SCIENTIFIC REPORTS
nature research

OPEN

A Regularized Stochastic Block Model for the robust community detection in complex networks

Received: 27 September 2018
Accepted: 28 August 2019
Published online: 13 September 2019

Xiaoyan Lu[1] & Boleslaw K. Szymanski[1,2]

The stochastic block model is able to generate random graphs with different types of network partitions, ranging from the traditional assortative structures to the disassortative structures. Since the stochastic block model does not specify which mixing pattern is desired, the inference algorithms discover the locally most likely nodes' partition, regardless of its type. Here we introduce a new model constraining nodes' internal degree ratios in the objective function to guide the inference algorithms to converge to the desired type of structure in the observed network data. We show experimentally that given the regularized model, the inference algorithms, such as Markov chain Monte Carlo, reliably and quickly find the assortative or disassortative structure as directed by the value of a single parameter. In contrast, when the sought-after assortative community structure is not strong in the observed network, the traditional inference algorithms using the degree-corrected stochastic block model tend to converge to undesired disassortative partitions.

The study of modular structure in networks has a long history in the literature[1–4]. The primary focus of this line of work has been the discovery of the assortative community structures in which the network nodes are partitioned into communities with edges more numerous inside them than across them. In complex networks, the modular structures include not only such assortative community structures but also other mixing patterns, like the core-periphery structures[5] and the bi-partite structures[6]. These various structures found in networks demand a more comprehensive class of community detection models than the assortative ones, targeted by the modularity-based approaches[2]. The modularity maximization is one of the most widely used approaches for assortative community detection. It aims at maximizing the modularity of the network partitions. The modularity is a broadly accepted quality metric that compares the number of edges observed in each community to the expected number of such edges in a random graph with the same degree sequence. As shown in[7], the modularity maximization is, in fact, equivalent to the maximum likelihood estimation of the degree-corrected planted partition model which is a special case of the degree-corrected stochastic block model[8]. Therefore, the inference of network partition from the stochastic block model can be considered a more general approach to network clustering than the traditional modularity-based community detection algorithms. The former can discover a variety of network structures, different from the traditional assortative community structures, such as the disassortative core-periphery structures[5].

The standard stochastic block model[9] is a generative graph model which assumes the probability of connecting two nodes in a graph is determined entirely by their block assignments. In the following, we will denote the block assignment of any node l by g_l. In the standard stochastic block model, the number of edges between nodes i and j follows the *Bernoulli* distribution with the mean $\omega_{g_i g_j}$. Hence, this model is fully specified by the block assignments of nodes $\{g_l\}$ and the mixing matrix $\Omega = \{\omega_{rs}\}$ governing the probabilities of observing one edge between each pairs of nodes from blocks r and s. If the diagonal elements ω_{rr} in the mixing matrix are larger than the off-diagonal elements, then the networks generated by the stochastic block model have the traditional assortative communities. When the off-diagonal elements ω_{rs} for $r \neq s$ are larger than the diagonal elements, the generated network contains disassortative mixing patterns, such as the structures observed in the core-periphery graphs[5] and in the bi-partite graphs[6]. In general, the inference using the stochastic block model aims at discovering Ω and $\{g_l\}$ which maximize the likelihood of generating the observed network. It does not impose any constraint on the assortativity of the mixing pattern Ω.

[1]Social and Cognitive Networks Academic Research Center and Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, 12180, USA. [2]Społeczna Akademia Nauk, Łódź, Poland. Correspondence and requests for materials should be addressed to B.K.S. (email: szymab@rpi.edu)

The standard stochastic block model assumes all nodes in a community have the same expected degree. However, a block in the realistic community structure often contains nodes with heterogeneous degrees. To address this issue, the degree-corrected stochastic block model[8] extends the standard stochastic block model by defining the expected number of edges between nodes i and j as $\lambda_{ij} = \omega_{g_i,g_j}\beta_i\beta_j$. Here the parameter λ_{ij} represents the expected number of edges, which is important because multi-edges (multiple edges between pair of nodes) are allowed in the model[8]. The model parameter β_l is associated with each node l. The same notation $\lambda_{i,j}$ is used for denoting probability of an edge between pair of nodes in the standard stochastic block model, so it is important to note that it has different meaning here. The expected node degrees in the degree-corrected stochastic block model, given the maximum likelihood estimates of these model parameters, are equal to the degrees observed in the input network. Hence, the degree-corrected stochastic block model allows a community to have a wide range of node degrees. This simple yet effective degree-correction modification improves the performance of statistical inference of community structures in complex networks.

As suggested in[10], both the traditional assortative community structures and the disassortative structures are potentially good fits to the degree-corrected stochastic block models. Hence, depending on the starting point, inference algorithms, e.g., the Markov chain Monte Carlo, tend to converge to the local maxima of the likelihood, which may correspond to the disassortative structures. To address this problem, in this paper, we propose a regularized stochastic block model which provides an extra regularization term to guide the discovery of assortative or disassortative structure by the statistical inference using the stochastic block model. Unlike the modularity maximization algorithm which always attempts to find traditional assortative communities, the inference using this regularized stochastic block model controls with a single parameter the mixing patterns discovered in the given network.

Results

The degree-corrected stochastic block model. The degree-corrected stochastic block model[8] is a generative model of graphs in which the edges are randomly placed between nodes. Let A be the adjacency matrix of an unweighted undirected multigraph, and let A_{ij} denote the number of edges between nodes i and j in this multigraph. The multi-edges and self-loop edges are practical in certain networks such as the World Wide Web where a web page may contain multiple hyperlinks pointing to other pages and to itself. Such edges are less common in social networks. However, most social networks are very sparse, so the impact of multi-edges and self-loop edges is negligible in such networks. For the convenience of notations, let a node i with k self-loop edges be represented by the diagonal adjacency matrix element $A_{ii} = 2k$.

The degree-corrected stochastic block model assumes the number of edges between two different nodes i and j follows the *Poisson* distribution with mean λ_{ij}, while the number of self-loop edges at node l follows the *Poisson* distribution with mean $\frac{1}{2}\lambda_{ll}$. Given the parameters $\{\lambda_{ij}\}$, the likelihood of generating A is

$$P(A|\{\lambda_{ij}\}) = \prod_i \frac{\left(\frac{1}{2}\lambda_{ii}\right)^{A_{ii}/2}}{\left(\frac{1}{2}A_{ii}\right)!} e^{-\lambda_{ii}/2} \prod_{i<j} \frac{\lambda_{ij}^{A_{ij}}}{A_{ij}!} e^{-\lambda_{ij}}. \tag{1}$$

Here, λ_{ij} defines the expected number of edges rather than the probability, because multi-edges are allowed in this model. In an unweighsted undirected network, after ignoring all terms independent of the λ_{ij}, the corresponding log-likelihood simplifies to

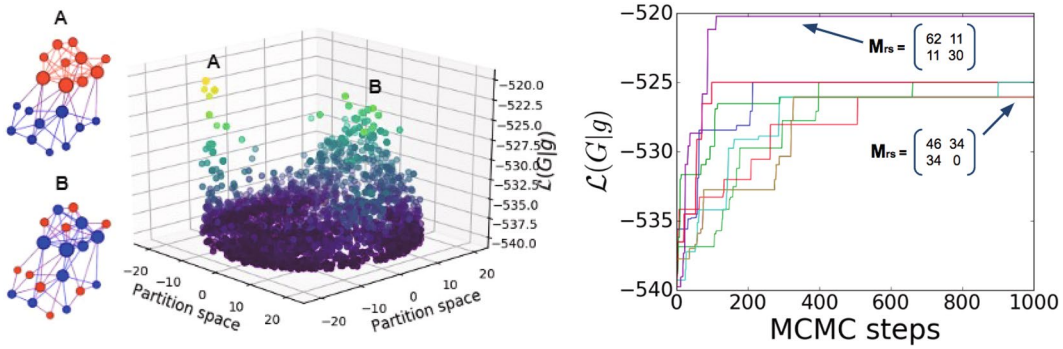
$$\log P(A|\{\lambda_{ij}\}) = \frac{1}{2}\sum_{ij}(A_{ij} log \lambda_{ij} - \lambda_{ij}). \tag{2}$$

In the degree-corrected stochastic block model, the model parameter λ_{ij} is defined as $\lambda_{ij} = \omega_{g_i,g_j}\beta_i\beta_j$ where for a node l, β_l denotes its parameter. Given the block assignments $\{g_i\}$, the authors of[8] obtain the maximum likelihood estimates of the model parameters as $\hat{\beta}_i = \frac{k_i}{\kappa_{g_i}}$ and $\hat{\omega}_{rs} = m_{rs}$, where k_i is the degree of node i, κ_r is the sum of the degrees of all nodes in a block r, and m_{rs} is the total number of edges between different blocks r and s, or, if $r = s$, twice the number of edges in the block r.

Assortative and disassortative structures. As defined in the literature, e.g.[1–3], assortative structures correspond to the traditional community structures, in which nodes are more frequently connected to each other inside communities than across them. The disassortative structures do not satisfy this condition. For example, a core-periphery structure[5] divides the networks nodes into a core part, in which nodes are often the hubs of the networks, and a periphery part with nodes of low-degree connecting to the core nodes. It can be seen as an advantage that the degree-corrected stochastic block model can generate the structures with different mixing patterns. However, when searching for the weak assortative community structures, the inference algorithm using this model may infer the disassortative structure instead[11].

We generate synthetic networks using the degree-corrected stochastic block model with the parameters ω_{rs} chosen for the assortative communities as follows

$$\omega_{rs} = \begin{cases} \gamma\omega_0 & \text{if } r = s, \\ \omega_0 & \text{if } r \neq s, \end{cases} \tag{3}$$



(a) Local optima of log-likelihood (b) Convergence

Figure 1. The convergence of 20 Markov chain Monte Carlo (MCMC) trials for the degree-corrected stochastic block model. The local maximum *A* found by MCMC represents the assortative communities, whereas there are other local maxima representing disassortative structures such as point B. (**a**) Multiple locally maximal partitions discovered by the MCMC inference using the degree-corrected stochastic block model; (**b**) Two out of the 20 MCMC trials find the most likely and sought-after assortative partition, while the other trials converge to the local maxima representing disassortative structures. The matrix M_{rs} indicates the number of edges between any node from block r and any node from block s, including cases when $r = s$.

where a large value of $\gamma > 1$ results in strong assortative structure while ω_0 controls the sparsity of the network. For the generation of synthetic networks, the degree sequence $\{k_i\}$ is drawn from a power-law distribution with exponent 2.5, and the two blocks are generated with equal size to which the nodes are randomly assigned.

For the example shown in Fig. 1, given the synthetic networks produced by the degree-corrected stochastic block model, we infer its block assignments $\{g_i\}$ using the Markov chain Monte Carlo (MCMC) algorithm[12,13]. Specifically, the number of communities is set as two for the process of generating it and for recovery of its parameters. To generate the samples, the degree-corrected stochastic block model uses $\omega_0 = 0.01$ and $\gamma = 10$. Each of the two blocks contains 10 nodes. We scatter the sampled partitions on the x-y plane in Fig. 1(a) while the z-axis indicates the log-likelihood of the corresponding partitions. The details of the MCMC algorithm are provided in the Supplementary Information.

Figure 1 shows the multiple local maxima of the log-likelihood of degree-corrected stochastic block model that exist for our sample network. The inference process finds three local maxima here: (i) partition A which corresponds to the assortative structure matching the ground truth block assignment used for its generation; (ii) partition B which corresponds to a disassortative structure; and (iii) another disassortative partition which is not explicitly marked in the Fig. 1(a). Under the degree-corrected stochastic block model, the MCMC inference starting from random initial partition finds the partition A matching the ground truth block assignment used for its generation in only two out of 20 trials. The other trials converge to the local maxima that represent disassortative structures, as shown in Fig. 1(b).

Since there are multiple local maxima in the log-likelihood of the degree-corrected stochastic block model, the inference algorithm may converge to any of them. In our experiments, the type of the discovered structure depends on the trial starting point and inference algorithm parameters. To avoid such undesired outcomes, we introduce a novel approach applicable to any inference algorithm that we named the Regularized Stochastic Block Model (RSBM). It constrains each node internal degree ratio used in the objective function. This ratio is defined as a fraction of the node's neighbors that are inside its community. The inference algorithm using the regularized model reliably finds assortative or disassortative structures as directed by the value of a single parameter.

Regularized stochastic block model. We extend the formulation of the expected number of edges between nodes i and j, determined by the *Poisson* rate λ_{ij}, in the degree-corrected stochastic block model by redefining it as

$$\lambda_{ij} = \begin{cases} \omega_{g_i g_j} I_i I_j & \text{if} \quad g_i = g_j \\ \omega_{g_i g_j} O_i O_j & \text{otherwise,} \end{cases} \tag{4}$$

where any node l has two associated parameters I_l and O_l. Given Eq. 2, the log-likelihood of generating graph G by this regularized stochastic block model can be written as

$$\mathcal{L}(G|g, \omega, \mathbf{I}, \mathbf{O}) = 2\sum_i (k_i^+ \log I_i + k_i^- \log O_i) + \sum_{rs} m_{rs} \log \omega_{rs} - \omega_{rs}\Lambda_{rs} \tag{5}$$

where k_i^+ is the number of neighbors of node i which are inside the same block given the block assignment g, $k_i^- = k_i - k_i^+$, and

$$\Lambda_{rs} = \begin{cases} \left(\sum_{i\in r} I_i\right)^2 & \text{if } r = s \\ \sum_{i\in r} O_i \sum_{i\in s} O_i & \text{if } r \neq s \end{cases} \tag{6}$$

To simplify, we write $i \in r$ if $g_i = r$. For block assignment g, the maximum-likelihood values of ω_{rs} are

$$\hat{\omega}_{rs} = \frac{m_{rs}}{\Lambda_{rs}}. \tag{7}$$

Dropping the constants and substituting using Eq. 5, we obtain

$$\mathcal{L}(A|g, \mathbf{I}, \mathbf{O}) = \sum_{rs} m_{rs} \log \frac{m_{rs}}{\Lambda_{rs}} + 2\sum_i (k_i^+ \log I_i + k_i^- \log O_i) \tag{8}$$

Note that if we set $I_i = O_i = 1$ here, the log-likelihood above reduces to the definition of standard stochastic block model with $\Lambda_{rs} = n_r n_s$ which is exactly the product of the sizes of two blocks r and s. When $I_i = O_i = k_i$, the second sum on the right hand side (RHS) becomes irrelevant to the maximum likelihood estimation (MLE) result. Hence, the log-likelihood reduces to the definition of degree-corrected stochastic block model in Eq. 8 with $\Lambda_{rs} = \kappa_r \kappa_s$, i.e., the product of the sums of degrees of nodes in two blocks r and s. Hence, by introducing here two sets of parameters $I = \{I_i\}$ and $O = \{O_i\}$ in the edge probability, we obtain a more generalized definition of the stochastic block model.

Regularization by prior in-degree ratios. In alternative formulation of our model, we define for each node i the prior in-degree ratios $f_i = I_i/(I_i + O_i)$ and $\theta_i = I_i + O_i$. By rewriting the second summation on the RHS of Eq. 8, we get

$$\mathcal{L}(G|g, \mathbf{I}, \mathbf{O}) = \sum_{rs} m_{rs} \log \frac{m_{rs}}{\Lambda_{rs}} - 2\sum_i k_i H\left(\frac{k_i^+}{k_i}, f_i\right) + 2\sum_i k_i \log \theta_i \tag{9}$$

where $H\left(\frac{k_i^+}{k_i}, f_i\right) = -\frac{k_i^+}{k_i} \log f_i - \frac{k_i^-}{k_i} \log(1 - f_i)$ represents the cross entropy between the observed in-degree ratio $\frac{k_i^+}{k_i}$ and the prior in-degree ratio f_i. By maximizing the log-likelihood defined in Eq. 9, the optimization process tends to find a partition with $\frac{k_i^+}{k_i} \approx f_i$ in an effort to reduce the sum of the cross entropy terms. Therefore, the prior in-degree ratios $\{f_i\}$ regularize the observed in-degree ratios $\left\{\frac{k_i^+}{k_i}\right\}$ in the resulting partition.

In the real networks with the traditional community structures, the low degree nodes are generally more likely to have neighbors inside the same block than the high degree nodes are. Suppose the prior in-degree ratio f_i depends only on the degree of node i, i.e. $f_i = f(k_i)$. Then, the function $f(k)$: $\mathbb{Z}_+ \to [0, 1]$ should be strictly decreasing. In an assortative partition of the network, we have

- $f(1) = 1$ because a node with degree one must connect to the community it belongs to;
- For $k \approx |V|$, $f(k) \ll 1$ because a super-hub eventually does not belong to any community as its degree is of the order of $|V|$, the number of nodes in the entire network.

A simple function $\{f(k)\}$ satisfying this requirement is of the form

$$f(k) = \alpha + \frac{(1 - \alpha)}{k}, \tag{10}$$

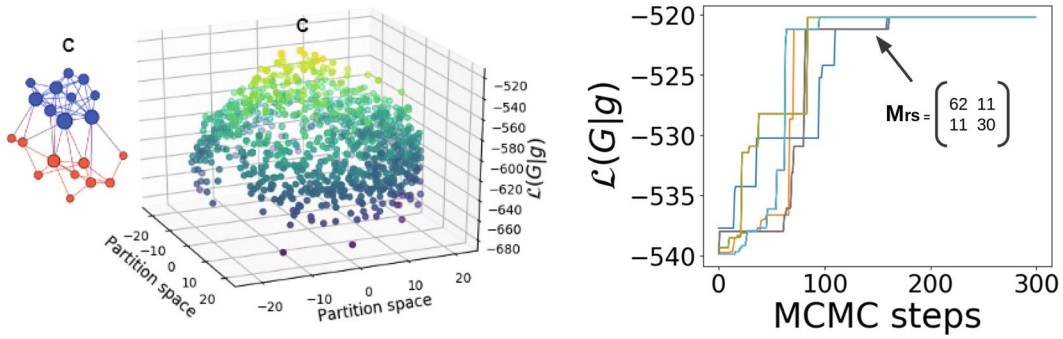
where α is the only extra parameter we introduce to the regularized stochastic block model (RSBM). Alternatively, we can select a constant $f \in (0, 1)$ such that

$$f(k) = \max\left(f, \frac{1}{k}\right), \tag{11}$$

where now f is the only extra parameter introduced for this form of the function f_i. The impact of different choices of f_i on the discovered block assignment is discussed in the following two subsections presenting experimental results.

Experimental results. We generate the synthetic networks with the assortative communities using the degree-corrected stochastic block model. The parameters ω_{rs} used in the network generation process are specified by Eq. 3. Selecting a small value of $\gamma > 1$ in Eq. 3 makes the generated community structures relatively weak, and therefore difficult to detect by the statistical inference using the degree-corrected stochastic block model.

We again adapt the Markov chain Monte Carlo (MCMC) algorithm[13] to infer the block assignments using our proposed regularized stochastic block model. Figure 2(a) shows there is only one unique local maximum, the partition C, found by 20 MCMC trials under the regularized stochastic block model. Therefore, all 20 MCMC trials converge to this unique local maximum. As shown in Fig. 2(b), the inference process finds the correct block assignment in at most 150 MCMC steps. The regularization terms enable the inference to converge to the suitable local maxima.



(a) Local optimum of log-likelihood (b) Convergence

Figure 2. The convergence of 20 Markov chain Monte Carlo (MCMC) trials for the regularized stochastic block model (RSBM) introduced here. All trials converge to the maximal point C which represents an assortative structure. (**a**) One maximal partition observed during the MCMC inference using our model. (**b**) All twenty MCMC trials find the sought-after assortative structure.

We use the real networks including the Karate club network of Zachary[14], the Dolphin social network of Lusseau *et al.*[15] and the network of fictional characters' interactions in the novel Les Miserables by Victor Hugo[2] to demonstrate the performance of the regularized model introduced here. The details of each network are presented in the Supplementary Information. For the Karate club network, we evaluate the effect of the regularization terms on the resulting partitions recovered using the Markov chain Monte Carlo (MCMC). For every node i in the network, we set the parameter $\theta_i = k_i$ and $f_i = max(f, 1/k_i)$ for our regularized stochastic block model. Figure 3 shows the most likely partition of the Karate club network found by MCMC using different f values. The color represents the block assignment, and the black dashed line divides the network into two parts in the ground truth partition. As shown in Fig. 3, when $f = 0.14$, the inference algorithm outputs a core-periphery structure which clusters high-degree nodes into the blue block and the remaining low-degree nodes into the red block. This is because the sum of cross entropy terms in Eq. 9 serves as a regularization term which penalizes partitions with the average in-degree ratio much larger than 0.14. As the value of f grows, the inference algorithm is becoming more likely to detect assortative structure. When $f = 0.85$, the inferred block assignment matches the ground truth partition of the Karate club network with the exception of one single red node. However, this node has only one connection to each block; thus, it is quite arguable to which block this node should belong.

The results in Fig. 4 show that the MCMC inference using our RSBM model on the three real networks finds partitions with higher *coverage* than the ones inferred by the degree-corrected stochastic block model. The *coverage* of a partition[1] is defined as the ratio of the number of edges with both endpoints in the same block to the total number of edges in the entire network

$$\text{coverage}(\mathbf{g}) = \frac{|\{(i,j) \in E | g_i = g_j\}|}{|E|}. \tag{12}$$

A low *coverage* indicates that the resulting partition is disassortative. An ideal assortative partition of the network, where all clusters are disconnected, yields a *coverage* of 1.

Figure 4 shows the *coverages* of the partitions found in the three real networks mentioned above. We randomly remove edges in these networks to further increase their sparsity. The numbers of blocks used in trials for these networks are set to their broadly accepted values in the literature. The results in Fig. 4 indicate that under the degree-corrected stochastic block model (DCSBM) the inference algorithm is likely to miss the assortative structures. It returns instead the disassortative partitions of the network, which also fit the model in such cases. In contrast, the MCMC inference using our regularized stochastic block model (RSBM) almost always produces the assortative structures. Interestingly, the inference using the degree-corrected stochastic block model produces the partitions with the *coverage* values distributed at two levels. This is similar to the case of the synthetic network in Fig. 1 where both the assortative communities and disassortative structures fit the degree-corrected stochastic block model. Which structure is found is determined by random sampling of the potential partitions. In other words, there is no way to guide whether a disassortative or an assortative structure is preferred. Hence, two MCMC trials may return completely different structures. In contrast, the inference using the regularized stochastic block model introduced here and using the prior in-degree ratio $f_i = 0.8 + 0.2/k_i$ is very robust. It only produces partitions with a high *coverage* distributed at the same level of the coverage as the assortative partitions found under the degree-corrected stochastic block model. Moreover, the sparsity of the networks does not have an obvious impact on the resulting partitions under our regularized model.

We evaluate the modularity[16] of resulting partitions of the real networks as a function of parameter f. A high modularity indicates the strong community structure. In our experiments, we use a constant f such that $f_i = f$ for each node i in the network with degree $k_i > 1/f$, otherwise $f_i = 1/k_i$. We start with a small f value and increase it in each iteration. The MCMC inference uses as a starting point a random partition of the original network initially and then uses the network partition found in the previous iteration. Figure 5 shows that, as the value of f increases, in general both the modularity and the *coverage* grow. When f is close to 0, the value of f does not have

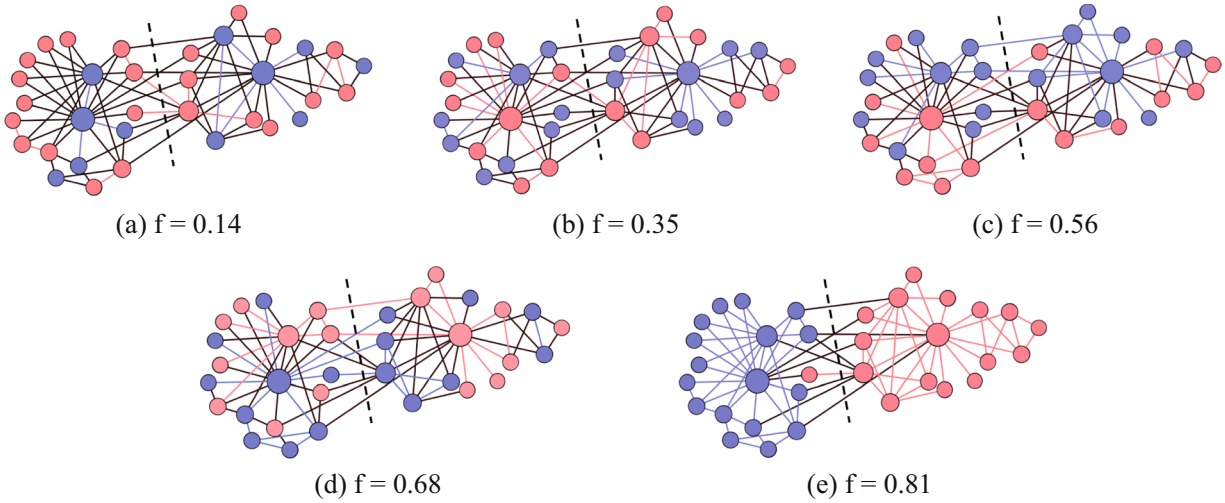


Figure 3. The Karate network partition inferred by Markov chain Monte Carlo under different parameter settings where nodes of the same color belong to the same partition. The black dotted line represents the ground truth partition. A small *f* results in a core-periphery partition of the network while a large *f* leads to an assortative partition. The values of *f* parameter are shown in the corresponding sub-figures captions.

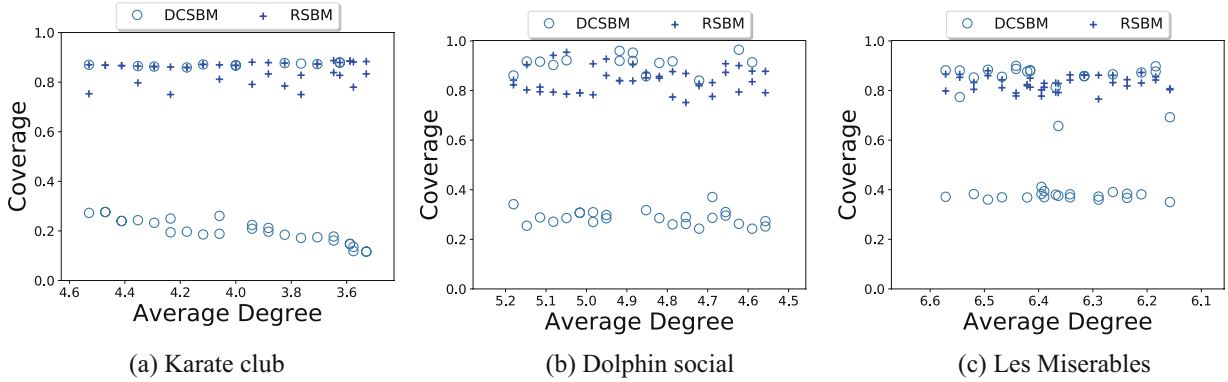


Figure 4. The *coverage* of partitions in the three real networks as a function of the average node degree. The maximal partitions are inferred by the Markov chain Monte Carlo algorithm under the degree-corrected stochastic block model (DCSBM), marked by circles, and its regularized extension (RSBM) introduce here, marked by crosses.

any effect on the network partition. However, as *f* becomes larger, then at a certain critical value of about 0.75, a critical transition arises at which the type of the recovered structure changes from disassortative partitions to assortative communities. Using *f* larger than the critical value does not further increase the modularity of the partition. These results indicate that, with one single parameter *f*, the user gains control over the type of structure to which the Markov chain Monte Carlo inference will converge. With a choice of high *f*, the inference algorithm is most likely to detect assortative communities.

Discussion

The stochastic block model is able to produce a wide variety of network structures, including traditional assortative structures and different from them disassortative structures. In theory, it should be possible to guide the inference algorithm which type of structures is preferred. Moreover, the existence of multiple local maxima of the log-likelihood may cause an inference algorithm to converge to the undesired type of structure. Although there were efforts to enable community detection at different levels of granularity, cf.[17,18], the need for controlling assertiveness of the solution has not been addressed so far. Here, we apply a simple yet effective constraint on nodes' internal degree ratios in the objective function. This approach is applicable to any inference algorithm. The resulting algorithm reliably finds assortative or disassortative structure as directed by the value of a single parameter *f*. We validate the model experimentally testing its performance on several real and synthetic networks. The experiments show that the inference using our regularized stochastic block model quickly converges to the sought-after assortative or disassortative structure. In contrast, the inference using the degree-corrected stochastic block model often converges to the local maximal partitions representing the undesired type of structure.

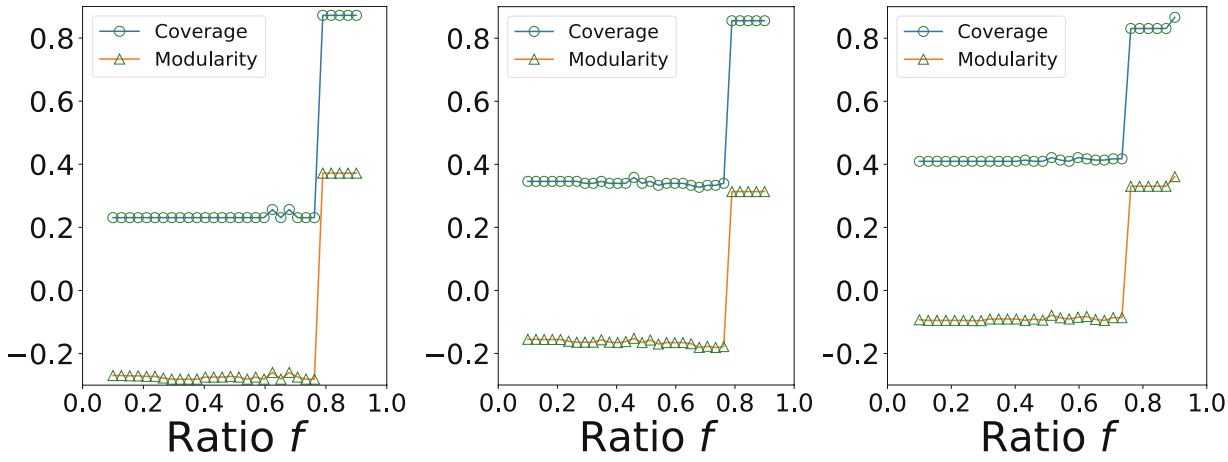


Figure 5. The *coverage* and modularity of partitions in the real networks as functions of parameter *f*. The maximal partitions are inferred by the Markov chain Monte Carlo algorithm under the degree-corrected stochastic block model (DCSBM) and its regularized extension (RSBM) introduced here.

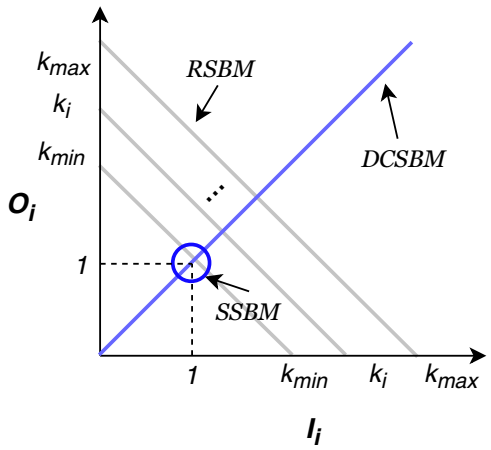


Figure 6. A unified view on the variants of stochastic block models. The standard stochastic block model (SSBM) is represented by the dark blue point (1, 1) since it sets $I_i = O_i = 1$ for each node i. The degree-corrected stochastic block model (DCSBM) requires $I_i = O_i$, which corresponds to the light blue line with a slope $f_i = 1/2$ in the 2D coordinate plane. The introduced here regularized stochastic block model (RSBM) allows choosing I_i and O_i on the line $I_i + O_i = k_i$ where k_i is the node degree. So our model is represented by a set of points on the parallel gray lines each of which intersects the x and y-axis at integer values corresponding to the degree of a node in the range $[k_{min}, k_{max}]$.

Methods

Proving properties of the regularized stochastic block model. Theorem 1 When $f_i = 1/2$ for each node i in the network, the MLE of our RSBM model defined by Eq. 9 becomes the MLE of degree-corrected stochastic block model.

According to Theorem 1 (the proofs of Theorem 1, 2 and 3 are provided in the Supplementary Information), if we add a constraint $I_i = O_i$ for every node i, the model introduced here reduces to the degree-corrected stochastic block. Figure 6 illustrates the relationship between the most commonly used variants of stochastic block models. If we plot the constraints of these models on a 2D plane with I_i and O_i as the x and y axes, respectively, the standard stochastic block model (SSBM) sets $I_i = O_i = 1$ for each i, which maps to the point (1, 1) in the 2D coordinate plane. The constraint $I_i = O_i$ which represents the degree-corrected stochastic block model (DCSBM) maps onto the line with a slope $f_i = 1/2$ in the 2D coordinate plane. Here, we extend the stochastic block model in two different directions: (i) using a constraint on $\theta_i = I_i + O_i$; (ii) choosing I_i and O_i on the line with a customized slope f_i. It turns out the latter, like the degree-corrected stochastic block model, preserves the degree sequence of the network.

Theorem 2 Given any customized $\{f_i\}$ and the corresponding maximum-likelihood estimator $\hat{\theta}_i$, our model defined by Eq. 9 preserves the degree sequence of a network, so the expected degree of node i generated by the RSBM model with $\hat{\theta}_i$ is

$$\sum_j \lambda_{ij} = k_i. \tag{13}$$

Information-theoretic interpretations. While the maximum likelihood estimator $\hat{\theta}_i$ preserves the degree sequence in the observed network, we find that the closed-form analytic expression of such estimator is hard to obtain. However, when we impose constraints that for each i, $\theta_i = k_i$ and look for the MLE of f_i for every i, the log-likelihood of the model introduced here has an interesting interpretation uncovered by the following theorem.

Theorem 3 When $\theta_i = k_i$ for every node i, maximizing the log-likelihood of Eq. 9 is equivalent to maximizing the target function

$$\mathcal{L} = \mathbb{D}_{KL}(p_{\text{degree}}(r, s)||p_{\text{null}}(r, s)) - 2\mathbb{E}_{k_i}\left[H\left(\frac{k_i^+}{k_i}\right)\right] \tag{14}$$

where the first term represents the Kullback-Leibler (KL) divergence between the edge distribution under the block model and the corresponding distribution under the null model which randomizes the edges inside and across blocks respectively, while the second term is twice the expectation of the binary entropy of in-degree ratio k_i^+/k_i.

Imposing the constraint $I_i + O_i = k_i$, we obtain a new model which involves a regularization term of the expected entropy of in-degree ratio k_i^+/k_i. According to Theorem 3, maximizing the log-likelihood of this model has a physical interpretation: the MLE of the RSBM model increases the KL divergence between the edge distribution under the block model and the corresponding edge distribution under the *null* model, and, at the same time, it decreases the expected entropy of the in-degree ratio k_i^+/k_i. Intuitively, the *null* model here separates the edges inside a block and the edges across different blocks. It assumes that all edges inside blocks are statistically equivalent, and so are all edges across different blocks. In contrast, the *null* model of the degree-corrected stochastic block model[8] mixes all the edges, assuming all the edges are statistically equivalent - the KL divergence term in its *null* model does not distinguish the edges inside the same block and those across different blocks. On the other hand, the sum of entropy function $H(k_i^+/k_i)$ controls the identification of the edges' types. In Eq. 14, $H(k_i^+/k_i)$ is the entropy function which achieves its minimum when either $k_i^+/k_i \to 0$ or $k_i^+/k_i \to 1$. Therefore, the MLE of this model is likely to classify the edges of a node as either all in-edges or all out-edges – the former case is likely to detect assortative communities and the latter case infers disassortative structure as the most probable block assignment.

For community detection problems, it is rarely observed that all neighbors of a node are not located in the same block, especially for nodes with low degrees. Unlike traditional community detection algorithms, the definitions of stochastic block model and its mentioned above variants do not explicitly control the fraction of neighbors in different blocks. Instead, they rely on the specific statistical inference to determine the block assignments. This observation inspires us to regularize the traditional stochastic block model on the in-degree ratio. Specifically, our Regularized Stochastic Block Model (RSBM) introduced here maximizes the objective function of Eq. 9 with $\theta_i = k_i$ and f_i defined by the prior in-degree ratio. As explained above, the KL divergence in RSBM plays the same role as the divergence does in the degree-corrected stochastic block model. The expected cross-entropy term serves as a regularization term to control the resulting partition. Intuitively, when f_i is close to one, then the inference algorithm tends to cluster nodes in a network into dense modules. Otherwise, when f_i is close to zero, then disassortative partitions such as the core-periphery structure are likely to be discovered because the regularized block model tends to assign adjacent nodes into different blocks to decrease k_i^+.

Data Availability

The network data sets used for the evaluation of our proposed model are available at http://konect.uni-koblenz.de/networks/.

References

1. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**(3–5), 75–174 (2010).
2. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E.* **69**(2), 026113 (2004).
3. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003).
4. Parnas, D., Clements, P. & Weiss, D. The modular structure of complex systems. *IEEE Trans. Softw. Eng.* **3**, 259–266 (1985).
5. Borgatti, S. & Everett, M. Models of core/periphery structures. *Social Netw.* **21**(4), 375–395 (2000).
6. Guillaume, J. L. & Latapy, M. Bipartite structure of all complex networks. *Inf. Proc. L.* **90**, 215–221 (2004).
7. Newman, M. E. J. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Phys. Rev. E.* **94**(5), 052315 (2016).
8. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E.* **83**(1), 016107 (2011).
9. Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: first steps. *Social Netw.* **5**, 109–137 (1983).
10. Peel, L., Larremore, D. & Clauset, A. The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**(5), e1602548 (2017).
11. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E.* **84**(6), 066106 (2011).
12. Peixoto, T. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Phys. Rev. E.* **89**(1), 012804 (2014).
13. Nasrabadi, N. Pattern recognition and machine learning. *J. Elect. Imag.* **16**(4), 049901 (2007).

14. Zachary, W. An information flow model for conflict and fission in small groups. *J. Anthro. Res.* **33**(4), 452–473 (1977).
15. Lusseau, D. *et al*. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. and Soc.* **54**(4), 396–405 (2003).
16. Newman, M. E. J. Modularity and community structure in networks. *Proc. Nat. Acad. Sci.* **103**(23), 8577–8582 (2006).
17. Jeub, L. G. S., Sporns, S. & Fortunato, S. Multiresolution Consensus Clustering in Networks. *Sci. Rep.* **8**, 3259 (2018).
18. Lu, X. & Szymanski, B. K. Asymptotic resolution bounds of generalized modularity and statistically significant community detection. arXiv, 1902.04243 (2019).

Acknowledgements

The authors express thanks to Dr. Santo Fortunato for helpful discussions of the degree-correlated SBM. This work was supported in part by the Army Research Laboratory (ARL) through the Cooperative Agreement (NS CTA) Number W911NF-09-2-0053, by DARPA and the Army Research Office (ARO) under Agreement No. W911NF-17-C-0099, and by the Office of Naval Research (ONR) under Grant N00014-15-1-2640. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies either expressed or implied of the Army Research Laboratory or the U.S. Government.

Author Contributions

Both authors contributed to the study design and manuscript preparation, X.L. conducted simulations and contributed the proof of the properties of the regularized block model introduced here, B.K.S. contributed the functional forms of the prior in-degree ratios. Both authors wrote, reviewed and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-49580-5.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.