Modelling and analysis of social network data from rank preferences

Pierre OSSELIN

St Peter's College

University of Oxford

A thesis submitted for the degree of

MSc in Mathematical Sciences

Trinity 2018

Chapter 1

Introduction

This paper aims at performing inference on social network through partial ranking of the individuals

Chapter 2

Community representation

2.1 Mathematic Formulation

Here we suppose that the strength between individuals are characterized by their similitude to certain communities.

We introduce p the number of communities and $(w_{i,k})$ the strength of affiliation of the individual i to the community p.

$\lambda_{i,j} = \sum_{k=1}^{p} w_{i,k} w_{j,k}$

This formulation is equivalent to the rank matrix approximation often encountered in the literature where we assume that:

$(\lambda) = WW^T$ where W is the matrix of the $(w_{i,k})$

We suppose here that the strength of the relations are only determined by the proximity with a certain community hence that the relations are reciprocal.

From the reordering inequality, we can observe that $\lambda_{i,j}$ is maximized when both i and j have the same community ranking preferences.

We can interpret the formula $\lambda_{i,j} = \sum_{k=1}^{p} w_{i,k} w_{j,k}$ also as a scalar product in the p-dimensional space of communities. In this space, each individual is a point represented by a vector $w_i \in (\mathbb{R}^+)^p$ the strength of affiliation is also computed by $\lambda_{i,j} = ||w_i||.||w_j||.cos(\theta)$ we can interpret the norm of the vector as the individual global influence or sociability, and the θ as the divergence in personality. Hence a given individual will appreciate another individual with great sociability but different personality or views as much as a low sociability individual with the same mindset.

In this interpretation, the prior function attributed to the $w_{i,k}$ is equivalent to putting a topography on the space, allowing points to lie more likely in a certain area.

In this 2.1

The main issue regarding this geometric approach is that having a bigger affiliation means that the community points have a greater scalar product, but doesn't mean those points are nearer. If we wanted to use a clustering approach to this problem, we should build a space in which the mapping to this space make it equivalent having great scalar product in the input space and being near in the output space

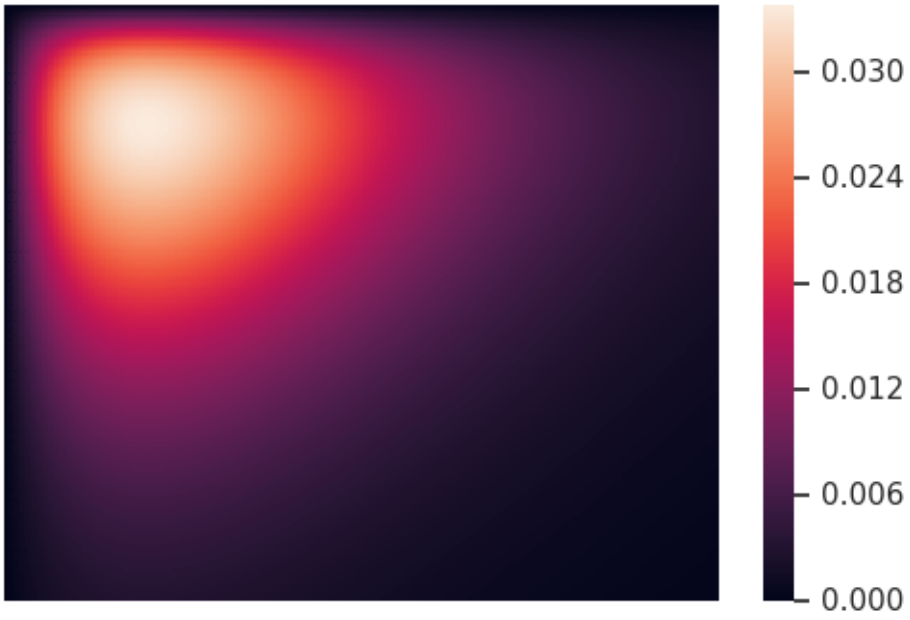

Figure 2.1: Topography of the community space for the same gamma distribution

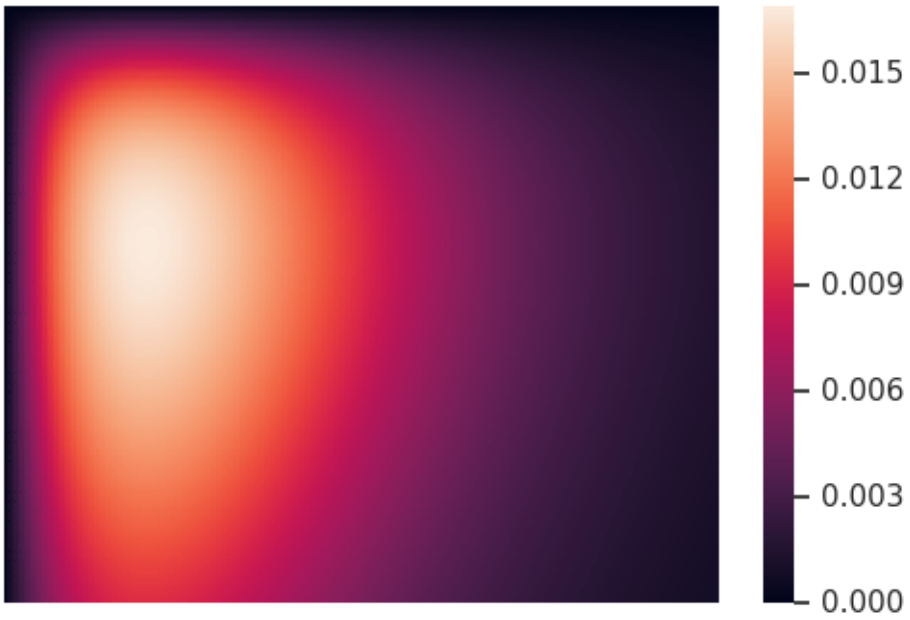

Figure 2.2: Topography of the community space for different gamma distribution

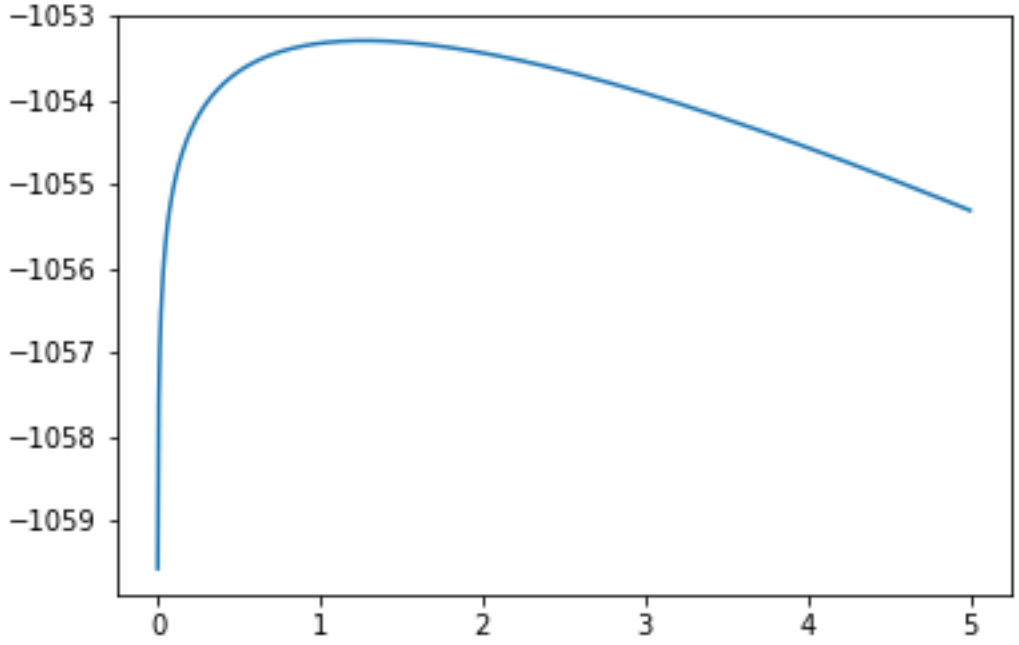

Figure 2.3: Posterior according to one coordinate

2.2 Computation of the posterior

Let's define the log-posterior of the community representation. Which is the sum of the log priors of the parameters and the log-likelihood.

$l((w);D) = \sum_{i=1}^{n}\sum_{k=1}^{p}(\alpha-1)ln(w_{i,k})-\beta w_{i,k}+\sum_{i=1}^{n}\sum_{j=1}^{K} ln(\lambda_{i,\rho_j})-ln(\sum_{j\neq i}\lambda_{i,j}-\sum_{l=1}^{j-1}\lambda_{i,\rho_l})$

To make the optimization easier, we can compute the gradient of the posterior according to the parameter $w_{r,s}$

With $\frac{\partial\lambda_{i,j}}{\partial w_{r,s}} = 1_{i=r}w_{j,s} + 1_{j=r}w_{i,s}$

We have : $\frac{\partial l((w);D)}{\partial w_{r,s}} = \frac{a-1}{w_{r,s}} - b + \sum_{j=1}^{K}\frac{1}{\lambda_{r,\rho_j^{(r)}}}w_{\rho_j^{(r)},s} + \sum_{i=1;i\neq r}^{n}\sum_{j=1}^{K}\frac{1}{\lambda_{i,r}}w_{i,s}1_{\rho_j^{(r)}=r} - \sum_{j=1}^{K}\frac{1}{\sum_{j\neq r}\lambda_{r,j}-\sum_{l=1}^{j-1}\lambda_{r,\rho_l^{(r)}}}\sum_{j\neq r}w_{j,s} - \sum_{i=1;i\neq r}^{n}\sum_{j=1}^{K}\frac{1}{\sum_{j\neq i}\lambda_{i,j}-\sum_{l=1}^{j-1}\lambda_{i,\rho_l^{(i)}}}w_{i,s}$
$+\sum_{j=1}^{K}\frac{1}{\sum_{j\neq r}\lambda_{r,j}-\sum_{l=1}^{j-1}\lambda_{r,\rho_l^{(r)}}}\sum_{l=1}^{j-1}w_{\rho_l^{(r)},s}+\sum_{i=1;i\neq r}^{n}\sum_{j=1}^{K}\frac{1}{\sum_{j\neq i}\lambda_{i,j}-\sum_{l=1}^{j-1}\lambda_{i,\rho_l^{(i)}}}\sum_{l=1}^{j-1}1_{\rho_l^{(i)}=r}w_{i,s}$

2.3 Simulations on simulated Data

2.3.1 Fixed parameters for the prior

We derived two optimization algorithm to find the MAP, one based on coordinate-wise optimization, given that for the a and b considered, the log-posterior seems

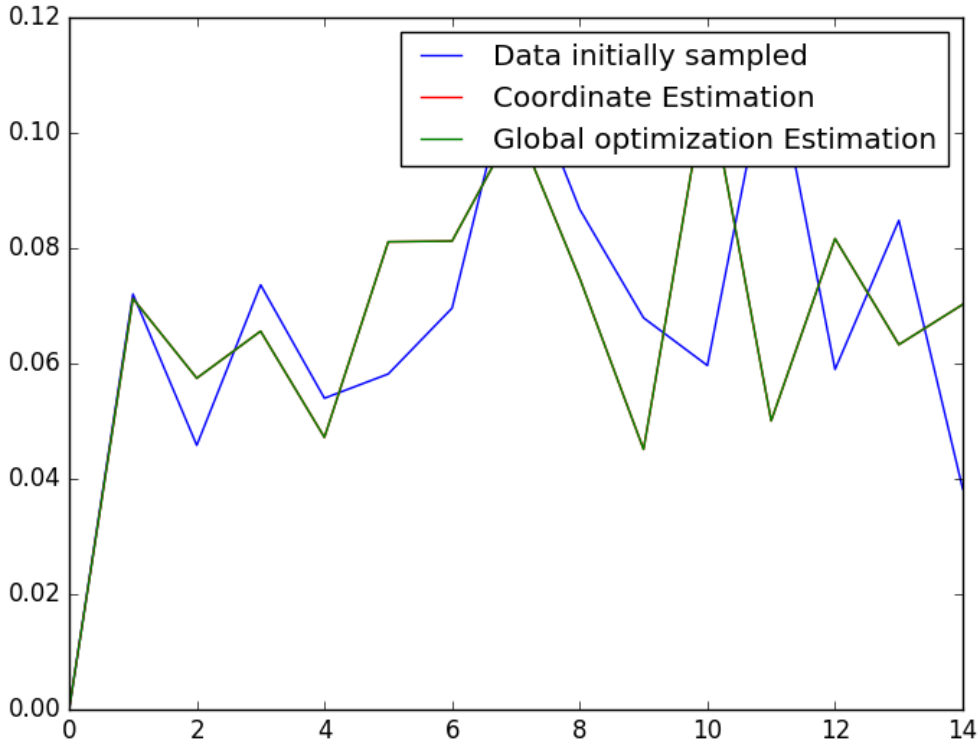

Figure 2.4: Comparison between the sampled and estimated Social affiliation of the first individual

coordinate-wise concave, hence assuring a more stable convergence. The other algorithm is based on the L-BFGS-B algorithm, popular in machine learning optimization.

For this simulation we took $a = 2$, $b = 1$ $n = 15$, $p = 3$, $k = 3$, drew a sample from w, computed λ and ρ. Then we maximized the posterior according to the two methods described above, in order to estimate the community strengths and then the individual mutual strength.

In figure 2.4 we can see that the coordinate and global optimization have converged to the same result. We obtain an absolute error of 0.02. Which is equivalent to 0.3 reported to the number of individual, and a ranking gap (mean absolute ranking error) of 2.13 for this individual.

It is interesting to compare the ranking gap to the expected ranking gap from a random permutation. We can manually compute with a tedious calculation that

$$\frac{1}{n}\mathbb{E}_{\sigma_n}\left[\sum_{k=1}^{n} |\sigma(k) - k|\right] = \frac{n^2-1}{3n}$$

For n = 18 we have a random expectation of the ranking gap of 6. Hence the model tends to preserve the "global" social network for each individual.

We can see that the inferred communities are quite similar to the original one. The original red and green communities are preserved except for the individual 6

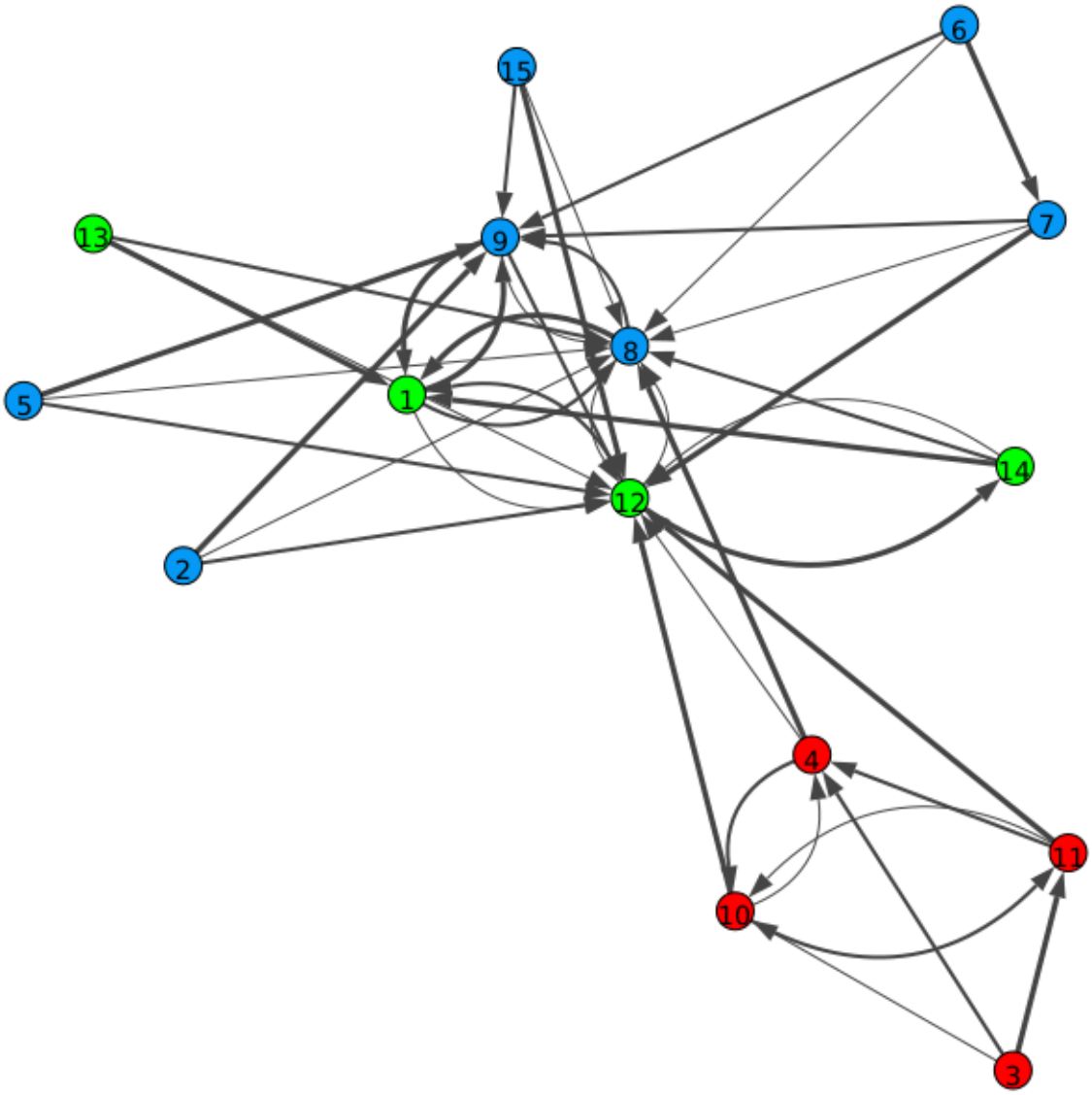

Figure 2.5: Original Community description

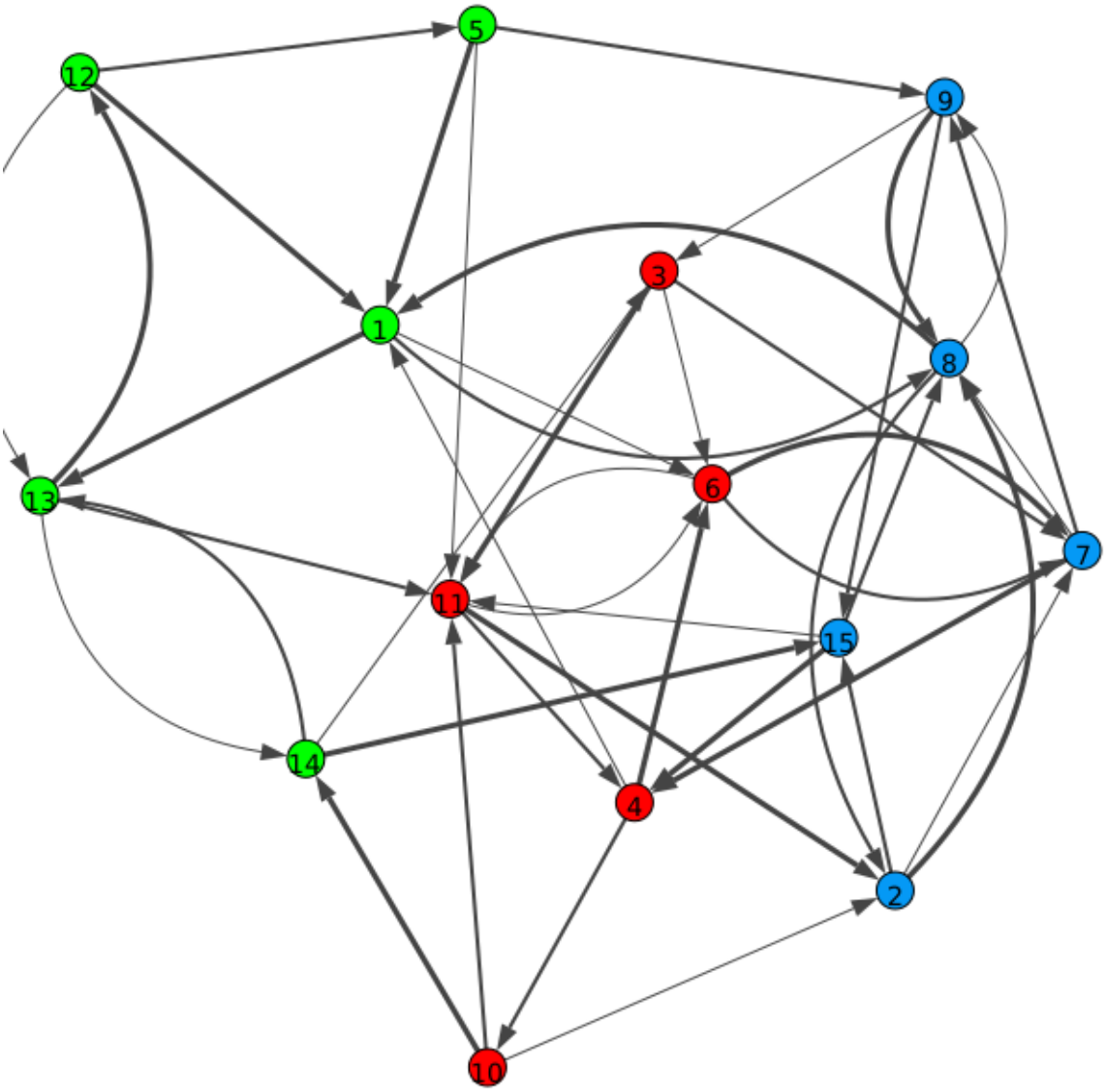

Figure 2.6: Inferred community

and 5 that has passed from the blue community to the red and green respectively. Here we have assigned an individual to his preferred community, the width of an edge represents the importance the son represents for the father.

2.3.2 Model with degree-correction

In this subsection, we try to capture the different popularity that a community can have over the other communities. In the previous model, the regularization terms $w_{i,j}$ in the posterior had the same weights, irrespective of the community concerned (Having a too big $w_{i,j}$ was equally improbable regardless of the community). In a model with degree-correction, i.e in which $w_{i,k} \sim \Gamma(\alpha, \beta_k)$, we do not prevent the possibility of a popular community in the regularization, which occurs by the emergence of popular individuals.

We can see in this example that the green community has been absorbed in an other community by the inference, and that 7 has been replaced by 3 in the other community.

2.3.3 Performance Comparison

To evaluate the performance of the models, we make prediction on the preferences for individuals who have not indicated their preferences, provide further preferences from individuals who did, or extract from the lists interpretable information about the structure of the underlying friendship network, e.g. in terms of latent communities. For this we will use different metrics. The ranking gap between latent appreciation, the absolute error in the strengths and the missed community. To do that, we will sample the strengths between individuals $\lambda_{i,j}$ and then sample full rankings from it. From this ranking we then hide a certain part of it, we feed the remaining parts to our model and then compute the error in terms of metrics previously defined to evaluate the performance.

2.4 Model with Covariates

The purpose of ERGMs, in a nutshell, is to describe parsimoniously the local selection forces that shape the global structure of a network. To this end, a network dataset, like those depicted in Figure 1, may be considered like the response in a regression model, where the predictors are things like propensity for individuals of the same sex to form partnerships or propensity for individuals to form triangles of partnerships. In Figure 1(b), for example, it is evident that the individual nodes appear to cluster in groups of the same numerical labels (which turn out to be students grades, 7 through 12); thus, an ERGM can help us quantify the strength of this intra-group effect. The information gleaned from use of an ERGM may then be used to understand a particular phenomenon or to simulate new random realizations of networks that retain the essential properties of the original. Handcock, Hunter, Butts, Goodreau,

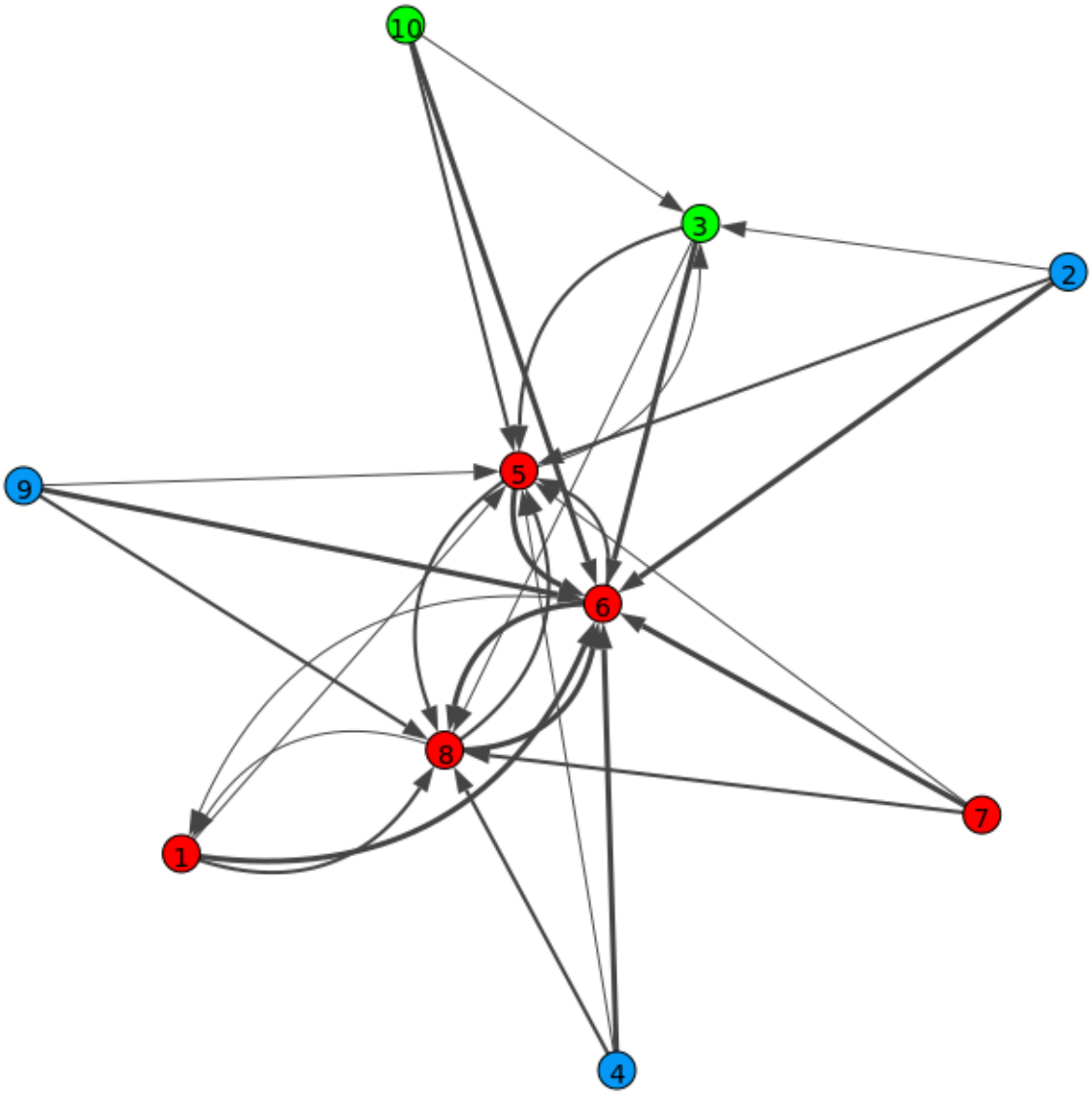

Figure 2.7: Original Community description for the degree Correction Model

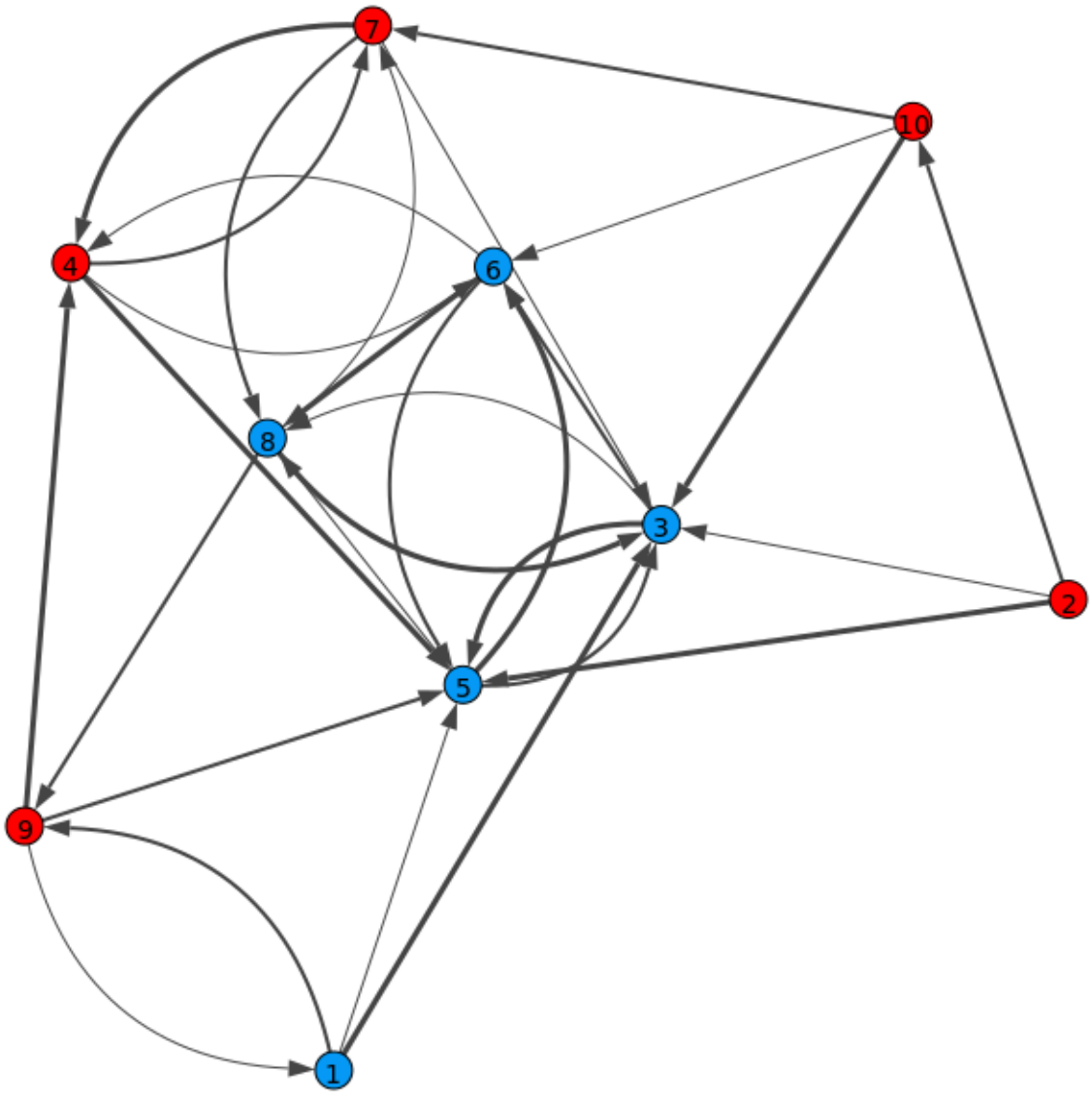

Figure 2.8: Inferred community for the degree correction Model

and Morris (2008) say more about the purpose of modeling with ERGMs; yet in this article, we focus primarily on technical details.

2.4.1 Notation

Let N be the set of nodes in the network of interest, indexed $\{1, ..., n\}$ The relationships in the network may be directed (e.g., friendship nominations, messages) or undirected (e.g., sexual partnerships, conversations). In the former case, we define the set of dyads (here used to refer to potential relationships) $\mathbb{Y}$ to be a subset of $N \times N$, the set of ordered pairs of nodes; in the latter case, it is a subset of $\{\{i, j\} : (i, j) \in N \times N\}$, the unordered pairs of nodes. (We will also use u($\mathbb{Y}$) to refer to an unordered version of $\mathbb{Y}$, i.e., u($\mathbb{Y}$) $\equiv \{\{i, j\} : (i, j) \in Y\}$.) Usually, $\mathbb{Y}$ is further constrained in that in most social networks studied, a node cannot have a relationship of interest with itself, excluding pairs of the form (i, i). For binary networks, in which the relationship of interest must be either present or absent, we use $\mathcal{Y} \in 2^{\mathbb{Y}}$, the set of subsets of $\mathbb{Y}$, to refer to the set of possible networks of interest, which may be further constrained (that is, $\mathbb{Y}$ may be a proper subset of $2^{\mathbb{Y}}$). We will use $\mathbf{Y}$ to refer to network random variables and $y \in \mathcal{Y}$ to refer to their realizations, and $y_{i,j}$ be an 0 - 1 indicator of whether a relationship of interest is present between i and j in a binary network context.

2.4.2 ERGM model

In the ERGM model (exponential-family random graph models), the distribution of $\mathbf{Y}$ can be parameterized in the form:

$P_\theta(Y = y) = \frac{exp((\eta(\theta))^T * z(x))}{\chi(\theta)}, y \in \mathcal{Y}$

where θ is a q-vector of model parameters, which are mapped to a p-vector of natural parameters by $\eta(.)$, and $g(.)$ is a p-vector of sufficient statistics, which capture network features of interest. These sufficient statistics typically embody the features of the network of interest that are believed to be significant to the social process which had produced it, such as degree distribution (e.g., propensity towards monogamy in sexual partnership networks), homophily (i.e., birds of a feather flock together), and triad-closure bias (i.e., a friend of a friend is a friend) . (Morris, Handcock and Hunter, 2008)

its postulated dependence structure, or both.

1) 1) ERGM models 2) Bernouilli model 3) p1 Model 4)

ERGM models As described, models work very poorly Learned parameters do not generate data that resembles the input Tend toward wholly connected or completely empty graphs

P1 Model : Original exponential family graph, with dyadic independence (facilitate the expression of the likelihood) and $P(Y_{i,j} = y_1, Y_{j,i} = y_2) \propto exp\{y_1(\mu + \alpha_i + \beta_j) + y_2(\mu + \alpha_j + \beta_i) + y_1 y_2 \rho\}$

P* Model : Generalization of the P1 model with $P(Y = y) = \frac{exp(\theta^T * z(x))}{\chi(\theta)}$ θ are

model parameters and $z(x)$ is a vector of network statistics. This Model is a generalization that relax the assumption of p1 of independent dyads.

Problem : Model degeneracy had proposed tend to, as the network size increases, concentrate an increasingly large probability mass on an increasingly small fraction of possible networks.

Models based on independence conditional on latent variables: random effects models and mixed effects models and extensions; stochastic block models and extensions, including mixed membership models; and latent space models and extensions.

Assume dyads conditional independent given some latent variables. i.e $P_\theta(Y = y|Z = z) = \prod_{(i,j)\in u(\mathbb{Y})} P_\theta(Y_{i,j} = y_{i,j}, Y_{j,i} = y_{j,i}|Z = z)$ Except the p2 model, we even assume $Y_{i,j}$ and $Y_{j,i}$ independant given Z. Those models can capture network dependencies of interest and are easier to work with in terms of computation.

Paper studied uses p2 model and add in a likelihood that take ranking into account. And compare it to other likelihood that do not and compare and draws conditions in wich ranking provide information.

These sufficient statistics typically embody the features of the network of interest that are believed to be significant to the social process which had produced it, such as degree distribution (e.g., propensity towards monogamy in sexual partnership networks), homophily (i.e., birds of a feather flock together), and triad-closure bias (i.e., a friend of a friend is a friend) . (Morris, Handcock and Hunter, 2008) A major limitation of ERGMs to date has been t

2.4.3 Random effects and mixed effects models

p2 modelp1 model can be represented as a generalized linear model and the p2 model as a generalized linear mixed model

2.4.4 Stochastic block models

2.4.5 Latent space models

2.5 Simulations on the Sampsons monk dataset

In this part we use the The data set collected by Sampson (1968) this is a classic data set in social network analysis. The data set summarizes relationships, observed at three distinct time points, among 18 monks who were about to enter a monastery when a conflict erupted. We use here the directed network where $y_{i,j} = 1$ denotes that monk i ranked monk j among his top three preferred individuals and $y_{i,j} = 0$ otherwise. The directed network is shown in Figure 2.9 , where circles represent monks and directed edges are oriented from i to j whenever $y_{i,j} = 1$. The monks are divided by Sampson into three groups: Loyal Opposition, Turks, and Outcasts.

We applied the same method as for the simulated data and obtained the following community. We can observe that our model identifies exactly the same clustering of communities as the ones observed by Sampson. Namely the Loyal Opposition (Red community), Turks (Green Community), and Outcasts (Blue Community). In this

graph the individuals are labeled according to their number provided with their names in the data set package and not according to their rank in the data-frame.

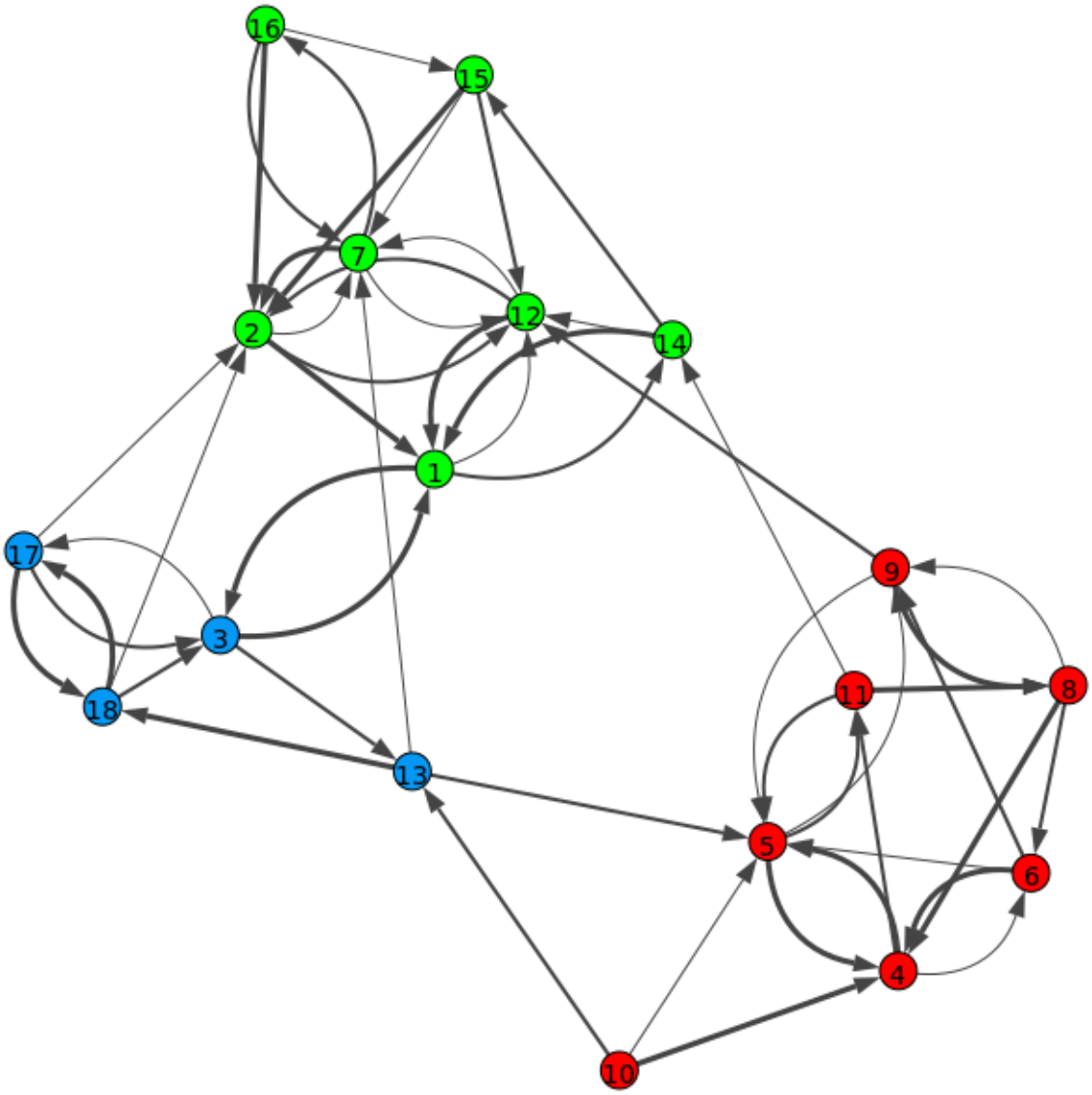

Figure 2.9: Inferred community for the Sampson DataSet