Politecnico di Milano

Bayesian inference of network structure from noisy data

Authors
Michele Bellomo
Paulina Moskwa
Noemi Rossi

Supervisor
Prof. Federico Bassetti

February 17, 2021

Introduction

Networks are a mathematical representation that is widely used to model relations between elements in a complex system. A network is characterized by its adjacency matrix $\mathbf{A}$. The adjacency matrix is a squared matrix which elements indicate whether pairs of nodes are connected or not in the graph. If $A_{ij} = 1$ then the nodes i and j are connected, if $A_{ij} = 0$ they are not. All the elements on the diagonal ($A_{ii} \, \forall i$) are equal to zero. If the graph is undirect, the matrix is symmetric. Once we obtain an estimate of the adjacency matrix we obtain the structure of a network. However, unfortunately, most of the network measurements are noisy and unreliable. The goal of our project is to make the best possible estimate of the network structure from the measurements.

The data we are provided with are collected in variables X_{ij}, where each variable represents the number of times we have an observation of the connection of the nodes i and j. For example, we may want to analyze a network of friendships. Every person is represented with a node. The data represents the number of times two people see each other in an arbitrary temporal interval. In this case X_{ij} is the number of times the persons i and j meet each other. We believe that the network structure of friendships affects these measurments. In fact, we believe that X_{ij} will be higher if i and j are friends. In these settings it will not be possible to tell exactly what that structure of the network is. That is the reason why we infer on a probability distribution over possible structures compatible with the observations.

To perform inference we need to make an assumption and some choices. The crucial assumption we need to make is that the observation X_{ij} must depend only on the adjacency matrix element A_{ij}. This means that observations of different node pairs are conditionally independent. The choices we have to make concern the way in which we think the data depend on the structure. More precisely, we have to take three choices.

1. **Data model**
   For a pair of nodes i and j we have to specify the expected distribution of the values X_{ij} for the case in which i and j have a connection and in the case in which they do not. If i and j have a connection we will write: $\mathbb{P}(X_{ij}|A_{ij} = 1, \theta)$. If i and j do not we will write: $\mathbb{P}(X_{ij}|A_{ij} = 0, \theta)$. In both cases θ represents all the other parameters of the distribution we are choosing.
2. **Network model**
   The second choice is the specification of the prior probability associated to the edge connecting i and j. As we have said, we assume that different edges are independent, and so we have to define only the probability of existing of each edge: $\mathbb{P}(A_{ij}|\theta)$. Obviously, the probability for an edge to not exist is given by $1 - \mathbb{P}(A_{ij}|\theta)$.
3. **Priors**
   The third model choice is the specification of the prior distribution on the parameters θ, namely $\mathbb{P}(\theta)$. There are no restriction about this probability.

Applying the Bayes theorem we can write the probabilities of the unknown quantity $\mathbf{A}$ and θ given the data $\mathbf{X}$. We use the formula

$$\mathbb{P}(\mathbf{A}, \theta|\mathbf{X}) = \frac{\mathbb{P}(\mathbf{X}|\mathbf{A}, \theta)\mathbb{P}(\mathbf{A}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\mathbf{X})}$$

to infer the probability distribution among the possible structures given the data.

Given our model choice and the measurements, through stan we compile and run the MCMC and we end up with many pairs $(\mathbf{A}_r, \theta_r)$, $r = 1, 2, 3, ..$ of networks and parameters compatible with the data. Any other network property we are interested in can be determined by inspecting these.

The estimate of the network is straightforward. Having multiple networks $\mathbf{A}_r$ we can do the mean element-wise, obtaining a matrix Q. The values of the matrix Q are in $[0, 1]$, hence Q is not an adjacency matrix. So, we take a threshold of 0.5: every value above 0.5 is replaced into 1 and every value below 0.5 is replaced into 0. This rounding allows us to obtain an adjacency matrix which is extremely more precise than any adjacency matrix $\mathbf{A}_r$. In fact, the matrices $\mathbf{A}_r$ are different among each other. We expect some edges to be (almost) always present. These will be the ones that we will mark as existent. We expect some edges to be present at all times. These will be the ones that will certainly be marked as existing. Instead, there will be some edges present in only a few samples. These are likely to be considered as outliers and hence marked as non-existent when we consider the average.

Other information that we can extract are the parameters contained in θ, namely the parameters that describe the distributions we are choosing. Also here we get a collection of θ's: θ_r. Once again, with the mean of all these θ_r we obtain our final estimates.

The data

The data [2] we decide to analyze regards contacts and friendship relations between students in a high school in Marseilles (France), in December 2013, as measured through several techniques.
We have four data sets that deals with different types of contact/friendship:

- **DATASET 1:** the data set gives the contacts of the students of nine classes during 5 days in Dec. 2013. The file contains a tab-separated list representing the active contacts during 20-second intervals of the data collection. Each line has the form "t i j Ci Cj", where i and j are the anonymous IDs of the persons in contact, Ci and Cj are their classes, and the interval during which this contact was active is [t – 20s, t]. If multiple contacts are active in a given interval, you will see multiple lines starting with the same value of t.
- **DATASET 2:** the data set corresponds to the directed network of contacts between students. Each line has the form "i j w", meaning that student i reported contacts with student j of aggregate duration of w. w takes values from 1 to 4 based one the time two students pass together. The more the value is high, to more time two students pass together.
- **DATASET 3:** the data set corresponds to the directed network of reported friendships. Each line has the form "i j", meaning that student i reported a friendship with student j.
- **DATASET 4:** the data set corresponds to the undirected network of Facebook friendship. Each line has the form "i j w", where w indicates if the friendship on Facebook is present of not.

In addition to these four data sets we have some information about the class and the gender of any student.
Dataset1 needed to be processed before doing analysis. We transformed the Dataset1 in order to have line in the form "i j int", where i and j are the anonymous IDs of the person and int is the number of interactions in the five days. An interaction is a meeting of two people, of any duration, that is at least 70 seconds away with respect to the interaction before. To out analysis we used this new data set.
For reduce the dimension of the problem, that originally has got 329 nodes, we decided to focus only on the Biology classes. In this way we analyze only 110 nodes.

Poisson - Erdős-Rényi Model

Our goal is to understand whether or not the data are conditioned by an underlying structure, which we represent with the adjacency matrix $\mathbf{A}$. We assume that two people meet more if they have a connection. But how can we detect these connections?

Our first approach is a Poisson model. We model the number of interactions as a Poisson random variable with mean λ_0 or λ_1 depending on whether there is or not a network connection. We assume that different edges are independent and, having no prior information, we assume that the probabilities of individual edges are the same. For ρ we assume a uniform prior, while for λ_0 and λ_1 we assume a semi-normal distribution. We can resume the model as:

1. Data model
$$\mathbb{P}(X_{ij}|A_{ij} = k, \lambda_k) = \frac{\lambda_k^{X_{ij}}}{X_{ij}!} e^{-\lambda_k}, \quad k = 0, 1$$

2. Erdős-Rényi Network model
$$\mathbb{P}(A_{ij} = 1|\rho) = \rho$$
$$\mathbb{P}(A_{ij} = 0|\rho) = 1 - \rho$$

3. Priors
$$\rho \sim \mathcal{U}([0, 1])$$
$$\lambda_k \sim Semi - Normal \quad k = 0, 1$$
$$\lambda_1 > \lambda_0$$

We tested the model on a simulation. We created a network of 20 nodes. Each edge exists with a probability $\rho = 0.2$. In the Figure 1 we can see the simulated data.

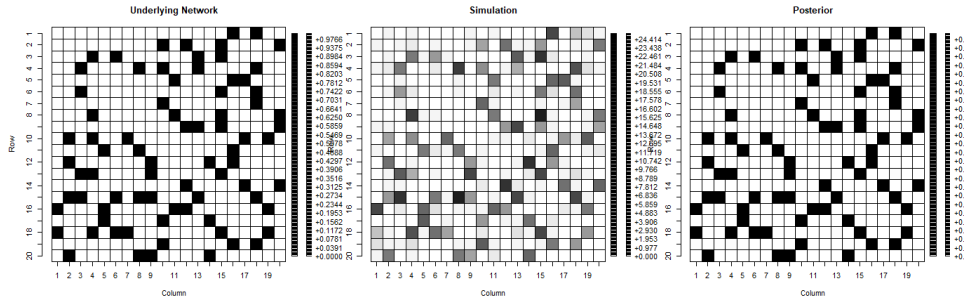


Figure 1: Simulated data for the Poisson model

On the left we see the true adjacency matrix, based on which we generated the central data set: if the edge exists, the data is simulated from a Poisson of $\lambda_1 = 14$, if not, from a Poisson of $\lambda_0 = 0.5$. The number of meetings is represented by an element in the matrix: the higher the number, the darker the element. Using Stan we compiled and ran the MCMC and we obtained the a Posteriori probability of existence of each edge. Taking 0.5 as threshold we obtained the last matrix. The reconstructed network

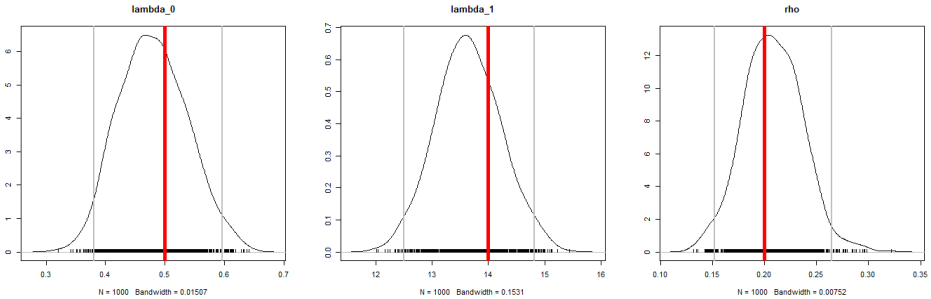


Figure 2: Parameters estimation of the simulated data for the Poisson model

and the original one look very similar. In fact, in the Figure 2 we can see that the true parameters, in red, fall always inside the 95% credible interval.

With the same model we moved on to the real data set. We can see the data in the Figure 3. On the left we see the data matrix representing the collected number of meetings for each pair of people. On the right, following the simulation procedure, we obtained the estimated adjacency matrix.

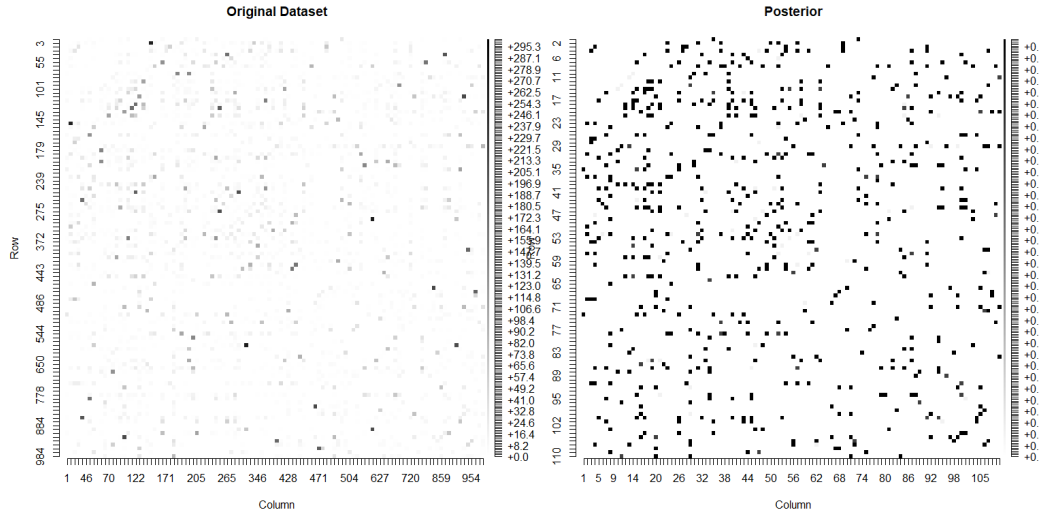


Figure 3: Original data set and estimated network with the Poisson model

We checked the traceplots and the accuracy plots, visible in Figure 4, and we consider them to be good. Moreover, in Figure 5 we can see that there is a substantial difference between the data coming from an existing edge and those coming from a non-existing one. In fact λ_1 is about 42 while λ_0 is about 0.68.

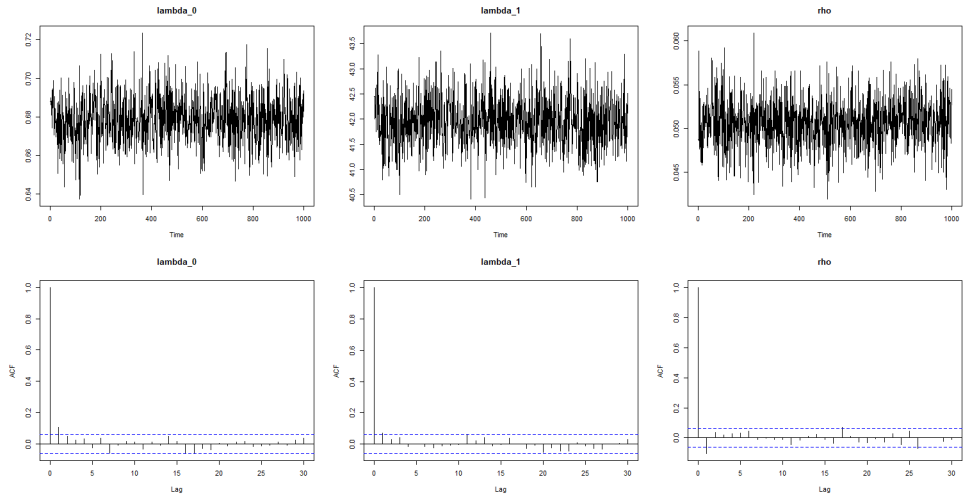


Figure 4: Traceplots and accuracy plots

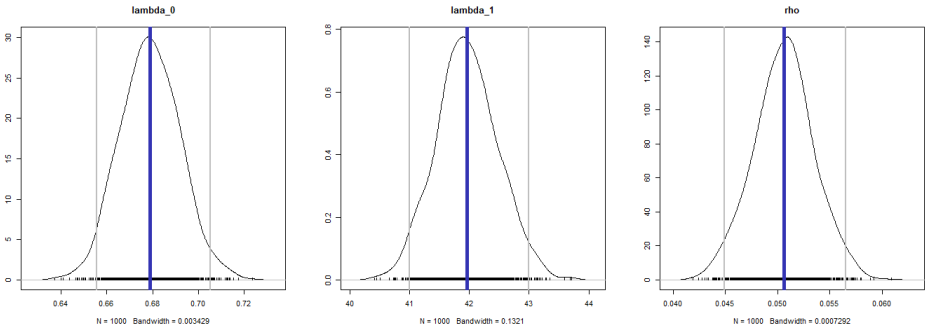


Figure 5: Parameters estimation of the real data in the case of Poisson model

Negative Binomial - Erdős-Rényi Model

As second approach we tried the negative binomial model. Again, the parameters change depending on the presence or absence of the underlying edge. We assume that the edges are independent and that the probability that they exist is equal to ρ. The prior we set for ρ is a uniform distribution, while for ϕ and μ we consider a semi-normal distribution. We can resume the model as:

1. Data model

$$\mathbb{P}(X_{ij}|A_{ij} = k, \mu_k, \phi_k) = \binom{X_{ij} + \phi_k - 1}{X_{ij}} (\frac{\mu_k}{\mu_k + \phi_k})^{X_{ij}} (\frac{\phi_k}{\mu_k + \phi_k})^{\phi_k}, \ k = 0, 1$$

2. Erdős-Rényi Network model

$$\mathbb{P}(A_{ij} = 1|\rho) = \rho$$
$$\mathbb{P}(A_{ij} = 0|\rho) = 1 - \rho$$

3. Priors

$$\rho \ \sim \mathcal{U}([0, 1])$$
$$\phi_k, \mu_k \sim Semi - Normal \quad k = 0, 1$$
$$\phi_1 > \phi_0, \quad \mu_1 > \mu_0$$

Once again, we first ran the model on a simulation. We generated a new network, produced a simulation based on the presence or absence of the edges and we obtained the final network. The results can be seen in Figure 6.

We can be satisfied with the final result. Also here, an analysis shows (in Figure 7)

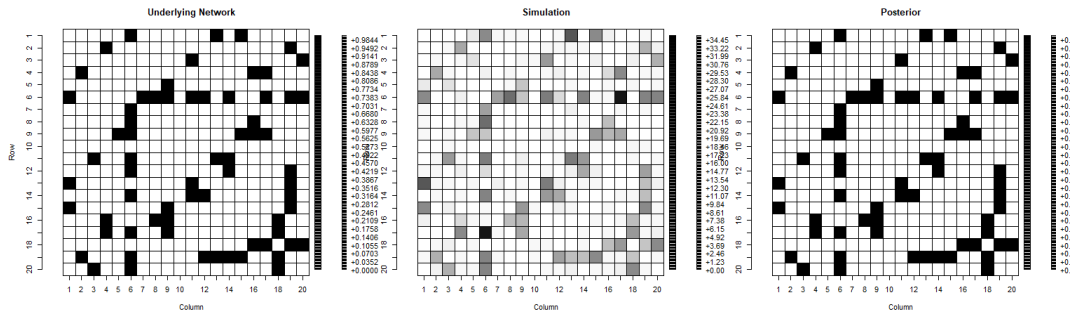


Figure 6: Simulated data for the negative binomial model

that the true parameters always fall inside the 95% credible interval.

We then switched to the real data set, shown in Figure 8. We see the real data set on the left and the obtained network on the right. The analysis of the traceplots and the autocorrelation plots (in Figure 9) is slightly different from the previous one. We see some signs of autocorrelation but overall we are satisfied. In Figure 10 it is shown the plot of the 95% credible intervals and the average of the parameters. Once again, we see that there is a substantial difference depending on the presence or absence of the edge.

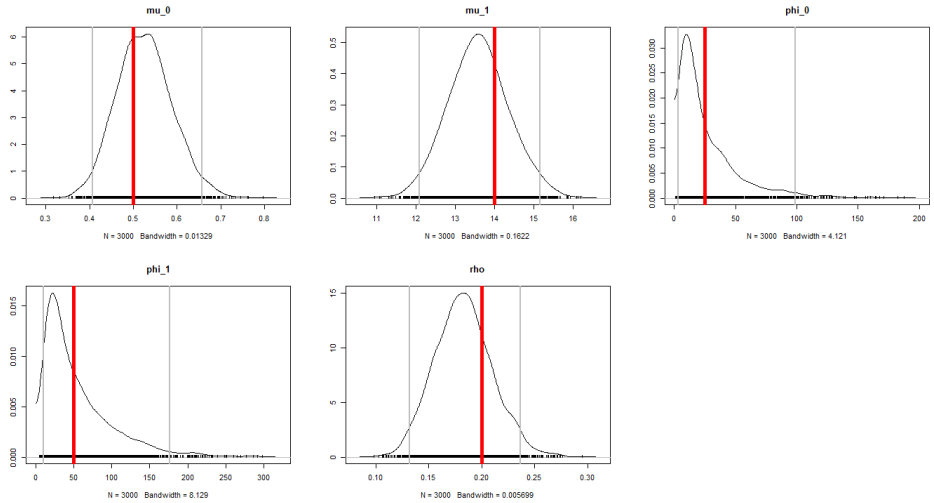


Figure 7: Parameters estimation of the simulated data for the negative binomial model

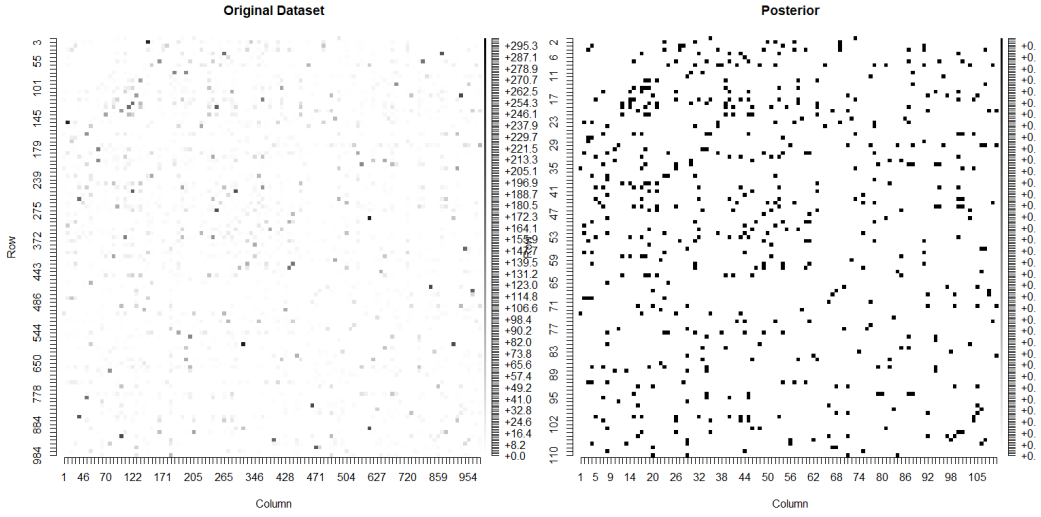


Figure 8: Original data set and estimated network with the negative binomial model

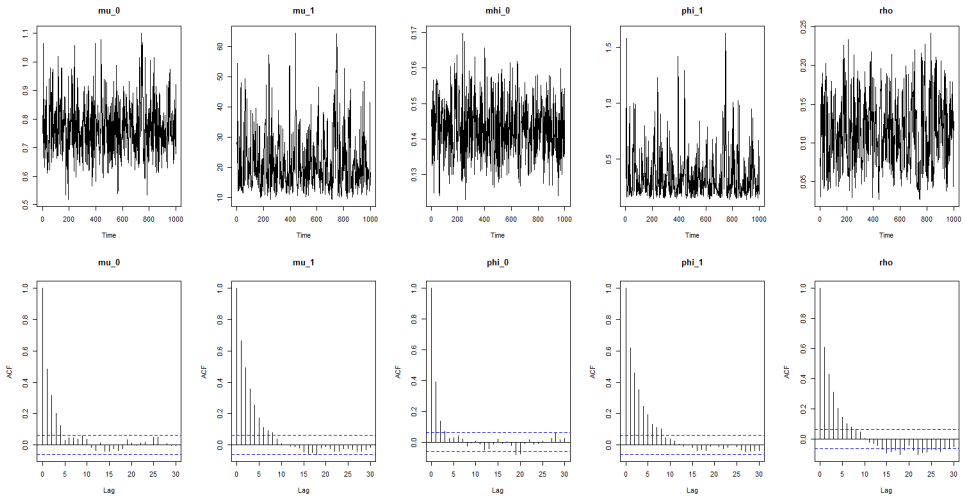


Figure 9: Traceplots and accuracy plots

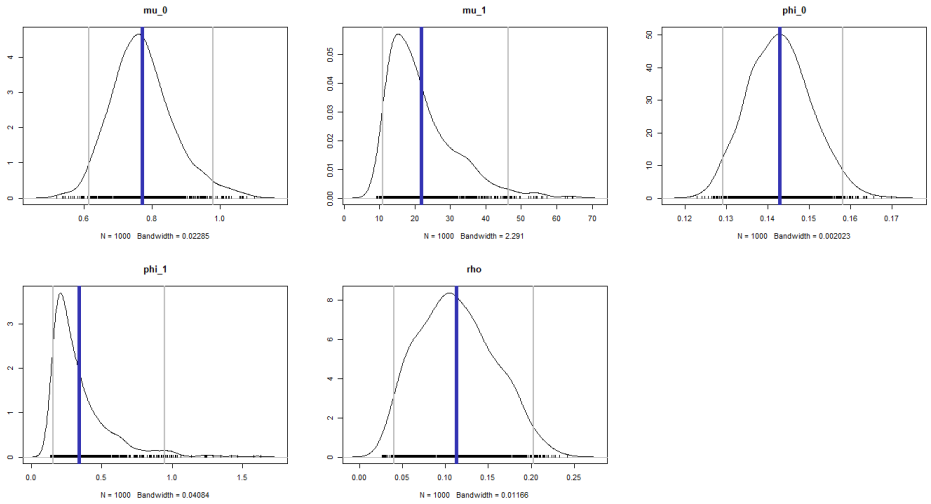


Figure 10: Parameters estimation of the real data in the case of negative binomial model

Negative Binomial - Stochastic Block Model

As a third approach we tried a stochastic block model. In this model we take into account the fact that students belong to different classes. In fact, we suppose that the probability of existence of an edge depends on the group to which the two nodes defining it belong. We assume all the edges are independent. We assume the probability of existence of an edge connecting i and j to be $\omega_{g_i g_j}$ $\forall i, j$. All these ω's are assumed to come from an uniform $\mathcal{U}[0,1]$. We suppose all the remaining parameters to come from a semi-normal distribution. We can resume the model as:

1. Data model

$$\mathbb{P}(X_{ij}|A_{ij}=k,\mu_k,\phi_k) = \binom{X_{ij}+\phi_k-1}{X_{ij}}(\frac{\mu_k}{\mu_k+\phi_k})^{X_{ij}}(\frac{\phi_k}{\mu_k+\phi_k})^{\phi_k},\ k=0,1$$

2. Stochastic block model

$$\mathbb{P}(A_{ij}=1|\omega_{g_i g_j}) = \omega_{g_i g_j}$$
$$\mathbb{P}(A_{ij}=0|\omega_{g_i g_j}) = 1-\omega_{g_i g_j}$$

3. Priors

$$\omega_{g_i g_j} \sim \mathcal{U}([0,1])\ \ \forall i,j$$
$$\phi_k, \mu_k \sim Semi-Normal \quad k=0,1$$

As previously, we started with a simulation. We simulated data from the parameters: $\mu_0 = 5$, $\mu_1 = 30$, $\phi_0 = 0.1$, $\phi_1 = 0.9$. The matrix we selected for the weights is:

```
     [,1] [,2] [,3]
[1,]  0.8  0.3  0.1
[2,]  0.3  0.7  0.2
[3,]  0.1  0.2  0.8
```

The underlying network, the simulation and the network provided by the posterior can be seen in Figure 11.

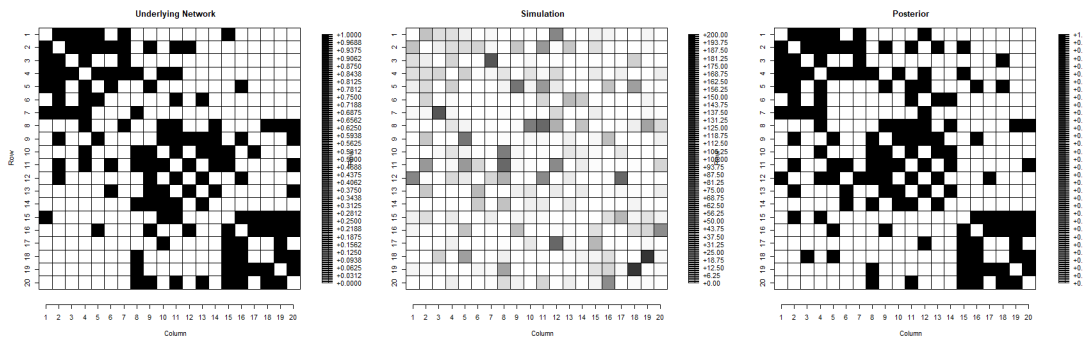


Figure 11: Simulated data for the negative binomial - stochastic block model

As shown in Figure 12, all the estimated confidence intervals of the parameters contain the true parameters (in red). This is a good start for the work on the real data.

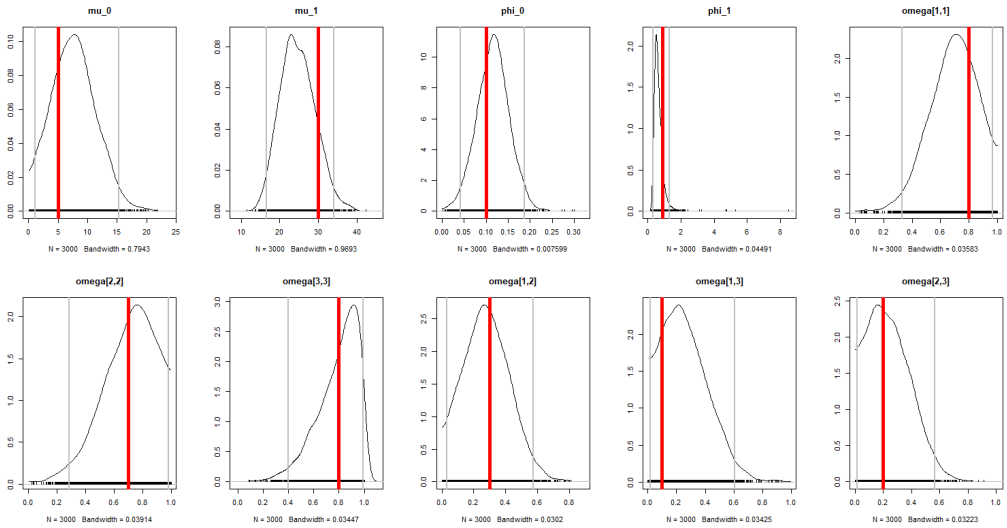


Figure 12: Parameters estimation of the simulated data for the negative binomial model - stochastic block model

We then switched to the real dataset, shown in Figure 13. We see the real dataset on the left and the obtained network on the right. The analysis of the traceplots and the autocorrelation plots (in Figure 14 and Figure 15, respectively) are very good. In Figure 16 it is shown the plot of the 95% credible intervals and the average of the parameters.

We can see that in this model intra-group edges have a probability of existence almost equal to 1. On the other hand, the probability of existence of edges that have nodes belonging to different groups is very low (close to 0).

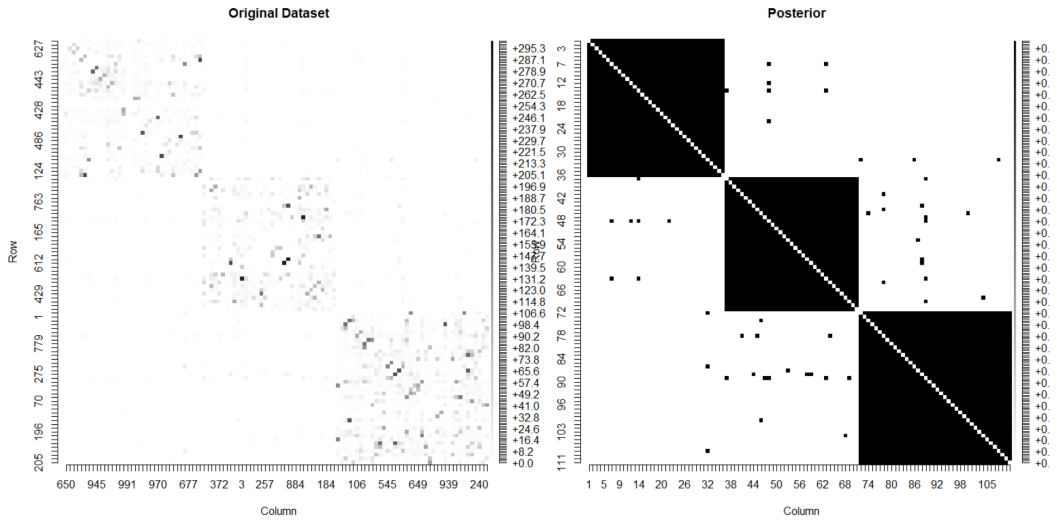


Figure 13: Original dataset and estimated network with the negative binomial model - stochastic block model

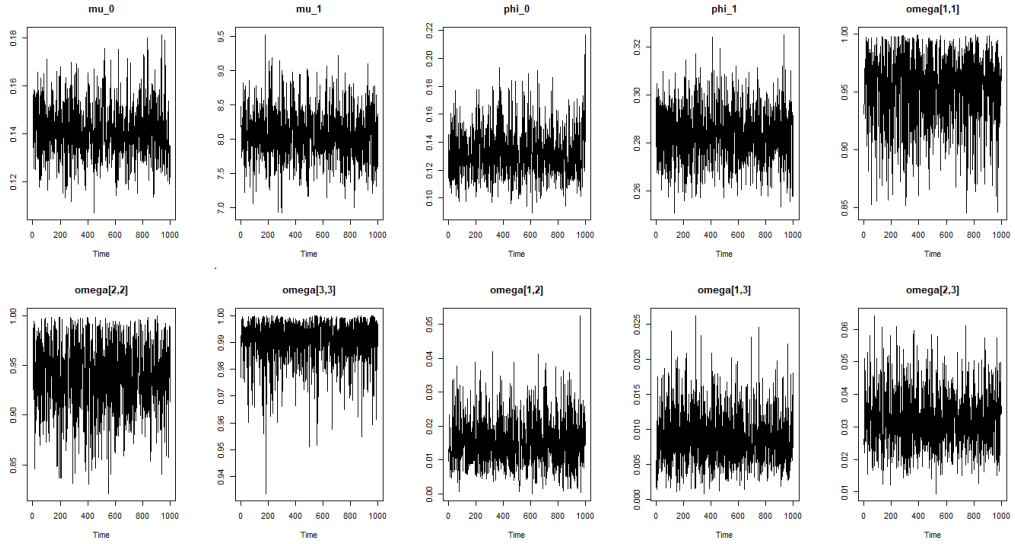


Figure 14: Traceplots

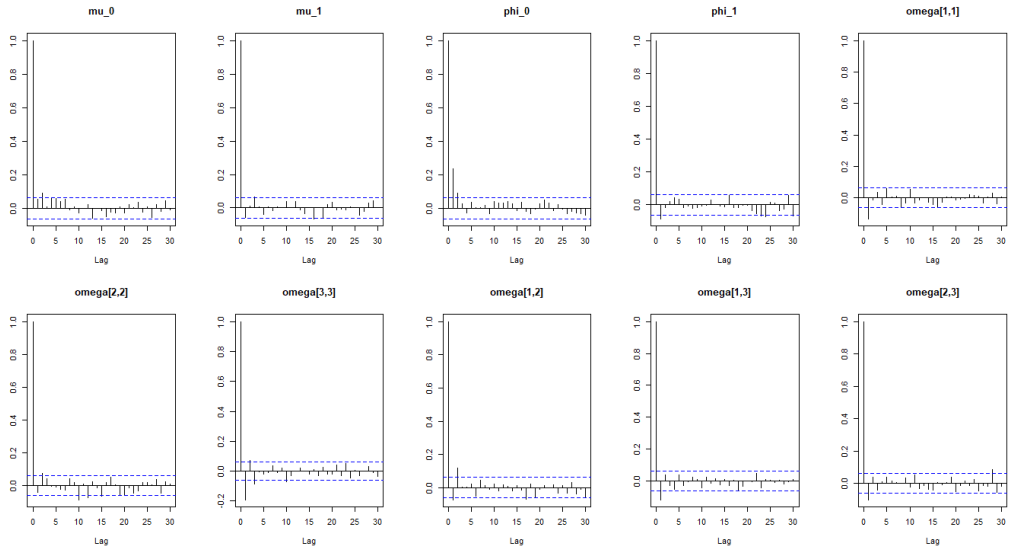


Figure 15: Accuracy plots

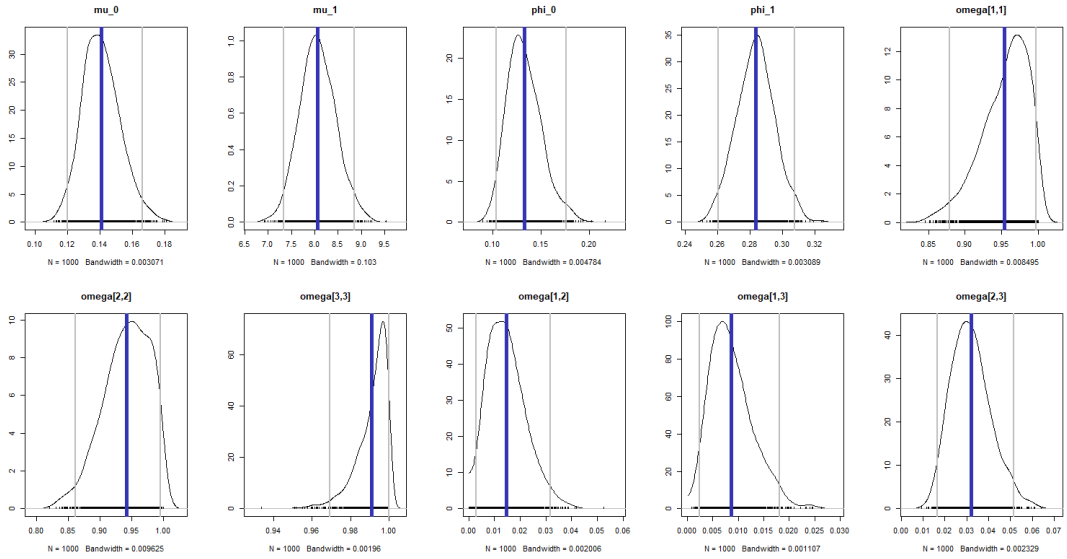


Figure 16: Parameters estimation of the real data in the case of negative binomial model - stochastic block model

Poisson - Multiedge Model

As a last approach we tried a multi-edge model. In this model we break the initial assumption. We no longer think that an edge may or may not exist, instead we believe that between two nodes there may be a connection of a different kind. In this model we are assuming that there are stronger friendships than others. We model this assumption by assuming that two nodes (students) are connected by a "strong" friendship with probability ρ_2, by a "weak" friendship with probability ρ_1, or unconnected with probability $1 - \rho_2 - \rho_1$. We assume that both ρ_1 and ρ_2 come from a uniform $\mathcal{U}[0, 1]$ and we impose that their sum is less or equal than one. We assume that, depending on a "strong" or "weak" connection or no connection, the number of meetings comes from a Poisson of variable parameter. We can resume the model as:

1. Data model
$$\mathbb{P}(X_{ij}|A_{ij} = k, \lambda_k) = \frac{\lambda_k^{X_{ij}}}{X_{ij}!} e^{-\lambda_k}, \quad k = 0, 1, 2$$

2. Multi-edge model
$$\mathbb{P}(A_{ij} = 2|\rho_2, \rho_1) = \rho_2$$
$$\mathbb{P}(A_{ij} = 1|\rho_2, \rho_1) = \rho_1$$
$$\mathbb{P}(A_{ij} = 0|\rho_2, \rho_1) = 1 - \rho_2 - \rho_1 := \rho_0$$

3. Priors
$$\rho_1, \rho_2 \sim \mathcal{U}([0, 1])$$
$$\rho_1 + \rho_2 \leq 1$$
$$\lambda_k \sim Semi - Normal \quad k = 0, 1, 2$$

We started with a simulation. We simulated data from the parameters: $\lambda_2 = 100$, $\lambda_1 = 20$, $\lambda_0 = 1$ and $\rho_2 = 0.1$, $\rho_1 = 0.3$, $\rho_0 = 1 - \rho_1 - \rho_2$. The underlying network, the simulation and the network provided by the posterior can be seen in Figure 17.

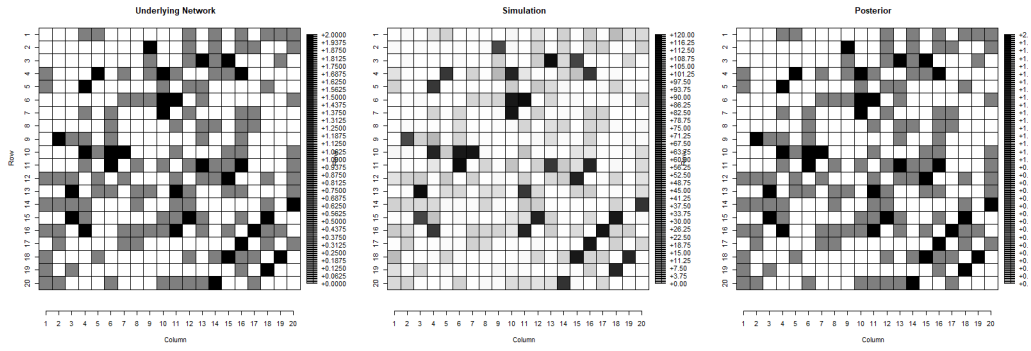


Figure 17: Simulated data for the poisson model - multiedge

As shown in Figure 18, all the estimated confidence intervals of the parameters contain the true parameters (in red).

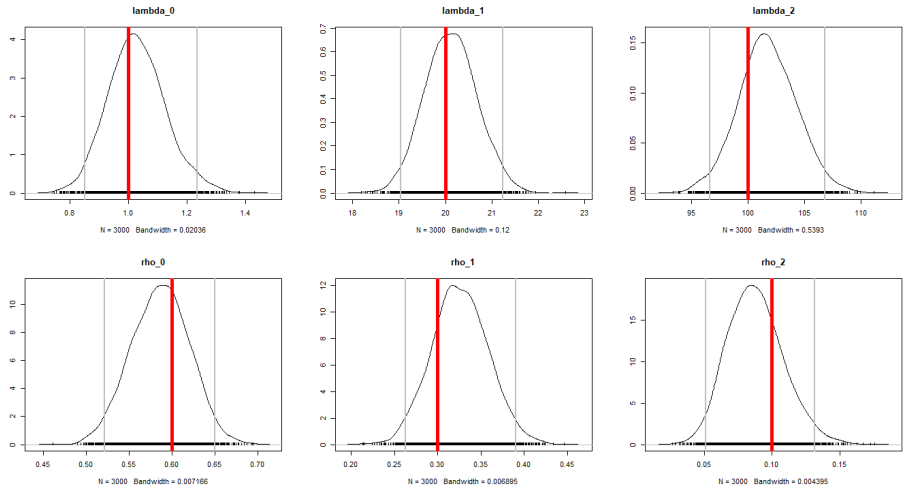


Figure 18: Parameters estimation of the simulated data for the poisson model - multiedge

We then switched to the real dataset, shown in Figure 19. We see the real dataset on the left and the obtained network on the right.

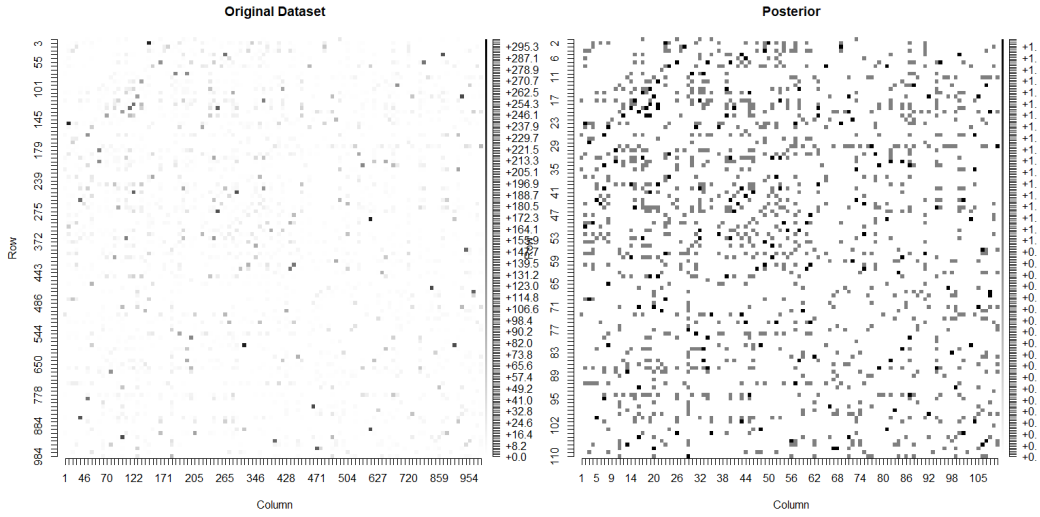


Figure 19: Original dataset and estimated network with the poisson model - multiedge

The analysis of the traceplots and the autocorrelation plots (in Figure 20 and Figure 21, respectively) are very good. In Figure 22 it is shown the plot of the 95% credible intervals and the average of the parameters.

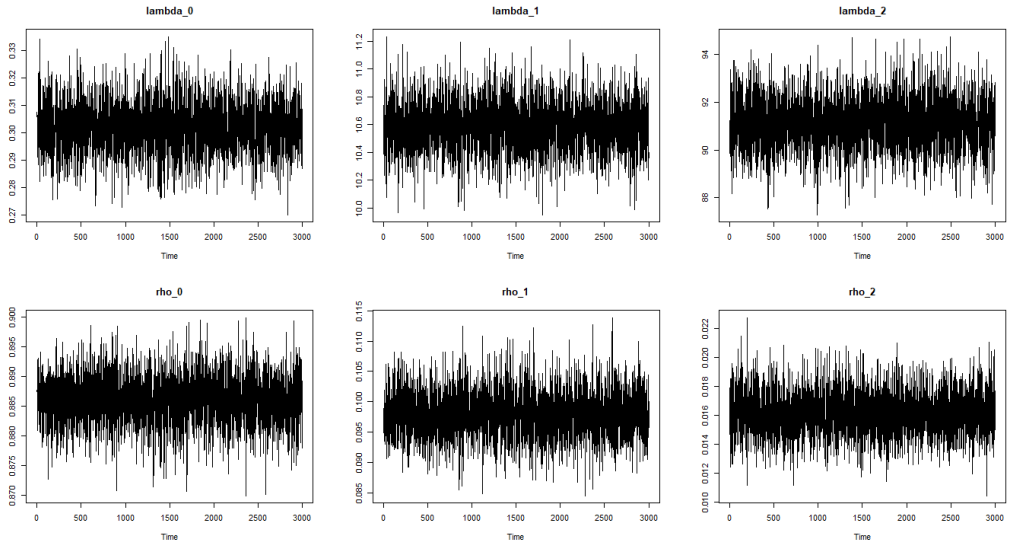


Figure 20: Traceplots

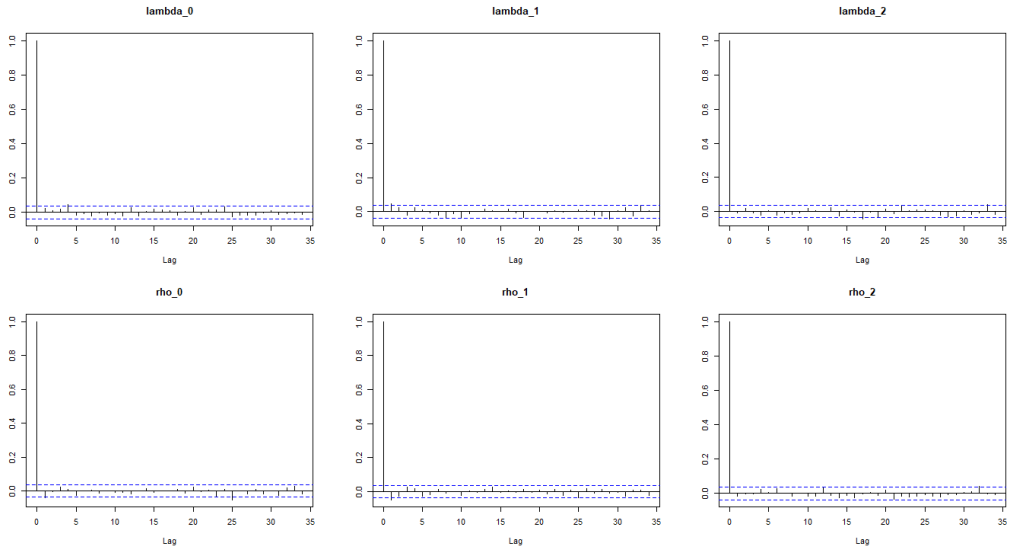


Figure 21: Accuracy plots

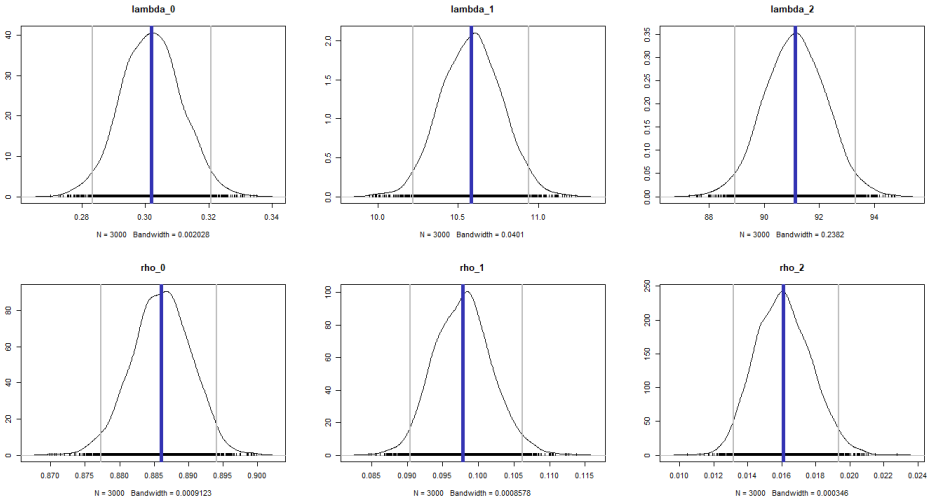


Figure 22: Parameters estimation of the real data in the case of Poisson model - multi-edge

Model comparison

In order to have a different view of the results, we plotted the networks obtained from the models. Shown in Figure 23, we represent each student with a point. Colors are given by the fact that our community is divided into three classes.

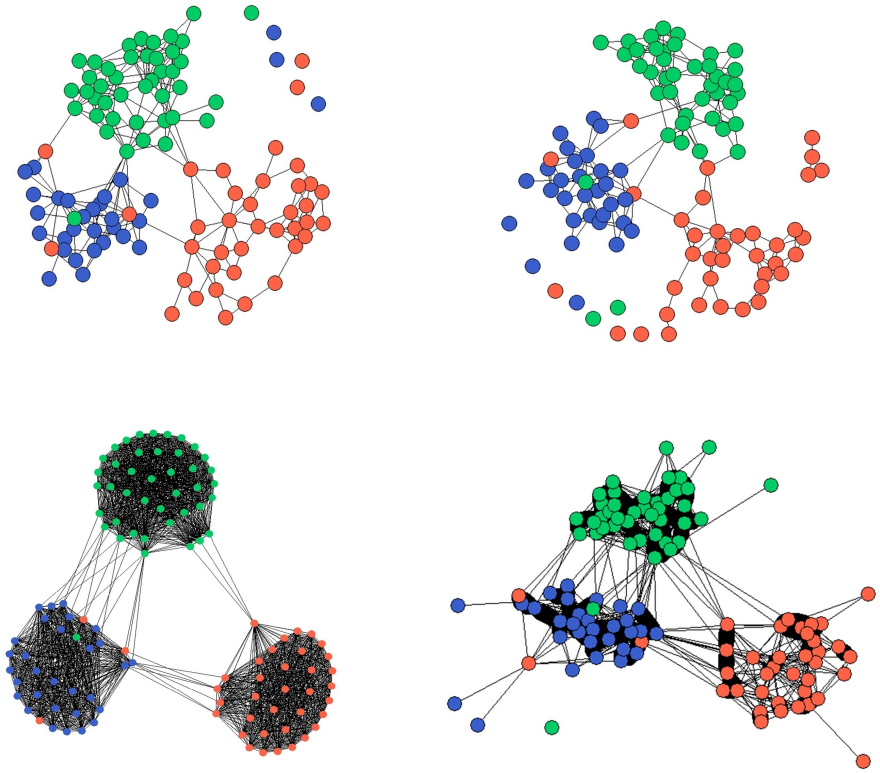


Figure 23: Estimated networks comparison: Poisson model ER (top-left), negative binomial model ER (top-right), negative binomial model - stochastic block model (bottom-left), Poisson model - multiedge (bottom-right)

Model Assessment

After building all these models we want to understand if they fit the data or not. To assess the fitness of any single model we performed Posterior Predictive Assessment [1]. The idea of Posterior Predictive Assessment is to compare the original observations against replicated observations, namely data that would appear if the experiment that produced the original data today were replicated tomorrow with the same model and the same unknown value of θ that produced original data.
These replicated observations are taken from the posterior predictive distribution, so the distribution of the replicated data jointly to θ is

$$\mathbb{P}(\mathbf{X}^{new}, \theta|\mathbf{X}) = \sum_{A} \mathbb{P}(\mathbf{X}^{new}|\theta, \mathbf{A})\mathbb{P}(\mathbf{A}|\theta, \mathbf{X})\mathbb{P}(\theta|\mathbf{X})$$

To quantify the distance between original data and the replicated ones we use the log-likelihood discrepancy, defined as follow:

$$D(\mathbf{X}, \theta) = \sum_{(i,j)} X_{ij} \log \frac{X_{ij}}{X^*_{ij}(\theta)}$$

$$D(\mathbf{X}^{new}, \theta) = \sum_{(i,j)} X^{new}_{ij} \log \frac{X^{new}_{ij}}{X^*_{ij}(\theta)}$$

$$\text{where} \quad X^*_{ij}(\theta) = \mathbb{E}[X^{new}_{ij}|\mathbf{X}, \theta] = \sum_{k} \mathbb{E}[X^{new}_{ij}|\theta, A_{ij} = k]\mathbb{P}(A_{ij} = k|\theta, \mathbf{X})$$

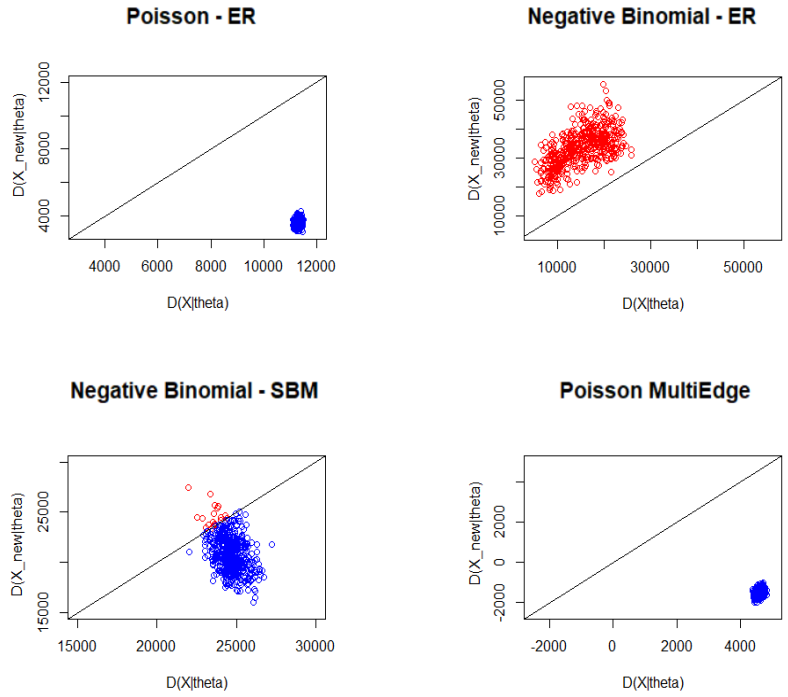
To quantify the difference between original and replicated data we compute the Posterior Predictive p-value p.

$$p = \mathbb{P}[D(\mathbf{X}^{new}, \theta) \geq D(\mathbf{X}, \theta)|\mathbf{X}]$$

In practice to compare the original data against replicated data and to compute the p-value we do the following:
Given a set of draws $\theta^{(j)}, j = 1, ..., J$ for each j:

1. Given $\theta^{(j)}$, draw the new data $\mathbf{X}^{new(j)}$ from the sampling distribution $\mathbb{P}(\mathbf{X}^{new(j)}|\theta^{(j)})$
2. Calculate $D(\mathbf{X}, \theta^{(j)})$ and $D(\mathbf{X}^{new(j)}, \theta^{(j)})$

In this way we obtain pairs $\{(D(\mathbf{X}, \theta^{(j)}), D(\mathbf{X}^{new(j)}, \theta^{(j)})), j = 1, ..., J\}$ to make a scatter plot for graphical assessment. The posterior predictive p-value is the percentage of pairs in which $D(\mathbf{X}^{new(j)}, \theta^{(j)})$ exceeds $D(\mathbf{X}, \theta^{(j)})$.



Scatterplot of the Discrepancy of the four models

References

[1] Andrew Gelman, Xiao-Li Meng, and Hal Stern. "Posterior predictive assessment of model fitness via realized discrepancies". In: *Statistica sinica* (1996), pp. 733–760.

[2] *High school contact and friendship networks*. URL: `http://www.sociopatterns.org/datasets/high-school-contact-and-friendship-networks/`.

[3] Mark EJ Newman. "Network structure from rich but noisy data". In: *Nature Physics* 14.6 (2018), pp. 542–545.

[4] Jean-Gabriel Young, George T Cantwell, and MEJ Newman. "Robust Bayesian inference of network structure from unreliable data". In: *arXiv preprint arXiv:2008.03334* (2020).