## Question 1 (3 points):

* Find out the node ID of

1. highest degree = S54
2. highest betweenness = S37
3. highest closeness = S37
4. highest eigenvector in the Highschool network = S110

* Afbeelding met diagram

  Automatisch gegenereerde beschrijvingAfbeelding met diagram

  Automatisch gegenereerde beschrijvingHighlight the above nodes in the Highschool network;

Afbeelding met grafiek, diagram

Automatisch gegenereerde beschrijving

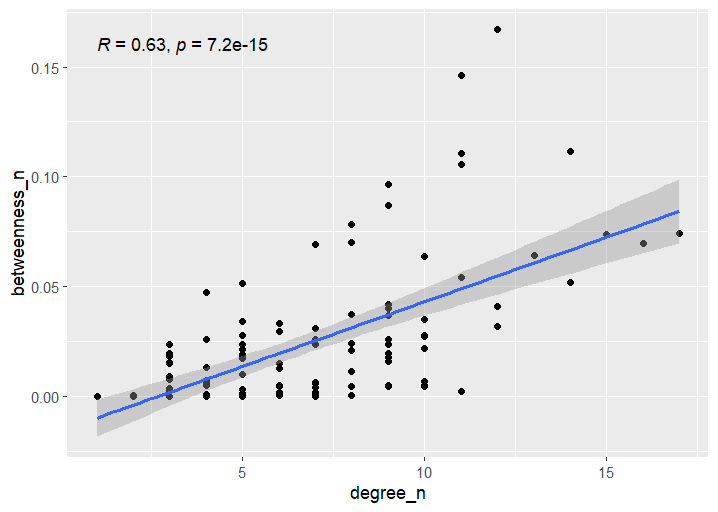
* Explain why these metrics identify the same node or different nodes as the most central one.

All these metrics are calculated in a different way which causes different outputs for the most central node. However according to the betweenness and closeness S37 is the most central. This is probably because these 2 methods both use the shortest path in deciding which node is the most central.

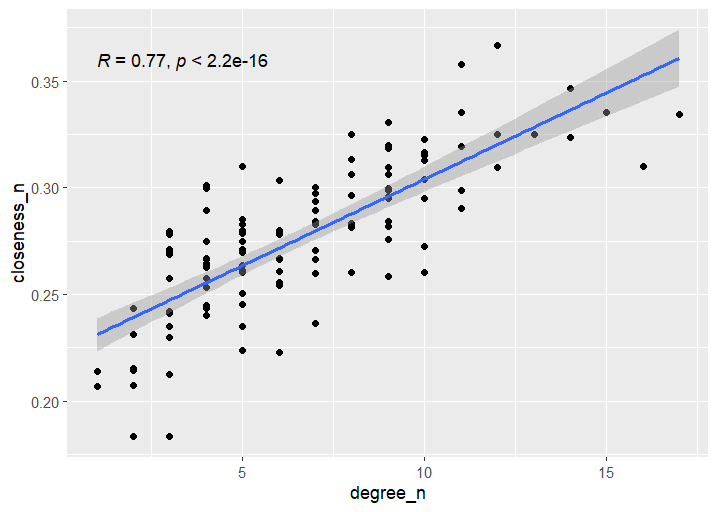
## Question 2 (5 points):

* Study the correlations for all the nodes in the Highschool network between

1. degree and betweenness,



1. degree and closeness,



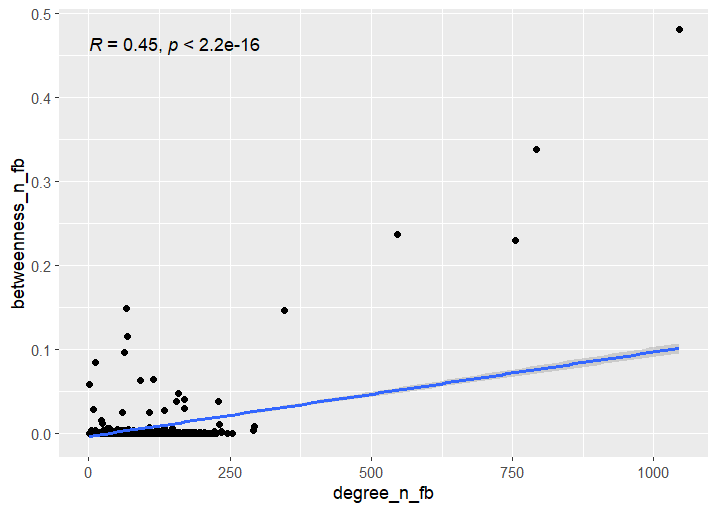
1. degree and eigenvector

Afbeelding met grafiek

Automatisch gegenereerde beschrijving

* Study the correlations for all the nodes in the Facebook network between;

1. degree and betweenness,

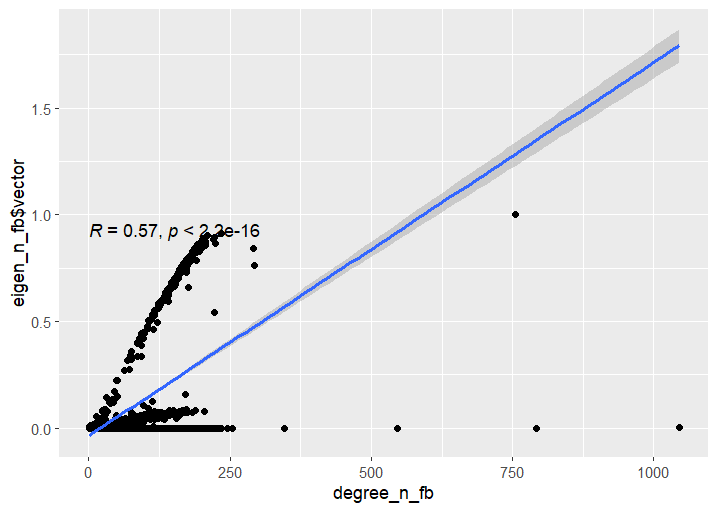


1. degree and closeness,

Afbeelding met grafiek

Automatisch gegenereerde beschrijving

1. degree and eigenvector



* From the above results, how well do different metrics correlate with each other? Which centrality metric will you use and why? Overall the correlations in the high school network is higher. For the high school network its best to use the closeness centrality since it correlates the best with the degree. For the Facebook network this would be the eigenvector centrality, however this might differ if the outliers are removed from the network.

## Question 3 (5 points):

* For both the Highschool and Facebook networks, calculate the shortest path lengths between every pair of two nodes. How many percentage of nodes can be reached within 6 path lengths? Does “six degree of separation” apply to each network?
* Study the degree distribution of these two networks, are they similar? Then use degree distribution to explain the degree of separation you answered above.

## Question 4, 4 points:

Now you will analyze the meso-scale structure of the Highschool network. First, check out the node attributes of Highschool network. You can find 1) the gender and 2) the residential hall of each student. A hypothesis can be formed as: If two students share some common characteristics, such as living in the same residential hall or of the same gender, their chance of being friends are higher. Test the above hypothesis by the following steps

1. Visualize the network and color the nodes by gender and residential hall, respectively.
2. Build 8 subgraphs of the original network according to gender and residential hall: 1 subgraph for female student, 1 subgraph for male student, 1 subgraph for students with unknown gender, and 5 subgraphs for students living in residential hall from 1501 to 1505, respectively. For example, to build a subgraph of all female students, you should keep all the nodes of female students and the edges between them. Other nodes and edges are removed.
3. Study the edge density of all the subgraph and compare them to the edge density of the original network. What is your conclusion for the hypothesis?

* Density for female friends is 0.0551786521935776
* Density for male friends is 0.0514285714285714
* Density for unknown friends is 0.1
* Density for 1501 friends is 0.12987012987013
* Density for 1502 friends is 0.0980392156862745
* Density for 1503 friends is 0.152046783625731
* Density for 1504 friends is 0.0758620689655173
* Density for 1505 friends is 0.0965909090909091
* Density for the whole network is 0.0536512667660209

Comparing the network density with all the subgraphs partly confirms the hypothesis for the study halls. However for gender this is not the case only for the unknown gender, but these are small in size.

Afbeelding met grafiek

Automatisch gegenereerde beschrijving

Figure 1 Unknow gender graph

## Question 5 (4 points):

To better understand the meso-scale structure, we will study community detection algorithm. One important stand of community detection algorithm is based on modularity, which tries to maximize the difference between the actual number of edges in a community and the expected number of edges in the community. However optimizing modularity in a network is NP-hard, therefore have to use heuristics.

1. Calculate the modularity of the Highschool network if community is merely identified by
   * 1. gender and
     2. residential hall, respectively.
2. Search the Louvain Community Detection and explain the algorithm in your own words.
3. Use the Louvain Community Detection to identify communities in the Highschool network. Compare the modularity value produced by the Louvain algorithm to those in 1), and explain the reasons for the differences.