Electronic Journal of Statistics
Vol. 6 (2012) 2574–2601
ISSN: 1935-7524
DOI: 10.1214/12-EJS753

Classification and estimation in the Stochastic Blockmodel based on the empirical degrees

Antoine Channarond, Jean-Jacques Daudin and Stéphane Robin

Département de Mathématiques et Informatique Appliquées
UMR 518 AgroParisTech/INRA
16 rue Claude Bernard, 75231 Paris Cedex 5, France
e-mail: {channarond; daudin; robin}@agroparistech.fr

Abstract: The Stochastic Blockmodel [16] is a mixture model for heterogeneous network data. Unlike the usual statistical framework, new nodes give additional information about the previous ones in this model. Thereby the distribution of the degrees concentrates in points conditionally on the node class. We show under a mild assumption that classification, estimation and model selection can actually be achieved with no more than the empirical degree data. We provide an algorithm able to process very large networks and consistent estimators based on it. In particular, we prove a bound of the probability of misclassification of at least one node, including when the number of classes grows.

AMS 2000 subject classifications: 62H30, 62H12.
Keywords and phrases: Stochastic Blockmodel, unsupervised classification, clustering, estimation, model selection.

Received November 2011.

1. Introduction

Strong attention has recently been paid to network models in many domains such as social sciences, biology or computer science. Networks are used to represent pairwise interactions between entities. For example, sociologists are interested in observing friendships, calls and collaboration between people, companies or countries. Genomicists wonder which gene regulates which other. But the most famous examples are undoubtedly the Internet, where data traffic involves millions of routers or computers, and the World Wide Web, containing millions of pages connected by hyperlinks. A lot of other examples of real-world networks are empirically treated in Albert and Barabási [1], and book Faust and Wasserman [12] gives a general introduction to mathematical modelling of networks, and especially to graph theory.

One of the main features expected from graph models is inhomogeneity. Some articles, e.g. Bollobás et al. [5] or Van Der Hofstad [26], address this question. In the Erdős-Rényi model introduced by Erdős and Rényi [11] and Gilbert [14], all nodes play the same role, while most real-world networks are definitely not homogeneous.

In this paper, we are interested in the Stochastic Blockmodel (SBM), introduced by Holland et al. [16] and inspired by Holland and Leinhardt [17] and Fienberg and Wasserman [13]. This model assumes discrete inhomogeneity in the underlying social structure of the observed population: n nodes are split into Q homogeneous classes, called blocks, or more generally clusters. Then it is assumed that the distribution of the edge between two nodes, depends only on the blocks to which they belong. Thereby, within each class, all nodes have the same connection behavior: they are said to be structurally equivalent [20]. When the class assignment is known, the social structure can possibly be visualized through the meta-graph [23], which emphasizes the role of each class. However the block structure is supposed to be not observed or *latent*. Thus the assignment Z and the model parameters must be estimated *a posteriori* through the observed graph X, which is a real challenge, especially in large networks.

Our main purpose in this paper is to present a consistent inference method under SBM, which can above all process very large graphs. Snijders and Nowicki [25] have proposed a maximum likelihood estimate based on the EM algorithm for very small graphs with $Q = 2$ blocks. They have also proposed a Bayesian approach based on Gibbs sampling for larger graphs (hundreds of nodes), which they have extended to arbitrary block numbers in Nowicki and Snijders [22]. However the usual techniques enables the processing of only relatively small graphs, because they suffer severely from the complexity of graph structure. In particular the EM algorithm deals with the conditional distribution of the labels Z given the observations X, whose dependency graph is actually a clique in the case of SBM (see paragraph 5.1 in Daudin et al. [9]). Inspired by Wainwright and Jordan [27], Daudin et al. [9] have developed approximate methods using variational techniques in the context of SBM. From a physical point of view, the variational paradigm amounts to mean-field approximation, see Jaakkola [18]. Thus thousands of nodes can be processed with this variational EM algorithm. Lastly, Celisse et al. [6] proves the variational method to be consistent precisely under SBM.

All previous methods treat both classification and parameter estimation directly and at the same time. They are alternatively updated at each step of EM-based algorithms. Yet those tasks are actually not symmetrical, and moreover estimators are quite simple when Z is known. The classification — remaining the main pitfall thus far — can be completed first, and then the latent assignment Z just replaced with this classification by plug-in in order to estimate the parameters.

Searching for clusters from a graph is computationally difficult and has different meanings. Many algorithms, especially coming from physics and computer science, aim at detecting highly connected clusters, which are self-defined as optimizing some objective function. See Lancichinetti et al. [19], Girvan and Newman [15] and methods based on modularity in Newman [21] and Bickel and Chen [3]. In contrast, the blocks under SBM have a model-based definition and do not necessarily have many inner connections (see examples in Daudin et al. [9]). Therefore, most algorithms designed for community detection are generally not suitable in this context.

Bickel and Chen [3], Choi et al. [7], Celisse et al. [6] and Rohe et al. [24] prove that it is asymptotically possible to uncover the latent structure of the graph Z. In this work, we additionally show under a separability assumption that it is possible to do so, just by utilizing degree data instead of the whole graph X. As a consequence, we can work with n variables instead of n^2, which makes classification computations much faster. The basic reason why so little information is needed — compared with other models with latent structure — is specific to SBM. The number of observed variables $(X_{ij})_{1\leq i,j\leq n}$ grows faster than the number of latent variables Z, therefore even marginal distributions of X concentrate very fast. Our algorithm actually expands the procedure introduced by Snijders and Nowicki [25] when $Q = 2$. Like Bickel and Chen [3], we provide probabilistic bounds for the occurrence of one error at least. Moreover we take the random assignment into account, even when the number of classes Q increases and the average degree vanishes. Related results are given in Choi et al. [7] and Rohe et al. [24]. Nevertheless the bounds in these papers concern the rate of misclassified nodes instead, and do not prevent the number of errors from growing to infinity. They also require the assignment Z to be fixed.

Furthermore a simulation study was carried out and shows that the method converges faster than expected from the theoretical bounds but slower than other existing methods. However it is much more computationally efficient, and does not require the storage of the whole adjacency matrix. For large networks, this trade-off might be necessary.

The paper is organized as follows. In Section 2, we begin by presenting the model we shall study and some notations are fixed. Above all a concentration property of the degree distribution is stated in paragraph 2.2, which will be very useful in proving the consistency of the method mentioned above. The classification algorithm (called LG) and the main results are presented in this section as well. In particular, Theorem 2.2 provides a bound of the error probability and Proposition 5.0.1 gives some convergence rates when the number of classes is allowed to grow. The consistency proof of the LG algorithm is provided in Section 3. Section 4 is devoted to deriving simple estimators of the parameters by plug-in and their consistency is also demonstrated. Section 5 adresses the issues related to the separability assumption and provides convergence rates of the LG algorithm as well. A simulation study in Section 6 illustrates the behavior of the LG algorithm, which is discussed afterwards. In Section 7, the model and the algorithm are more accurately studied. As an application, it is lastly proved that it is likewise possible to find out asymptotically the right number Q of blocks of the model. That completes the method relying just on degrees.

2. The Stochastic Blockmodel

2.1. Model

We first recall the SBM. For all integers $n \geq 1$, $[n]$ denotes the set $\{1, \ldots, n\}$. The undirected binary graphs with n nodes are defined by the pair $([n], X)$ where X is a symmetric binary square matrix of size n. X is called the adjacency matrix of the graph. Let $Q \geq 1$ be the number of blocks.

- $Z = (Z_i)_{i\in[n]}$ denotes the *latent* vector of $[Q]^n$ such that $Z_i = q$ if the node i is q-labeled. Let $\alpha = (\alpha_1, \ldots, \alpha_Q)$ be the vector of the block proportions in the whole population.

$$Z = (Z_i)_i \text{ i.i.d.} \sim \mathcal{M}(1; \alpha)$$

- Conditionally on the labels Z, the variables $\{X_{ij}, i, j \in [n]\}$ are independent Bernoulli variables. Conditionally on $\{Z_i = q, Z_j = r\}$, the parameter of X_{ij} is π_{qr}.

$$(X_{ij}|Z_i = q, Z_j = r) \sim \mathcal{B}(\pi_{qr})$$

π_{qr} is the connection probability between any q-labeled node and any r-labeled node. Noting $\pi = (\pi_{qr})_{q,r\in[Q]}$ the connection matrix, the parameters of the model are (α, π). This model will be denoted by $\mathcal{G}(n, \pi, \alpha)$. Note that in the sequel n will be often removed in the notations for the sake of simplicity.

This is a classical problem in mixture models: the block labeling is naturally not identifiable. The content of the blocks remains unchanged by permutating labels. But equivalence classes are identifiable as soon as $n \geq 2Q$, see Celisse et al. [6].

2.2. Degree distribution

For all $i \in [n]$, let $D_i^n = \sum_{j\neq i} X_{ij}$ the degree of the node i, that is the number of neighbors of this node.

Proposition 2.0.1. *For all $q \in [Q]$, let $\overline{\pi}_q = \sum_{r\in[Q]} \alpha_l \pi_{qr}$. D_i^n is a binomial distributed random variable conditionally on $Z_i = q$ with parameters $(n-1, \overline{\pi}_q)$.*

$(D_i^n)_{i\in[n]}$ is therefore a sample of a mixture of binomial distributed random variables with parameters $(n-1, \overline{\pi}_q)_{q\in[Q]}$ and proportions $(\alpha_q)_{q\in[Q]}$.

These variables are correlated. Thus we are not in the validity range of the usual algorithms for mixtures like EM. But there is only one edge shared by any pair of nodes and the degrees are consequently not heavily correlated. Using the EM algorithm would make sense for practical purposes. Nevertheless we have chosen to use a faster one-step algorithm, unlike EM which is iterative.

A concentration inequality for binomial random variables

The following inequality will be useful throughout the article. This will especially account for the fast concentration of the degree distribution. It is a straightforward consequence of Hoeffding's inequality for bounded variables.

Theorem 2.1 (Hoeffding)**.** *Let $n \geq 1$, $p \in]0, 1[$ and $(Y_i)_{i\in[n]}$ a sequence of independent identically distributed Bernoulli random variables with parameter p. Let $S_n = \sum_{i=1}^n Y_i$. Then for all $t > 0$:*

$$P\left(\left|\frac{S_n}{n} - p\right| > t\right) \leq 2e^{-2nt^2} \qquad \text{(CCT)}$$

Concentration property of the normalized degrees

Define the normalized degree of node $i \in [n]$:

$$T_i^n = \frac{D_i^n}{n-1}$$

$(T_i^n)_{i\in[n]}$ cluster around their average conditionally on the node class when n is increasing, according to (CCT):

$$P(|T_i^n - \overline{\pi}_q| > t | Z_i = q) \leq 2e^{-2nt^2} \tag{1}$$

Hence normalized degrees corresponding to q-labeled nodes gather around $\overline{\pi}_q$. Consequently, in the degree distribution, nodes from different classes split up into groups centered around $\overline{\pi}_q$, provided that all conditional averages $(\overline{\pi}_q)_{q\in[Q]}$ are different. From now on, we will assume that they are:

Assumption

$$\forall q, r \in [Q] \quad q \neq r \Rightarrow \overline{\pi}_q \neq \overline{\pi}_r \tag{H}$$

Also define δ the size of the smallest gap between two distinct conditional averages (Assumption (H) amounts to $\delta > 0$):

Definition 1.

$$\delta = \min_{q\neq r} |\overline{\pi}_q - \overline{\pi}_r|$$

Because of the concentration, a larger gap is expected between normalized degrees of nodes from different classes than nodes from the same class. The LG algorithm relies on this remark. It consists in building Q blocks by finding the $Q-1$ largest gaps formed by two consecutive normalized degrees.

The smaller δ is, the closer the degrees are and so the harder the separation of the classes between them is: δ can be regarded as separability parameter of the model. Given δ, n must be large enough so that the classes are clearly separated. This issue is explicitly discussed in Section 5.

Note that this assumption rules out some models, for example the case of π_{qq} equal for all q and π_{qr} equal for all $q \neq r$ and equal proportions which was studied in Decelle et al. [10] with a physical point of view.

2.3. Largest Gaps algorithm

If $(u_i)_{i\in[n]}$ is a sequence of real numbers, $(u_{(i)})_{i\in[n]}$ denotes the same sequence but sorted in increasing order.

Algorithm

- Sort the sequence of the normalized degrees in increasing order:

$$T_{(1)} \leq \cdots \leq T_{(n)}$$

- Calculate every gap between consecutive normalized degrees:

$$T_{(i+1)} - T_{(i)} \text{ for all } i \in [n-1]$$

- Find the indexes of the $Q-1$ largest gaps: $i_1 < \cdots < i_{Q-1}$, such that for all $k \in [Q-1]$ and for all $i \in [n] \setminus \{i_1, \ldots, i_{Q-1}\}$:

$$T_{(i_{k+1})} - T_{(i_k)} \geq T_{(i+1)} - T_{(i)}$$

- Noting $(i_0) = 0$ and $(i_Q) = n$, associate with each index (i) a class number: $i \mapsto k$ such that $(i_{k-1}) < (i) \leq (i_k)$.

Example

On the figure below, the largest gaps correspond to the intervals $[T_{(2)}, T_{(3)}[$, denoted by ①, and $[T_{(9)}, T_{(10)}[$, denoted by ②. Nodes (1) and (2) are therefore classified in class 1, nodes from (3) to (9) in 2, nodes (10) and (11) in 3.

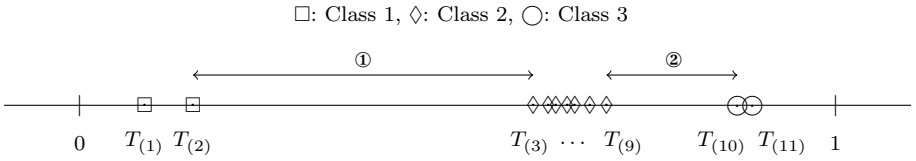

FIG 1. *Repartition of the normalized degrees.*

This algorithm has all the qualities mentioned in Introduction and makes good use of the concentration, which makes the consistency easy to prove. Whereas variational EM algorithms runs as many quadratic steps as needed to reach convergence and classical spectral clustering runs in cubic time, this algorithm is especially fast. Indeed the sorting runs in quasilinear time and although the computation of the degrees is quadratic, this is a very basic operation which is very quickly performed. Note that Condon and Karp [8] gave an algorithm running in linear time and consistent under SBM — called planted ℓ-partition model in this paper —, but provided that the weights of the blocks are equal.

2.4. Main result

The true (respectively estimated) partition of $[n]$ in classes is denoted by the set $\{\mathcal{C}_q^n\}_{q\in[Q]}$, (resp. by $\{\widehat{\mathcal{C}}_q^n\}_{q\in[Q]}$) and the cardinality of the true q-labeled class by N_q^n (resp. by $\widehat{N}_q^n$). We expect the estimated partition to be almost surely

the true partition when n is large enough. Define E_n as the event "The LG algorithm makes at least one mistake", that is:

$$E_n = \left\{ \{\widehat{\mathcal{C}}_q^n\}_q \neq \{\mathcal{C}_q^n\}_q \right\}$$

Definition 2. $\{\widehat{\mathcal{C}}_q^n\}_{q\in[Q]}$ is said to be consistent if

$$P_{\alpha,\pi}^n(E_n) \xrightarrow[n\to\infty]{} 0$$

Let us define α_0 the smallest proportion of the model:

$$\alpha_0 = \min_{q\in[Q]} \alpha_q$$

Theorem 2.2. *Under Assumption* (H),

$$P_{\alpha,\pi}^n(E_n) \leq 2ne^{-\frac{1}{8}n\delta^2} + Q(1-\alpha_0)^{n+1}$$

The proof of this theorem is given in the paragraph 3.3. Note that the bound is uniform over all models with the same δ, even though these do not behave exactly the same way. In particular the intraclass variability has a certain effect on the concentration of the node degrees of the class. Sparse models concentrate faster than models with a medium density for example.

3. Consistency proof of the LG algorithm

3.1. An ideal event for the algorithm

The LG algorithm delivers the true partition especially when none of the classes is empty, and the spreading of the normalized degrees is small compared with the minimal gap δ. A_n denotes the event "No true class is empty", that is

$$A_n = \bigcap_{q\in[Q]} \{\mathcal{C}_q^n \neq \varnothing\} = \bigcap_{q\in[Q]} \{N_q^n = 0\}$$

Definition 3. We call maximal intraclass distance (or spreading) the random variable d_n defined by:

$$d_n = \max_{q\in[Q]} \sup_{i\in\mathcal{C}_q^n} |T_i^n - \overline{\pi}_q|$$

This is the maximal distance between the normalized degree of a node and its own conditional mean, over all nodes and all classes. This is basically a measurement of the within-class spreading of the normalized degrees.

Proposition 3.0.1. *Under Assumption* (H), *the following inclusion holds for all* $\varepsilon > 0$*:*

$$A_n \cap \left\{ d_n \leq \frac{\delta}{4+\varepsilon} \right\} \subset \overline{E}_n$$

Proof. Suppose that $A_n \cap \{d_n \leq \frac{\delta}{4+\varepsilon}\}$ is true. For all $i, j \in [n]$ and $q, r \in [Q]$:

- If nodes i and j have label q, then:

$$|T_i - T_j| \leq |T_i - \overline{\pi}_q| + |T_j - \overline{\pi}_q| \leq \frac{2\delta}{4+\varepsilon}$$

- Inversely, if they have different labels, respectively q and r, then:

$$\begin{aligned} |T_i - T_j| &\geq |T_j - \overline{\pi}_q| - |T_i - \overline{\pi}_q| \\ &\geq |T_j - \overline{\pi}_q| - \frac{\delta}{4+\varepsilon} \\ &\geq |\overline{\pi}_l - \overline{\pi}_q| - |T_j - \overline{\pi}_l| - \frac{\delta}{4+\varepsilon} \\ &\geq \delta - \frac{\delta}{4+\varepsilon} - \frac{\delta}{4+\varepsilon} = \frac{2+\varepsilon}{4+\varepsilon}\delta > \frac{2\delta}{4+\varepsilon} \end{aligned}$$

As a conclusion of this alternative, i and j are in the same class if and only if $|T_i - T_j| \leq \frac{2\delta}{4+\varepsilon}$. Notice moreover that there exists exactly $Q-1$ intervals among the set $([T_i, T_j[)_{i,j}$ strictly greater than $\frac{2\delta}{4+\varepsilon}$ on this event. Hence the $Q-1$ largest intervals lie between groups of normalized degrees from different classes; whereas all others lie between degrees of the same class. In this case the algorithm returns the true partition. □

3.2. Bound of the probability of large spreading

In this paragraph we shall show that the dispersion d_n converges to 0 thanks to the subgaussian tail of the binomial distributions. This is a basic result of this article, because all others require controlling the dispersion.

Proposition 3.0.2. *For all $t > 0$:*

$$P(d_n > t) \leq 2ne^{-2nt^2}$$

Proof. It consists in conditioning by the class of each node, in order to apply the concentration inequality (CCT), and of a union bound. Since $D_i^n \sim \mathcal{B}(n, \overline{\pi}_q)$, (CCT) gave the inequality (1):

$$P(|T_i - \overline{\pi}_q| > t | Z_i = q) \leq 2e^{-2nt^2}$$

Hence:

$$\begin{aligned} P(d_n > t) &= \mathbb{E}\left(P(d_n > t|Z)\right) \\ &= \mathbb{E}\left(P\left(\cup_{q\in[Q]} \cup_{i\in\mathcal{C}_q} \{|T_i - \overline{\pi}_q| > t\}|Z\right)\right) \\ &\leq \mathbb{E}\left(\sum_{q\in[Q]}\sum_{i\in\mathcal{C}_q} P(|T_i - \overline{\pi}_q| > t|Z)\right) \\ &\leq \mathbb{E}\left(\sum_{q\in[Q]}\sum_{i\in\mathcal{C}_q} P(|T_i - \overline{\pi}_q| > t|Z_i = q)\right) \\ &\leq 2ne^{-2nt^2} \end{aligned}$$

□

Remark. Furthermore d_n almost surely converges to 0 because the upper bound is summable, by applying a usual consequence of the Borel-Cantelli lemma.

3.3. Bound of the error probability (proof of Theorem 2.2)

Thanks to the bound of the probability of large spreading, one can easily conclude that the ideal event $A_n \cap \{d \leq \frac{\delta}{4+\varepsilon}\}$ is actually strongly likely for n large enough and for all $\varepsilon > 0$:

Proof. First we have $A_n \cap \{d \leq \frac{\delta}{4+\varepsilon}\} \subset \overline{E}_n$ according to Proposition 3.0.1, hence:

$$P(E_n) \leq P\left(\overline{A_n \cap \{d_n \leq \frac{\delta}{4+\varepsilon}\}}\right) \leq P\left(d_n > \frac{\delta}{4+\varepsilon}\right) + P(\overline{A_n})$$

On the one hand, Proposition 3.0.2 implies that:

$$P\left(d_n > \frac{\delta}{4+\varepsilon}\right) \leq 2\exp\left(-2n\left(\frac{\delta}{4+\varepsilon}\right)^2\right)$$

On the other hand $\overline{A}_n$ corresponds to "There exists an empty class". For all $q \in [Q]$, $N_q \sim \mathcal{B}(n, \alpha_q)$, hence:

$$\begin{aligned} P(\overline{A}_n) &= P\left(\cup_{q\in[Q]}\{N_q = 0\}\right) \\ &\leq \sum_{q\in[Q]} P(N_q = 0) = \sum_{q\in[Q]} (1-\alpha_q)^n \leq Q(1-\alpha_0)^n. \end{aligned}$$

Once the both previous inequalities have been put together, we have an upper bound of $P(E_n)$ which depends on ε. The limit of the upper bound when ε tends to zero yields the bound of the Theorem. □

4. Consistency of the plug-in estimators

If the true classes were known, the usual moment estimators would be enough to estimate (α, π). Indeed the empirical proportions estimate α and the connection frequencies estimate the connection probabilities. We first prove that if we knew the classes, we would obtain a consistent estimate. However those variables are not observed but latent. That is why we plug the partition delivered by any consistent classification algorithm into these estimators. Notice that it does not depend on the choice of the consistent algorithm.

Notations For all q, r in $[Q]$, $\mathcal{C}_{qr}$ denotes $\mathcal{C}_q \times \mathcal{C}_r$, and N_{qr} its cardinality. If $q \neq r$, $N_{qr} = N_q N_r$ and if $q = r$, $N_{qq} = \frac{N_q(N_q-1)}{2}$. We define the following estimators:

$$\widetilde{\alpha}_q = \frac{N_q}{n} \text{ and } \widetilde{\pi}_{qr} = \frac{1}{N_{qr}} \sum_{(i,j)\in\mathcal{C}_{qr}} X_{ij}$$

Recall that all of these variables are hidden thus far.

4.1. Estimation with revealed classes

Theorem 4.1. $(\widetilde{\alpha}, \widetilde{\pi})$ *is a consistent estimator of* (α, π).

Proof. For all $q \in [Q]$, N_q is the sum of n independent Bernoulli random variables with parameter α_q. Applying directly the concentration inequality, we get for all $t > 0$ and $q \in [Q]$: $P\left(\left|\frac{N_q}{n} - \alpha_q\right| > t\right) \leq 2e^{-2nt^2}$. Applying the concentration inequality (CCT) conditionally on N_{qr} and then taking the expectation, we get for all $t > 0$:

$$P\left(|\widetilde{\pi}_{qr} - \pi_{qr}| > t\right) = \mathbb{E}\left[P\left(|\widetilde{\pi}_{qr} - \pi_{qr}| > t | N_{qr}\right)\right] \leq 2\mathbb{E}\left(e^{-2N_{qr}t^2}\right)$$

Define:

$$\alpha_{qr} = \alpha_q \alpha_r \text{ if } q \neq r \text{ and } \alpha_{qq} = \frac{\alpha_q^2}{2} \text{ if } q = r.$$

Let (r_n) be a non-negative sequence tending to infinity. We split up the support of the expectation into two pieces, depending on the values of N_{qr}. On the one hand the exponential term inside the expectation is bounded on the first piece of the support by a deterministic sequence. On the other hand, the probability of the support of the second piece of the expectation $\left\{|N_{qr} - \alpha_{qr}n^2| > r_n\right\}$ is accurately controlled by using the concentration inequality derived from (CCT) in Appendix A.

$$\begin{aligned}
\mathbb{E}\left[\exp(-2N_{qr}t^2)\right] &= \mathbb{E}\left[\exp(-2N_{qr}t^2)\mathbb{1}_{\{|N_{qr}-\alpha_{qr}n^2|\leq r_n\}}\right. \\
&\qquad \left. + \exp(-2N_{qr}t^2)\mathbb{1}_{\{|N_{qr}-\alpha_{qr}n^2|>r_n\}}\right] \\
&\leq \mathbb{E}\left[\exp(-2t^2(\alpha_{qr}n^2 - r_n))\right] + P(|N_{qr} - \alpha_{qr}n^2| > r_n) \\
&\leq \exp(-2t^2(\alpha_{qr}n^2 - r_n)) + P\left(\left|\frac{N_{qr}}{n^2} - \alpha_{qr}\right| > \frac{r_n}{n^2}\right) \\
&\leq \exp\left[-r_n t^2\left(\frac{n^2\alpha_0^2}{r_n} - 1\right)\right] + 4\exp\left(-\frac{1}{2}\frac{r_n^2}{n^3}\right) \qquad \text{(B)}
\end{aligned}$$

In order to have a vanishing bound (B), we just have to choose (r_n) such that:

$$\underline{\lim}_{n\to+\infty} \frac{\alpha_0^2 n^2}{r_n} > 1 \text{ and } \frac{r_n^2}{n^3} \xrightarrow[n\to+\infty]{} +\infty$$

For example, $r_n = n^{7/4}$, hence:

$$\mathbb{E}\left[\exp(-2N_{qr}t^2)\right] \leq \exp\left[-n^{7/4}t^2\left(n^{1/4}\alpha_0^2 - 1\right)\right] + 4\exp\left(-\frac{1}{2}\sqrt{n}\right)$$

Then we conclude with a union bound:

$$P(\|\widetilde{\pi} - \pi\|_\infty > t) \leq 2Q^2\left(e^{-n^{7/4}t^2\left(n^{1/4}\alpha_0^2 - 1\right)} + 4e^{-\frac{1}{2}\sqrt{n}}\right)$$

Finally we conclude for all parameters:

$$P(\|(\widetilde{\pi}, \widetilde{\alpha}) - (\alpha, \pi)\|_\infty > t) \leq 2Q^2\left(e^{-n^{7/4}t^2\left(n^{1/4}\alpha_0^2 - 1\right)} + 4e^{-\frac{1}{2}\sqrt{n}}\right) + 2Qe^{-2nt^2}$$

□

4.2. Estimation with hidden classes

We now assume that we have got a partition of the nodes $\{\widehat{\mathcal{C}}_q\}_q$ returned by any classification algorithm. The estimators $\widehat{\alpha}$ and $\widehat{\pi}$ are defined by plug-in with the estimated partition $\{\widehat{\mathcal{C}}_q\}_q$ instead of the true one $\{\mathcal{C}_q\}_q$. If the classification is right, then estimators both with hat and with tilde are equal.

$$\widehat{\alpha}_q = \frac{\widehat{N}_q}{n} \text{ and } \widehat{\pi}_{qr} = \frac{1}{\widehat{N}_{qr}} \sum_{(i,j)\in\widehat{\mathcal{C}}_{qr}} X_{ij}$$

Theorem 4.2. *If $\{\widehat{\mathcal{C}}_q\}_q$ is consistent, then $(\widehat{\alpha},\widehat{\pi})$ is a consistent estimator of (α,π).*

Proof. For all $t>0$, let $B_t^n = \{\|(\widehat{\alpha},\widehat{\pi})-(\alpha,\pi)\|>t\}$.

$$\begin{aligned}\forall t>0 \quad P(B_t^n) &= P(B_t^n \cap \overline{E}_n) + P(B_t^n \cap E_n)\\ &\leq P(B_t^n \cap \overline{E}_n) + P(E_n)\end{aligned}$$

On the event $\overline{E}_n$, the equality $(\widehat{\alpha},\widehat{\pi}) = (\widetilde{\alpha},\widetilde{\pi})$ holds, hence:

$$\forall t>0\ P(B_t^n) \leq P\left(\|(\widetilde{\alpha},\widetilde{\pi})-(\alpha,\pi)\|>t\right) + P(E_n).$$

The first term converges to 0 according to Theorem 4.1 and the second one as well, provided the algorithm is consistent (see Theorem 2.2). □

4.3. Conclusions

The previous paragraphs did not depend on the algorithm chosen. Now putting together the results of the previous section and the results concerning the LG algorithm, we get:

Theorem 4.3. *For all $t>0$*

$$\begin{aligned}P(\|(\widehat{\pi},\widehat{\alpha})-(\alpha,\pi)\|_\infty>t) \leq 2Q^2\left(e^{-n^2t^2\left(\alpha_0^2-n^{-1/4}\right)}+4e^{-\frac{1}{2}\sqrt{n}}\right)+2Qe^{-2nt^2}\\ +2ne^{-\frac{1}{8}n\delta^2}+Q(1-\alpha_0)^n\end{aligned}$$

Note that the estimation procedure requires larger graphs to achieve consistency than does the classification procedure with the LG algorithm alone. This is basically due to the variability of the empirical proportions.

Since the upper bound is summable, a usual consequence of the Borel-Cantelli lemma implies the strong consistency of these estimators.

5. Using LG algorithm under weak separability

The case of a weak separation of the classes is now considered, that is when δ vanishes or is exactly zero.

5.1. Convergence rates of the LG algorithm

Here the separability parameter δ is supposed to be vanishing when n is increasing. This amounts to remove asymptotically the assumption (H). Moreover the number of classes Q is supposed to be growing with n. It is actually connected because if Q is growing and all of the $\overline{\pi}_q$ are distinct at the same time, then δ is necessarily vanishing. Convergence rates ensuring LG to be consistent are provided for δ, Q and α_0, in order to illustrate up to where, at least, the algorithm theoretically works.

In this subsection only, another asymptotic framework is chosen. The parameters (α, π) are assumed to be functions of n. Consistency does not mean convergence under the distribution of $\mathcal{G}(n, \alpha, \pi)$ anymore, but under $\mathcal{G}(n, \alpha^n, \pi^n)$, with $\alpha^n = (\alpha_1^n, \ldots, \alpha_{Q_n}^n)$ and $\pi^n = (\pi_{qr}^n)_{1 \leq q,r \leq Q_n}$. It is assumed that:

$$\delta_n \xrightarrow[n\to\infty]{} 0,\ \alpha_0^n \xrightarrow[n\to\infty]{} 0 \text{ and } Q_n \xrightarrow[n\to\infty]{} +\infty$$

Proposition 5.0.1. *The classification procedure with LG algorithm is still consistent under the following assumptions:*

(a) $\underline{\lim}_{n\to+\infty} \delta_n \sqrt{\dfrac{n}{\ln n}} > 2\sqrt{2}$, *implying* $Q_n = O\left(\sqrt{\dfrac{n}{\ln n}}\right)$

(b) $\underline{\lim}_{n\to+\infty} -\dfrac{n \ln(1-\alpha_0^n)}{\ln Q_n} > 1$

For example, if $Q_n = 1 + \left\lfloor \sqrt{\frac{n}{\ln n}} \right\rfloor$, *it is sufficient that:* $\alpha_0^n \geq \frac{\ln n}{2n}$.

Proof. Assumption (a) implies that there exists $C > 2\sqrt{2}$ such that for n large enough:

$$\delta_n \sqrt{\frac{n}{\ln n}} \geq C \text{ and then } \frac{n\delta_n^2}{\ln n} - 8 \geq C^2 - 8 > 0$$

Therefore

$$\begin{aligned} n \exp\left(-\frac{1}{8} n\delta_n^2\right) &= \exp\left[-\frac{1}{8} \ln n \left(\frac{n\delta_n^2}{\ln n} - 8\right)\right] \\ &\leq \exp\left[-\frac{1}{8} \ln n \left(C^2 - 8\right)\right] \xrightarrow[n\to+\infty]{} 0 \end{aligned}$$

Secondly the model requires $(Q_n - 1)\delta_n \leq 1$ as a necessary condition. Hence, applying (a):

$$Q_n \leq 1 + \frac{1}{\delta_n} = O\left(\sqrt{\frac{n}{\ln n}}\right)$$

According to Assumption (b), there exists $C' > 1$ such that for n large enough:

$$-\frac{n \ln(1-\alpha_0^n)}{\ln Q_n} > C', \text{ so that:}$$

$$\begin{aligned} Q_n(1-\alpha_0^n)^n &= \exp\left[\ln Q_n + n\ln(1-\alpha_0^n)\right] \\ &= \exp\left[-\ln Q_n\left(\frac{-n\ln(1-\alpha_0^n)}{\ln Q_n} - 1\right)\right] \\ &\leq \exp\left(-\ln Q_n\,(C'-1)\right) \xrightarrow[n\to+\infty]{} 0 \end{aligned}$$

Thus it has been just proved that the two terms of the bound of the theorem 2.2 were vanishing, which finishes the proof. □

Large graphs are more and more sparse as n increases, which results in the decrease in the connectivity defined by $\overline{\overline{\pi}}_n = \mathbb{E}_{\alpha^n,\pi^n}(T_1^n)$. Convergence rates are now given when sparsity increases.

Proposition 5.0.2. *The LG algorithm is still consistent in the following cases:*

- $\overline{\overline{\pi}}_n = O\left(\left(\frac{\ln n}{n}\right)^{3/2}\right)$, *if* Q_n *is bounded.*
- $\overline{\overline{\pi}}_n = O\left(\sqrt{\frac{\ln n}{n}}\right)$, *if* $Q_n \sim \sqrt{\frac{n}{\ln n}}$.

Proof. We sketch the proof with the following inequality, where the right hand side estimates the connectivity of the sparsest model:

$$\overline{\overline{\pi}}_n = \sum_{q=1}^{Q_n} \alpha_q \overline{\pi}_q \geq \sum_{q=1}^{Q_n} \alpha_q^n (q-1)\delta_n \geq \alpha_0^n \frac{Q_n(Q_n-1)}{2}\delta_n$$

□

5.2. Separation of mixed classes

In this paragraph, it is supposed to be known that two average normalized degrees are equal, so that $\delta = 0$. there are Q classes and $\overline{\pi}_q = \overline{\pi}_r$ for some q and r. For the sake of simplicity, all other conditional averages are assumed to be pairwise distinct.

The LG algorithm can be previously applied to the graph with the input parameter $Q-1$. The $Q-1$ groups returned by LG are asymptotically the true classes, except classes q and r, which are mixed together in one group of nodes, denoted by $M \subset [n]$.

We shall briefly explain a procedure to separate this group, using the concentration of some additional binomial variables, namely the number of common neighbors of each pair of nodes (or number of paths of length 2 between each pair of nodes). Since there is a quadratic number of node pairs, this is not as fast as our procedure using degrees only.

Note that the paths of length 2 have been considered in the stochastic block model in some papers, for spectral clustering in Rohe et al. [24] or for parameter estimates in Ambroise and Matias [2]. More general motifs are also studied in Bickel et al. [4].

Notation. Define $\underline{\alpha}$ the diagonal matrix the diagonal coefficients of which are $(\alpha_q)_{q\in[Q]}$ and the bilinear map on $\mathbb{R}^Q$:

$$\langle\cdot,\cdot\rangle_\alpha : (X,Y) \mapsto {}^tX\underline{\alpha}Y$$

which is a scalar product, as soon as α_q is non-negative for all q. $\|\cdot\|_\alpha$ denotes the associated norm.

For all pairs of nodes $(i,j) \in M \times M$, define

$$D_{ij} = \sum_{k\neq i,j} Y_{ijk}, \text{ where } Y_{ijk} = X_{ik}X_{jk}.$$

Y_{ijk} is a Bernoulli distributed variable, that equals one if and only if i *and* j are both connected to k. Its parameter conditionally depends on each class of nodes i and j:

- If i and j both belong to the q-labeled class:

$$P(Y_{ijk} = 1 | Z_i = Z_j = q) = \sum_{l=1} \alpha_l \pi_{ql}^2 = \|\pi_q\|_\alpha^2$$

  where π_q is the row vector $(\pi_{ql})_l$. Symmetrically, if they both belong to the r-labeled class, the parameter is $\|\pi_r\|_\alpha^2$.
- Otherwise, if they belong to distinct classes $q \neq r$:

$$P(Y_{ijk} = 1 | Z_i = q, Z_j = r) = \sum_{l=1}^{Q} \alpha_l \pi_{ql}\pi_{rl} = \langle \pi_q, \pi_r\rangle_\alpha$$

The behavior of the new variables D_{ij} looks like that of the degrees; they once more quickly concentrate around their average value as a consequence of the concentration of binomial variables. There are three groups of node pairs, concentrating around $\|\pi_q\|_\alpha^2$, $\|\pi_r\|_\alpha^2$, or $\langle\pi_q,\pi_r\rangle_\alpha$. The first two contain only pairs of nodes of the same membership, whereas the last one is made up of pairs of nodes of different memberships.

Up to a label switch, it can be supposed that $\|\pi_q\|_\alpha \leq \|\pi_r\|_\alpha$. The following lemma shows that the group with pairs of nodes of different memberships is well separated from one of the other two. This will be sufficient to separate classes q and r.

Lemma 5.1.

$$0 \leq \langle\pi_q,\pi_r\rangle_\alpha < \|\pi_r\|_\alpha^2$$

Proof. First of all $\langle\pi_q,\pi_r\rangle_\alpha \geq 0$ because this was defined as a probability. Then, by applying the Cauchy-Schwarz inequality:

$$\langle\pi_q,\pi_r\rangle_\alpha \leq \|\pi_q\|_\alpha\|\pi_r\|_\alpha \leq \|\pi_r\|_\alpha^2$$

The case of equality in the Cauchy-Schwarz inequality cannot arise; if it did, then π_q and π_r would be collinear vectors. Noting c the constant of collinearity, it would yield $\overline{\pi}_q = c\overline{\pi}_r$. But $\overline{\pi}_q$ and $\overline{\pi}_r$ are assumed to be equal in this section; hence $c = 1$. π_q and π_r would be equal. This is not allowed by the model for identifiability reasons. The inequality is eventually strict. □

Now the LG algorithm is applied to the set of variables $(D_{ij})_{i,j\in M}$ with $Q = 2$ as input parameter. Define W as the set of the pairs which are returned in the second group — the groups being sorted in increasing order — and F as the set of nodes, which are involved in those pairs. Let K be the graph defined by (F, W).

Note that K has no obvious relation to the observed graph X. An edge between $i \in M$ and $j \in M$ just means that the pair (i, j) has been classified in the second group by LG.

Proposition 5.1.1. *In the graph K there are edges only between nodes from the same class with high probability when n is large enough. As a consequence K is asymptotically made of one or two cliques and each clique of K is made of all nodes from either class q or class r.*

Proof. There are two major cases, depending on the relative position of $\|\pi_q\|_\alpha^2$, $\|\pi_r\|_\alpha^2$ and $\langle \pi_q, \pi_r \rangle_\alpha$.

- If $\|\pi_q\|_\alpha^2 \leq \langle \pi_q, \pi_r \rangle_\alpha < \|\pi_r\|_\alpha^2$, the gap between $\|\pi_q\|_\alpha^2$ and $\langle \pi_q, \pi_r \rangle_\alpha$ is actually strictly smaller than the gap between $\langle \pi_q, \pi_r \rangle_\alpha$ and $\|\pi_q\|_\alpha^2$:

$$\begin{aligned}\|\pi_r\|_\alpha^2 - \langle \pi_q, \pi_r \rangle_\alpha - \left(\langle \pi_q, \pi_r \rangle_\alpha - \|\pi_q\|_\alpha^2\right) \\ = \|\pi_q\|_\alpha^2 + \|\pi_r\|_\alpha^2 - 2\langle \pi_q, \pi_r \rangle_\alpha = \|\pi_q - \pi_r\|_\alpha^2 > 0\end{aligned}$$

  As a consequence, LG selects asymptotically the gap between $\langle \pi_q, \pi_r \rangle_\alpha$ and $\|\pi_r\|_\alpha^2$ as the largest one. Then the second group returned by LG is asymptotically made up of the node pairs concentrated around $\|\pi_r\|_\alpha^2$, i.e. the pairs of nodes from class r. K forms asymptotically one clique, which is made up of all nodes from class r.
- If $\langle \pi_q, \pi_r \rangle_\alpha < \|\pi_q\|_\alpha^2 \leq \|\pi_r\|_\alpha^2$, LG selects asymptotically either the gap between $\|\pi_q\|_\alpha^2$ and $\|\pi_r\|_\alpha^2$ and then there is only one clique as in the previous case, or the gap between $\langle \pi_q, \pi_r \rangle_\alpha$ and $\|\pi_q\|_\alpha^2$ and then the second group returned is made up of the node pairs concentrated around $\|\pi_q\|_\alpha^2$ and $\|\pi_r\|_\alpha^2$. There are two cliques and each one corresponds to one class.

□

Since the content of one of the two classes is known, the node group M which contains nodes with mixed memberships can be separated for large enough n.

Remark. Here it is supposed to be known that $\overline{\pi}_q = \overline{\pi}_r$. However we do not provide here any procedure to know if the averages degrees are really equal. Further developments would be needed to test this hypothesis, using the size of the tail of the observed distribution of the variables $(D_{ij})_{ij\in M}$ for instance.

Indeed these variables concentrate around only one value when there is only one class, and around several values when there are more than one class.

6. Simulation study

Our main purpose in this study is to figure out how the LG algorithm behaves in practice, and above all, to check whether the bounds of Theorem 2.2 are pessimistic or not. The empirical frequency of the graphs with no error would be of great interest, because that is the quantity the bound concerns. But actually this frequency has no smooth evolution: it suddenly shifts from 0 to almost 1. We shall use two types of classification error rates: a global one and one for each class, so as to examine more accurately the results given by the algorithm.

Moreover the results of LG are compared with these of the variational method [9], which is available online in the packages MixNet[1], MixeR[2] and WMixnet[3]. The latter has been chosen in the current simulation study. In WMixnet the variational EM-algorithm (VEM) is initialized by a spectral clustering algorithm [24]. VEM can be additionally run several times with multiple reinitializations in order to prevent from getting caught in a local maximum. WMixnet also proproses a smoothing option working the following way. VEM algorithm is run with several values of Q. As soon as the likelihood is nonincreasing or the ICL criterion is not convex with respect to Q, the VEM is run once more for the problematic values of Q. It is basically reinitialized with the classification returned by VEM either for $Q-1$ classes after having split one class or for $Q+1$ classes after having merged two classes.

The results will be given with and without smoothing.

6.1. Simulation design

The parameters used in the simulation are:

$$\alpha = (0.3\ 0.55\ 0.15) \quad \pi = \begin{pmatrix} 0.03 & 0.02 & 0.045 \\ 0.02 & 0.05 & 0.09 \\ 0.045 & 0.09 & 0.25 \end{pmatrix}$$

Hence $\overline{\pi} = (0.0267\ 0.047\ 0.1005)$ and $\delta = 0.0203$. The parameters have been chosen so that the graphs are relatively sparse.

200 graphs are drawn from the model $\mathcal{G}(n, \alpha, \pi)$ for n from 200 to 11000. Then both LG and WMixnet are applied to each graph so as to obtain the node classification and the parameter estimators. Above 3000 nodes (respectively 5600) WMixnet with smoothing (resp. without) turned out too slow to be run in reasonable time. However it has already converged from $n = 2200$ nodes (resp. $n = 5200$).

[1]See at http://stat.genopole.cnrs.fr/logiciels/mixnet.
[2]See at http://cran.r-project.org/web/packages/mixer/index.html.
[3]See at http://ssbgroup.fr/mixnet/wmixnet.html.

The evolutions of the classification error rates and the estimators with respect to the number of nodes n are averaged over the 200 graphs and displayed from 200 to 11000 nodes for the LG algorithm, and to 5600 nodes for the variational method.

Error rates First of all, the global error rate g_n is defined as the proportion of node pairs (i, j), either classified in distinct classes whereas their true labels are identical, or classified together whereas their true labels are different. That is, denoting $\widehat{Z}$ the label vector returned by the classification algorithm:

$$g_n(Z, \widehat{Z}) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \left(\mathbb{1}_{Z_i = Z_j} \mathbb{1}_{\widehat{Z}_i \neq \widehat{Z}_j} + \mathbb{1}_{Z_i \neq Z_j} \mathbb{1}_{\widehat{Z}_i = \widehat{Z}_j} \right)$$

Secondly, we also propose error rates per class. Define I_q, resp. M_q, the rate of intruders (or false positive rate) in the class q predicted by the algorithm, resp. the rate of missing nodes of the true class q (or false negative rate):

$$I_q^n(Z, \widehat{Z}) = \frac{1}{\widehat{N}_q} \sum_{i \in \widehat{\mathcal{C}}_q} \mathbb{1}_{Z_i \neq q} \text{ and } M_q^n(Z, \widehat{Z}) = \frac{1}{N_q} \sum_{i \in \mathcal{C}_q} \mathbb{1}_{\widehat{Z}_i \neq q}$$

Labels will be allocated to the nodes in order of increasing degree in the classification algorithms. Indeed the true labels are expected to be sorted this way, because $\overline{\pi}_1 < \overline{\pi}_2 < \overline{\pi}_3$. This partially solves the label switching problem which arises when trying to identify the true labels instead of the equivalence classes.

6.2. Results

The evolution is quite satisfactory because the global error rate g_n of LG completely vanishes from $n = 8600$ nodes, which is even earlier than expected from the bound of Theorem 2.2 (see Figures 2 and 3). Indeed this bound predicted

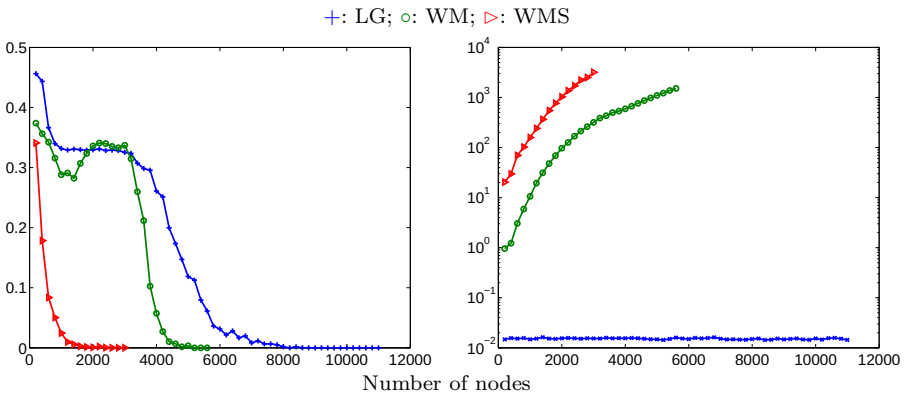

FIG 2. *Error rates g_n and running time as functions of the graph size n.*

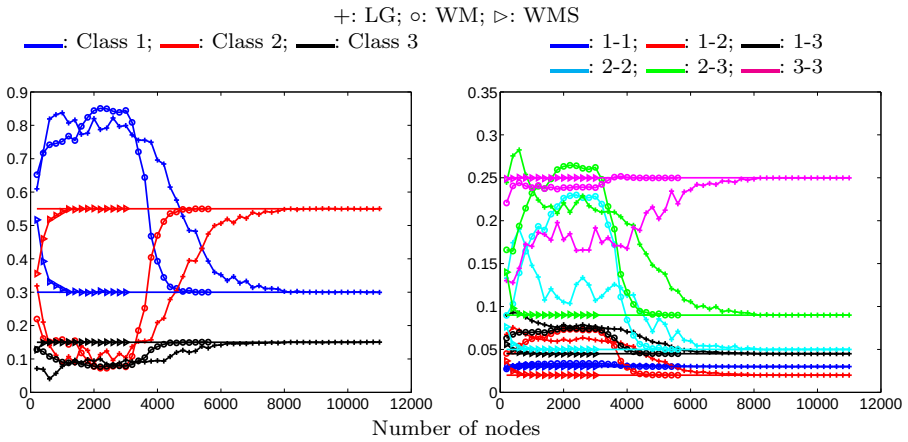

FIG 3. *Means of the estimators*

that the probability of at least one error would not be less than 0.05 earlier than $n = 300000$. The bound seems to be pessimistic, basically because of the union bound, used in the proof of Proposition 3.0.2. Note nevertheless that (see also the remark following the theorem 2.2) since the model is relatively sparse, the classification is not as intricate as for models with medium density. For instance we have also tried a model with $\delta = 0.02$ and average normalized degrees close to 0.6, and the global error rate vanished only from $n = 40000$ nodes (not shown here).

WMixnet with smoothing (WMS) converges so fast that its error rate completely vanished from $n = 2400$ nodes, much earlier than LG. Up to $n = 3000$, both WMixnet without smoothing (WM) and LG return poor and very similar results. Then the error rate of WM suddenly vanishes from $n = 5200$ nodes. Thus there is a gap between $n = 5600$ and $n = 8600$ where WMixnet is hardly usable and LG does not provide good results.

The running time of LG seems to be constant with respect to n, because the asymptotical regime (quasilinear) has not been reached yet, whereas these of the WMixnet algorithms are dramatically increasing.

Transitional phase of LG Now the behavior of LG alone is more accurately discussed. After a dramatic decrease of the error rate of LG at the beginning, its evolution encounters a slight stagnation between $n = 1000$ and $n = 3000$ nodes (see Figure 4). An interpretation of this transitional phase of LG is given using the error rates per class.

The third class is much better detected even at small graph sizes, unlike class 1 and class 2. Indeed it is sufficient that the maximal intraclass distance d_n is less than $(\overline{\pi}_3 - \overline{\pi}_2)/4$ to detect this class, whereas the other two are not supposed to be separated before

$$d_n < \frac{\overline{\pi}_2 - \overline{\pi}_1}{4} = \frac{\delta}{4} < \frac{\overline{\pi}_3 - \overline{\pi}_2}{4}$$

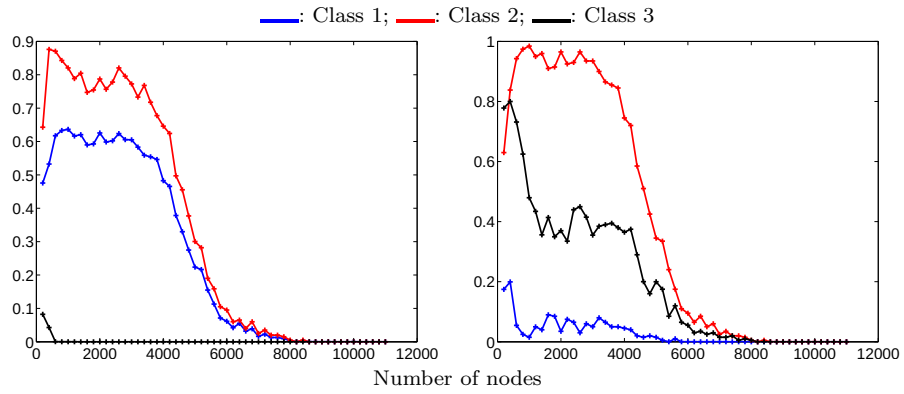

FIG 4. *Error rates I_q^n and M_q^n as functions of n for LG.*

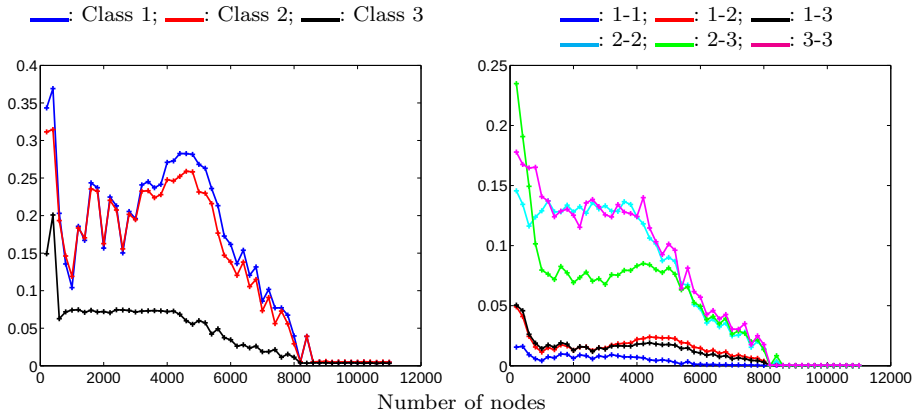

FIG 5. *Standard deviations of the estimators for LG.*

according to our previous study. That is the reason why the global error rate dramatically decreases until reaching $n = 1000$ nodes, and why it does not decrease anymore before reaching $n = 3000$. Note that the bound of Theorems 3.0.2 and 2.2 had not predicted this before reaching $n = 39000$ and $n = 317000$ respectively.

In short, as long as the tails of the normalized degree distribution are overlapping, the classes are mixed and cannot be properly detected. The curves show in particular that most nodes of class 2 seem to be caught by class 1, since there are many missing nodes in class 2 (the biggest class) and many intruders in class 1. Thus the proportion of class 2 is underestimated in the transitional phase, whereas the proportion of class 1 is overestimated. This inversion is clearly shown again on the graphics of the estimates (see Figure 3) and results in a high variability of the estimators (see Figure 5) related to these classes.

Moreover missing nodes of class 3 must be caught by class 2 as well. Therefore interconnectivity 2-3 is overestimated, close to 3-3, and the estimator has a

strong variability. Moreover the graphics related to the connectivities estimates (on Figure 3 and Figure 5) illustrates that as long as there is a lot of missing nodes in a class, the intraconnectivity estimator of the class is not good and has a high variability (see the curves of 2-2 and 3-3, unlike 1-1).

6.3. Conclusions

As a conclusion of this practical comparison, the LG algorithm should be used only for very large graphs, when nothing else is computationally feasible. LG can deal with millions of nodes on the same computers we used for the current simulation. For small graphs, other techniques provide better results.

This algorithm lacks robustness because it takes every normalized degree into account and each one carries the same weight, even if it is isolated and not statistically representative. In the worst case, one untypical node is sufficient to trick the algorithm, making the classification wrong by a majority. This often arises at small graph size in generated data and may occur at any size also in real data.

7. Model selection

Up to this section, the number of classes was supposed to be known and was an input parameter of the LG algorithm. Our main purpose hereafter is to examine more accurately the sequence of the gaps sorted in increasing order and then the sequence of the intervals between the means of the groups given by the LG algorithm, depending on the selected number of classes Q for the model. As an application of this study, we finally show that degrees are once more sufficient to select the right number of classes for large enough n.

7.1. Study of the gap sequence

We will use the same notations as in the last section. Moreover Q_0 denotes the true number of classes, and Q the current input parameter of the LG algorithm. We will often use the event $B_n = A_n \cap \{d_n \leq \frac{\delta}{5}\}$, where no class is empty and the dispersion d_n is so small that the $Q_0 - 1$ largest intervals separate the true classes (see Proposition 3.0.1 with $\varepsilon = 1$). Then we can affirm that two normalized degrees are in the same class if and only if their distance is less than $2d_n$.

Let $(G_q^n)_{q\in[n-1]}$ be the sequence of the distances between consecutive normalized degrees $(T_{(i+1)}^n - T_{(i)}^n)_{i\in[n-1]}$, but sorted in decreasing order:

$$G_1^n \geq G_2^n \geq \cdots \geq G_{n-1}^n$$

The Q_0-1 largest gaps in the LG algorithm have lengths $G_1, \ldots, G_{Q_0-1}$. Define also $(\gamma_q)_{q\in[Q_0-1]}$ the sequence $(\overline{\pi}_{(q+1)} - \overline{\pi}_{(q)})_{q\in[Q_0-1]}$, sorted in decreasing order. This is called the sequence of the theoretical gaps. The following theorem states

that largest empirical gaps converge to the corresponding theoretical gaps, which enforces our intuition about the model.

Theorem 7.1. *For all $q < Q_0$, $G_q \xrightarrow[n\to+\infty]{} \gamma_q$ a.s.*

Refer to Appendix B to see the proof. One can easily realize that the only gap (among the $Q_0 - 1$ largest) lying between $\overline{\pi}_{(q)}$ and $\overline{\pi}_{(q+1)}$ converges to $\overline{\pi}_{(q+1)} - \overline{\pi}_{(q)}$. However the index of this interval is random and depends on n. This interesting but technical problem is solved in the second part of the proof. For the moment we provide a weaker version of this theorem, the proof of which is much simpler. Its conclusion is sufficient for our purposes.

Theorem 7.2. *For all $q < Q_0$, $\underline{\lim}_{n\to+\infty} G_q > 0$*

Proof. If $q < Q_0$: on the event B_n, the $Q_0 - 1$ largest intervals necessarily lie between normalized degrees from different classes. There exists $i \in \mathcal{C}_r$ and $j \in \mathcal{C}_s$, where $s \neq r$ such that $G_q = |T_i - T_j|$. But $|T_i - \overline{\pi}_r| \leq d_n$ and $|T_j - \overline{\pi}_s| \leq d_n$, hence

$$G_q \geq |\overline{\pi}_r - \overline{\pi}_s| - 2d_n \geq \delta - \frac{2}{5}\delta = \frac{3}{5}\delta > 0$$

Namely $B_n \subset \{G_q \geq \frac{3}{5}\delta\}$.

$$P\left(G_q < \frac{3}{5}\delta\right) \leq P\left(\overline{B}_n\right) \leq 2e^{-\frac{2}{25}n\delta^2} + Q_0(1-\alpha_0)^n$$

As the upper bound is summable, according to the Borel-Cantelli lemma,

$$P\left(\overline{\lim_{n\to+\infty}}\{G_q < \frac{3}{5}\delta\}\right) = 0$$

Therefore $\underline{\lim}_{n\to+\infty} G_q \geq \frac{3}{5}\delta > 0$ almost surely. □

All further gaps lie between degrees of nodes of the same class and then converge to zero. The next theorem gives an estimation of the convergence rate.

Theorem 7.3. *For all $\beta \in]0,1[$, the triangular array*

$$\{n^{\frac{1-\beta}{2}} G_q^n; Q_0 \leq q \leq n-1\}$$

converges uniformly w.r.t. q and a.s. to zero when n tends to infinity.

Proof. First of all, recall that for all n,

$$G_{Q_0}^n \geq G_{Q_0+1}^n \geq \cdots \geq G_{n-1}^n \geq 0$$

Therefore we can just prove that $n^{\frac{1-\beta}{2}} G_{Q_0} \xrightarrow[n\to+\infty]{} 0$, and the uniform convergence will follow.

On the event B_n, the $Q_0 - 1$ largest intervals lie between normalized degrees from different classes. The next intervals lie between degrees from the same

class, and the distance to their corresponding conditional mean is at most d_n. As G_{Q_0} is one of these, $G_{Q_0} \leq 2d_n$. Hence, for all $0 < t < \frac{\delta}{5}$:

$$\begin{aligned} P\left(n^{\frac{1-\beta}{2}} G_{Q_0} > t\right) &= P(n^{\frac{1-\beta}{2}} G_{Q_0} > t \cap B_n) + P\left(n^{\frac{1-\beta}{2}} G_{Q_0} > t \cap \overline{B}_n\right) \\ &\leq P(2n^{\frac{1-\beta}{2}} d_n > t) + P\left(\overline{B}_n\right) \\ &\leq 2(e^{-\frac{1}{2}n^{\beta}t^2} + e^{-\frac{2}{25}n\delta^2}) + Q_0(1-\alpha_0)^n \end{aligned}$$

□

7.2. Study of the intervals between estimated classes

By distances between estimated classes, we mean distances between empirical averages of the normalized degrees of each class, provided by the LG algorithm. Define m_q to be the average of the normalized degrees of the q-labeled class estimated by the algorithm:

$$m_q = \frac{1}{N_q} \sum_{i \in \widehat{\mathcal{C}}_q} T_i$$

The sequence of the gaps between consecutive averages $(m_{(q+1)} - m_{(q)})_{q \in [Q-1]}$ is sorted in order of decreasing length, just as the sequence of the gaps $(T_{(i+1)} - T_{(i)})_{i \in [n-1]}$ is in the previous paragraph. This new sequence is denoted by $(H_q^n)_{q \in [Q-1]}$. Of course it depends on the current Q, whereas $(G_q)_q$ does not.

When $Q = Q_0$, H_q and G_q are very close for all $q \leq Q_0 - 1$. On the contrary, when $Q < Q_0$, some of the $(H_q)_{q \in [Q_0 - 1]}$ stretch over several classes and include more than one of the G_q. As a result, there is at least one q such that H_q differs from G_q for large enough n.

Theorem 7.4.

1. *If* $Q = Q_0$, *then* $\sum_{q=1}^{Q-1} (H_q - G_q) \xrightarrow[n \to +\infty]{a.s.} 0$
2. *If* $Q < Q_0$, *then* $\varliminf_{n \to +\infty} \sum_{q=1}^{Q-1} (H_q - G_q) > 0$ *a.s.*

Proof. Let $(J_q)_{q \in [Q_0 - 1]}$ the $Q_0 - 1$ largest intervals between consecutive normalized degrees, hence for all q, $|J_q| = G_q$. Define also $J'_0 = [0, \min_{i \in [n]} T_i[$ and $J'_Q = [\max_{i \in [n]} T_i, 1[$. The union of $J'_0, J_1, \ldots, J_{Q-1}, J'_Q$ partially covers the interval $[0, 1[$. These intervals are separated and the distance between the bounds of consecutive intervals is at most $2d_n$. As a result:

$$1 - 2Q_0 d_n \leq \sum_{q=1}^{Q_0 - 1} G_q + H_0 + H_Q \leq 1 = \sum_{q=0}^{Q} H_q$$

1. If $Q = Q_0$, subtracting the right-hand side (which actually equals 1), we deduce from both previous inequalities that:

$$-2Q_0 d_n \leq \sum_{q=1}^{Q_0-1} (G_q - H_q) \leq 0$$

The first assertion follows directly from this inequality; for all $t > 0$:

$$P\left(\left|\sum_{q=1}^{Q_0-1} (H_q - G_q)\right| > t\right) \leq P(2Q_0 d_n > t)$$
$$\leq 2\exp\left(-2n\left(\frac{t}{2Q_0}\right)^2\right) = 2\exp\left(-\frac{1}{2Q_0^2} n t^2\right)$$

2. If $Q < Q_0$, subtracting the right-hand side from the second inequality yields:

$$\sum_{q=Q}^{Q_0-1} G_q \leq \sum_{q=1}^{Q-1} (H_q - G_q)$$

But as shown in Theorem 7.2, the lower limit of G_q is non-negative for all $q \leq Q_0 - 1$. *A fortiori*, the second assertion of the theorem 7.4 follows.

□

7.3. Application to model selection

To sum up the previous paragraph, when Q is the right number of classes, the quantity $\sum_{q=1}^{Q-1}(H_q - G_q)$ converges to zero, and when Q is too small, it converges to a non-negative value, because one of the H_q does not match G_q. Thus this quantity measures the risk of underestimating the number of classes.

However, its minimization over all $Q \in \{2, \ldots, n\}$ yields the unexpected solution $Q = n$, for all Q_0. Therefore we have to penalize overly small gaps. We chose to use an *ad hoc* penalty, that can be easily inferred from our previous study. Define for all $Q \in \{2, \ldots, n\}$:

$$f_Q = \sum_{q=1}^{Q-1} (H_q - G_q) + \frac{1}{n^{\frac{1-\beta}{2}} G_{Q-1}} \in [0, +\infty] \text{ where } \beta \in]0, 1[.$$

Theorem 7.5.

1. *If* $Q = Q_0$, *then* $f_Q \xrightarrow[n\to+\infty]{a.s.} 0$
2. *If* $Q < Q_0$, *then* $\underline{\lim}_{n\to+\infty} f_Q > 0$ *a.s.*
3. *If* $Q > Q_0$, *then* $f_Q \xrightarrow[n\to+\infty]{a.s.} +\infty$ *uniformly w.r.t.* Q

It follows that $\widehat{Q} = \arg\min_{2\leq Q\leq n} f_Q \xrightarrow[n\to+\infty]{} Q_0$ *a.s.*

Proof.

1. If $Q = Q_0$, applying Theorem 7.4, the sum $\sum_{q=1}^{Q-1}(H_q - G_q)$ converges a.s. to 0. According to Theorem 7.2, $\underline{\lim}_{n\to+\infty} G_{Q_0-1} > 0$ almost surely. Therefore:

$$\frac{1}{n^{\frac{1-\beta}{2}} G_{Q-1}} \xrightarrow[n\to+\infty]{a.s.} 0, \text{ and then } f_Q \xrightarrow[n\to+\infty]{a.s.} 0$$

2. If $Q < Q_0$, according to the second assertion of Theorem 7.4, the lower limit of the first term is non-negative. There is no change by adding the second term, because it is positive. Hence:

$$\underline{\lim}_{n\to+\infty} f_Q > 0$$

3. If $Q > Q_0$, the sum $\sum_{q=1}^{Q-1} H_q - G_q$ is lower bounded by -1 (notice that it is even positive), and according to the second assertion of Theorem 7.3, $(n^{\frac{1-\beta}{2}} G_{Q-1})_n$ uniformly converges to 0 w.r.t. $Q > Q_0$. The last assertion follows.

□

8. Conclusions

Unlike most of the methods known thus far, the LG algorithm is able to process very large graphs. In fact it provides good results only for such graphs. Nevertheless, according to the simulation study, the algorithm is efficient even for smaller graphs than theoretically expected. Moreover it is self-sufficient: it provides consistent methods for node clustering, parameter estimation and model selection. It performs every task using the degree data alone. Lastly, this algorithm is free from any preliminary setting. There is need neither for any prior knowledge nor for multiple runnings of the algorithm. Thus it can quickly provide initialization values for other algorithms which depend severely on them.

However other techniques provide better results for small graph size and it does not seem to be a practical method for real data, because of the lack of robustness above all.

As a conclusion, the LG algorithm is a good theoretical tool which proves this statement: for large enough n, when the average degrees are separated enough, the degree data alone is a sufficient statistics to achieve all of the statistical inference under SBM.

Acknowledgements

We thank the reviewers for their helpful comments and remarks, and we also thank Jean-Benoist Léger for the help to use his package WMixnet.

Appendix A: Concentration inequality for products of binomial distributed variables

Proposition A.0.1. *Let X (respectively Y) be a sum of n independent bernoulli distributed variables with parameter p, respectively q. Then for all $t > 0$*

$$P\left(\left|\frac{XY}{n^2} - pq\right| > t\right) \leq 4\exp(-\frac{1}{2}nt^2)$$

Proof.

$$\begin{aligned} P\left(\left|\frac{XY}{n^2} - pq\right| > t\right) &= P\left(\left|\left(\frac{X}{n} - p\right)\frac{Y}{n} + \left(\frac{Y}{n} - q\right)p\right| > t\right) \\ &\leq P\left(\left|\frac{X}{n} - p\right|\frac{Y}{n} > \frac{t}{2}\right) + P\left(\left|\frac{Y}{n} - q\right|p > \frac{t}{2}\right) \\ &\leq P\left(\left|\frac{X}{n} - p\right| > \frac{t}{2}\right) + P\left(\left|\frac{Y}{n} - q\right| > \frac{t}{2}\right) \\ &\leq 2 \times 2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) = 4\exp(-\frac{1}{2}nt^2) \end{aligned}$$

The last line is obtained by applying the usual concentration inequality (CCT) to both X and Y. □

With a similar proof, we prove that for all $t \in]0, 1/4]$:

$$P\left(\left|\frac{X(X-1)}{2n^2} - \frac{\alpha^2}{2}\right| > t\right) \leq 4\exp\left(-2nt^2\right)$$

Appendix B: Proof of Theorem 7.1

Let us define $(J_i)_{i\in[n]}$ the sequence of the intervals $[T_{(i)}, T_{(i+1)}[$ sorted in order of decreasing length, hence for all $i \in [n]$, $|J_i| = G_i$. We suppose hereafter that the sequence $(\overline{\pi}_q)_q$ is sorted in increasing order: $\overline{\pi}_1 < \cdots < \overline{\pi}_Q$.

Proof. On the event B_n, among the $Q_0 - 1$ largest intervals, we can associate with each $\overline{\pi}_q$ the only one lying between $\overline{\pi}_q$ and $\overline{\pi}_{q+1}$. Namely the only J_i with $i \in [Q_0 - 1]$ such that $J_i \cap]\overline{\pi}_q, \overline{\pi}_{q+1}[\neq \varnothing$. $S(q)$ denotes the index in $[Q_0 - 1]$ corresponding to this unique interval.

Moreover, $s(q)$ denotes one of the indexes $s \in [Q_0 - 1]$ such that $\gamma_s = \overline{\pi}_{q+1} - \overline{\pi}_q$, chosen so that s is injective. Let us point out that S is a random permutation whereas s is deterministic. In order to simplify notations, we silently make the deterministic index change $r = s(q)$. Thereby $(\gamma_q)_q$ still denotes the sequence $(\gamma_{s(q)})_q$, and S the permutation $S \circ s^{-1}$.

Notice that on B_n and especially when $d_n \leq \frac{\delta}{5}$:

$$[\overline{\pi}_q + d_n, \overline{\pi}_{q+1} - d_n] \subset J_{S(q)} \subset [\overline{\pi}_q - d_n, \overline{\pi}_{q+1} + d_n]$$

$$\text{Hence } |G_{S(q)} - \gamma_q| \leq 2d_n. \tag{2}$$

1. We first prove that the gap $G_{S(q)}$ converges to the theoretical gap γ_q. For all $t > 0$:

$$\begin{aligned} P(|G_{S(q)} - \gamma_q| > t) &= P(|G_{S(q)} - \gamma_q| > t \cap B_n) + P(|G_{S(q)} - \gamma_q| > t \cap \overline{B}_n) \\ &\leq P(2d_n > t) + P(\overline{B}_n) \\ &\leq 2(e^{-\frac{1}{2}nt^2} + e^{-\frac{2}{25}n\delta^2}) + Q_0(1-\alpha_0)^n \end{aligned} \tag{3}$$

2. Secondly, none of the $Q_0 - 1$ largest intervals permute anymore expect for those having the same theoretical values. It follows from the inequality (2) that for all $q, r \in [Q_0 - 1]$,

$$\gamma_q - \gamma_r - 4d_n \leq G_{S(q)} - G_{S(r)} \leq \gamma_q - \gamma_r + 4d_n$$

Define $\eta = \frac{1}{5}(\min_{q\in[Q]}(\gamma_q - \gamma_{q+1}) \wedge \delta)$, a threshold designed to distinguish distances converging to one value from those converging to another. On the event $d_n \leq \eta$, the previous inequality yields:

$$\gamma_q - \gamma_r - 4\eta \leq G_{S(q)} - G_{S(r)} \leq \gamma_q - \gamma_r + 4\eta$$

- If $\gamma_q - \gamma_r < 0$, then $\gamma_q - \gamma_r + 4\eta < 0$ is also true by the definition of η. As a result of the inequality just above, $G_{S(q)} - G_{S(r)} < 0$.
- If $\gamma_q - \gamma_r > 0$, then $\gamma_q - \gamma_r - 4\eta > 0$, and $G_{S(q)} - G_{S(r)} > 0$.

If $(u_i)_{1\leq i\leq m}$ is a sequence, we write $i \sim_u j$ if and only if $u_i = u_j$. $\sim_u$ is an equivalence relation. Applying the Lemma B.1 stated and proved afterwards, if $d_n \leq \eta$, there exists $r \sim_\gamma q$ such that $q = S(r)$. Notice furthermore that the sequence $(\gamma_q)_{q\in[Q_0-1]}$ is constant on the $\sim_\gamma$-equivalence classes. The term $|G_q - \gamma_q|$ is necessarily in the sum $\sum_{r\sim q}|G_{S(r)} - \gamma_r|$. Finally, define

$$\begin{aligned} P(|G_q - \gamma_q| > t) &= P(|G_q - \gamma_q| > t \cap B_n) + P(|G_q - \gamma_q| > t \cap \overline{B}_n) \\ &\leq P\left(\sum_{r\sim q}|G_{S(r)} - \gamma_r| > t\right) + P(\overline{B}_n) \\ &\leq \sum_{r\sim q} P\left(|G_{S(r)} - \gamma_r| > \frac{t}{Q_0}\right) + P(\overline{B}_n) \\ &\leq 2Q_0(e^{-\frac{1}{2Q_0^2}nt^2} + e^{-\frac{2}{25}n\delta^2}) + 2e^{-2n\eta^2} \text{ according to (3).} \end{aligned}$$

□

Lemma B.1. *Let $(u_i)_{1\leq i\leq m}, (v_i)_{1\leq i\leq m}$ be two real decreasing sequences. Let p be the number of $\sim_u$-equivalence classes and σ one permutation of $\{1, \dots, m\}$. We especially assume that for all $i, j \in \{1, \dots, m\}$,*

- $u_i < u_j \Rightarrow v_{\sigma(i)} < v_{\sigma(j)}$
- $u_i > u_j \Rightarrow v_{\sigma(i)} > v_{\sigma(j)}$

Then $\sigma = \sigma_1 \circ \cdots \circ \sigma_p$ where the support of σ_i is the i^{th} $\sim_u$-equivalence class.

Proof. Since u is decreasing, the $\sim_u$-equivalence classes are just sets of consecutive natural integers. Define recursively $(r_i)_{1\leq i\leq p}$ the increasing sequence of indexes j when the value of u_j changes:

- Let $r_1 = 1$.
- For $i \geq 1$, let r_{i+1} be the smallest integer $j > r_i$ such that $u_{r_i} = \cdots = u_{j-1} > u_j$.

The construction of $(r_i)_i$ implies that for all $j < r_i$, all $r_i \leq l < r_{i+1}$ and all $k \geq r_{i+1}$: $u_j < u_k < u_l$, and furthermore $v_{\sigma(j)} < v_{\sigma(k)} < v_{\sigma(l)}$ as well. As v decreases, $\sigma(\{r_i, \ldots, r_{i+1} - 1\}) = \{r_i, \ldots, r_{i+1} - 1\}$. The result follows directly from this. □

References

[1] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002. MR1895096

[2] C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011. MR2885838

[3] P.J. Bickel and A. Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106 (50):21068, 2009.

[4] P.J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):38–59, 2011. MR2906868

[5] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1): 3–122, 2007. ISSN 1098-2418. MR2337396

[6] A. Celisse, J.-J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statist.*, 6:1847–1899, 2012. ISSN 1935-7524.

[7] D.S. Choi, P.J. Wolfe, and E.M. Airoldi. Stochastic blockmodels with growing number of classes. *Arxiv preprint arXiv:1011.4644*, 2010.

[8] A. Condon and R.M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2): 116–140, 2001. MR1809718

[9] J.J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008. MR2390817

[10] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[11] P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959. MR0120167

[12] K. Faust and S. Wasserman. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.

[13] S.E. Fienberg and S.S. Wasserman. Categorical data analysis of single sociometric relations. *Sociological methodology*, 12:156–192, 1981. ISSN 0081-1750.

[14] E.N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30 (4):1141–1144, 1959. ISSN 0003-4851. MR0108839

[15] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821, 2002. MR1908073

[16] P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. MR0718088

[17] P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76 (373):33–50, 1981. ISSN 0162-1459. MR0608176

[18] T.S. Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, pages 129–159, 2000.

[19] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11:033015, 2009.

[20] F. Lorrain and H.C. White. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1 (1):49–80, 1971.

[21] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103 (23):8577, 2006.

[22] K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96 (455):1077–1087, 2001. MR1947255

[23] F. Picard, V. Miele, J.J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using MixNet. *BMC bioinformatics*, 10(Suppl 6):S17, 2009.

[24] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional Stochastic Block Model. *Arxiv preprint arXiv:1007.1684*, 2010. MR2893856

[25] T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997. MR1449742

[26] R. Van Der Hofstad. Random graphs and complex networks. *Available on* `http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf`, 2009.

[27] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.