Community detection in graphs

Julián Alvarez de Giorgi [1] Borachhun You [2] Nicolás Enrique Cecchi [2]

[1]Institut Polytechnique de Paris

[2]ENS Paris-Saclay

Overview

Community detection in graphs is the task of, given a network $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ where $\mathcal{E}$ is the set of edges and $\mathcal{V}$ is the set of vertices, assigning each vertex $v_i \in \mathcal{V}$ to at least one of multiple classes $\mathcal{K} = \{1, \dots, K\}$.

In this presented work, the task has been achieved by fitting a mixture model to the network, particularly we have used the Stochastic Block Model (SBM), presented in [1], and a mixture model proposed by Newman and Leicht.

The model parameters are found by statistical learning. We study and implement a set of community detection algorithms and test their performance on real and synthetic datasets.

Stochastic block model

The Stochastic Block Model (SBM) models a community network as follows:

- Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a pair of a set of N vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$ connecting them.
- X the $N \times N$
- Let the nodes belong to $\mathcal{K}$ (unobserved) clusters with probability distribution $\alpha = (\alpha_1, ..., \alpha_K)$, s.t. $\sum_{k=1}^{K} \alpha_k = 1$.
- Let g_i be the group to which node i belongs and the hidden variable $Z \in \{0,1\}^{N \times K}$ whose entries $z_{ik} = 1$ if the node i belongs to cluster k, and 0 otherwise.
- Let the parameter π_{ql} be the probability of a node from community q to link with a node from community l., defining the **cluster connectivity matrix** $\Pi = (\pi_{ql})$.

Then the objective is **finding the membership indicators** Z

This parametrization allows to sample a graph the SBM by

- Sampling the membership of each node independently: $Z_i \sim \mathcal{M}(1, \alpha)$ where $\mathcal{M}$ is the multinomial distribution.
- Sampling each edge independently given the node's membership: $X_{ij}|\{i \in q\}\{j \in l\} \sim \mathcal{B}(\pi_{ql})$, with $\mathcal{B}$ the Bernoulli distribution.

Newman-Leicht mixture model

Newman and Leicht presented a mixture model [2] that generalizes SBM by replacing Π by a matrix $\theta = \theta_{ri}$, whose elements represent the probability that a directed link from a vertex in cluster r is drawn with vertex i, hence $\theta \in \mathrm{R_+}^{K \times N}$, which satisfies the normalization condition $\sum_{i=1}^{N} \theta_{ri} = 1$. No assumption on the communities of i.

Statistical inference

The framework usually adopted for fitting a model to data is that of likelihood maximization:

$$\max_{\beta} \mathbf{P}(\mathcal{X}, \mathcal{Z}|\beta) \tag{1}$$

Where $\mathcal{X}$ is the observed data, $\mathcal{Z}$ the classes assignation and β are the parameters of the model.

The difficulty of this is that $\mathcal{Z}, \beta$ are usually unknown quantities that are dependent on each other, hence the likelihood is untractable.

In the case of the NMM, we use the **Expectation Maximization** algorithm on the expected value of the log-likelihood over the latent variable z:

$$\bar{\mathcal{L}} = \sum_i \sum_j \tau_{ir} \Big(\log \alpha_r + \sum_j x_{ij} \log \theta_{rj} \Big) \tag{2}$$

In the case of the SBM, the authors take a **Variational EM** approach, we approximate the likelihood by a family of simpler distribution, and maximize the posteriors by maximizing an Evidence Lower Bound (ELBO):

$$\mathcal{J}(\mathcal{R}_{\mathcal{X}}) = \log \mathcal{L}(\mathcal{X}) - KL(\mathcal{R}_{\mathcal{X}(\cdot)}; \mathbb{P}(\cdot|\mathcal{X})) \tag{3}$$

Experiments

The following figures show the results of the implementation of the studied models tested on various datasets. In each of the figures, the positions of the nodes indicate the real communities while the colors of the nodes represent the predicted communities.

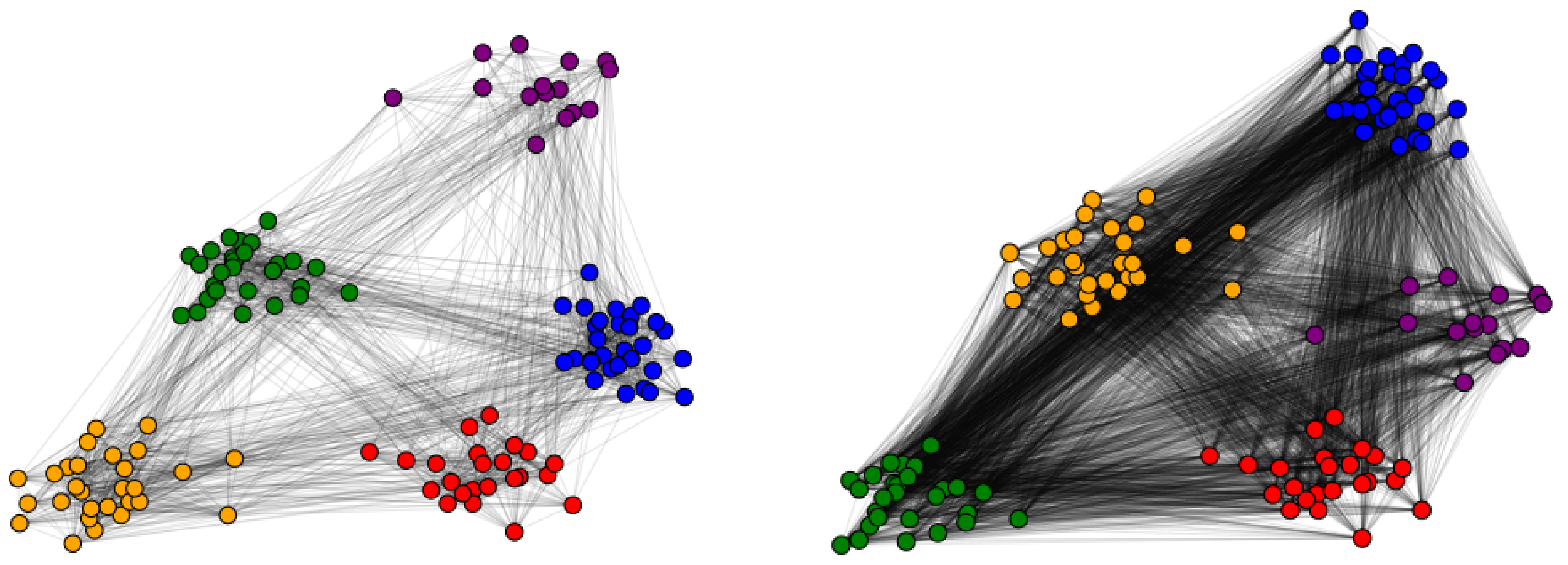

Figure 1. Result of NMM on 2 simulated graphs with: assortative mixing, more edges inside communities (left), and disassortative mixing, more edges between communities (right)

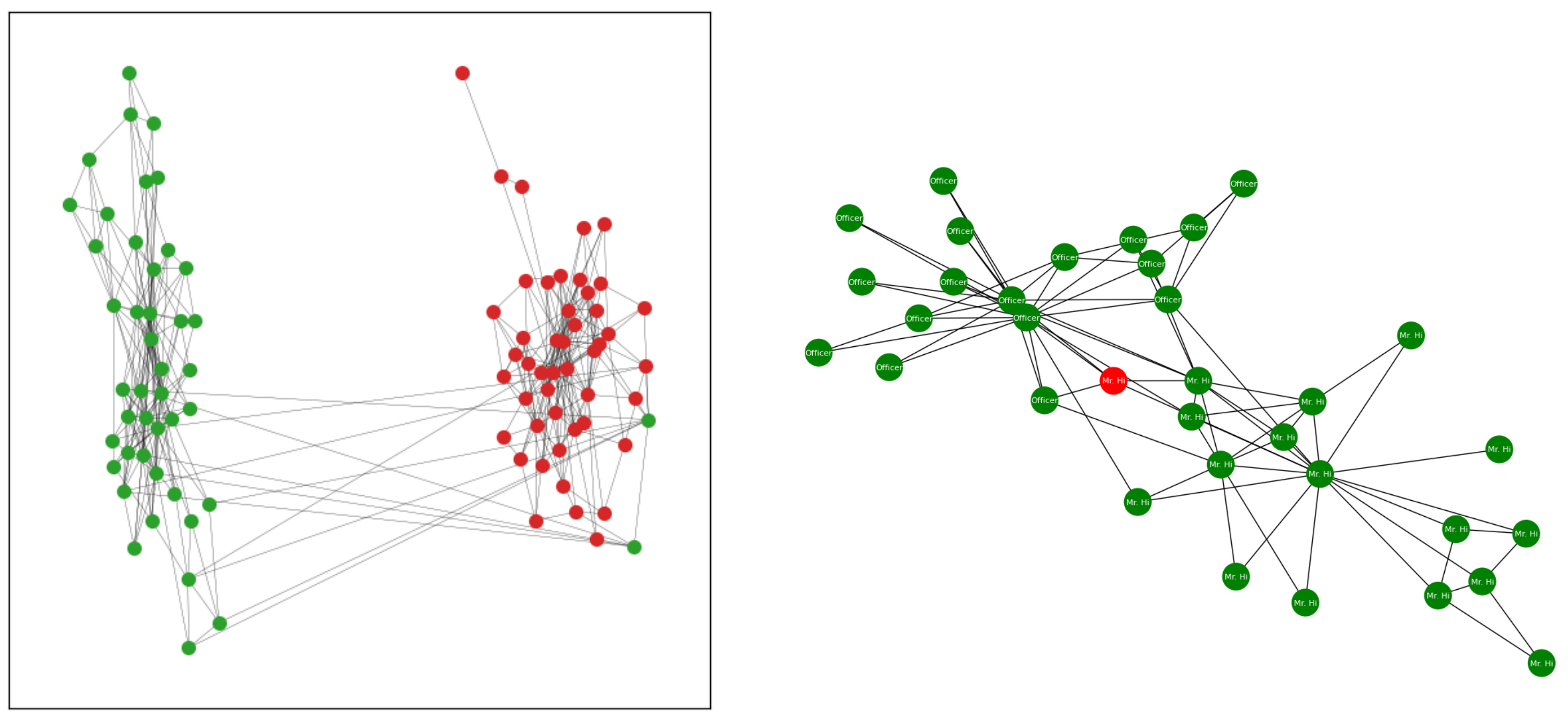

Figure 2. Result of NMM on two communities datasets. On the left, the network of co-purchasing US politics books [3], the left cluster corresponds to "liberal" books and the right cluster corresponds to "conservative" books. The image on the right shows the well known Zachary's Karate Club [4].

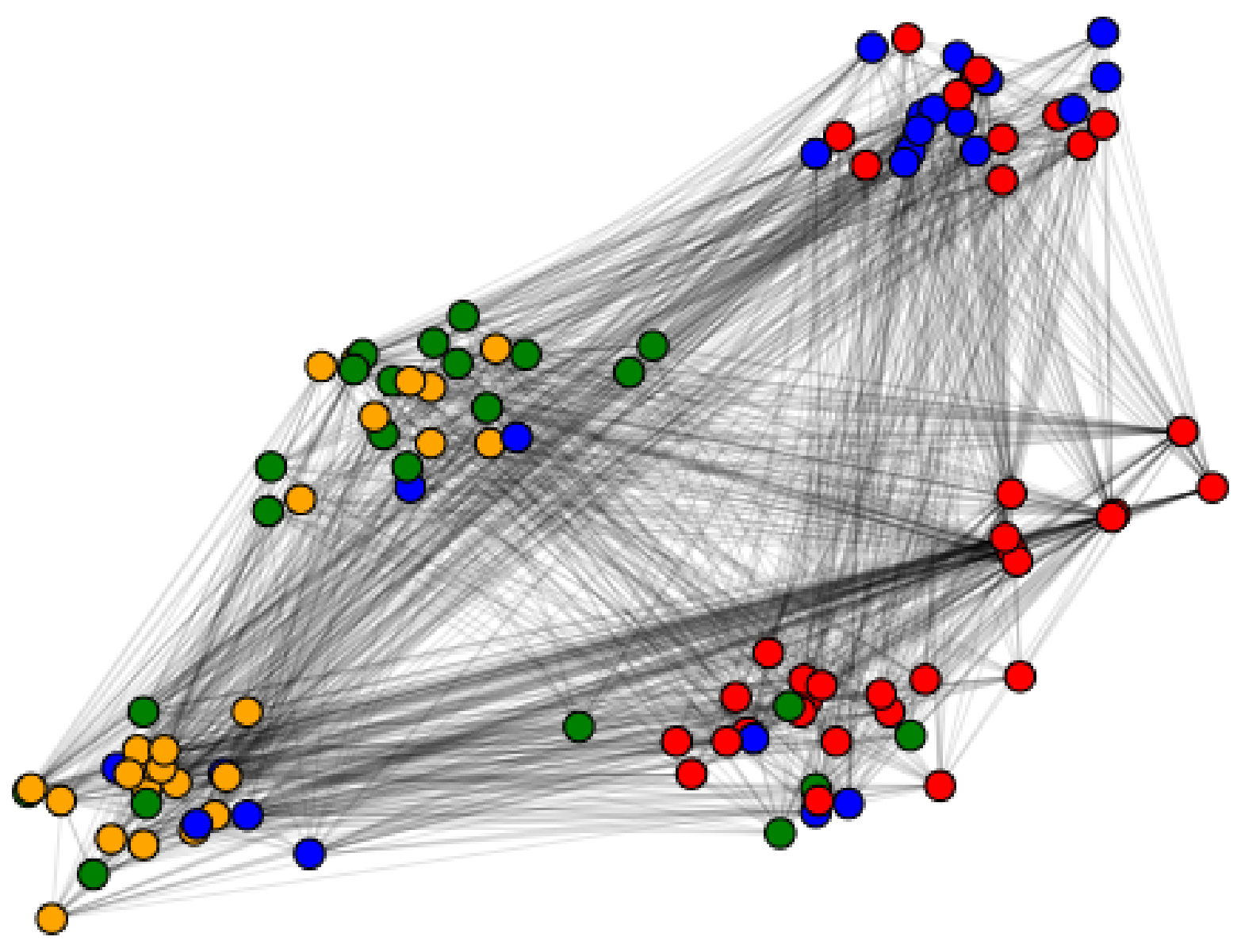

Figure 3. Result of NMM on Keystone's network

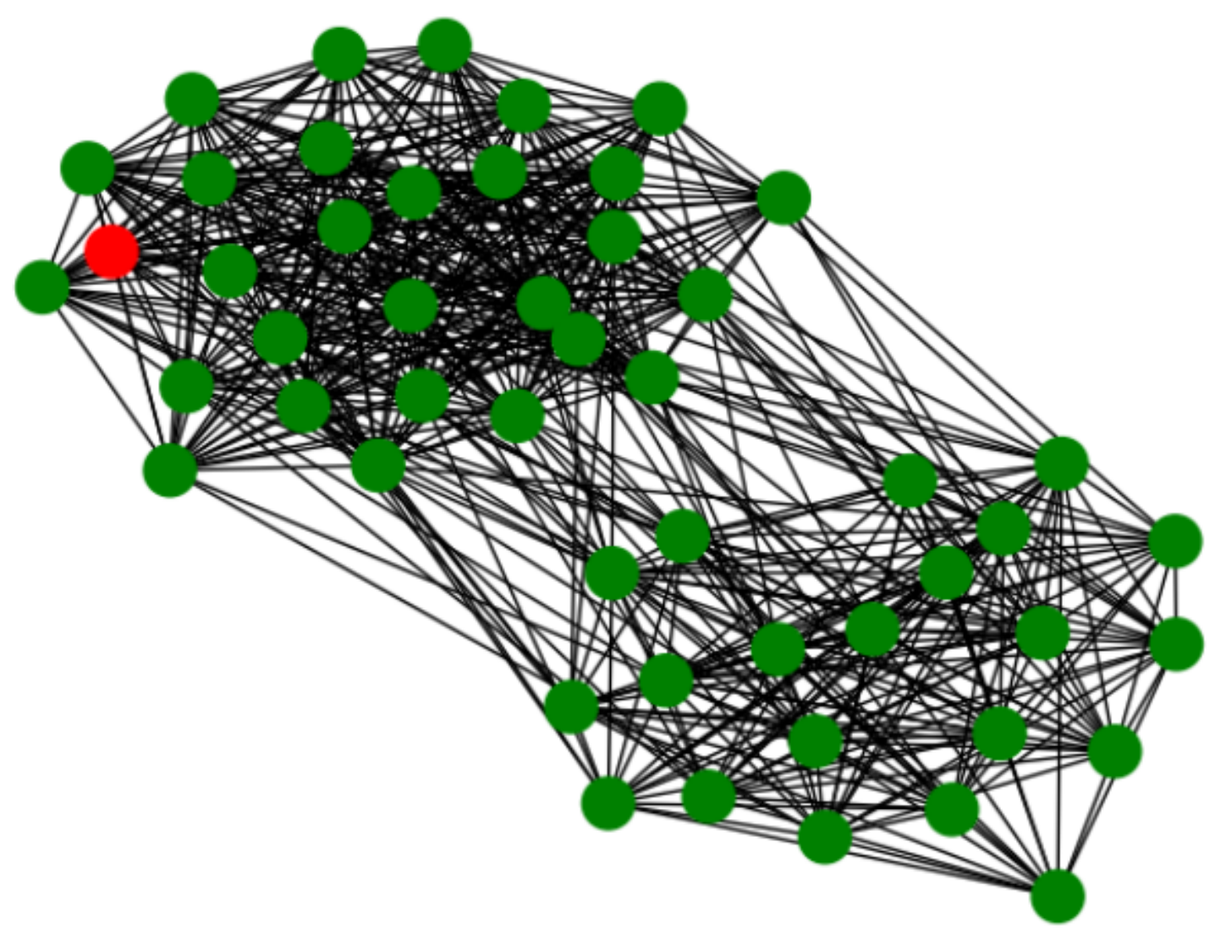

Figure 4. Result of SBM on a simple two communities network with assortative mixing.

References

[1] J-J Daudin, Franck Picard, and Stéphane Robin. A mixture model for random graphs. Statistics and computing, 18(2):173–183, 2008.

[2] M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences, 104(23):9564–9569, June 2007.

[3] Valdis Krebs. Mark newman network datasets.

[4] Wayne Zachary. An information flow model for conflict and fission in small groups. Journal of anthropological research, 33, 11 1976