Community detection in graphs

Julián Alvarez de Giorgi
Institut Polytechnique de Paris
Palaiseau, France
julian.alvarez@telecom-paris.fr

Borachhun You
ENS Paris-Saclay
Gif-sur-Yvette, France
borachhun.you@ens-paris-saclay.fr

Nicolás Enrique Cecchi
ENS Paris-Saclay
Gif-sur-Yvette, France
nicolas.cecchi1@ens-paris-saclay.fr

Abstract

This report explores the field of community detection with Stochastic Block Model and Newman-Leicht Mixture Model. Different inference methods, which are used to estimate the parameters of the models, are also presented, along with an experiment that applies the methods on various network datasets.

1 Introduction

The beginning of graph theory is usually pin-pointed at the famous work by Euler on the problem of the Seven bridges of Königsberg [8]. A lot has happened in the field since then. A graph is a network-like object, composed of a set of items called vertices (or nodes), with connections between them, called edges. Systems taking the form of networks abound in the world [17]. Today, a canonical example is that of online social networks [14, 15, 22], but not only. Network science can also be applied to epidemiology [16, 21], gene-disease relations [2], energy distribution [20], air or maritime port networks [1, 27] or online political discourse [12].

The analysis of these network systems can have different objectives. One may focus on the robustness of the networks to determinate threats [6, 10], understand its dynamics [5] or finding communities (also called *clusters*) [4, 13, 23], which is the focus of this report.

Community detection refers to finding groups of 'similar' nodes, based on some similarity measure [23]. Many techniques exist for finding communities, such as: modularity based methods, spectral clustering, methods based on statistical inference [9] and also based on deep learning [25].

In this work, we've implemented a class to handle a Stochastic Block Model (SBM), which is a widely known mixture model employed for vertex clustering in networks. This model has been proven to fit well to a wide variety of real-life networks, and also incorporates the membership of vectors. Many inference methods based on different assumptions have been proposed to estimate the parameters of the model, as the in [4], where they propose a Variational EM method on the conditional expectation of the complete data log-likelihood, or in [3] where they propose a greedy algorithm to estimate simultaneously the number of clusters and the parameters of the model through the EM algorithm on the exact integrated complete data log likelihood. Both methods were tried to be implemented, the first one working under simple networks, and the second one was implemented without success.

Furthermore, we've also implemented what we called the Newman Mixture Model (NMM), in which the parameterization slightly differs from the SBM models. We have implemented the EM inference method described in [18], and we have tested it successfully on different networks. Moreover, a link between SBM and NMM was made, allowing to generate a SBM from the parametrization of the NMM. https://doi.org/10.7551/mitpress/1100.003.0014

In this report, we present a brief introduction to the SBM and NMM, alongside the inference methods implemented in section 2, then in section 3 we present the insights and calculations of the inference methods, continuing with section 4, where the results of the experiments are presented, and finally in section 5, some conclusions are drawn.

2 Graph models

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. We denote X the $N \times N$ *adjacency matrix* whose x_{ij} entry takes the value 1 if there is an edge from node i to node j, 0 otherwise. If the graph is undirected, the matrix will be symmetric, such that $x_{ij} = x_{ji}, \forall (i, j) \in N \times N$. We will consider that no self-loops exist, and therefore $x_{ii} = 0$.

Furthermore, consider that the nodes belong to K (unobserved) clusters with probability distribution $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$, s.t. $\sum_{k=1}^{K} \alpha_k = 1$. Let's note g_i the group to which node i belongs, this information is captured with 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. When performing a clustering task, this membership indicator Z is the objective.

2.1 Stochastic Block Model

Besides the general considerations stated previously, SBM considers the parameter π_{ql}: the probability of a node from community q to link with a node from community l. This allows to construct the cluster-connectivity matrix $\mathbf{\Pi} = (\pi_{ql})$. It's important to highlight, that for undirected graphs, $\mathbf{\Pi}$ is symmetric.

This parametrization allows us to sample a graph of any size from the SBM, by sampling the membership of each node independently, that is, $\boldsymbol{Z_i} \sim \mathcal{M}(1, \boldsymbol{\alpha})$ where $\mathcal{M}$ is the multinomial distribution, and by sampling each edge independently given the node's membership:

$$X_{ij} | \{i \in q\}\{j \in l\} \sim \mathcal{B}(\pi_{ql}) \tag{1}$$

With $\mathcal{B}$ the Bernoulli distribution.

2.2 Newman-Leicht Mixture Model (NMM)

Newman and Leicht presented a mixture model in [18], whose difference with respect to SBM is that NMM considers, instead of the cluster-connectivity matrix Π, a matrix $\boldsymbol{\theta}$, whose elements θ_{ri} represent the probability that a directed link from a vertex in cluster r is drawn with vertex i, hence $\boldsymbol{\theta} \in \mathrm{R}_{+}{}^{K\times N}$, which satisfies the normalization condition $\sum_{i=1}^{N} \theta_{ri} = 1$. It's important to observe that there's no assumption on the communities of i.

Notice that this model has more parameters than the SBM since it considers the probability of a link between a cluster and each point, instead of just between clusters.

We can notice that the NMM generalizes the SBM, which 'forces' all the θ_{ri} be equal for all the vertices i in a certain cluster. We have implemented this relation, by parametrizing an SBM model from the parameters of the NMM.

3 Methods

The framework usually adopted for fitting a model to data is that of likelihood maximization. That is maximizing $\mathbf{P}(\mathcal{X}, \mathcal{Z}|\beta)$ w.r.t. β, being the parameters of the model, $\beta = \{\boldsymbol{\alpha}, \boldsymbol{\Pi}\}$ for the SBM, and $\beta = \{\boldsymbol{\alpha}, \boldsymbol{\theta}\}$ on the NMM framework.

The difficulty of this is that $\mathcal{Z}, \beta$ are unknown quantities that are dependent on each other. To overcome this, the EM algorithm [7] is used to iteratively improve the approximations. Newman and Leicht [18] followed this approach. A similar method, variational EM, was proposed by [4] to maximize a lower bound of the log-likelihood.

Finally, in [3], they directly focus on the integrated classification likelihood (ICL) criteria, they calculate this quantity exactly, and propose a greedy method to estimate the number of cluster alongside the parameters by switching each vertex to all different clusters and evaluating the ICL changes.

3.1 Fitting the NMM to data

In this setting, we have to maximize $\mathbb{P}(X, g|\alpha, \theta)$ with respect to α, θ, which can be done by writing:

$$\mathbb{P}(X, g|\alpha, \theta) = \underbrace{\mathbb{P}(X|g, \alpha, \theta)}_{\prod_{ij} \theta^{x_{ij}}_{g_i,j}} \underbrace{\mathbb{P}(g|\alpha, \theta)}_{\prod_i \alpha_{g_i}} \tag{2}$$

$$\mathbb{P}(X, g|\alpha, \theta) = \prod_i \alpha_{g_i} \Big(\prod_j \theta^{x_{ij}}_{g_i,j} \Big) \tag{3}$$

$$\tag{4}$$

Usually, one does not work with the likelihood but the log-likelihood:

$$\mathcal{L} = \sum_i \Big(\log \alpha_{g_i} + \sum_j x_{ij} \log \theta_{g_i,j} \Big) \tag{5}$$

Since the g_i are unknowns, we cannot compute that value. However, one can compute the expected value by averaging over g, as follows:

$$\bar{\mathcal{L}} = \sum_i \sum_j \tau_{ir} \Big(\log \alpha_r + \sum_j x_{ij} \log \theta_{rj} \Big) \tag{6}$$

Where $\tau_{ir} = \mathbb{P}(g_i = r|X, \alpha, \theta)$, which is the probability of vertex i belonging to group r.

Finding this expected value is the best estimate that we have, but computing it still posses a problem because of the mutual dependence between the unknowns. As explained above, the EM algorithm can be applied.

One begins therefore by computing the expected memberships, given an initialization of α and θ, together with the observed data:

$$\tau_{ir} = \frac{\mathbb{P}\big(X, g_i = r|\alpha, \theta\big)}{\mathbb{P}\big(A|\alpha, \theta\big)} \tag{7}$$

$$\tau_{ir} = \frac{\alpha_r \prod_j \theta_{rj}^{x_i j}}{\sum_t \alpha_t \prod_j \theta_{rj}^{x_i j}} \tag{8}$$

$$\tag{9}$$

After which, one can update the other parameters:

$$\alpha_r = \frac{1}{N} \sum_i \tau_{ir} \tag{10}$$

$$\theta_{rj} = \frac{\sum_i x_{ij} \tau_{ir}}{\sum_i \tau_{ir} \sum_j x_{ij}} \tag{11}$$

$$\tag{12}$$

That have the natural interpretations of the average nodes that belong to a class and the proportion of outbound edges of class r that go into node j.

Extending this model to the undirected case is straightforward. The probability that an edge exists between node i from class g_i and node j from class g_j is given by $\theta_{g_i i} \theta_{g_j j}$, which also satisfies the normalization conditions.

At the end of the estimation procedure, one has estimations of the following quantities: $\theta \sim \hat{\theta}$, a K by N matrix with the probabilities of a node from cluster k connecting to a specific node n; $\alpha \sim \hat{\alpha} \in \mathbb{R}^K$, a vector of class probabilities; $Z \sim \hat{\tau} \in \mathbb{R}^{N\times K}$, the matrix of cluster assignments.

3.2 Inference of stochastic block model parameters

The estimation of stochastic block models was thoroughly developed by Nowicki and Snijders [19, 24] assumes that vertices are partitioned into (unobserved) classes and that the probability distribution of the relation between two vertices depends only in the classes that they belong to.

As in *Newman-Leicht Mixture Model*, consider the nodes can belong to K classes with prior probabilities $\alpha = [\alpha_1, \ldots, \alpha_K]$ and indicator variables $\{z_{ik}\}$ for denoting if node i belongs to class k, with the usual constraint of $\sum_k z_{ik} = 1$:

$$\alpha_k = \mathbb{P}(z_{ik} = 1) = \mathbb{P}(i \in k), \text{with} \sum_k \alpha_k = 1$$

Then π_{ql} denotes the probability of a vertex from class q to be connected to one of class l. In the undirected setting, we have that $\pi_{ql} = \pi_{lq}$.

A final assumption of the model is that edges x_{ij} are conditionally independent given their classes:

$$\begin{cases} x_{ij} | \{i \in q, j \in l\} \sim \mathcal{B}(\pi_{ql}) \text{if } i \neq j \\ x_{ii} = 0 \end{cases}$$

Since EM requires the computation of the conditional $\mathbb{P}(\mathcal{Z}|\mathcal{X})$, which is not tractable, Daudin and Robin [4] propose a variational approach. that optimizes a lower bound of the log likelihood:

$$\mathcal{J}(\mathcal{R}_{\mathcal{X}}) = \log \mathcal{L}(\mathcal{X}) - KL(\mathcal{R}_{\mathcal{X}(\cdot)}; \mathbb{P}(\cdot|\mathcal{X})) \tag{13}$$

where KL is the Kullback–Leibler divergence, $\mathbb{P}(Z|X)$ is the true conditional distribution of the indicator variables Z given the data, and $\mathcal{R}_{\mathcal{X}}$ an approximation of this conditional distribution. We get the real log-likelihood if the approximated distribution is equal to the true one. The proposed algorithm optimizes $\mathcal{J}$ alternatively with respect to $\mathcal{R}_{\mathcal{X}}$ and then with respect to $(\boldsymbol{\alpha}, \boldsymbol{\pi})$.

Approximating $\mathcal{R}_{\mathcal{X}}$ as multinomials such that:

$$\mathcal{R}_{\mathcal{X}}(\mathcal{Z}) = \prod_i \mathcal{M}(\mathcal{Z}, \boldsymbol{\tau_i}) \tag{14}$$

Where $\boldsymbol{\tau_i} = (\tau_{i1}, \ldots, \tau_{iK})$, τ_{iq} can be interpreted as $\mathbb{P}(Z_{iq} = 1|\mathcal{X})$.

With this modeling, the parameters that maximize $\mathcal{J}$ at each step are:

$$\begin{cases} \hat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_l b(X_{ij}; \pi_{ql})^{\hat{\tau}_{jl}} \\ \end{cases} \tag{15}$$

With $b(x, \pi) = \pi^x (1 - \pi)^{(1-x)}$.

$$\begin{cases} \hat{\alpha}_q = \frac{1}{n} \sum_i \hat{\tau}_{iq} \\ \hat{\pi}_{ql} = \frac{\sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jl} X_{ij}}{\sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jl}} \end{cases} \tag{16}$$

After an initialization of $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\pi}}$ one can iterate between 15 and 16 until convergence.

3.3 From Newman-Leicht to SBM

Considering the resemblance between the two models, an intuitive additional step allows one to estimate the parameters of the SBM from the results of Newman-Leicht. Considering that the θ_{ri} are normalized such that $\sum_i \theta_{ri} = 1$, one can compute π_{ql}:

$$\pi_{ql} \sim \sum_{i \in l} \theta_{ri} \tag{17}$$

Which is the sum of the probability of a link between cluster r over all the nodes in cluster l.

4 Experiments

We've tested our SBM model with the variational inference EM for a simply 2 classes synthetic networks, the results are shown in Fig. 1

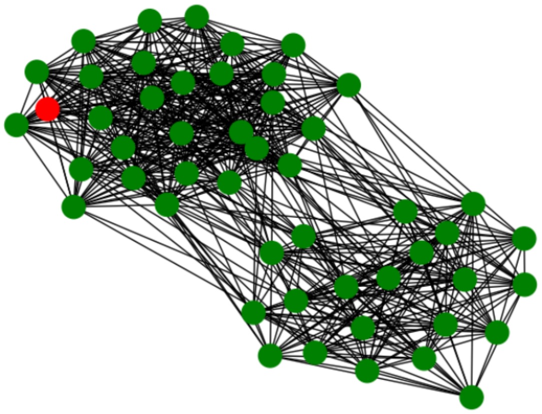


Figure 1. Variational EM for an easy 2-cluster network. In green the nodes well predicted, in red the mistakes

However, for more clusters, or more complex networks, the method behaves badly, which led us to think that the success in this simply dataset is only due to the initialization, which in our case we are doing a spectral clustering as initialization.

We've further tried to implement this method from scratch in different manners and using different Python tools and debugging with pdb, but did not manage to find the bug.

On the other hand, we have tested our implementation of the NMM in a variety of networks.

4.1 Zachary's Karate Club

Zachary's Karate Club dataset, collected by Wayne W. Zachary in 1977 [26], stands as a well-known example in network analysis and social sciences. The dataset captures the dynamics of friendships among 34 members of a university karate club. Each member is represented as a node in the graph, with edges denoting friendships or interactions. What makes this data set particularly intriguing is an event that took place within the club. Following a conflict between the club administrator and the instructor, the karate club splintered into two factions, offering a unique insight into the community dynamics.

We've applied the model with good results, an accuracy of 97, 06%, which is only one mistake in a node that acts as a nexus between both factions. The results are shown in Fig. 2.

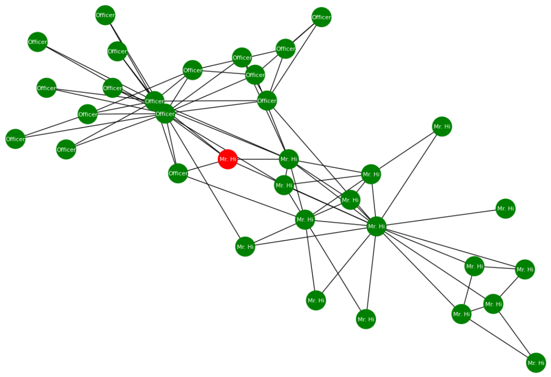
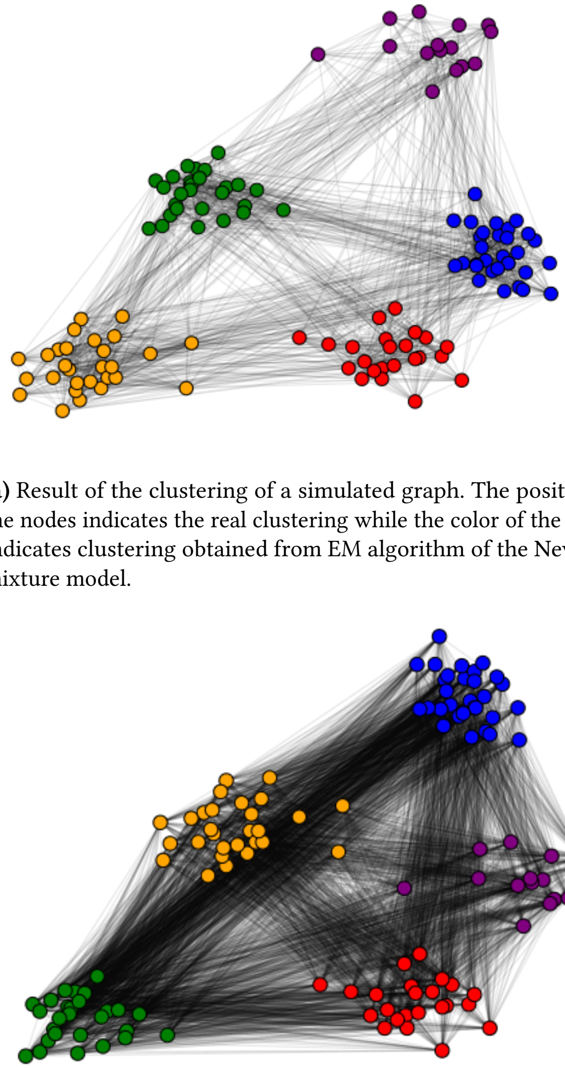


Figure 2. Result of the clustering obtain with the EM inference on the Newman mixture model for the Karatee club dataset. In green are the nodes clustered in the correct communities, and in red the ones clustered in the wrong community.



(a) Result of the clustering of a simulated graph. The position of the nodes indicates the real clustering while the color of the nodes indicates clustering obtained from EM algorithm of the Newman mixture model.

(b) Simulated graph with nodes colored acording to prediction of the model.

Figure 3

4.2 Simulated graph

We have developed a Python class called CommunitiesGraph, that allows the creation of synthetic graphs with communities structure following the SBM structure.

To test the algorithm, we generate 2 graphs with 5 communities each, one with assortative mixing, that is more edges inside the communities, and the other with disassortative mixing (more edges between different communities). Both graphs with 125 nodes, with an inter-class edge probability of 0.8 and intra-class edge probability, and 0.1 inter-class edge probability for the assortative graph, and switched probabilities for the disassortative graph.

Figure 3 presents the result of the clustering obtained from the algorithm for both graph structures. The grouped positions of the nodes represent the real clusters whereas the colors of the nodes correspond to the clustering given by the algorithm. We can see that all nodes are correctly clustered.

4.3 Books about US politics

"Books about US politics" network [11] represents co-purchase relationships among books on the topic of US politics, which were published around the time of the 2004 US presidential election and sold by Amazon. Two different types of books are included in this dataset: "liberal" and "conservative", which are represented by the nodes. Furthermore, edges between the nodes indicate that those books are frequently purchased together by customers.

Figure 4 shows the structure of the network as well as the clustering obtained from the algorithm. The left cluster corresponds to liberal books whereas the right cluster corresponds to the conservative books. We can observe that most edges exist within clusters rather that between them, which indicates that if one book is purchased, the next book to be purchased is more likely to be the type as the previous book. As for the clustering obtained from the algorithm, it can be seen that there are 2 conservative nodes that are missed classified as liberal. Other than that, the other nodes are correctly clustered, showing that the utilized model works well for this specific dataset.

4.4 Keystone's network

Following the experiments taken in [18], a KeystonGraph class was implemented to handle a network with 108 nodes, which is directed. The first 100 nodes are divided into 4 classes, with 25 nodes each, and the edges are randomly drawn uniformly with all the nodes with the constraint that

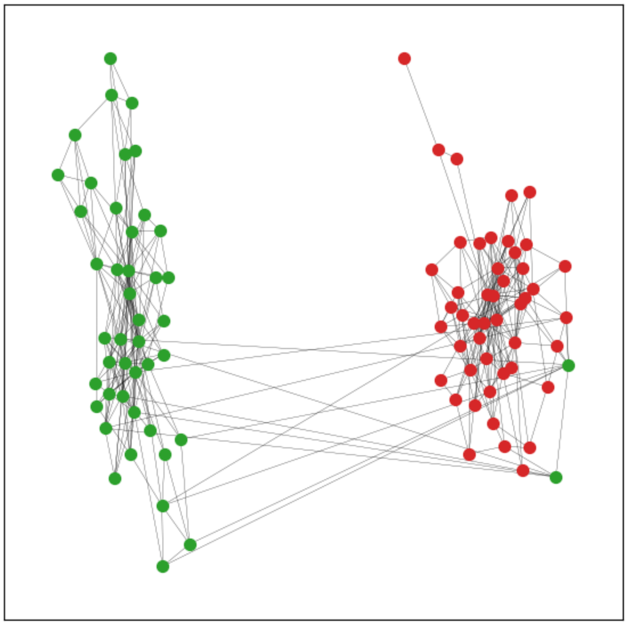


Figure 4. Result of the clustering of the US politics books. The left cluster corresponds to "liberal" books and the right cluster corresponds to "conservative" books. The colors of the nodes are the result from the result from EM algorithm of the Newman mixture model.

the in-degree and out-degree of each node is 10. Then the resting 8 nodes are called keystones and are such that all nodes of cluster A, cluster B, cluster C, and cluster D are connected to the keystones $\{1, 2, 3, 4\}$, $\{3, 4, 5, 6\}$, $\{5, 6, 7, 8\}$, and $\{7, 8, 1, 2\}$ respectively. The clustering was done for 4 classes, and we see the results in Fig. 5, where the positions are the original clusters with the keystones and the colors are the 4 predicted communities of the model.

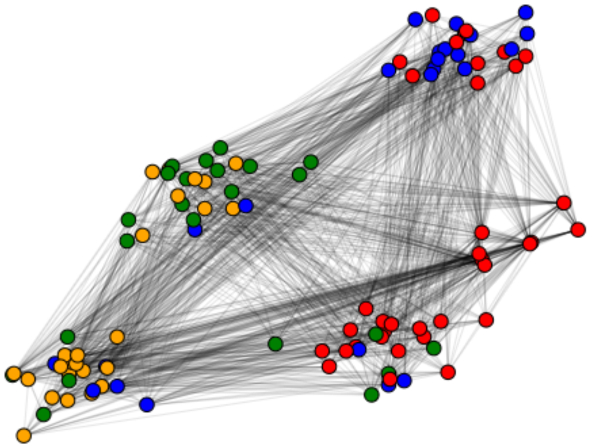


Figure 5. Keystone dataset model prediction.

We can see in Fig.5, that even though each cluster has a majority well predicted, the model does not perform that well for this topology, contrary to the result shown in the author's paper.

5 Conclusion

Given the wide variety of applications, network science, and community detection in particular, are very active research fields.

In this report we analyze different methods for clustering on networks. The clustering is performed via the computation of parameters that characterize different models for communities that emerge in graphs. We have studied in depth two particular models: Stochastic Block Model (SBM) and Newman-Leicht Mixture Model (NMM), both of them without considering overlap of communities.

Additionally, we have seen the inference methods that exists to compute these parameters. The methods were implemented in Python and are made available in the attached code.

Not all the implementations worked correctly, and we were not able to make every model converge, but some of them did work as expected and the results are satisfactory.We believe, however, that our implementations are in the good direction and with some changes should be able to perform as expected.

The models were applied to both real and simulated datasets. The results obtained from the experiment were promising since the algorithms perform almost perfectly for all except one dataset, which was the Keystone's network.

References

[1] Ganesh Bagler. 2008. Analysis of the airport network of India as a complex weighted network. *Physica A: Statistical Mechanics and its Applications* 387, 12 (2008), 2972–2980.

[2] Anna Bauer-Mehren, Markus Bundschus, Michael Rautschka, Miguel A Mayer, Ferran Sanz, and Laura I Furlong. 2011. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PloS one* 6, 6 (2011), e20284.

[3] E. Côme and P. Latouche. 2014. Model selection and clustering in stochastic block models with the exact integrated complete data likelihood. arXiv:1303.2962 [stat.ME]

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

[5] M Argollo De Menezes and A-L Barabási. 2004. Fluctuations in network dynamics. *Physical review letters* 92, 2 (2004), 028701.

[6] Anthony H Dekker and Bernard D Colbert. 2004. Network robustness and graph topology. In *Proceedings of the 27th Australasian conference on Computer science-Volume 26*. 359–368.

[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)* 39, 1 (1977), 1–22.

[8] Leonard Euler. 1726. *Solutio problematis ad geometriam situs pertinentis.* Petropolis, Typis Academiae.

[9] Santo Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5 (2010), 75–174.

[10] Michele Garetto, Weibo Gong, and Don Towsley. 2003. Modeling malware spreading dynamics. In *IEEE INFOCOM 2003. Twenty-Second*

annual joint conference of the IEEE computer and communications societies (IEEE Cat. No. 03CH37428), Vol. 3. IEEE, 1869–1879.

[11] Valdis Krebs. [n. d.]. Mark Newman network datasets. https://public.websites.umich.edu/~mejn/netdata/

[12] Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. 2011. Overlapping stochastic block models with application to the french political blogosphere. (2011).

[13] Pierre Latouche, Etienne Birmele, and Christophe Ambroise. 2012. Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling* 12, 1 (2012), 93–115.

[14] Abdul Majeed and Ibtisam Rauf. 2020. Graph theory: A comprehensive survey about graph theory applications in computer science and social networks. *Inventions* 5, 1 (2020), 10.

[15] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. 2002. Random graph models of social networks. *Proceedings of the national academy of sciences* 99, suppl_1 (2002), 2566–2572.

[16] M. E. J. Newman. 2002. Spread of epidemic disease on networks. *Physical Review E* 66, 1 (July 2002). https://doi.org/10.1103/physreve.66.016128

[17] M. E. J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Rev.* 45, 2 (2003), 167–256. https://doi.org/10.1137/S003614450342480 arXiv:https://doi.org/10.1137/S003614450342480

[18] M. E. J. Newman and E. A. Leicht. 2007. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences* 104, 23 (June 2007), 9564–9569. https://doi.org/10.1073/pnas.0610537104

[19] Krzysztof Nowicki and Tom A B Snijders. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association* 96, 455 (2001), 1077–1087.

[20] Giuliano Andrea Pagani and Marco Aiello. 2013. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications* 392, 11 (2013), 2688–2700.

[21] Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters* 86, 14 (April 2001), 3200–3203. https://doi.org/10.1103/physrevlett.86.3200

[22] Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. 2007. Recent developments in exponential random graph (p*) models for social networks. *Social networks* 29, 2 (2007), 192–215.

[23] Satu Elisa Schaeffer. 2007. Graph clustering. *Computer science review* 1, 1 (2007), 27–64.

[24] Tom AB Snijders and Krzysztof Nowicki. 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification* 14, 1 (1997), 75–100.

[25] Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu. 2014. Learning deep representations for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 28.

[26] Wayne Zachary. 1976. An Information Flow Model for Conflict and Fission in Small Groups1. *Journal of anthropological research* 33 (11 1976). https://doi.org/10.1086/jar.33.4.3629752

[27] Rawya Zreik, César Ducruet, Charles Bouveyron, and Pierre Latouche. 2017. Cluster 19 dynamics in the collapsing Soviet shipping network. *Advances in Shipping Data Analysis and Modeling: Tracking and Mapping Maritime Flows in the Age of Big Data* (2017), 317.

Received 22 December 2023