An introductory analysis of Italian statisticians co-authorship network

A. Barbieri, A. Pederzani Samarati, D. Sancin

July 2023

1 Exploratory Analysis

The networks used in this project are centered around Italian academic statisticians, reporting all the collaborations (in terms of papers published together) inside the population itself and academics from different disciplines during two different periods, 2012-2015 and 2016-2019.

1.1 General information and preprocessing

We start by looking at broad features of the networks: they both are undirected weighted networks, where weights specify the number of papers published together by two academics. For each node also a series of attributes are present: `name` (scopus ID to identify the scholar), `affiliation_id` (scopus ID to identify the university), `department/university`, `history` (year of first publication), `academic_range` (years of activity from first published paper), `h_index`, `country_code`, `role` (type of relation with the university) and `SSD` (*Settore Scientifico Disciplinare*, sub-field of affiliation). The first network (referring to the period 2012-2015) has 5,740 nodes and 25,358 edges, while the second (2016-2019) counts 8,129 nodes and 38,513 edges.

Before moving on with the analysis, some preprocessing steps are needed:

1. **Drop nodes containing NAs**: some nodes show some missing values that may interfere with successive operation, thus we decide to just remove them, since they represent a microscopic fraction of the total (33 nodes were dropped in the first network and 31 from the second).

2. **Drop useless attributes:** some attributes are not usable or they represent redundant data so we decided to drop them. In particular `affiliation_id, department/university, history` were removed.

3. **Remove nodes with no connections to statisticians**: We noticed that some components of the graph are made out only of scholars not belonging to the statistics field. Given that the overall goal of the project is to analyze Italian statisticians and their relations we remove all the nodes not belonging to the statistician class that do not share a direct connection to a statistician, dropping 1,247 nodes on the first period and 13 in the second.

1.2 Full networks analysis

After the preprocessing phase the descriptive analysis can be started; in figures 1 and 2 it is possible to observe the networks for 2012/2015 and 2016/2019 respectively, from which it can be seen that most of the academics are not statisticians, actually in 2012/2015 only 12.6% of the academics are statisticians while in 2016/2019 they account for only 8.4% of the total. Moreover, both networks are very sparse as the density, defined as the ratio of the actual number of edges and the largest possible number of edges in the graph, is just 0.001 for both networks.
The first thing to be checked is the distribution of the degrees, edge weights and vertex strength (sum of the weights of the adjacent edges for each vertex), then also the global transitivity (ratio of the count of triangles and connected triples in the graph) and the mean distance (average shortest path) are checked. In the table 1 we can see the computed values. For the network 12/15 the minimum degree is 0, but there is only one isolated node. As it can be seen in figure 3 and 4, the degrees distributions are skewed towards lower degrees: actually this is a common situation in real networks,

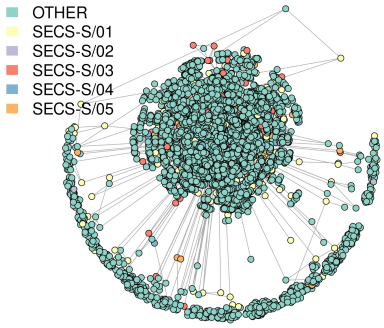

Figure 1: 2012/2015 network

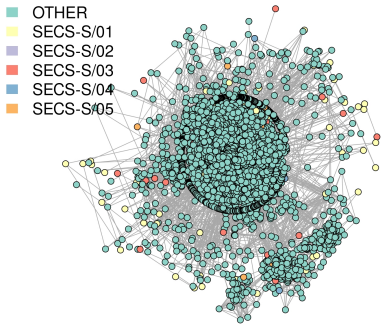

Figure 2: 2016/2019 network

	network 2012/2015	network 2016/2019
mean degree	8.35	9.44
mean weight	1.30	1.27
mean strength	10.93	12.03
global transitivity	0.66	0.50
mean distance	11.67	9.16

Table 1: Descriptive statistics about the networks.

where the degree distribution follows a power law; the same holds also for the weights and strengths distributions.
After this some centrality indexes are considered, in particular the degree centrality, which is simply defined as the number of incident edges upon a node, and the node betweenness centrality, which is based on the number of shortest paths passing through a node. In both networks the ten most central nodes are the same according to the two centrality indexes considered; the most central node in the 2012/2015 network is a full professor belonging to sector S/01, who worked for 24 years and has a h-index of 24, the most central node in the 2016/2019 network is a full professor belonging to sector S/01, who worked for 21 years and has a h-index of 47. Moreover in the 2012/2015 network the ten most central nodes are all statisticians, and also in 2016/2019 all except one are statisticians.
The average h-index for the ten most central nodes is 24.4 in 2012/2015 and 31.1 in 2016/2019, and the average academic range is 27.6 in 2012/2015 and 21.7 in 2016/2019; so it is expected that a higher h-index and a longer career might be both indicative of node popularity.
Two nodes are present in the top ten most central nodes of both networks, the third most central in 2012/2015 becomes the seventh in 2016/2019, and the fifth most central in 2012/2015 becomes the sixth in 2016/2019. In figures 5 and 6 it is possible to see how the ego network of the third most central node of 2012/2015 evolved in 2016/2019.

There are 14 more nodes in the 2016/2019 ego network. Noticeably the few statisticians present are all of the sector S/01, the local clustering coefficient of the central node is 0.09 in both networks.

Next the possibility of decomposing the graph is studied, the 2012/2015 network has 88 components, and a giant component of 3604 nodes (80% of the total) is present; the 2016/2019 network has 82 components and a giant component of 7146 nodes (88% of the total) is present.

Finally the ei index of some attributes is considered: the ei index is defined as the number of ties external to the groups minus the number of ties that are internal to the groups divided by the total number of ties, the index range from -1(homophily) to 1(heterophily). For multiple groups it is based on the mixing matrix, in which the element a_{ij} is the number of ties between groups i and j. In particular the attributes considered are the SSD, the h-index and the academic range; the h-index

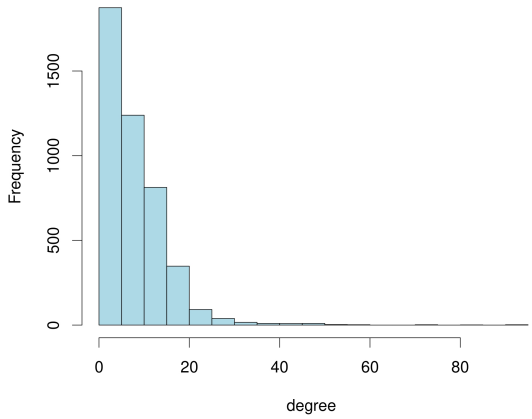

Figure 3: degree distribution 2012/2015 network

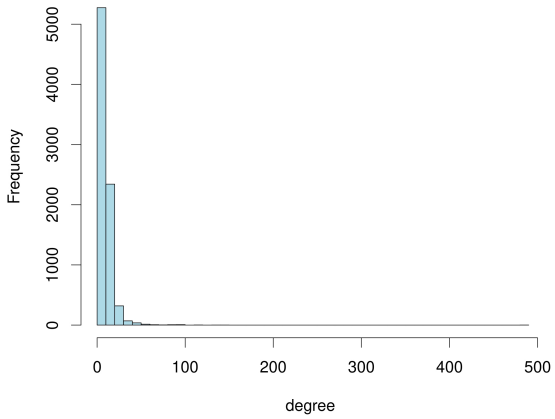

Figure 4: degree distribution 2016/2019 network

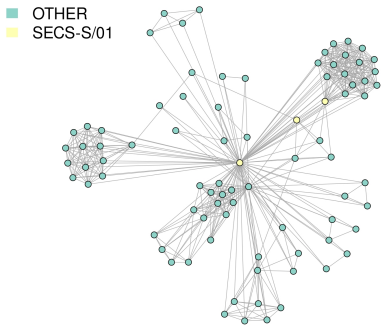

Figure 5: ego network in 2012/2015

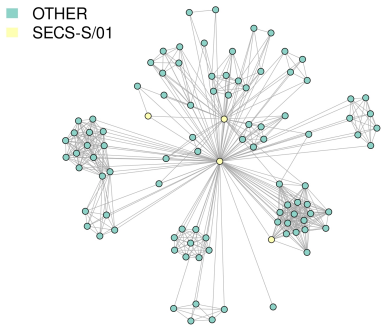

Figure 6: ego network in 2016/2019

and the academic range are factorized in four levels based on their quartiles. In table 2 the ei indexes value are shown. in both periods we can see that there is homophily for the SSD, and heterophily for the h-index and academic factor.

	network 2012/2015	network 2016/2019
ei SSD	-0.46	-0.51
ei h-index	0.36	0.36
ei academic range	0.44	0.45

Table 2: ei indexes

1.3 Statisticians collaborations

We now refine the analysis and look more closely at the collaborations between statisticians. In figure 7 and 8 the networks of only statisticians are shown, while table 3 shows the indexes previously discussed, computed on the two subgraphs. There is a high number of isolated nodes because many statisticians collaborated only with non-statisticians, which are not considered in this part of the analysis. The lack of non-statisticians probably also explain the low global transitivity and the higher mean distance, since many pairs of statisticians are connected through a non-statistician. As

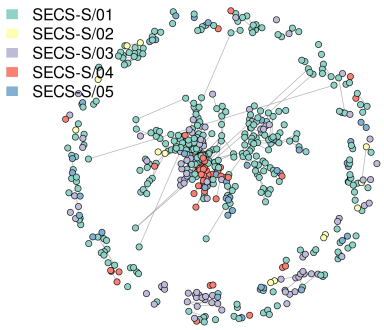

Figure 7: 2012/2015 network of statisticians

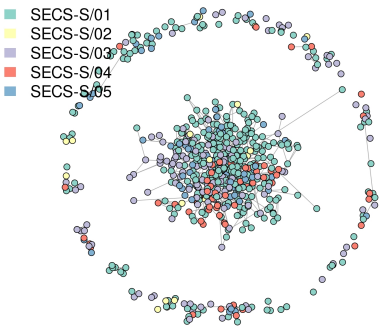

Figure 8: 2016/2019 network of statisticians

	network 2012/2015	network 2016/2019
number of nodes	562	681
number of edges	573	915
edge density	0.003	0.003
mean degree	2.03	2.68
mean weight	2.38	2.09
mean strength	4.26	5.63
global transitivity	0.37	0.36
mean distance	17.66	9.70
isolated nodes	108	123
number of components	165	154
giant component(%)	50	67
ei SSD	-0.63	-0.44
ei h-index	0.20	0.27
ei academic range	0.32	0.34

Table 3: various statistics about the statistician networks

in the full networks the degree, weight and strength distribution all follow a power law distribution. Finally, as for the full network, there is homophily for the SSD and heterophily for the h-index and the academic range.
An interesting analysis comes by looking at the specific SSD-induced populations and how they relate each other. In this step of the analysis we collapse the nodes belonging to the same SSD into single SSD representative nodes, summing all the edges to obtain a single edge between each pair of such nodes (if edges are present). To get a clearer idea of the "strength" of the connection, we used a normalized weight, defined as follows. Let $W_{i,j}$ be the weight of the edge connecting two generic SSD nodes, and let n_i be the number of scholars inside the i-th node. We define the normalized weight as

$$\hat{W}_{i,j} = \frac{W_{i,j}}{n_i + n_j}. \tag{1}$$

With this definition we can look at the strength of a connection without effects due to the size of the nodes.

We start from the 2012-2015 net by creating the nodes according to the SSDs present: Other (3898), S/01 (355), S/02 (15), S/03 (100), S/04 (48) and S/05 (44). The computed normalized weights can be seen in Table 4.

We can see that the strongest relation happens inside the S/01 group itself (normalized weight: 1.335), while the biggest relation between different nodes is between OTHER and S/01 statisticians

	S/01	S/02	S/03	S/04	S/05
OTHER	1.156	0.08	0.21	0.125	0.147
S/01	1.335	0.07	0.119	0.067	0.129
S/02		0.433	-	-	0.017
S/03			0.73	0.02	0.027
S/04				0.667	0.141
S/05					0.204

Table 4: SSD normalized weights for the period 2012-2015.

	S/01	S/02	S/03	S/04	S/05
OTHER	1.119	0.107	0.233	0.077	0.164
S/01	1.301	0.111	0.243	0.091	0.279
S/02		0.35	0.032	0.0125	-
S/03			0.978	0.061	0.151
S/04				0.775	0.377
S/05					0.419

Table 5: SSD normalized weights for the period 2016-2019.

(1.156). One could notice that actually each group shows a strong relation within itself (i.e. among statisticians with the same SSD). In terms of pure degree (that is not-null weights) we can appreciate how every node interacts with every other one, the only exception are S/02 statistician that did not collaborate with S/03 and S/04. Looking also at the weights we can see that there are some preferential attachments mechanisms: some examples are the cited OTHER-S/01, S/03 collaborates more with OTHER and S/01 while S/04 statisticians prefer collaborating with OTHER or S/05 statisticians. Despite this, we can see that these collaborations do not have weights as high as the inter-SSD collaborations. The period 2016-2019 is represented in Table 5 and we can see a very different situation: first, there are more people inside each node (following the order presented before we have 7404, 403, 20, 136, 60 and 62 people). The strongest connection is once again inside the S/01 group, however with a lower weight with respect to the previous period: actually we can see that almost all the other connections have a higher weight, meaning that all of the statisticians have been more productive, also with a higher number of collaborations with different SSDs. Interestingly, the group with the lowest number of connections is once again S/02, however this time the only connection missing is with the S/05 group. As in the previous period, some preferential attachment dynamics are present, especially in the group S/05, showing connections above 0.15 with all the nodes connected and a connection with group S/04 almost as high as the internal one.
Given what was found in the data exploration the following analyses will be done considering the two networks of only statisticians, moreover the homophily for the SSD group will be useful both to establish a way to assess the quality of the found communities and to fit the ERGM.

2 Community Detection analysis

In this part of the analysis we focus on the statisticians during the period 2016-2019. In order to perform a community detection analysis and thus finding the clusters in the graph, we firstly have selected the giant component of the subgraph (the subset of nodes with a path between any two nodes that contains the majority of the nodes), counting 463 researchers (68% of the statisticians).

A preliminary step in community detection analysis is the definition of communities/clusters, which has a meaningful impact on the choice of the centric approach and subsequently on the algorithm to use to perform the analysis.

Considering the results obtained in the explorative analysis, we want to explore the disjoint communities of statisticians implicitly formed and check if these cohesive groups reflect the results obtained by the ei index of the SSD attribute. Indeed, the latter is the node attribute with the lowest ei index, suggesting a great homophily. This implies that communities composed by statisticians with homogeneous SSD should be dense and they should have few connections with the other groups with

almost completely different SSD.

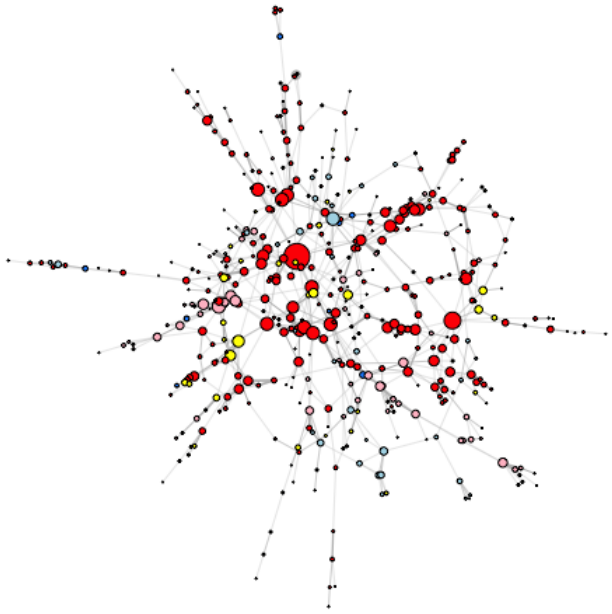

Figure 9: Giant component of statisticians for the period 2016-2019. The nodes size is proportional to the degree of the nodes while the color represents the SSD attribute. The width of the edges is proportional to the weight of the edges.

According to our aim, we have opted for a network centric approach. Since we are still interested in dense communities without imposing too restrictive rules and due to the dimension of the giant component we have adopted the Louvain method. This is a hierarchical clustering algorithm which allows to obtain a network community structure with high modularity. The only parameter to tune for the Louvain method is the resolution which has an implication on the number and sizes of the clusters. Due to our primary objective, we have tuned the resolution parameter by choosing the value of the latter which maximises the mean Gini index computed on the SSD attribute over clusters obtained. Indeed, the largest the Gini index the largest is the cluster homogeneity according to a given node attribute. SSD The value that the resolution can take is not bounded, but it has a strong impact on the result of the algorithm. After many tests, we have identified as meaningful the range of values between 0.5 and 1.5. Indeed, as we decrease the resolution the number of clusters produced by the algorithm decreases, while the modularity increases. On the other hand, by increasing too much the resolution the number of clusters increases, while the modularity decreases. As the number of clusters increases we get smaller and smaller communities, which have a high, but not meaningful Gini index. The range of values of resolution selected allows us to perform a fair parameter tuning, while obtaining results that make sense with our network, both in terms of communities size, modularity and homogeneity in the SSDs.

After a grid search over different resolution parameters, the best found one is resolution=1.3, producing 32 communities with modularity=0.86, as shown in Figure 10.

These results are in line with the results obtained in the exploratory analysis. As shown in Table 6, the mean Gini Index of the communities is high, meaning that the nodes inside communities are homogeneous with respect to the SSD attribute. The mean intra-cluster density is 0.29, this value is quite high compared to the density of the giant component (0.0077). Finally, the mean inter-cluster edges is 0.28. Indeed most of the clusters do not have edges that connect the communities each other.

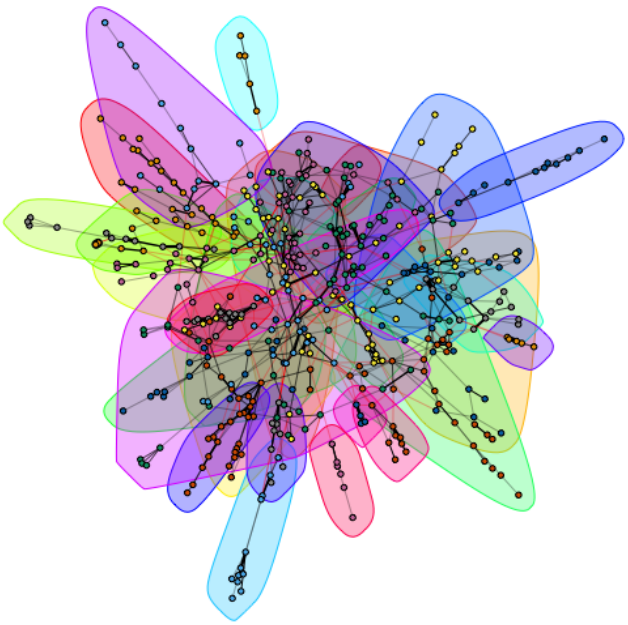

Figure 10: Communities obtained with the Louvain method.

	mean	standard deviation
Communities size	14.46	6.60
Gini Index	0.67	0.23
Intra-cluster density	0.29	0.20
Inter-cluster edges	0.28	0.77
Diameter	11.47	3.8

Table 6: Descriptive statistics of the communities.

The Louvain algorithm is very dependant on the order of the nodes, influencing the clusters produced and the modularity. To verify the stability of our output we have run the algorithm 1000 times (also during the tuning of the resolution parameter), each time shuffling the order of the nodes. Figure 11 shows the distribution of the number of communities over the trials, while Figure 12 shows the distribution of the modularity. Both Figure 11 and Figure 12 prove that the results that we obtain are stable.

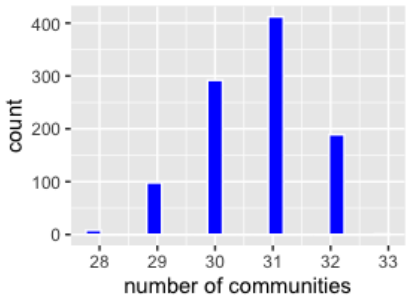

Figure 11: Distribution of number of communities over 1000 trials.

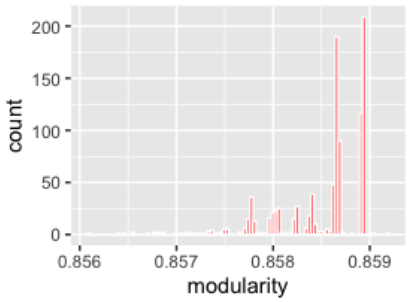

Figure 12: Distribution of modularity over 1000 trials.

3 ERGM models

In this last section we are going to present some possible ERGM models for the two periods. As presented in section 1, we are dealing with two large networks, and trying to model them as a whole will likely result in a big computational failure. The idea is then to model only the relations between statisticians by keeping only the relative induced subgraphs, further filtering the networks by keeping only the giant components, accounting for 50% and 67.9% of the total respectively. A final note regards the weights of the edges: given their distribution, highly skewed to the left with the majority of the edges showing a weight of 1, we decide to maintain all the connections, whatever value they take.

We start by examining the 2016-2019 period. The best model, according to the BIC score, takes into consideration the number of collaborations, the count of researchers of all the possible combinations of SSDs (excluding SECS-S/02 group, since all of the coefficients including it were not significant) and factorized h-index (excluding the first group with lowest h-index always for low-significance). The results of the fitting process can be seen in Figure 13.

```
Maximum Likelihood Results:

                          Estimate Std. Error MCMC % z value Pr(>|z|)
edges                     -4.78853    0.05791      0 -82.684  < 1e-04 ***
mix.ssd.SECS-S/01.SECS-S/03 -1.27224    0.13190      0  -9.645  < 1e-04 ***
mix.ssd.SECS-S/03.SECS-S/03  1.08022    0.11653      0   9.270  < 1e-04 ***
mix.ssd.SECS-S/01.SECS-S/04 -1.67906    0.21855      0  -7.683  < 1e-04 ***
mix.ssd.SECS-S/03.SECS-S/04 -2.37385    0.57960      0  -4.096  < 1e-04 ***
mix.ssd.SECS-S/04.SECS-S/04  1.77316    0.16478      0  10.761  < 1e-04 ***
mix.ssd.SECS-S/01.SECS-S/05 -0.53978    0.12820      0  -4.210  < 1e-04 ***
mix.ssd.SECS-S/03.SECS-S/05 -0.87866    0.27235      0  -3.226 0.001255 **
mix.ssd.SECS-S/04.SECS-S/05  0.28795    0.22013      0   1.308 0.190847
mix.ssd.SECS-S/05.SECS-S/05  0.92232    0.22659      0   4.071  < 1e-04 ***
mix.h_fact.2.3               0.42741    0.08257      0   5.176  < 1e-04 ***
mix.h_fact.3.3               0.75542    0.12798      0   5.903  < 1e-04 ***
mix.h_fact.2.4               0.73195    0.18873      0   3.878 0.000105 ***
mix.h_fact.3.4               1.25014    0.21006      0   5.951  < 1e-04 ***
mix.h_fact.4.4               2.17426    0.60688      0   3.583 0.000340 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 148268  on 106953  degrees of freedom
 Residual Deviance:   9157  on 106938  degrees of freedom

AIC: 9187  BIC: 9331  (Smaller is better. MC Std. Err. = 0)
```

Figure 13: Coefficients for the 2016-2019 model.

We can see that all the coefficients are significant, excluding the "mix" between S/04 and S/05. Furthermore, the conclusion we made in Section 1 are partially confirmed by the parameters of the model: we can see that these are positive (i.e. a higher probability to form a connection) for connections between researchers of the same group and also for connection between group S/04 and S/05 (even though the coefficient is not statistically different from 0). However, connections between different SSDs almost always lead to negative coefficients, meaning that heterogeneous connections SSD-wise are more unlikely. The same reasoning applies to the h-index mix variables, showing a positive effect in the probability of forming connections and growing as the h-index grows. Also homophily effects were included in the testing phase, however they were always not significantly different from 0 and lowered the BIC score, thus we removed them in the final model. We also tried adding some dyad-dependent terms, however most of the times they led to computational failures due to the size of the network or high MCMC errors leading to unstable estimates.

We now take a look at the goodness-of-fit graphs to assess the model quality, based on 100 simulated networks, which are reported in Figure 14. We can see that the median of the statistics used are always close to the observed data, as well as the degree distribution which almost perfectly follows the observed one but for the isolated nodes and degrees 1 and 2. The flaws of the model emerge in the last two graphs: on the left one we can see that the number of edge-wise shared partners is way underestimated (especially for the numbers 1 and 2), while in the right one we can see that the model produces networks with low minimum path lengths.

Lastly, we briefly talk about the 2012-2015 model, which shares most of the flaws with the model just presented above. The best model found is smaller in terms of parameters with respect to the previous one and takes into consideration the overall number of connections, homophily effects in the SSDs and the h-index of the nodes. Estimated coefficients are reported in Figure 15 and once again match the conclusions of Section 1: there's a strong homophily in the SSDs, characterized by

positive coefficients (i.e. higher probability of forming a connection) and an increase in probability is also due to the h-index, where the higher the more likely to form connections. However, the GOF plots, showed in Figure 16, highlight the same problems presented before, with an underestimation of edge-wise shared partners and a minimum geodesic distance shifted to the left and too narrow if compared to the actual one.

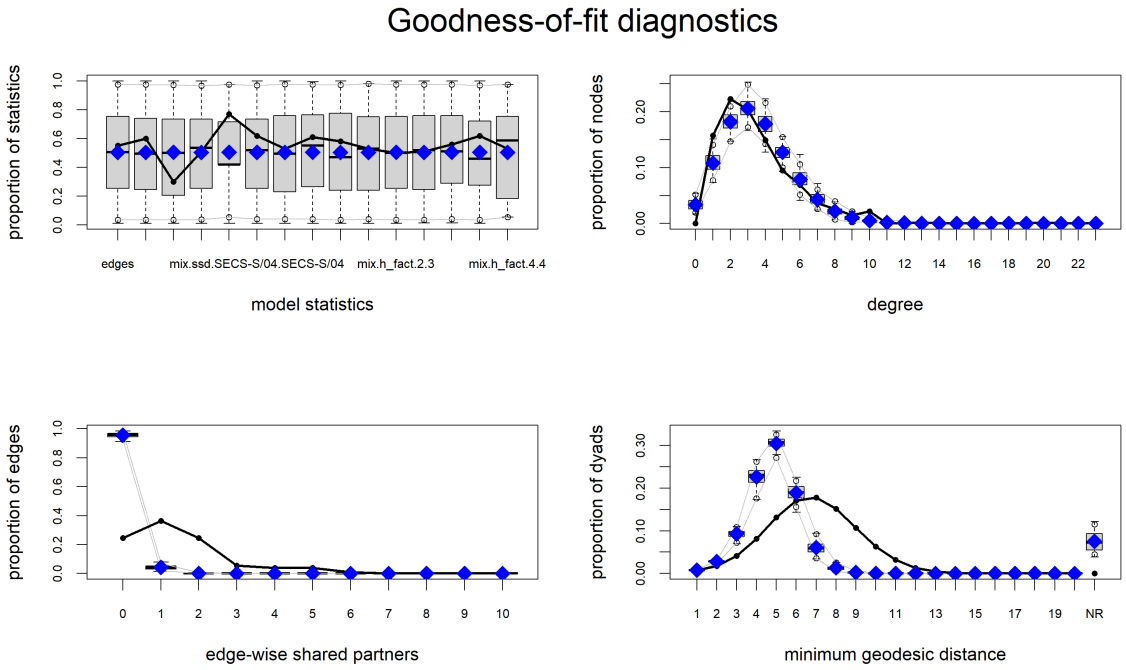

Figure 14: GOF plots for the 2016-2019 model.

```
Maximum Likelihood Results:

                       Estimate Std. Error MCMC % z value Pr(>|z|)
edges                 -5.937885   0.162063      0 -36.639  < 1e-04 ***
nodematch.ssd.SECS-S/01  1.132200   0.122989      0   9.206  < 1e-04 ***
nodematch.ssd.SECS-S/02  4.496173   1.231657      0   3.651 0.000262 ***
nodematch.ssd.SECS-S/03  2.257593   0.223679      0  10.093  < 1e-04 ***
nodematch.ssd.SECS-S/04  2.920986   0.206498      0  14.145  < 1e-04 ***
nodematch.ssd.SECS-S/05  2.159626   0.468607      0   4.609  < 1e-04 ***
nodecov.h_index          0.024882   0.005285      0   4.708  < 1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null Deviance: 54537  on 39340  degrees of freedom
 Residual Deviance:  4428  on 39333  degrees of freedom

AIC: 4442  BIC: 4502  (Smaller is better. MC Std. Err. = 0)
```

Figure 15: Coefficients for the 2012-2015 model.

Some further improvements that could be done in the models is the inclusion of statistics regarding also the relations outside the statisticians' sphere, such as the number of collaborations with researchers not affiliated with the statistics field, or to jointly model statisticians and other at the same time: this may reduce the diameter issue seen in the models. Furthermore, a key attribute that we neglected (due to issues in the quality of the data as explained in Section 1) is the affiliation of the researchers: the geographic location of the researcher and its affiliation university could, in our view, greatly improve the model, since in general professors in the same department are more likely to collaborate.

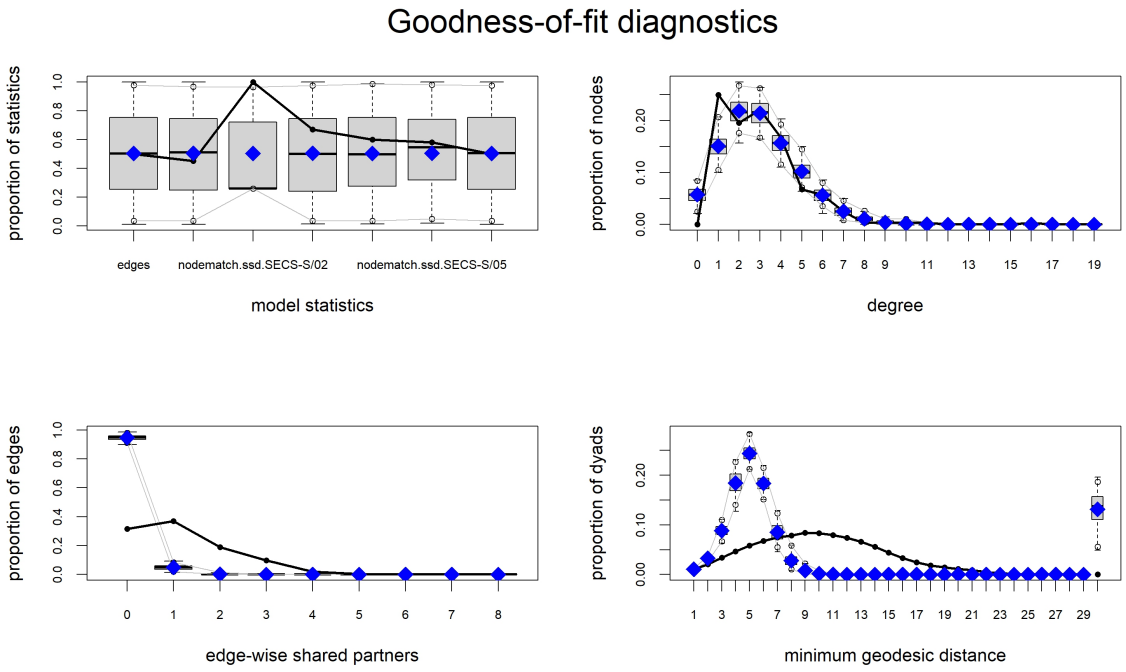

Figure 16: GOF plots for the 2012-2015 model.