Lab 2 - ERGM (Exponential Random Graph Models)

CompSci 396-0: Social Networking Analysis ***Win 2022***

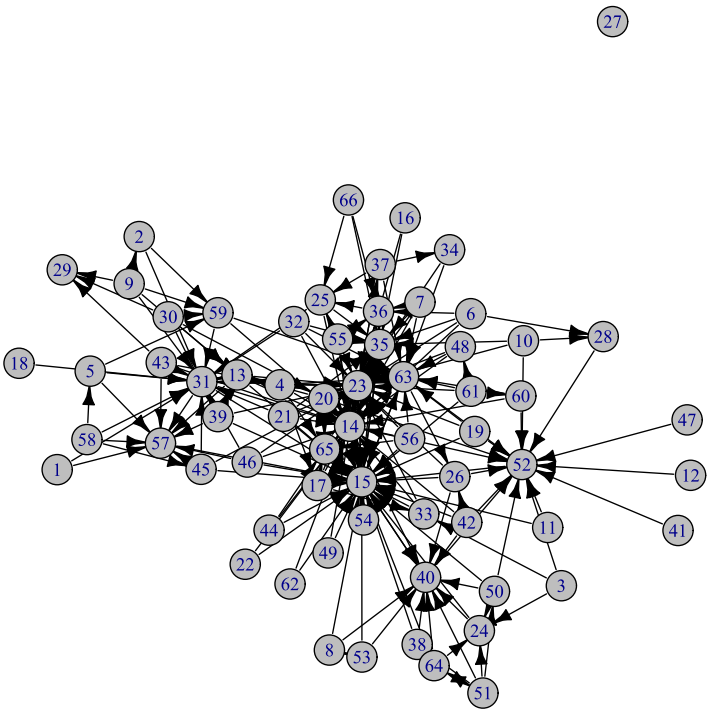
Student Name: Jiaqi Guo NetID: JGR9647

● Responses to Question

■ Part I: Building and Visualizing the Networks (15 pts)

1. (5 points) Plot the base (Advice) network and include it in your report. Explain whether this plot seems, at a glance, to match what you would expect to see if hypothesis 1 were true. Think of this as just a basic descriptive check – we will perform a more rigorous statistical test in part II of the lab.

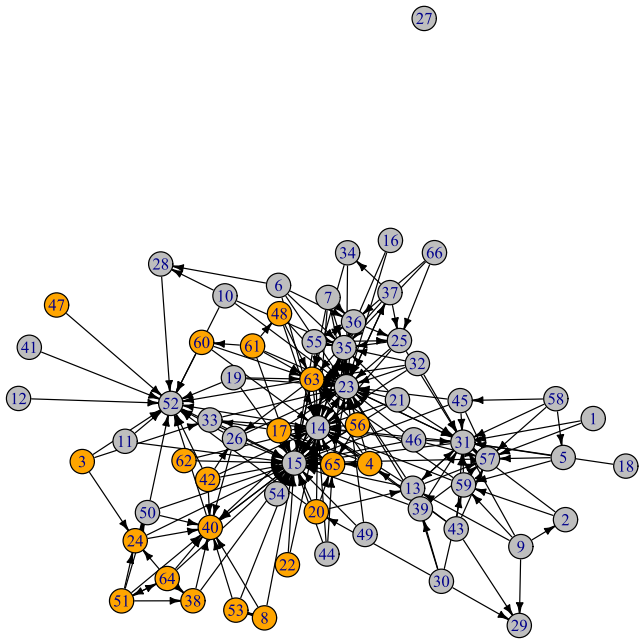
   Hypothesis 1: There will be indegree popularity effects (tendency of a small number of nodes to receive many ties) in who people go to for advice.



We can easily observe that there are several nodes with high in-degree centrality in the figure, such as **node 14, 15, 23**, etc. This suggests that there is a popularity effect on who people go to for advice, which meet our expectations.

2. (5 points) Plot the base network with the nodes now colored based on sex and include it in your report. Explain whether this plot seems to match what you would expect to see if hypothesis 3 were true.

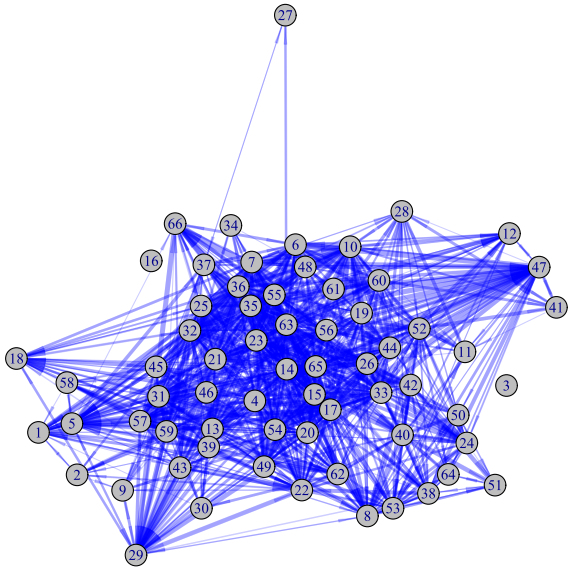
   Hypothesis 3: There will be homophily based upon the sex of individuals, in terms of who employees go to for advice.



From the graph above, it is clear that individual gender can influence the choice of whom an employee seeks advice from. A few typical nodes are Node 40 and node 31, and most of the nodes pointing to them are **homogenous**, which meet our expectations in **hypothesis 3**.

3. (5 points) Plot the covariate network and include it in your report. Comparing the network plots, explain whether this plot seems to match what you would expect to see if hypothesis 4 were true.

Hypothesis 4: Employees who message someone more frequently on ESM will be more likely to report going to that person for advice.



Although the specific information about the number of connections cannot be accurately obtained from the image, we can still observe that the connection density near nodes with higher in-degree centrality is also higher (low transparency), which means that these nodes receive more information through ESM. This phenomenon matches our expectations on **Hypotheses 4**.

■ Part II: Model Estimation (55 pts)

1. (3 points) Build two ERGM models to test the hypotheses using the different network statistics described below and include the results (screenshot of the model output tables from the R console) in your report. Fit model 1 (simple model) and model 2 (complex model) using the terms already specified in the R script provided to you.

Model 1:

```
> summary(model1)
Call:
ergm(formula = advice ~ edges + mutual + edgecov(hundreds_messages) +
    nodemix("leader", base = 3), constraints = ~bd(maxout = 5))

Monte Carlo Maximum Likelihood Results:

                        Estimate Std. Error MCMC % z value Pr(>|z|)
edges                  -0.67509    0.13572      0  -4.974  < 1e-04 ***
mutual                  1.12232    0.41113      0   2.730 0.006336 **
edgecov.hundreds_messages  0.34519    0.09096      0   3.795 0.000148 ***
mix.leader.0.0         -2.79212    0.15354      0 -18.185  < 1e-04 ***
mix.leader.1.0         -3.44660    0.48934      0  -7.043  < 1e-04 ***
mix.leader.1.1         -0.13560    0.43024      0  -0.315 0.752625
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance:    0.0  on 4290  degrees of freedom
 Residual Deviance: -724.6  on 4284  degrees of freedom

Note that the null model likelihood and deviance are defined to be 0. This means that all
likelihood-based inference (LRT, Analysis of Deviance, AIC, BIC, etc.) is only valid between
models with the same reference distribution and constraints.

AIC: -712.6  BIC: -674.4  (Smaller is better. MC Std. Err. = 1.381)
```

Model 2:

```
> summary(model2)
Call:
ergm(formula = advice ~ mutual + gwidegree(log(2), fixed = T) +
    dgwodegree(2, fixed = T, cutoff = 5) + dgwesp(log(2), type = "OTP",
    fixed = T, cutoff = 5) + nodematch("female") + nodemix("leader",
    base = 3) + nodematch("department") + nodeicov("office") +
    nodeocov("office") + diff("tenure") + edgecov(hundreds_messages),
    constraints = ~bd(maxout = 5), control = control.ergm(MCMC.effectiveSize = 50))

Monte Carlo Maximum Likelihood Results:

                                Estimate Std. Error MCMC % z value Pr(>|z|)
mutual                         -0.44651    0.49772      0  -0.897  0.36965
gwideg.fixed.0.693147180559945 -2.42481    0.33522      0  -7.234  < 1e-04 ***
gwodeg.fixed.2                 -3.63308    0.28283      0 -12.846  < 1e-04 ***
gwesp.OTP.fixed.0.693147180559945  0.64558    0.09751      0   6.620  < 1e-04 ***
nodematch.female                0.19826    0.14116      0   1.405  0.16015
mix.leader.0.0                 -1.09335    0.21443      0  -5.099  < 1e-04 ***
mix.leader.1.0                 -1.53896    0.55758      0  -2.760  0.00578 **
mix.leader.1.1                 -0.48848    0.54084      0  -0.903  0.36642
nodematch.department            2.10831    0.18035      0  11.690  < 1e-04 ***
nodeicov.office                -0.24562    0.14366      0  -1.710  0.08730 .
nodeocov.office                 0.02697    0.19668      0   0.137  0.89093
diff.t-h.tenure                -0.14442    0.02390      0  -6.042  < 1e-04 ***
edgecov.hundreds_messages       0.41135    0.09248      0   4.448  < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance:     0  on 4290  degrees of freedom
 Residual Deviance: -1125  on 4277  degrees of freedom

Note that the null model likelihood and deviance are defined to be 0. This means that all
likelihood-based inference (LRT, Analysis of Deviance, AIC, BIC, etc.) is only valid between
models with the same reference distribution and constraints.

AIC: -1099  BIC: -1017  (Smaller is better. MC Std. Err. = 1.168)
```

Comparison:

```
> screenreg(list("model1"=model1,"model2"=model2))

=============================================================
                                  model1       model2
-------------------------------------------------------------
edges                            -0.68 ***
                                 (0.14)
mutual                            1.12 **      -0.45
                                 (0.41)       (0.50)
edgecov.hundreds_messages         0.35 ***     0.41 ***
                                 (0.09)       (0.09)
mix.leader.0.0                   -2.79 ***    -1.09 ***
                                 (0.15)       (0.21)
mix.leader.1.0                   -3.45 ***    -1.54 **
                                 (0.49)       (0.56)
mix.leader.1.1                   -0.14        -0.49
                                 (0.43)       (0.54)
gwideg.fixed.0.693147180559945                -2.42 ***
                                              (0.34)
gwodeg.fixed.2                                -3.63 ***
                                              (0.28)
gwesp.OTP.fixed.0.693147180559945              0.65 ***
                                              (0.10)
nodematch.female                               0.20
                                              (0.14)
nodematch.department                           2.11 ***
                                              (0.18)
nodeicov.office                               -0.25
                                              (0.14)
nodeocov.office                                0.03
                                              (0.20)
diff.t-h.tenure                               -0.14 ***
                                              (0.02)
-------------------------------------------------------------
AIC                              -712.62      -1099.25
BIC                              -674.44      -1016.51
Log Likelihood                    362.31        562.62
=============================================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

2. (45 points) For each of the hypotheses, interpret the results from your models and state whether that hypothesis was supported. To determine whether a hypothesis is supported, look at whether there is a p-value < 0.05 and the directionality (positive/negative) of the effect. When you interpret the results, convert the model coefficients, which are given by R as conditional log-odds, into odds ratios.

Hypothesis 1: There will be indegree popularity effects (tendency of a small number of nodes to receive many ties) in who people go to for advice.

For model 2, *gwidegree*: There is a negative estimate(0.09 in odd ratio), and it is statistically significant ($p<0.05$).Which suggest incoming ties are 0.09 times less likely to be directed towards nodes that have no other incoming ties(more likely to be directed towards nodes that already have other incoming ties).

This hypothesis is supported.

Hypothesis 2: Individuals will be more likely to report go to advice from people in their own department, as opposed to other departments.

For model 2, *nodematch.department*: There is a positive estimate(8.23 in odd ratio), and it is statistically significant($p<0.05$), which suggests that compared with people in other department, individual are 8.23 times more likely to tie with people in the same department.

This hypothesis is supported.

Hypothesis 3: There will be homophily based upon the sex of individuals, in terms of who employees go to for advice.

For model 2, *nodematch.female*: There is a positive estimate(1.22 in odd ratio), but it is not statistically significant($p>0.05$), which suggests individuals are less likely to ask people of the same sex for advice.

This hypothesis is not supported

Hypothesis 4: Employees who message someone more frequently on ESM will be more likely to report going to that person for advice.

For both model 1 and 2, *edgecov.hundreds_message*: There is a positive estimate, and it is statistically significant($p<0.05$), which suggests that individuals are about 1.5 times more likely to ask people they frequently talked with on ESM for advice.

This hypothesis is supported.

Hypothesis 5: If an employee *i* goes to another employee *j* for advice, it will be more likely that *j* also goes to employee *i* for advice.

For model 1, *mutual*: There is a positive estimate(3.06 in odd ratio), and it is statistically significant ($p<0.05$), which suggests employee i will be 3.06 times more likely to connect with j, if j is connected to i.

For model2, there is a negative estimate(0.63 in odd ratio), but it is not statistically significant($p>0.05$). So, the conclusion will be the same as model 1.

This hypothesis is supported.

Hypothesis 6: There will be indegree preferential attachment effects – That is, a tendency for a small number of employees to be sought out for advice from many others (as opposed to advice seeking behaviors being spread evenly amongst all employees).

For model 2, *gwidegree*: There is a negative estimate (0.09 in odd ratio, which is inequal to 1), and it is statistically significant ($p<0.05$). Which suggests, at least, that incoming ties are unlikely to be equally distributed among all employees. Therefore, we can take this hypothesis as **true**.

This hypothesis is supported.

Hypothesis 7: Employees who work in the main office will be more likely to go to others for advice than employees from the secondary office.

For model 2, *nodeocov.office*: There is a positive estimate (1.30 in odd ratio), but it is not statistically significant ($p>0.05$). Which suggests employees work in the main office are less or equal likely to go to others for help.

This hypothesis is not supported

Hypothesis 8: Employees who work in the main office will be more likely to be sought after for advices than employees from the secondary office.

For model 2, *nodeicov.office*: There is a negative estimate (0.78 in odd ratio), but it is not statistically significant ($p>0.05$). Which means employees work in the main office are more or equal likely to be sought after for advice than employees from the secondary office.

This hypothesis is supported.

Hypothesis 9: Advice seeking relationships tend to be transitive - That is, if individual i goes to an individual k for advice, and k goes to an individual j for advice, then i is more likely to go to j for advice as well.

For model 2, *gwesp.OTP*: There is a positive estimate (1.90 in odd ratio), and it is statistically significant ($p<0.05$). Which means if the relationship $i \rightarrow k \rightarrow j$ exist, then i will be 1.9 times more likely to have direct relationship with j.

This hypothesis is supported.

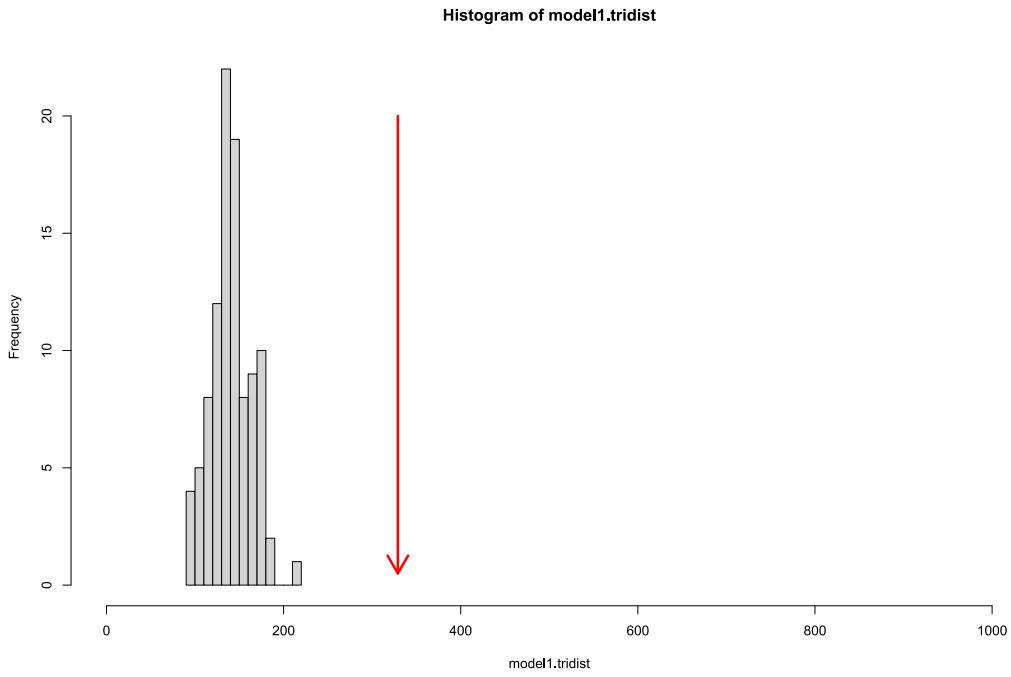
■ Part III: Model Estimation (55 pts)

1. (10 points) Attach the model diagnostics for model 1 and 2 in your report (you should submit a single PDF file) and interpret the plots. Has the MCMC process converged to a desired state? for every term in the model

According to the Diagnostics plot (both model 1 & 2) in the appendix, we can observe that for each term in our model, the data in all graphs on the left only fluctuate up and down around a certain central value, indicating that our model(both model 1 &2) has both converged to a relatively stable state.

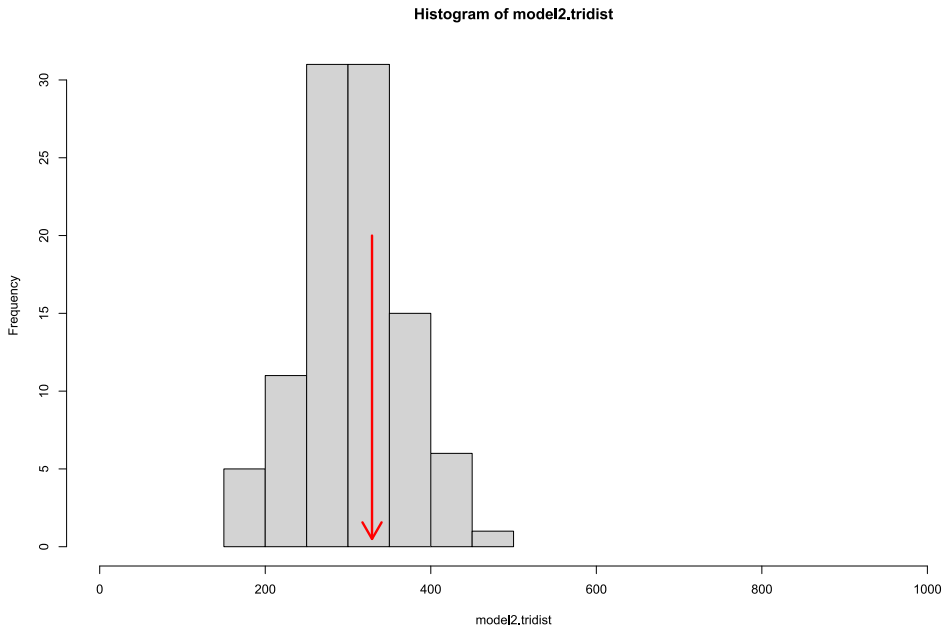
2. (10 points) Perform Goodness of Fit test to check how well the estimated model captures certain statistical features of the observed network for both model 1 and 2
   a. To do so, simulate many networks from the estimated model and extract 100 samples from the simulation process. Please note, this may take 2 minutes or more to compute.
   b. Extract the number of triangles from each of the 100 samples.
   c. Compare the distribution of triangles in the sampled networks with the observed network by generating a histogram of the triangles. Interpret your result -- is the estimated model a good one in terms of triangle measure?

■ Model 1:



The red arrow is the number of triangles on the actual advice network, here the number is 329. For model 1, The quantity distribution of triangle structures in our simulated model is actually quite different from the number of triangles in the actual advice network, which means our model may not be a good reflection of reality.

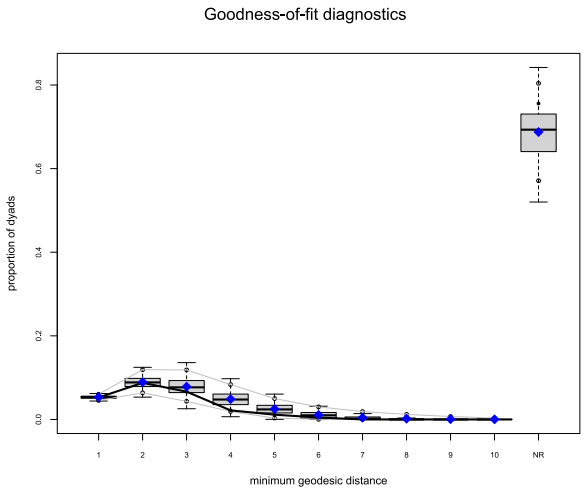
■ Model 2:



In model 2 histogram, we can easily observe that the number of triangles in our simulated network matches the number of triangles in the real network, which suggests that our model 2 did a good job of replication the number of triangles and the actual advice network.

3. (10 points) Repeat this goodness-of-fit evaluation process for a variety of other network statistics just for model 2 (for example, degree distribution, distribution of edgewise shared partners, and the distribution of geodesics). Simulate networks as we did above, compile statistics for these simulations as well as the observed network, and calculate p-values of all of the aforementioned values to evaluate the correspondence between the networks simulated by the model and the observed network. Report the p-values for the simulation and interpret them.

■ Distribution of Geodesics



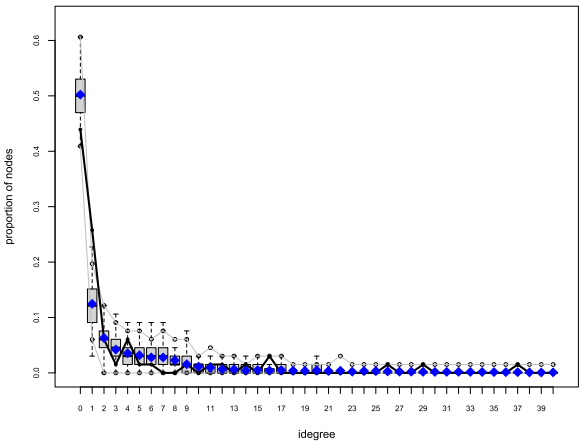
Goodness-of-fit for minimum geodesic distance

	obs	min	mean	max	MC p-value
1	225	184	228.455	266	0.81
2	376	206	382.535	553	0.92
3	285	110	336.890	593	0.56
4	93	28	208.565	418	0.11
5	49	1	107.080	260	0.28
6	20	0	47.030	199	0.60
7	0	0	18.595	144	0.51
8	0	0	7.490	94	1.00
9	0	0	2.780	63	1.00
10	0	0	0.960	40	1.00
11	0	0	0.155	10	1.00
12	0	0	0.005	1	1.00
Inf	3242	2231	2949.460	3612	0.25

Generally speaking, the closer our P value is to 1, the more similar our simulated network is to the observed network(the better performance of our simulated model). If the p-value is less than 0.05, it mean that our simulation s very different from the actual data(bad performance).

From the perspective of geodesic distance, we can see that our simulation model has a high consistency with the actual observed model, with p-values not less than 0.05. This indicates that our model can well simulate the min-geodesic-distance property of the real observed networks.

■ In-degree

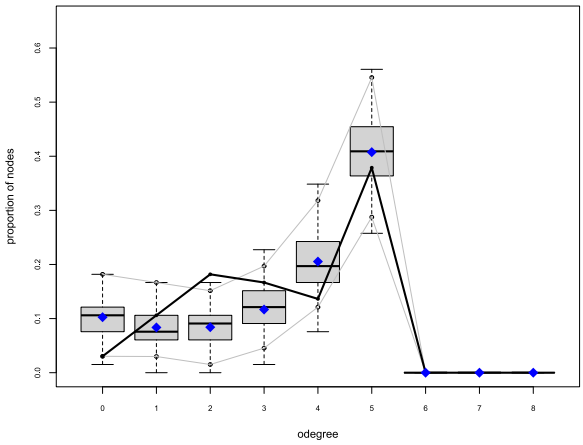


Goodness-of-fit for in-degree

	obs	min	mean	max	MC p-value
idegree0	29	24	33.150	42	0.18
idegree1	17	2	8.225	15	0.00
idegree2	4	0	4.160	10	1.00
idegree3	1	0	2.785	7	0.47
idegree4	4	0	2.340	7	0.48
idegree5	1	0	2.105	6	0.72
idegree6	1	0	1.840	6	0.87
idegree7	0	0	1.835	6	0.34
idegree8	0	0	1.505	6	0.43
idegree9	1	0	1.030	5	1.00
idegree10	0	0	0.740	3	0.82
idegree11	1	0	0.610	4	0.91
idegree12	1	0	0.430	2	0.73
idegree13	0	0	0.420	3	1.00
idegree14	1	0	0.290	3	0.53
idegree15	0	0	0.340	3	1.00
idegree16	2	0	0.280	2	0.06
idegree17	0	0	0.330	2	1.00
idegree18	0	0	0.265	2	1.00
idegree19	0	0	0.250	2	1.00
idegree20	0	0	0.295	2	1.00
idegree21	0	0	0.235	2	1.00
idegree22	0	0	0.255	2	1.00
idegree23	0	0	0.145	2	1.00
idegree24	0	0	0.225	2	1.00
idegree25	0	0	0.220	2	1.00
idegree26	1	0	0.180	2	0.35
idegree27	0	0	0.150	2	1.00
idegree28	0	0	0.150	1	1.00
idegree29	1	0	0.130	2	0.25
idegree30	0	0	0.120	3	1.00
idegree31	0	0	0.120	1	1.00
idegree32	0	0	0.125	2	1.00
idegree33	0	0	0.115	2	1.00
idegree34	0	0	0.105	2	1.00
idegree35	0	0	0.095	1	1.00
idegree36	0	0	0.085	2	1.00
idegree37	1	0	0.085	1	0.17
idegree38	0	0	0.060	1	1.00
idegree39	0	0	0.050	1	1.00
idegree40	0	0	0.060	1	1.00
idegree41	0	0	0.020	1	1.00
idegree42	0	0	0.015	1	1.00
idegree43	0	0	0.020	1	1.00
idegree44	0	0	0.010	1	1.00

From the perspective of network in-degree, we can see that our simulation model is highly consistent with the actual observed model, although there is one defects(when In-degree = 1, p-value is less than 0.05). This indicates that our model can well simulate the in-degree property of the real observed networks.

■ Out-degree

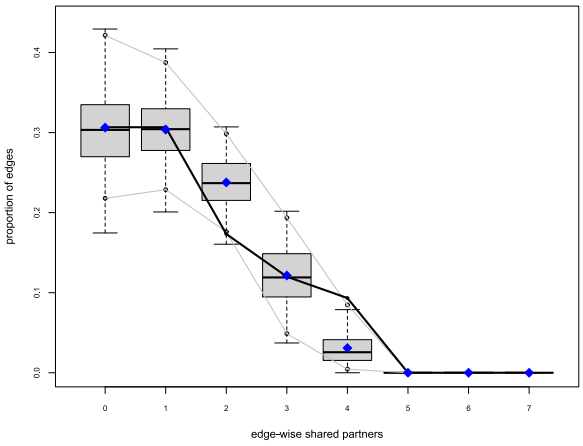


Goodness-of-fit for out-degree

	obs	min	mean	max	MC p-value
odegree0	2	1	6.770	13	0.06
odegree1	7	0	5.520	13	0.62
odegree2	12	0	5.550	13	0.02
odegree3	11	1	7.705	15	0.34
odegree4	9	5	13.555	25	0.17
odegree5	25	17	26.900	43	0.79

From the perspective of network out-degree, our simulation model is basically consistent with the actual observed model, except when the out-degree is equal to 2, the P value is less than 0.05. This indicates that our model can well simulate the out-degree property of the real observed networks.

■ Edgewise shared partners

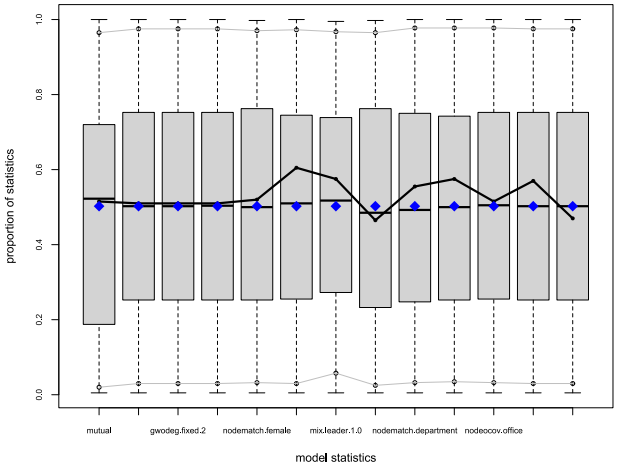


Goodness-of-fit for edgewise shared partner

	obs	min	mean	max	MC p-value
esp0	69	44	69.440	96	1.00
esp1	69	39	69.285	101	1.00
esp2	39	28	54.520	79	0.12
esp3	27	8	28.035	58	1.00
esp4	21	0	7.175	30	0.05

From the perspective of network edgewise shared partners, our simulation model is consistent with the actual observed model, with p-values not less than 0.05. This indicates that our model can well simulate the out-degree property of the real observed networks.

■ Model Statistic



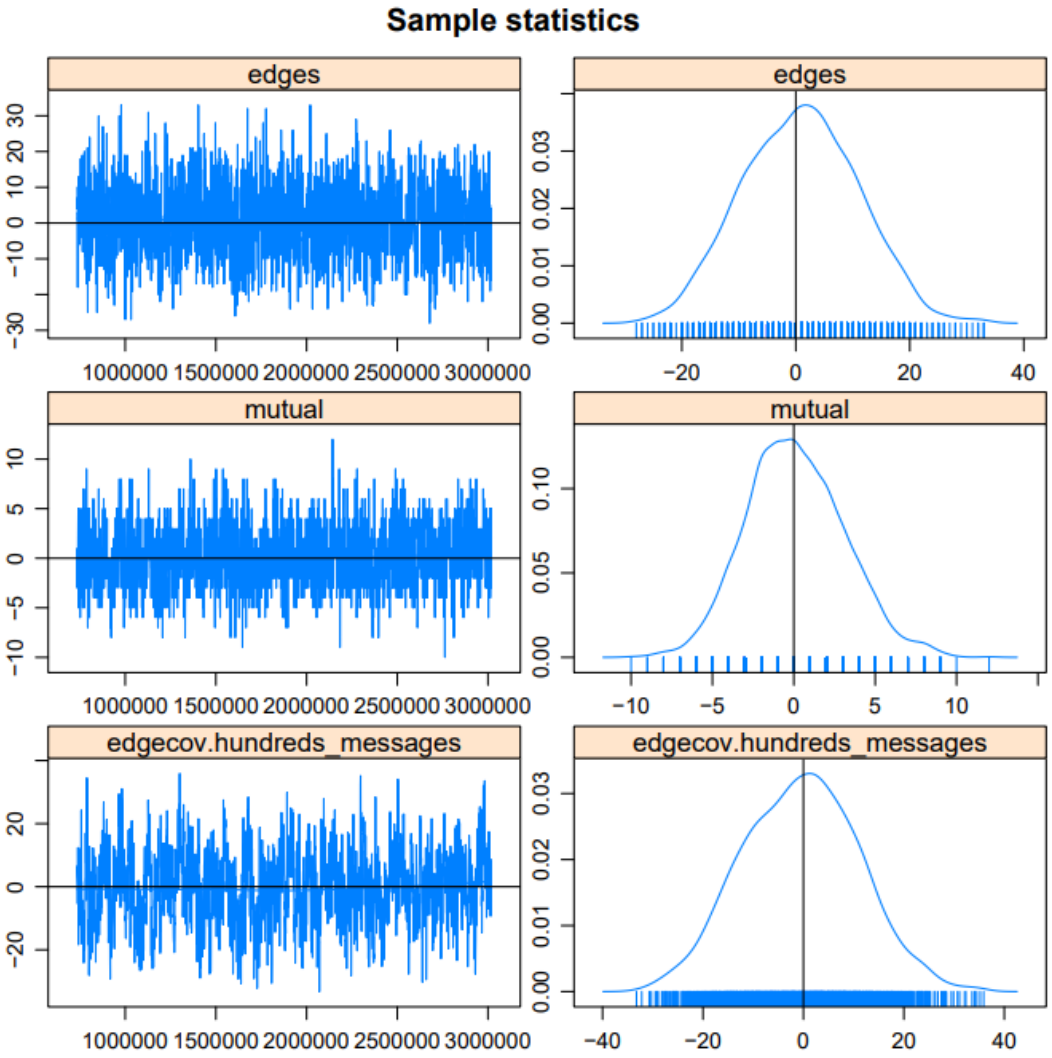
Goodness-of-fit for model statistics

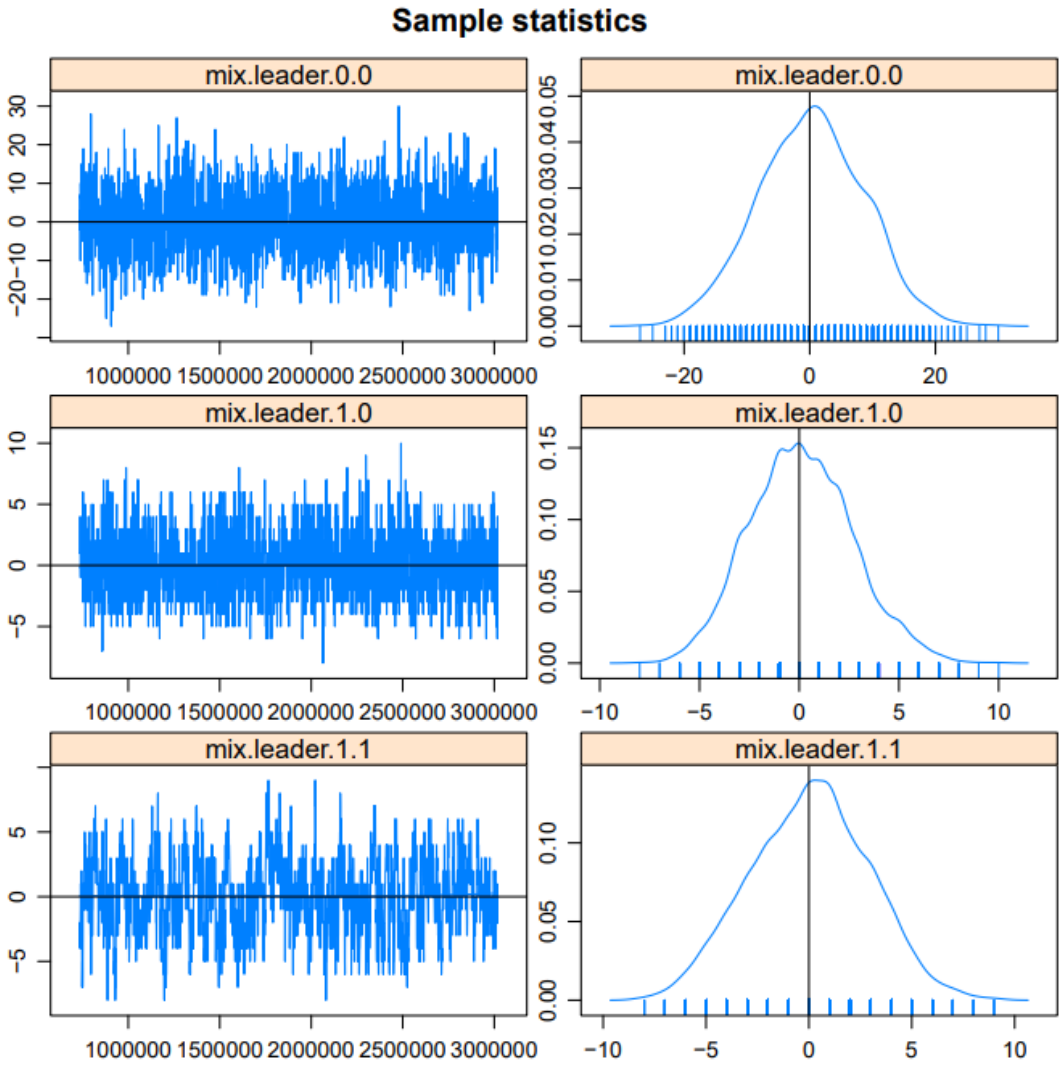
	obs	min	mean	max	MC p-value
mutual	12.0000	4.00000	12.02500	21.00000	1.00
gwideg.fixed.0.693147180559945	54.1507	41.66965	54.17031	66.94322	0.98
gwodeg.fixed.2	182.8840	150.51704	182.86674	209.88846	0.98
gwesp.OTP.fixed.0.693147180559945	214.1250	128.87500	213.57937	288.37500	0.98
nodematch.female	162.0000	127.00000	161.47000	188.00000	1.00
mix.leader.0.0	94.0000	61.00000	96.75000	134.00000	0.81
mix.leader.1.0	8.0000	1.00000	7.91000	15.00000	1.00
mix.leader.1.1	12.0000	4.00000	11.34000	21.00000	0.93
nodematch.department	112.0000	83.00000	114.66500	155.00000	0.94
nodeicov.office	182.0000	148.00000	182.83500	220.00000	0.96
nodeocov.office	173.0000	146.00000	173.29500	207.00000	1.00
diff.t-h.tenure	-737.2575	-981.37808	-729.59323	-538.56986	0.86
edgecov.hundreds_messages	56.6900	31.09000	56.47605	95.00000	0.94

In terms of other statistical properties we have mentioned before, we can see that our simulation model has a high consistency with the actual observed model, with all p-values closer to 1. This means our simulated model 2 can generally reflect all the necessary properties in actual advice network.

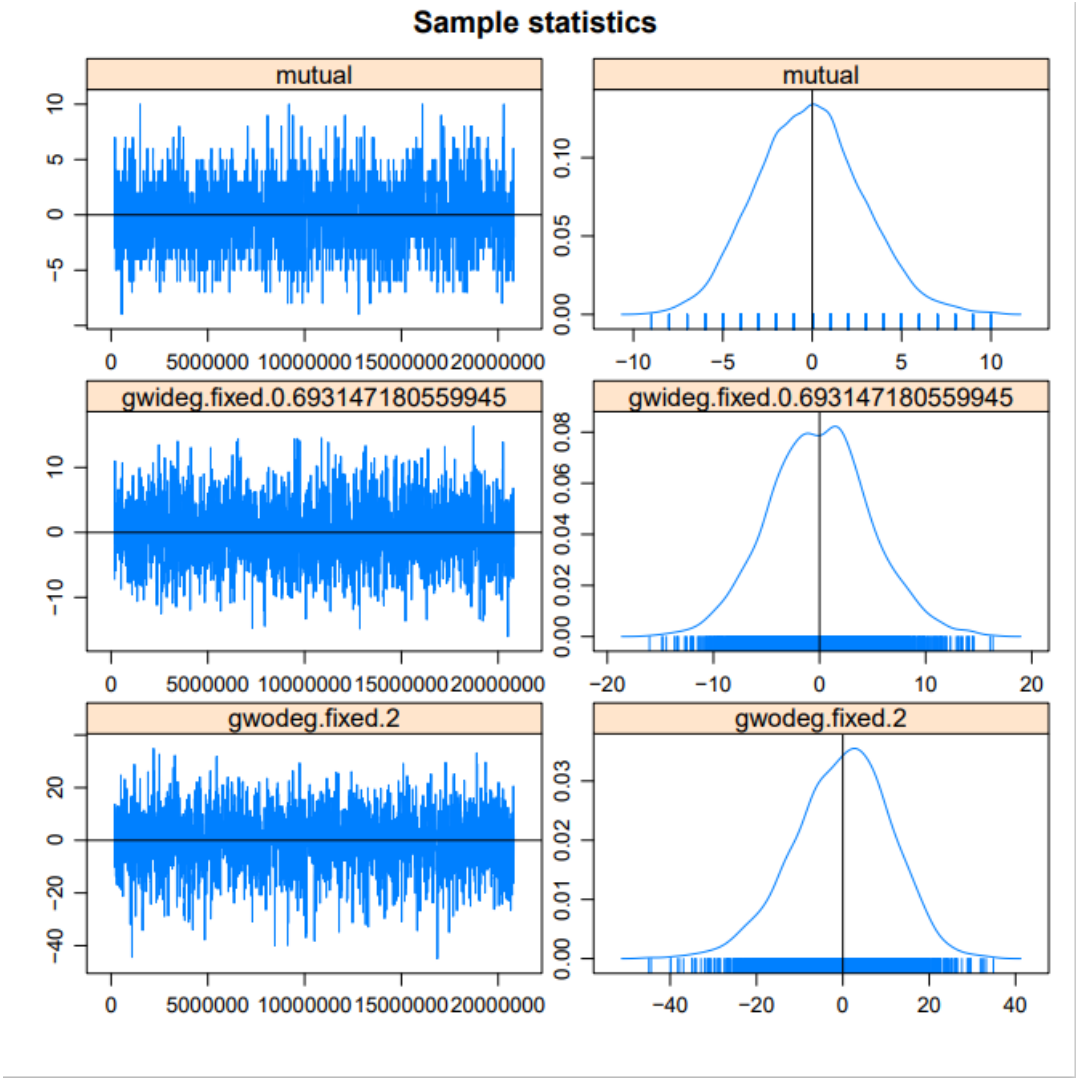
Appendix

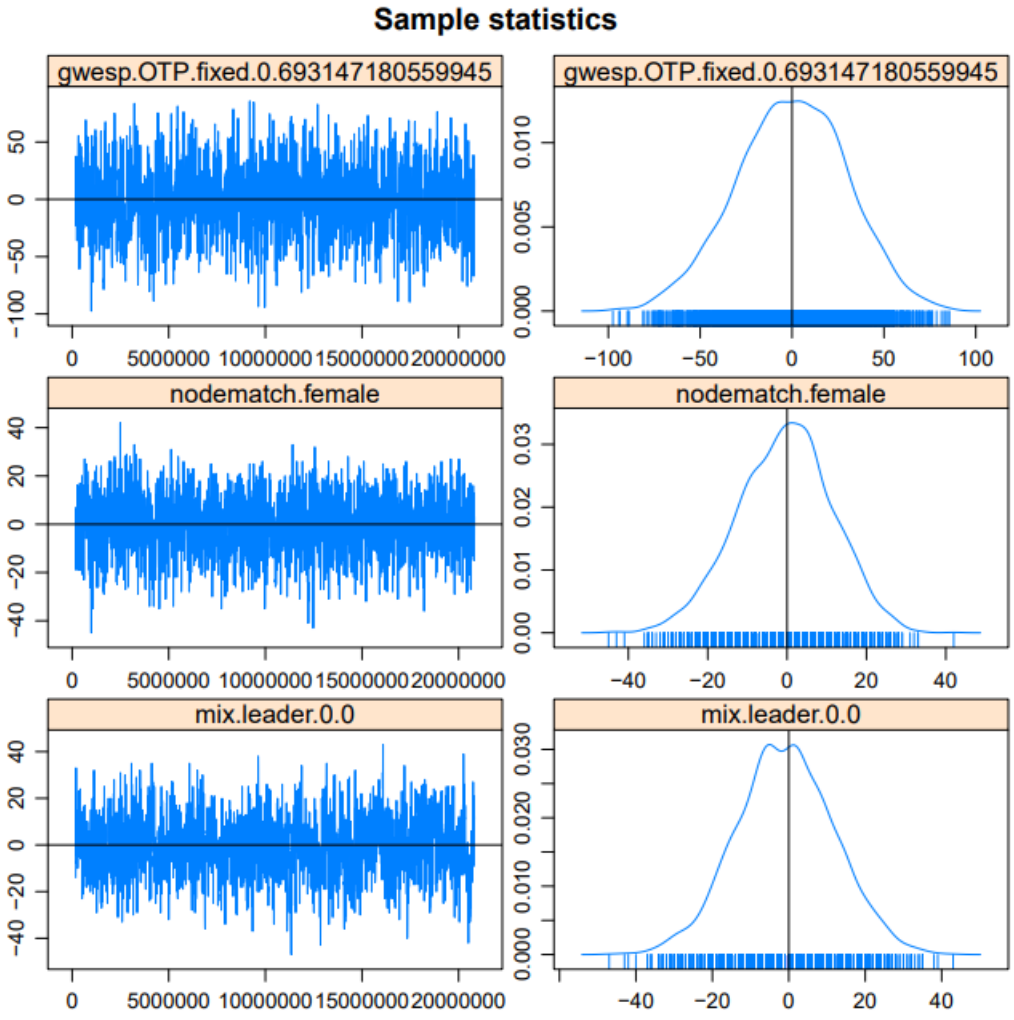
Model 1: Sample Diagnostics

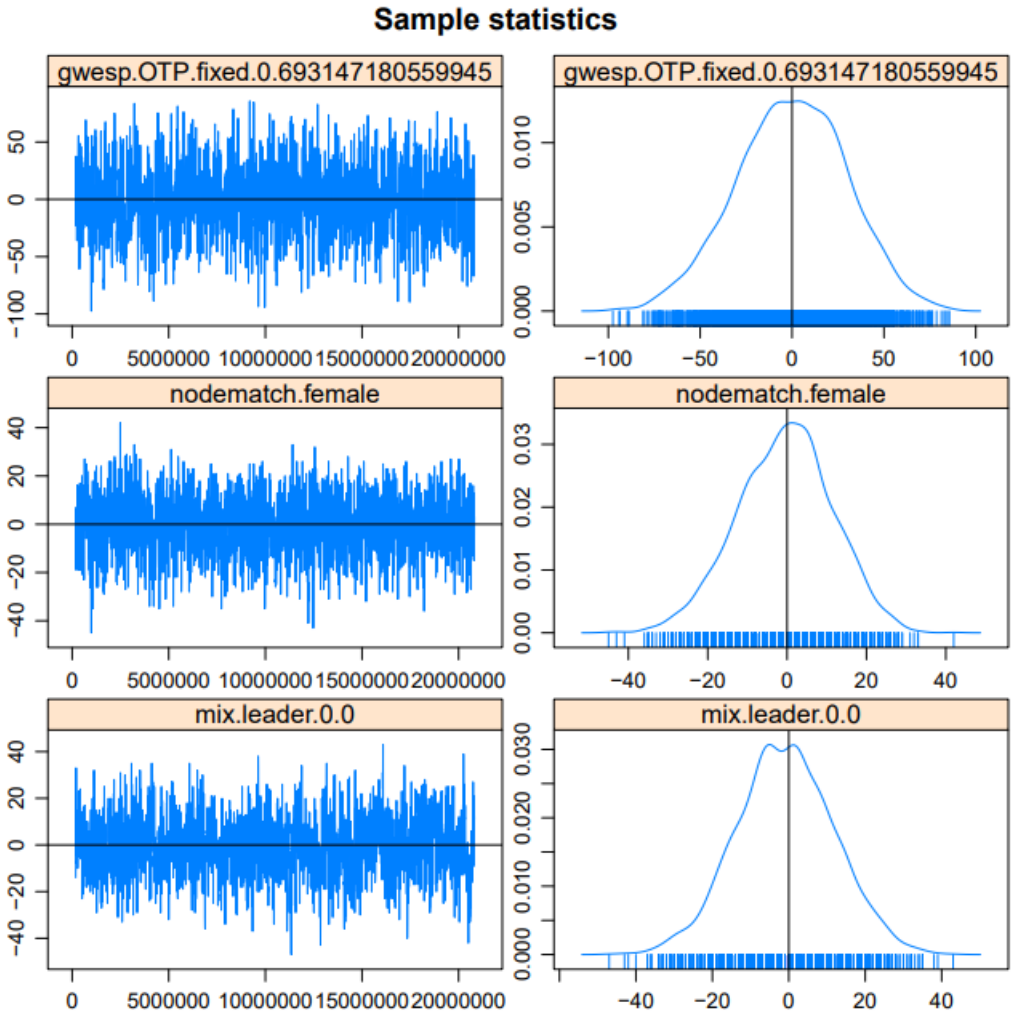


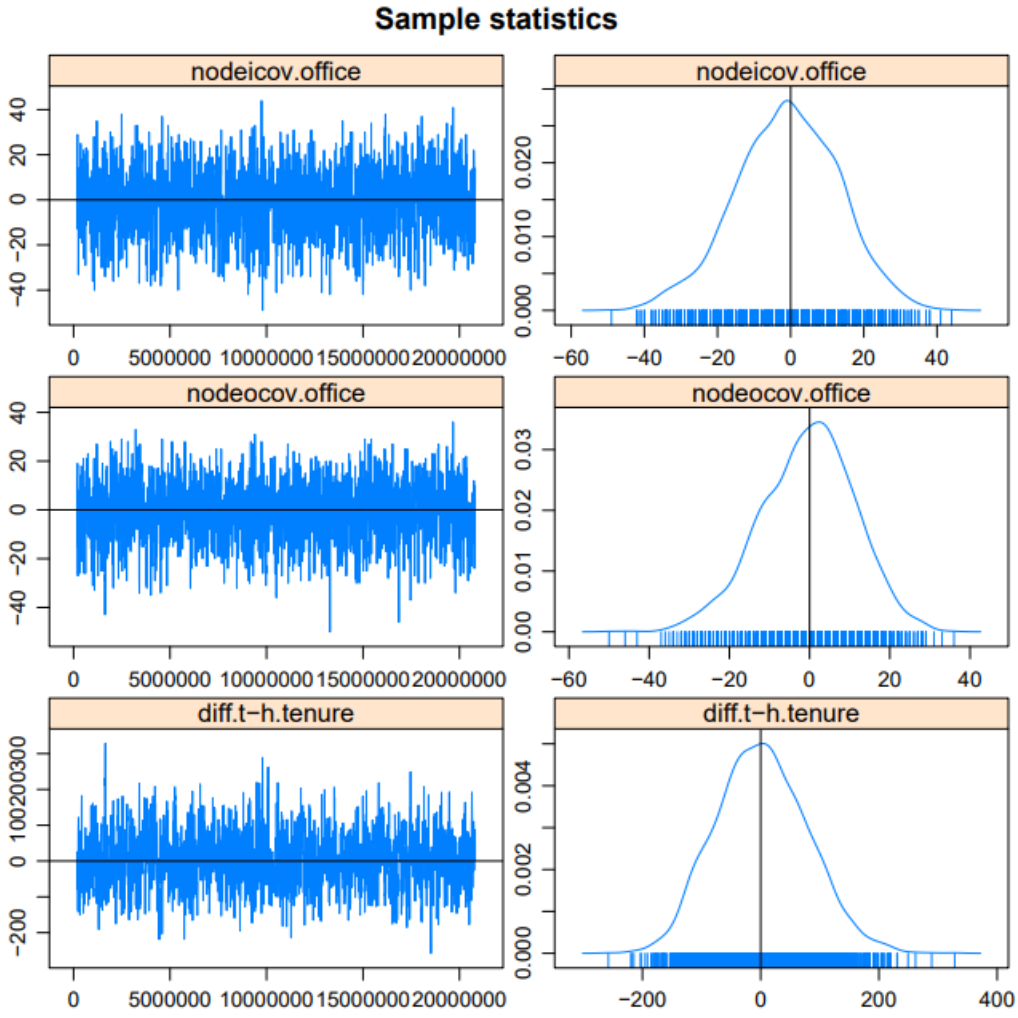

Sample statistics
mix.leader.0.0
mix.leader.0.0
mix.leader.1.0
mix.leader.1.0
mix.leader.1.1
mix.leader.1.1
1000000 1500000 2000000 2500000 3000000

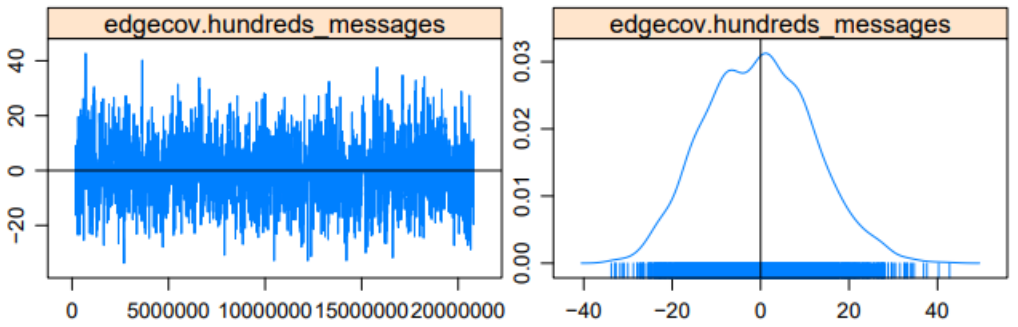
Model 2: Sample Diagnostics




Sample statistics
gwesp.OTP.fixed.0.693147180559945
gwesp.OTP.fixed.0.693147180559945
nodematch.female
nodematch.female
mix.leader.0.0
mix.leader.0.0
0
5000000
10000000
15000000
20000000


Sample statistics
gwesp.OTP.fixed.0.693147180559945
gwesp.OTP.fixed.0.693147180559945
nodematch.female
nodematch.female
mix.leader.0.0
mix.leader.0.0
0
5000000
10000000
15000000
20000000


Sample statistics
nodeicov.office
nodeocov.office
diff.t−h.tenure
0
5000000
10000000
15000000
20000000
−60
−40
−20
0
20
40
−200
200
400


Sample statistics
edgecov.hundreds_messages
0
5000000
10000000
15000000
20000000
−40
−20
0
20
40