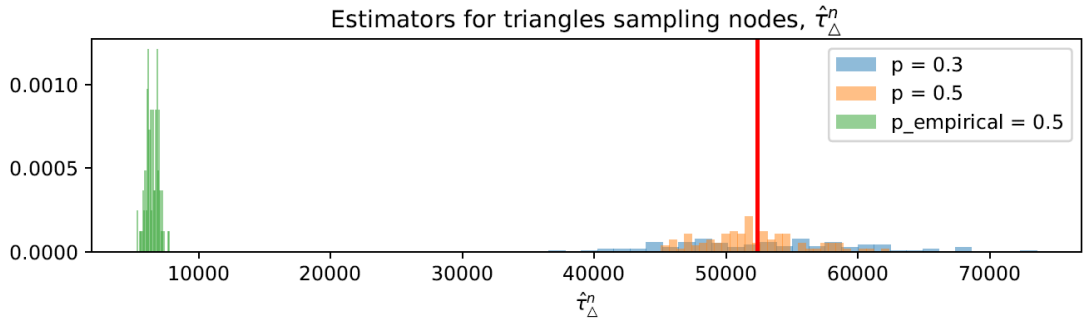
$p = 0.5$ data on triangles and two-stars are very close to the expected ones, given their theoretical probabilities. Let us recall that $p = \frac{1}{2}$, $p^2 = \frac{1}{4}$ and $p^3 = \frac{1}{8}$. The number of triangles are $\frac{1}{8}$ of the real number in node sampling and edge sampling, $\frac{1}{2}$ in star sampling. Two-star count confirms the $\frac{1}{8}$ ratio with node sampling, $\frac{1}{4}$ in edge sampling and $\frac{1}{2}$ in star sampling. Transitivity is rather correct in node and star sampling, due to the equal values of probabilities (of triangles and two-stars).

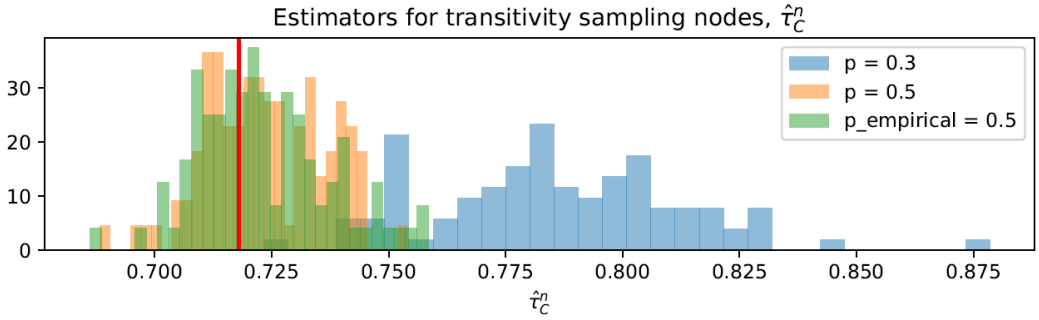
$p = 0.3$ The comment done for $p = 0.5$ can be replicated, with the only difference that the probabilities of triangles and two-stars only match in node sampling, resulting in a correct transitivity estimation.

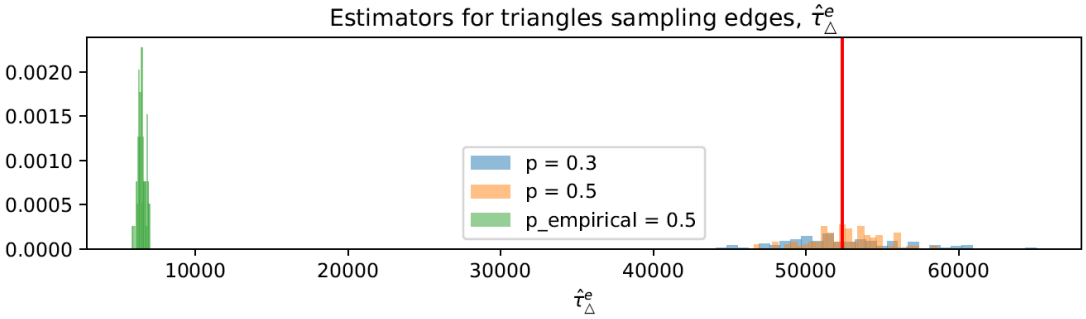
$p = 0.1$ probability seems to low to draw reliable conclusions in node and edge sampling, whereas they seem closer to the expected ones in star sampling.

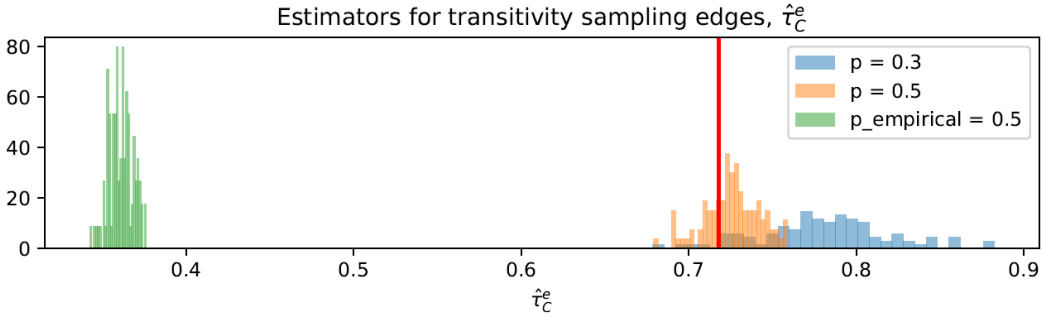
f) In this point, for each $p = 0.3$ and $p = 0.5$ (also the one with empirical estimator), and for each sampling scheme, I obtained $n = 100$ samples from the original network, I calculated the HT estimator for the number triangles and for transitivity and plotted the results in six plots. In these, the red bar represents the real values of number of triangles and two-stars obtained with the full network.

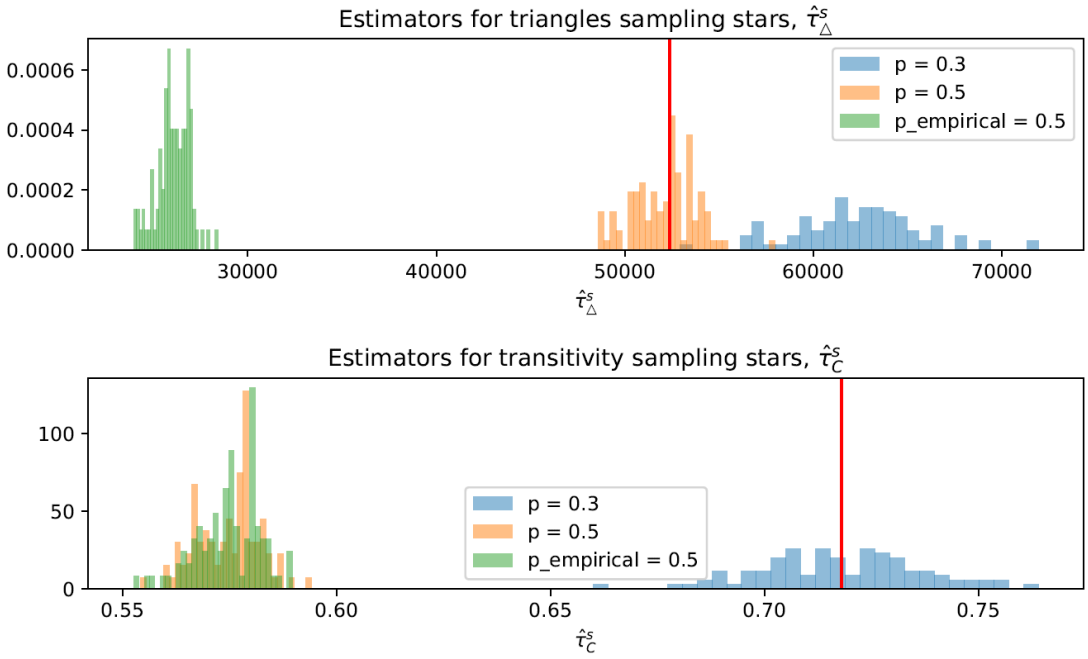
Do the HT estimators lie near the true value? Both in node and edge sampling HT estimators with the 0.5 probability are very close to the real values, differences increase with 0.3 probability and empirical value. In the star plot, transitivity seems a bit underestimated, probably due to the fact that the simplified formula used for two-star probability is a bit lower than the exact one.


Estimators for triangles sampling nodes, $\hat{\tau}^n_\triangle$
p = 0.3
p = 0.5
p_empirical = 0.5
0.0010
0.0005
0.0000
10000
20000
30000
40000
50000
60000
70000
$\hat{\tau}^n_\triangle$


Estimators for transitivity sampling nodes, $\hat{\tau}^n_C$
p = 0.3
p = 0.5
p_empirical = 0.5
30
20
10
0
0.700
0.725
0.750
0.775
0.800
0.825
0.850
0.875
$\hat{\tau}^n_C$


Estimators for triangles sampling edges, $\hat{\tau}^e_\triangle$
p = 0.3
p = 0.5
p_empirical = 0.5
0.0020
0.0015
0.0010
0.0005
0.0000
10000
20000
30000
40000
50000
60000
$\hat{\tau}^e_\triangle$


Estimators for transitivity sampling edges, $\hat{\tau}^e_C$
p = 0.3
p = 0.5
p_empirical = 0.5
80
60
40
20
0
0.4
0.5
0.6
0.7
0.8
0.9
$\hat{\tau}^e_C$




Here are some comments divided by sampling scheme:

nodes The estimators for triangles are distributed quite normally around the true value. On the other hand, the empirical estimation does not lie close to the first two and so neither to the true value. Concerning the estimators for transitivity, the ones with $p = 0.5$ (including the empirical) follow a similar trend and are almost centered on the true value. Quite the opposite, the one with $p = 0.3$ is almost normally distributed, but with a much higher variance and mean compared to the other two.

edges The considerations on the estimators for triangles mirror those made for the nodes sampling. However, for the ones for transitivity the situation is different. The distribution closer to the true value is the one with $p = 0.5$. The one with $p = 0.3$ is close, but centered on a higher mean. Concerning the empirical one, it is very far away from the other two with a mean around 0.2.

stars The estimators for triangles are centered on different means and have increasing variance (from left to right). The distribution with $p = 0.5$ is the one almost centered on the true value, while the empirical one is far away on the left and the one with $p = 0.3$ is slightly shifted on the right.
The estimators for transitivity follows quite the same distributions, but this time the one centered on the true value is the one with $p = 0.3$. The other two are quite similar and have a definitely lower mean and so shifted on the left of the true value.

g) 1. Let us call $N_{OP}$, the number of customers of the operator: $N_{OP} = 6,421,148$. Given the 20% market share, we can estimate the overall number of customers in the country $N_C = 5 * N_{OP} = 32,105,740$.
The total number of customers of competitor companies in the country is $N_C - N_{OP}$. The reason why the network considered has more than 75 million nodes is that it includes international mobile phone numbers (customers of other countries). The network could be considered as a star sampling from the global (national + international) mobile communication network, where customers in the country have been sampled with probability $\frac{1}{5}$ (leading to the given company's customers) and international mobile phones have probability 0. As star sampling also adds nodes adjacent the ones picked, the final network includes both customers from competitor companies in the country and international phone numbers. So:

$$N_1 = N_{OP} + \alpha(N_C - N_{OP}) + N_{INT} = 75,048,105$$

Where

* $N_1$ is the total number of nodes in the network
* $\alpha < 1$ represents the fraction of customers from competitor companies in the country
* $N_{INT}$ is the number of international phone numbers included

Let us now compute the requested estimations:

a) Empirical. We simply use the sampled network and evaluate transitivities:

$$C = \frac{\hat{\tau}_\triangle}{\hat{\tau}_\angle} = \frac{153,328,324}{15,856,481,566} = 9.7 * 10^{-3} \approx 0.01$$

b) using HT. We need to estimate the probability of a triangle/two-star to be included in the final network. A triangle in the network always includes at least two nodes of the company, but could include one international or from a competitor node. A two-star includes at least one node of the company. As we don't know the actual number of international nodes, let us use the probabilities computed in ex 1c:

$$\pi_\triangle = p^2 = \left(\frac{1}{5}\right)^2 = \frac{1}{25}$$

$$\pi_\angle = p = \frac{1}{5}$$

$$C = \frac{\frac{153,328,324}{\pi_\triangle}}{\frac{15,856,481,566}{\pi_\angle}} = \frac{5 * 153,328,324}{15,856,481,566} = 0,048$$

2. The network can be seen as a form of edge sampling from the network of real connections between customers of the company/operator we consider. The edge probability is $\frac{1}{10}$, so we can apply HT probabilities as done in ex. 1b.

   a) Empirical. We simply use the sampled network and evaluate transitivities:

$$C = \frac{\hat{\tau}_{\triangle}}{\hat{\tau}_{\angle}} = \frac{11,794,486}{634,259,263} = 1.85 * 10^{-2} \approx 0.019$$

   b) using HT. Let us use the probabilities computed in ex 1b:

$$\pi_{\triangle} = p^3 = \left(\frac{1}{5}\right)^3$$

$$\pi_{\angle} = p^2 = \left(\frac{1}{5}\right)^2$$

$$C = \frac{\frac{11,794,486}{\pi_{\triangle}}}{\frac{634,259,263}{\pi_{\angle}}} = \frac{5 * 11,794,486}{634,259,263} = 0,093$$

# Problem 2

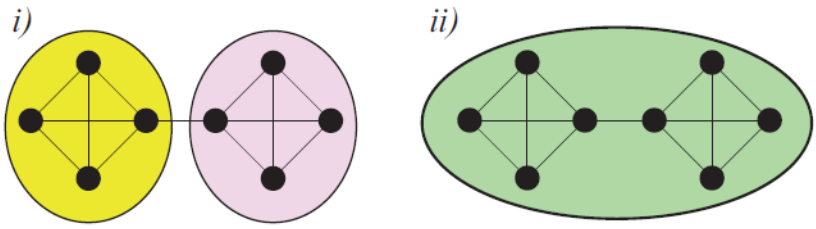



Figure 2: Two module configurations for a)

a) In this exercise I calculate the modularity for the two partitions shown above. The modularity $Q$ of a partition into communities can be written as

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

where:

- $c$ is the single cluster
- $P$ is the set of all the clusters in the network
- $L$ is the number of links in the whole network
- $l_c$ is the number of links internal to cluster $c$
- $d_c$ is the total degree of nodes in cluster $c$

Both networks have $L = 13$. The first one has 2 clusters, while the second one has just one. Let's calculate the modularities:

*i)*

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

$$= 2 * \left[ \frac{6}{13} - \left( \frac{13}{26} \right)^2 \right] =$$

$$= 2 * \left[ \frac{6}{13} - \left( \frac{1}{4} \right) \right] =$$

$$= 0.423$$

NOTE: I multiplied by 2 because the clusters have the same parameters.

*ii)*

$$Q = \sum_{c \in P} \left[ \frac{l_c}{L} - \left( \frac{d_c}{2L} \right)^2 \right] =$$
$$= \left[ \frac{13}{13} - \left( \frac{26}{26} \right)^2 \right] =$$
$$= [1 - 1] = 0$$

b) In order to write $\Delta Q = Q_2 - Q_1$ in terms of $l_i$ and $d_i$ I rewrite the single terms and the formula as follows:

$$Q_1 = Q_a + Q_b + Q_R$$
$$Q_2 = Q_{ab} + Q_R$$
$$\Delta Q = Q_2 - Q_1 = Q_{ab} - (Q_a + Q_b)$$
$$\Delta Q = \frac{l_{ab}}{L} - \left( \frac{d_{ab}}{2L} \right)^2 - \frac{l_a}{L} + \left( \frac{d_a}{2L} \right)^2 - \frac{l_b}{L} + \left( \frac{d_b}{2L} \right)^2 =$$
$$= \frac{l_{ab} - l_a - l_b}{L} - \frac{d_{ab}^2 - d_a^2 - d_b^2}{4L^2}$$

c) The results of the rewrites are the following

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

For $\Delta Q > 0$, then:

$$1 - \frac{d_a d_b}{2L} > 0$$
$$L > \frac{d_a d_b}{2}$$

We are working under the assumption that degrees ($d_a$, $d_b$, $d_{ab}$) are constant values, whereas $L$ increases. As the total degree can be considered as a rough estimation of the network size, this means that we are considering communities/clusters a, b (and

their possible union ab) with given/fixed size, while the rest of the network increases in size. On the one hand, the formula $L > \frac{d_a d_b}{2}$ can be seen as a minimum size for $L$, to allow merging a and b. On the other hand, the same formula that it is not convenient to merge a and b if they are big enough with respect to $L$: as seen in the answer to the next question, this also provides a minimal size for a cluster/community, given $L$, to guarantee this it is not convenient to further merge it.

d) Let us consider how $d_a$ is related to the size of a community ($n$ is the number of nodes): if a is a clique (plus one outside edge), the cumulative degree $d_a$ is twice the number of edges plus one: $d_a = n*(n-1)+1 \approx n2$ The previous formula becomes $L > \frac{n^2 * d_b}{2}$, which could be rewritten as $n < \sqrt{\frac{2L}{d_b}}$. Let's take another community/cluster b in the network with minimal $d_b$ value ($d_b$ does not depend on $L$), and introduce a constant $C = \sqrt{\frac{2}{d_b}}$, then the condition for a to be merged conveniently with b is $n < C * \sqrt{L}$. This gives the condition for a community to be merged, so the condition not to be merged is that n has a minimum value of $C * \sqrt{L}$, so $n \propto \sqrt{L}$.