

SCIENTIFIC REPORTS



OPEN

Cross-validation estimate of the number of clusters in a network

Tatsuro Kawamoto¹ & Yoshiyuki Kabashima²

Network science investigates methodologies that summarise relational data to obtain better interpretability. Identifying modular structures is a fundamental task, and assessment of the coarse-grain level is its crucial step. Here, we propose principled, scalable, and widely applicable assessment criteria to determine the number of clusters in modular networks based on the leave-one-out cross-validation estimate of the edge prediction error.

Mathematical tools for graph or network analysis have wide applicability in various disciplines of science. In fact, many datasets, e.g., biological, information, and social datasets, represent interactions or relations among elements and have been successfully studied as networks^{1,2} using the approaches of machine learning, computer science, and statistical physics. In a broad sense, a major goal is to identify macroscopic structures, including temporal structures, that are hidden in the data. To accomplish this goal, for example, degree sequences, community and core-periphery structures, and various centralities have been extensively studied. Here, we focus on identifying modular structures, namely, graph clustering^{3,4}. Popular modular structures are the community structure (assortative structure) and disassortative structure^{5–8}, although any structure that has a macroscopic law of connectivity can be regarded as a modular structure. The Bayesian approach using the stochastic block model⁹, which we will describe later, is a powerful tool for graph clustering. In general, graph clustering consists of two steps: selecting the number of clusters and determining the cluster assignment of each vertex. These steps can be performed repeatedly. Some methods require the number of clusters to be an input, whereas other methods determine it automatically. The former step is called model selection in statistical frameworks, and this step is our major focus.

The selection of the number of clusters is not an obvious task. For example, as shown in Fig. 1, more complex structures can be resolved by using a model that allows a larger number of clusters. However, because the whole purpose of graph clustering is to coarse grain the graph, the partition with excessively high resolution is not desirable. This issue is also related to the level of resolution that is required in practice. The role of the model selection frameworks and criteria is to suggest a possible candidate, or candidates, for the numbers of clusters to be analysed.

Which framework and criterion to use for model selection and its assessment is an inevitable problem that all practitioners face, as multiple candidates have been proposed in the literature. A classical prescription is to optimise an objective function that defines the strength of a modular structure, such as the modularity^{10,11} and the map equation^{12,13}; combinatorial optimisation among all of the possible partitions yields both the number of clusters and the cluster assignments. Although it might seem to be unrelated to statistical frameworks, it should be noted that modularity maximisation has been shown to be equivalent to likelihood maximisation of a certain type of stochastic block model^{11,14}. One can also use the spectral method and count the number of eigenvalues outside of the spectral band^{15–17}; the reason for this prescription is the following: while the spectral band stems from the random nature of the network, the isolated eigenvalues and corresponding eigenvectors possess distinct structural information about the network. As another agnostic approach, an algorithm¹⁸ that has a theoretical guarantee of the consistency for the stochastic block model in sparse regime with sufficiently large average degree was recently proposed; a network is called sparse when its average degree is kept constant as the size of the network grows to infinity, and typically, it is locally tree-like. Finally, in the Bayesian framework, a commonly used principle is to select a model that maximises the model's posterior probability^{19–23} or the model that have the minimum description length^{24–26}, which leads to Bayesian Information Criterion (BIC)-like criteria.

Minimisation of the prediction error is also a well-accepted principle for model selection, and cross-validation estimates it adequately^{27,28}. This approach has been applied to a number of statistical models, including those

¹Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, 2-3-26 Aomi, Koto-ku, Tokyo, Japan. ²Department of Mathematical and Computing Science, Tokyo Institute of Technology, 4259-G5-22, Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa, 226-8502, Japan. Correspondence and requests for materials should be addressed to T.K. (email: kawamoto.tatsuro@gmail.com)

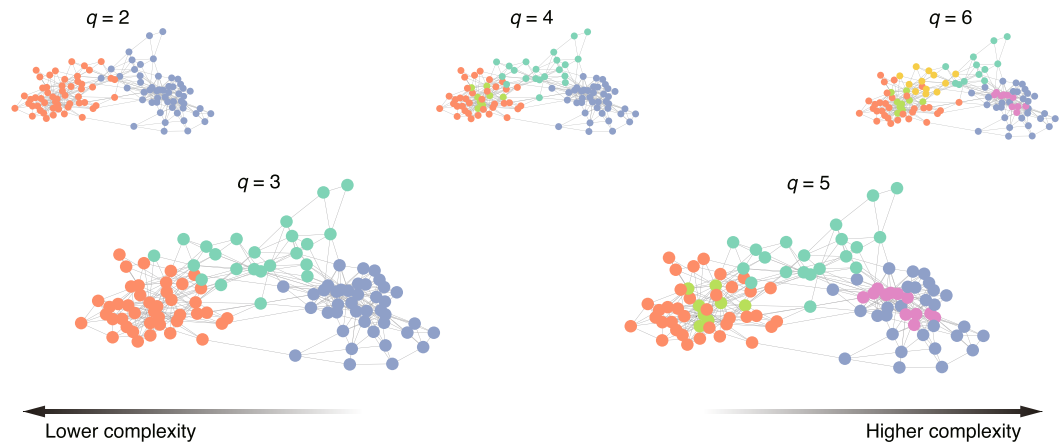


Figure 1. Partitions of the network of books about US politics. The inferred cluster assignments for various input numbers of clusters q . The vertices with the same colour belong to the same cluster. The partition with large q has a higher resolution and is obtained by fitting a model that has a higher complexity.

models that have hidden variables^{29,30}. Although the cross-validation model assessments for the stochastic block model are also considered in refs 31–33, they are either not of the Bayesian framework or performed by a brute-force approach. A notable advantage of performing model selection using prediction error is that the assumed model is not required to be consistent with the generative model of the actual data, whereas the penalty term in the BIC is derived by explicitly using the assumption of model consistency. In this study, we propose an efficient cross-validation model assessment in the Bayesian framework using the belief propagation (BP)-based algorithm.

Generally speaking, the prediction error can be used as a measure for quantifying the generalisation ability of the model for unobserved data. In a data-rich situation, we can split the dataset into two parts, the training and test sets, where the training set is used to learn the model parameters and the test set is used to measure the model's prediction accuracy. We conduct this process for models that have different complexities, and we select the model that has the least prediction error or the model that is the most parsimonious if the least prediction error is shared by multiple models. In practice, however, the dataset is often insufficient to be used to conduct this process reliably. The cross-validation estimate is a way to overcome such a data-scarce situation. In K -fold cross-validation, for example, we randomly split the dataset into K (>1) subsets, keep one of them as a test set while the remaining subsets are used as the training sets, and measure the prediction error. Then, we switch the role of the test set to the training set, pick one of the training sets as a new test set, and measure the prediction error again. We repeat this process until every subset is used as a test set; then, we repeat the whole process for different random K -splits. The average over all the cases is the final estimate of the prediction error of a given model's complexity.

For a dataset with N elements, the N -fold cross-validation is called the *leave-one-out cross-validation* (LOOCV); hereafter, we focus on this approach. Note that the method of splitting is unique in the LOOCV. In the context of a network, the dataset to be split is the set of edges and non-edges, and the data-splitting procedure is illustrated in Fig. 2: for each pair of vertices, we learn the parameters while using the data without the information of edge existence between the vertex pair; then, we measure whether we can correctly predict the existence or non-existence of an edge. At first glance, this approach appears to be inefficient and redundant because we perform the parameter learning $N(N-1)/2$ times for similar training sets, which is true when the LOOCV is implemented in a straightforward manner. Nevertheless, we show that such a redundant process is not necessary and that the prediction error can be estimated efficiently using a BP-based algorithm. It is possible to extend the LOOCV estimate that we propose to the K -fold cross-validation. As we mention in the Methods section, however, such an estimate has both conceptual and computational issues, i.e., the LOOCV is an exceptionally convenient case.

The BP-based algorithm that we use to infer the cluster assignment was introduced by Decelle *et al.*^{21,34}. This algorithm is scalable and performs well for sparse networks; this is favourable, because real-world networks are typically sparse. Indeed, it is known that the BP-algorithm asymptotically achieves the information-theoretic bound^{35,36} (when the number of clusters is 2) in the sense of accuracy when the network is actually generated from the assumed model, the stochastic block model. In the original algorithm, the number of clusters q^* is determined by the Bethe free energy, which is equivalent to the approximate negative marginalised log-likelihood; among the models of different (maximum) numbers of clusters, the most parsimonious model with low Bethe free energy is selected. Despite its plausibility, however, it is known that this prescription performs poorly in practice. One of the reasons is that the estimate using the Bethe free energy relies excessively on the assumption that the network is generated from the stochastic block model, which is almost always not precisely correct. We show that this issue can be substantially improved by evaluating cross-validation errors instead of the Bethe free energy while keeping the other parts of the algorithm the same. With this improvement, we can conclude that the BP-based algorithm is not only an excellent tool in the ideal case but is also useful in practice.

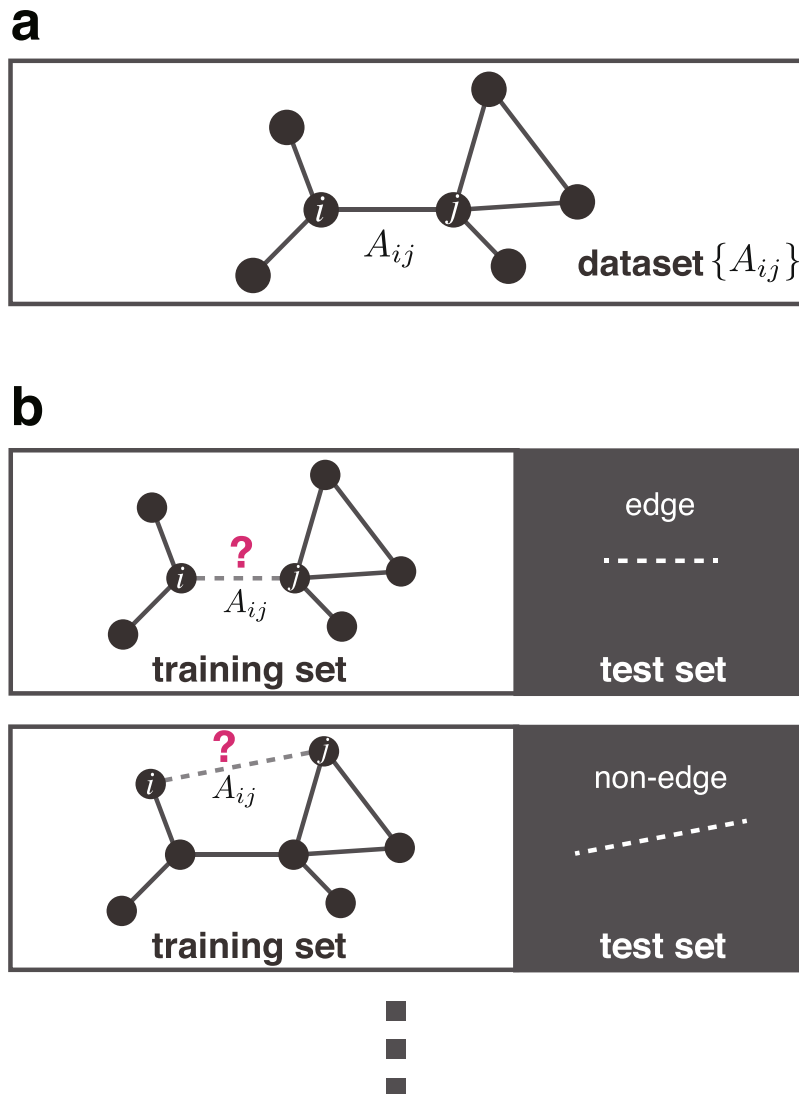


Figure 2. The LOOCV error estimate of the network data. **(a)** The original dataset is given as the adjacency matrix A of a network. **(b)** In each prediction, we hide a single piece of edge or non-edge information from A : this unobserved edge or unobserved non-edge is the test set, and the adjacency matrix $A^{(i,j)}$ without the information on that edge or non-edge is the training set.

We denote the set of vertices and edges in the networks as V and E , and their cardinalities as N and L , respectively. Throughout the current study, we consider undirected sparse networks, and we ignore multi-edges and self-loops for simplicity.

Results

Stochastic block model. Let us first explain how an instance of the stochastic block model is generated. The stochastic block model is a random graph model that has a planted modular structure. The fraction of vertices that belong to cluster σ is specified by γ_σ ; accordingly, each vertex i has its planted cluster assignment $\sigma_i \in \{1, \dots, q^*\}$, where q^* is the number of clusters. The set of edges E is generated by connecting pairs of vertices independently and randomly according to their cluster assignments. In other words, vertices i and j with cluster assignments $\sigma_i = \sigma$ and $\sigma_j = \sigma'$ are connected with probability $\omega_{\sigma\sigma'}$; the matrix ω that specifies the connection probabilities within and between the clusters is called the affinity matrix. Note that the affinity matrix ω is of $O(N^{-1})$ when the network is sparse.

As the output, we obtain the adjacency matrix A . While the generation of network instances is a forward problem, graph clustering is its inverse problem. In other words, we infer the cluster assignments of vertices σ or their probability distributions as we learn the parameters (γ, ω) , given the adjacency matrix A and the input (maximum) number of clusters q . After we have obtained the results for various values of q , we select q^* . Note here that the input value q is the maximum number of clusters allowed and that the actual number of clusters that appears for a given q can be less than q .

Cross-validation errors. We consider four types of cross-validation errors. All of the errors are calculated using the results of inferences based on the stochastic block model. We denote $A^{(i,j)}$ as the adjacency matrix of a network in which A_{ij} is unobserved, i.e., in which it is unknown whether $A_{ij} = 0$ or $A_{ij} = 1$. We hereafter generally denote $p(X|Y)$ as the probability of X that is conditioned on Y or the likelihood of Y when X is observed.

The process of edge prediction is two-fold: first, we estimate the cluster assignments of a vertex pair; then, we predict whether an edge exists. Thus, the posterior predictive distribution $p(A_{ij} = 1|A^{(i,j)})$ of the model in which the vertices i and j are connected given dataset $A^{(i,j)}$, or the marginal likelihood of the learned model for the vertex pair, is the following:

$$\begin{aligned} p(A_{ij} = 1|A^{(i,j)}) &= \sum_{\sigma_i, \sigma_j} p(A_{ij} = 1|\sigma_i, \sigma_j) p(\sigma_i, \sigma_j|A^{(i,j)}) \\ &= \left\langle p(A_{ij} = 1|\sigma_i, \sigma_j) \right\rangle_{A^{(i,j)}}, \end{aligned} \quad (1)$$

where $\langle \dots \rangle_{A^{(i,j)}}$ is the average over $p(\sigma_i, \sigma_j|A^{(i,j)})$ and we have omitted some of the conditioned parameters. The error should be small when the prediction of edge existence for every vertex pair is accurate. In other words, the probability distribution in which each element has probability of equation (1) is close, in the sense of the Kullback–Leibler (KL) divergence, to the actual distribution, which is given as the empirical distribution. Therefore, it is natural to employ, as a measure of the prediction error, the cross-entropy error function³⁷

$$\begin{aligned} E_{\text{Bayes}}(q) &= -\overline{\log p(A_{ij}|A^{(i,j)})} \\ &= -\frac{1}{L} \sum_{i < j} \left[A_{ij} \log p(A_{ij} = 1|A^{(i,j)}) \right. \\ &\quad \left. + (1 - A_{ij}) \log p(A_{ij} = 0|A^{(i,j)}) \right], \end{aligned} \quad (2)$$

where $\bar{X} \equiv \sum_{i < j} X(A_{ij})/L$. Note that we have chosen the normalisation in such a way that E_{Bayes} is typically $O(1)$ in sparse networks. We refer to equation (2) as the Bayes prediction error, which corresponds to the LOOCV estimate of the *stochastic complexity*³⁸. As long as the model that we use is consistent with the data, the posterior predictive distribution is optimal as an element of the prediction error because the intermediate dependence (σ_i, σ_j) is fully marginalised and gives the smallest error.

Unfortunately, the assumption that the model that we use is consistent with the data is often invalid in practice. In that case, the Bayes prediction error E_{Bayes} may no longer be optimal for prediction. In equation (2), we employed $-\log p(A_{ij}|A^{(i,j)})$ as the error of a vertex pair. Instead, we can consider the log-likelihood of cluster assignments $-\log p(A_{ij}|\sigma_i, \sigma_j)$ to be a fundamental element and measure $\langle -\log p(A_{ij}|\sigma_i, \sigma_j) \rangle_{A^{(i,j)}}$ as the prediction error of a vertex pair. In other words, the cluster assignments (σ_i, σ_j) are drawn from the posterior distribution, and the error of the vertex pair is measured with respect to those fixed assignments. Then, the corresponding cross-entropy error function is

$$E_{\text{Gibbs}}(q) = \overline{\langle -\log p(A_{ij}|\sigma_i, \sigma_j) \rangle_{A^{(i,j)}}}. \quad (3)$$

We refer to equation (3) as the Gibbs prediction error. If the probability distribution of the cluster assignments is highly peaked, then E_{Gibbs} will be close to E_{Bayes} , and both will be relatively small if those assignments predict the actual network well. Alternatively, we can measure the *maximum a posteriori* (MAP) estimate of equation (3). In other words, instead of taking the average over $p(\sigma_i, \sigma_j|A^{(i,j)})$, we select the most likely assignments to measure the error. We refer to $E_{\text{MAP}}(q)$ as the prediction error.

Finally, we define the Gibbs training error, which we refer to as E_{training} . This error can be obtained by taking the average over $p(\sigma_i, \sigma_j|A)$ instead in equation (3), i.e.,

$$E_{\text{training}}(q) = \overline{\langle -\log p(A_{ij}|\sigma_i, \sigma_j) \rangle_A}. \quad (4)$$

Because we make use of the information regarding the edge existence when we take the average, the result is not a prediction error, but is the goodness of fit of the assumed model to the given data.

At first glance, the cross-validation errors that we presented above might appear computationally infeasible because we must know $p(\sigma_i, \sigma_j|A^{(i,j)})$ with respect to every vertex pair. In the Methods section, however, we show that we have analytical expressions of the cross-validation errors in terms of the outputs of BP; therefore, the model assessment for the sparse networks is very efficient.

In the following subsections, we show the performances of these cross-validation errors for various networks relative to the performance of the Bethe free energy.

Real-world networks. First, and most importantly, we show the performance of the Bethe free energy and cross-validation errors on real-world networks. Figures 1 and 3 show the results on the network of books on US politics³⁹ (which we refer to as *political books*). This network is a copurchase network whose vertices are books sold on [Amazon.com](https://www.amazon.com), and each edge represents a pair of books that was purchased by the same buyer; the meta-data of the dataset has the labels “conservative”, “liberal”, and “neutral.”

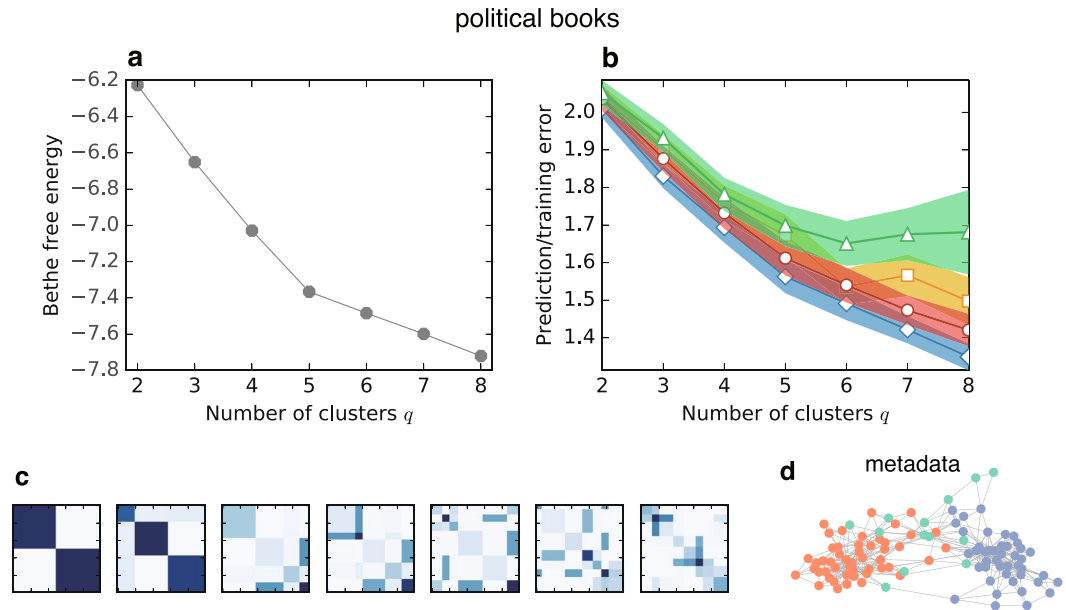


Figure 3. Model assessments and inferred clusters in the network of books about US politics. **(a)** Bethe free energy and **(b)** prediction and training errors as functions of the number of clusters q . The four data in **(b)** are, from top to bottom, Gibbs prediction errors E_{Gibbs} (green triangles), MAP estimates E_{MAP} of E_{Gibbs} (yellow squares), Bayes prediction errors E_{Bayes} (red circles), and Gibbs training errors E_{training} (blue diamonds). For each error, the constant term is neglected and the standard error is shown in shadow. **(c)** The learned parameters, ω and γ , are visualised from $q=2$ to 8. **(d)** The cluster assignments that are indicated in the metadata of the dataset; the vertices with the same colour belong to the same cluster.

Figure 3b shows that the cross-validation estimate of the Gibbs prediction error E_{Gibbs} saturates favourably, while the Bethe free energy (Fig. 3a), Bayes prediction error E_{Bayes} , and Gibbs training error E_{training} keep decreasing as we increase the input number of clusters q . Although the MAP estimate of the Gibbs prediction error E_{MAP} often exhibits similar behaviour to the Gibbs prediction error E_{Gibbs} , we observe in the results of other datasets that it does not appear to be distinctively superior to E_{Gibbs} .

Compared with the Gibbs prediction error E_{Gibbs} , the Bayes prediction error E_{Bayes} appears to be more sensitive to the assumption that a network is generated by the stochastic block model, so that we rarely observe the clear saturation of the LOOCV error. In a subsection below, we show how overfitting and underfitting occur by deriving analytical expressions for the differences among the cross-validation errors.

How should we select the number of clusters q^* from the obtained plots of cross-validation errors? Although this problem is well defined when we select the best model, i.e., the model that has the least error, it is common to select the most parsimonious model instead of the best model in practice. Then, how do we determine the “most parsimonious model” from the results? This problem is not well defined, and there is no principled prescription obtained by consensus.

However, there is an empirical rule called the “one-standard error” rule²⁷ that has often been used. Recall that each cross-validation estimate is given as an average error per edge; thus, we can also measure its standard error. The “one-standard error” rule suggests selecting the simplest model whose estimate is no more than one standard error above the estimate of the best model. In the case of the *political books* network, the best model is the model in which $q=6$. Because the simplest model within the range of one standard error is $q=5$, it is our final choice.

The actual partition for each value of q is shown in Figs 1 and 3c. The cluster assignments indicated in the metadata is presented in Fig. 3d for reference. In Fig. 3c, the learned values of the parameters ω and γ are visualised: a higher value element of the affinity matrix ω is indicated in a darker colour, and the size of an element reflects the fraction of the cluster size γ_{σ} . For $q=3$, we can identify two communities and a cluster that is connected evenly to those communities, as presumed in the metadata. For $q=5$, in addition, the sets of core vertices in each community are also detected. Note that recovering the labels in the metadata is not our goal. The metadata are not determined based on the network structure and are not the ground-truth partition⁴⁰.

It should also be noted that we minimise the Bethe free energy in the cluster inference step in every case. When we select the number of clusters q^* , i.e., in the model selection step, we propose to use the cross-validation errors instead of the Bethe free energy.

The results of other networks are shown in Fig. 4. They are the friendship network of a karate club⁴¹, coauthorship network of scientists who work on networks⁷ (see ref. 42 for details of the datasets), and the metabolic network of *C. elegans*⁴³. We refer to those as *karate club*, *network science*, and *C. elegans*, respectively. The results are qualitatively similar to the *political books* network. Note that the initial state dependency can be sensitive when the input number of clusters q is large. Therefore, the results can be unstable in such a region.

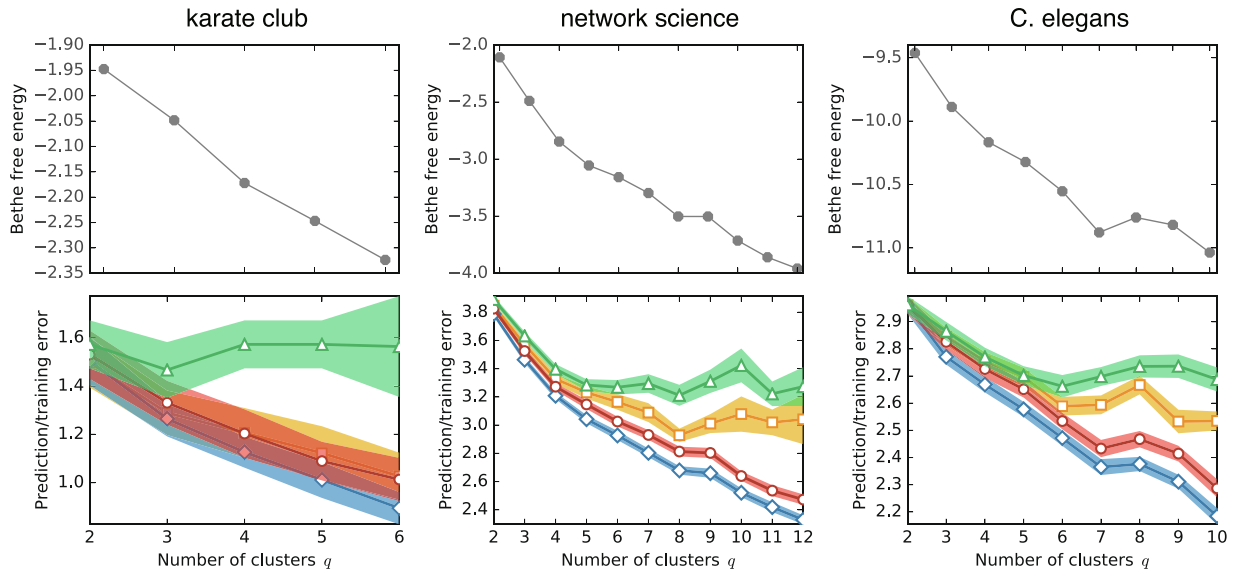


Figure 4. Results of model assessments for various real-world networks. They are plotted in the same manner as Fig. 3a,b.

From the results in Figs 3 and 4, one might conclude that the Gibbs prediction error E_{Gibbs} is the only useful criterion in practice. However, for example, when the holdout method is used instead of the LOOCV (see the Methods section), it can be confirmed that the Bayes prediction error E_{Bayes} also behaves reasonably.

The error for $q=1$. For the *karate club* network in Fig. 4, we see that the errors keep decreasing except for the Gibbs prediction error E_{Gibbs} . Recall that, for the stochastic block model with large q and low average degree to be detectable, the network is required to exhibit sufficiently strong modular structure²¹. Thus, although we do not a priori know which prediction error should be referred to, because the network has average degree smaller than 5, it is hard to believe that the errors other than E_{Gibbs} are appropriate. The Gibbs prediction error E_{Gibbs} has the smallest error at $q=3$, and the model in which $q=2$ has an error within the range of one standard error of $q=3$. However, one might suspect that the most parsimonious model is the model with $q=1$, i.e., the model that we should assume is a uniform random graph, and there is no statistically significant modular structure. In this case, the connection probability for an arbitrary vertex pair is determined by a single parameter, i.e., $p(A_{ij}|\sigma_i\sigma_j) = \omega$, and it is simply $\omega = 2L/N(N-1)$. Moreover, there is no difference among the errors that we listed above. Hence,

$$E_{\text{Bayes}} = E_{\text{Gibbs}} = E_{\text{MAP}} = E_{\text{training}} \approx 1 - \log \omega. \quad (5)$$

Here, we used the fact that $\omega = O(N^{-1})$ because we consider sparse networks. In every plot of cross-validation errors, the constant and $O(N^{-1})$ terms are neglected. The number of vertices and edges in the *karate club* network are $N=34$ and $L=78$, respectively, i.e., $-\log \omega \simeq 1.97$, which is much larger than the errors for $q=2$; we thus conclude that the number of clusters $q^*=2$.

Degree-corrected stochastic block model. It has been noted that the standard stochastic block model that we explained above is often not suitable for inferring clusters in real-world networks because many of them have a scale-free degree distribution, while the standard stochastic block model is restricted to having a Poisson degree distribution. This inconsistency affects both the cluster inference for a given q^* and the model selection. For example, as shown in Fig. 5a, we found that all of the criteria largely overfit for the *political blogs* network when the standard stochastic block model is used. The *political blogs* network is a network of hyperlinks between blogs on US politics (we neglect the directions of the edges), which is expected to have two clusters.

The degree-corrected stochastic block model^{44,45} is often used as an alternative. This model has the degree-correction parameter θ_i for each vertex in addition to γ and ω , where θ allows a cluster to have a heterogeneous degree distribution. (See ref. 44 for details of the model.) The cross-validation model assessment can be straightforwardly extended to the degree-corrected stochastic block model: for the inference of cluster assignments, the BP algorithm can be found in ref. 46; for the cross-validation errors, we only need to replace $\omega_{\sigma_i\sigma_j}$ for the probability $p(A_{ij} = 1|\sigma_i, \sigma_j)$ with $\theta_i\omega_{\sigma_i\sigma_j}\theta_j$ in each case. As shown in Fig. 5b, the assessment is reasonable when a degree-corrected stochastic block model is used. Although the error drops for $q \geq 5$, it is better to discard this part because the numbers of iterations until convergence become relatively large there; we regard it as the “wrong solution” that is mentioned in ref. 47.

When $q=1$, we do not assume the Erdős-Rényi random graph, but instead assume the random graph that has the same expected degree sequence as the actual dataset⁷, which is the model often used as the null model in

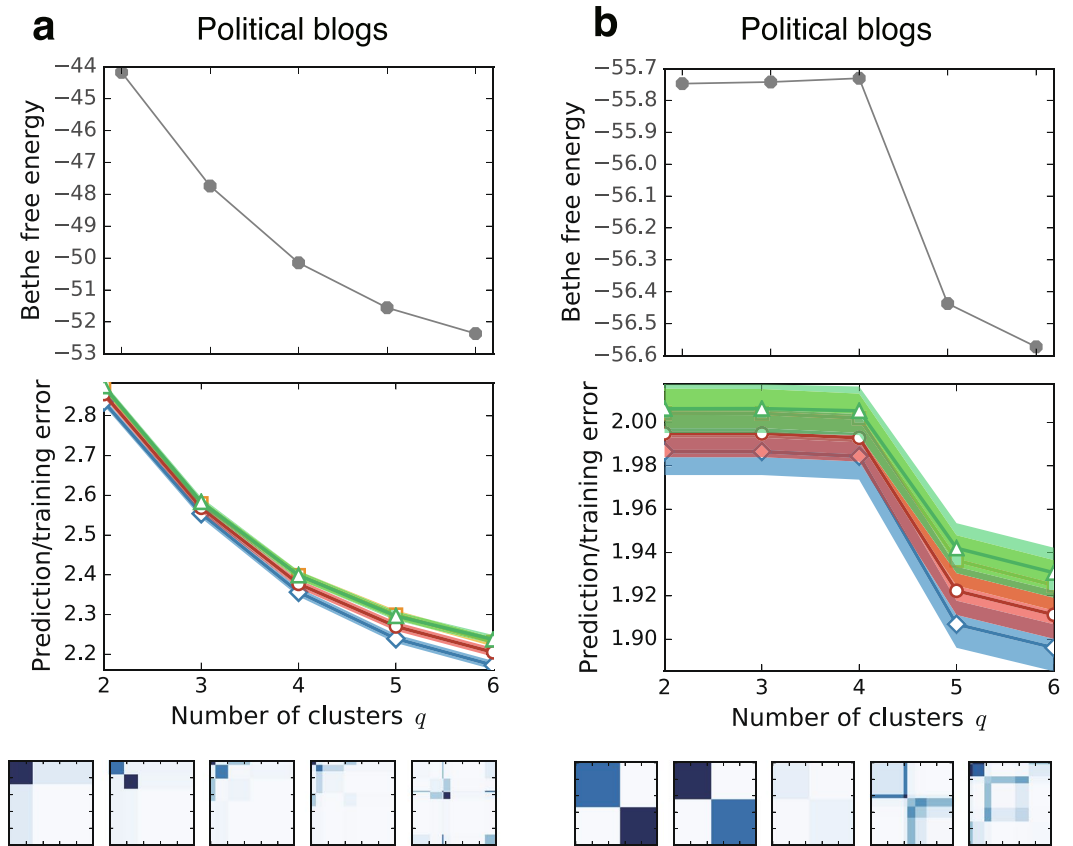


Figure 5. Results of the model assessments for the network of hyperlinks between blogs on US politics. **(a)** The results of the standard model and the learned parameters. **(b)** The results of the degree-corrected model and the learned parameters.

modularity. Thus, the probability that vertices i and j are connected is $d_i d_j / 2L$. After some algebra, assuming that $d_i d_j / 2L \ll 1$, the error for $q = 1$ is approximately

$$1 - \frac{1}{L} \sum_{i=1}^N d_i \log d_i + \log(2L), \quad (6)$$

which is equal to equation (5) when the network is regular. In the case of the *political blogs* network, the cross-validation error for $q = 1$ is approximately 3.42. (It is 2.42 in the plot because the constant term is neglected.) Thus, we obtain $q^* = 2$.

Synthetic networks and the detectability threshold. It is important to investigate how well the cross-validation errors perform in a critical situation, i.e., the case in which the network is close to the uniform random graph. To accomplish this goal, we observe the performance on the stochastic block model, the model that we assume for the inference itself. As explained in the subsection of the stochastic block model, the closeness to the uniform random graph is specified with the affinity matrix ω . Here, we consider a simple community structure: we set ω_{in} for the diagonal elements, and the remaining elements are set to ω_{out} , where $\omega_{\text{in}} \geq \omega_{\text{out}}$. We parametrise the closeness to the uniform random graph, with $\epsilon = \omega_{\text{out}} / \omega_{\text{in}}$; $\epsilon = 0$ represents that the network consists of completely disconnected clusters, and $\epsilon = 1$ represents the uniform random graph. There is a phase transition point ϵ^* above which it is statistically impossible to infer a planted structure. This point is called the detectability threshold. Our interest here is to determine whether the planted number of clusters q^* can be correctly identified using the cross-validation errors when ϵ is set to be close to ϵ^* .

Figure 6 shows the Bethe free energy and cross-validation errors for the stochastic block models. In each plot, we set $q^* = 4$, each cluster has 1,000 vertices, and the average degree c is set to 8. From bottom to top, the values of ϵ are 0.1, 0.15, 0.2, and 0.25, while the detectability threshold is $\epsilon^* \simeq 0.31^{21}$. Note that when the value of ϵ is close to ϵ^* , the inferred cluster assignments are barely correlated with the planted assignments; thus, the result is close to a random guess anyway.

In all of the plots, the curves of validation become smoother as we increase the value of ϵ , which indicates that it is difficult to identify the most parsimonious model. It is clear that the Bethe free energy and Bayes prediction error E_{Bayes} perform better than the Gibbs prediction error E_{Gibbs} in the present case because the networks we analyse correspond exactly to the model we assume. Figure 6 shows that the Gibbs prediction error E_{Gibbs} and its

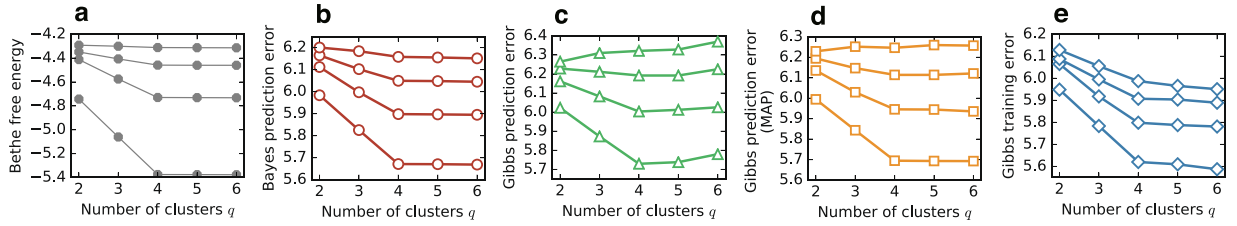


Figure 6. Performance of the model assessment criteria on various stochastic block models. The results of (a) Bethe free energy, (b) Bayes prediction error E_{Bayes} , (c) Gibbs prediction error E_{Gibbs} , (d) MAP estimate of the Gibbs prediction error E_{MAP} and (e) Gibbs training error E_{training} are shown. The lines in each plot represent the results of various values of ϵ .

MAP estimate E_{MAP} underestimate q^* near the detectability threshold. This finding is consistent with the results that we obtained for the real-world networks that the Bethe free energy and Bayes prediction error E_{Bayes} tend to overestimate. Indeed, under certain assumptions, we can derive that the Bayes prediction error E_{Bayes} identifies the planted number of clusters all the way down to the detectability threshold, while the Gibbs prediction error E_{Gibbs} strictly underfits near the detectability threshold. (See the Methods section for the derivation.) Therefore, there is a trade-off between the Bayes prediction error E_{Bayes} and the Gibbs prediction error E_{Gibbs} ; their superiority depends on the accuracy of the assumption of the stochastic block model and the fuzziness of the network.

Relation among the cross-validation errors. We show how the model assessment criteria that we consider in this study are related. First, we derive the relation among the errors E_{Bayes} , E_{Gibbs} , and E_{training} . By exploiting Bayes' rule, we have

$$\begin{aligned} \log p(A_{ij}|A^{(i,j)}) &= \log p(A_{ij}|\sigma_i, \sigma_j, A^{(i,j)}) \\ &+ \log \frac{p(\sigma_i, \sigma_j|A^{(i,j)})}{p(\sigma_i, \sigma_j|A)}. \end{aligned} \quad (7)$$

Note here that the left-hand side does not depend on σ_i and σ_j . If we take the average with respect to $p(\sigma_i, \sigma_j|A^{(i,j)})$ on both sides,

$$\begin{aligned} \log p(A_{ij}|A^{(i,j)}) &= \left\langle \log p(A_{ij}|\sigma_i, \sigma_j, A^{(i,j)}) \right\rangle \\ &+ D_{\text{KL}}\left(p(\sigma_i, \sigma_j|A^{(i,j)}) \parallel p(\sigma_i, \sigma_j|A)\right), \end{aligned} \quad (8)$$

where $D_{\text{KL}}(p||q)$ is the KL divergence. Taking the sample average of the edges, we obtain

$$E_{\text{Bayes}} = E_{\text{Gibbs}} - D_{\text{KL}}\left(p(\sigma_i, \sigma_j|A^{(i,j)}) \parallel p(\sigma_i, \sigma_j|A)\right). \quad (9)$$

If we take the average over $p(\sigma_i, \sigma_j|A)$ in equation (8) instead,

$$E_{\text{Bayes}} = E_{\text{training}} + D_{\text{KL}}\left(p(\sigma_i, \sigma_j|A) \parallel p(\sigma_i, \sigma_j|A^{(i,j)})\right). \quad (10)$$

Because the KL divergence is non-negative, we have

$$E_{\text{training}} \leq E_{\text{Bayes}} \leq E_{\text{Gibbs}}. \quad (11)$$

This inequality follows directly from Bayes' rule and applies broadly. Note, however, that the amounts of the errors do not indicate the relationships among the numbers of clusters selected.

Under a natural assumption, we can also derive the inequality for the number of clusters selected. Let q be the trial number of clusters. If the cluster assignment distributions for different values of q constitute a hierarchical structure, i.e., a result with a small q can be regarded as the coarse graining of a result with a larger q , then the information monotonicity^{48,49} of the KL divergence ensures that

$$q_{\text{training}}^* \geq q_{\text{Bayes}}^* \geq q_{\text{Gibbs}}^*. \quad (12)$$

In other words, the Bayes prediction error E_{Bayes} can only overfit when the Gibbs prediction error E_{Gibbs} estimates q^* correctly, while the Gibbs prediction error E_{Gibbs} can only underfit when the Bayes prediction error E_{Bayes} estimates q^* correctly. In addition, when the model predicts the edge existence relatively accurately, E_{MAP} is biased in such a way that the error becomes small. Therefore, E_{MAP} tends to be smaller than E_{Gibbs} . As observed for real-world networks, E_{Gibbs} typically performs well while E_{Bayes} often overfits in practice; equation (9) implies that

detailed information about the difference in the cluster assignment distribution is often not relevant and simply causes overfitting.

In addition to the relationship among the cross-validation errors, it is also fruitful to seek relationships between the Bethe free energy and the prediction errors. We found that it is possible to express the Bethe free energy in terms of the prediction errors. (See the Methods section for the derivation).

Discussion

To determine the number of clusters, we both proposed LOOCV estimates of prediction errors using BP and revealed some of the theoretical relationships among the errors. They are principled and scalable and, as far as we examined, perform well in practice. Unlike the BIC-like criteria, the prediction errors do not require the model consistency to be applicable. Moreover, although we only treated the standard and degree-corrected stochastic block models, the applicability of our LOOCV estimates is not limited to these models. With an appropriate choice of block models, we expect that the present framework can also be used in the vast majority of real-world networks, such as directed, weighted, and bipartite networks and networks that have positive and negative edges. This is in contrast to some criteria that are limited to specific modular structures.

The code for the results of the current study can be found at ref. 50. (See the description therein for more details of the algorithm).

The selection of the number of clusters q^* can sometimes be subtle. For example, as we briefly mentioned above, when the inference algorithm depends sensitively on the initial condition, the results of the LOOCV can be unstable; all of the algorithms share this type of difficulty as long as the non-convex optimisation problem is considered. Sometimes the LOOCV curve can be bumpy in such a way that it is difficult to determine the most parsimonious model. For this problem, we note that there is much information other than the prediction errors that help us to determine the number of clusters q^* , and we should take them into account. For example, we can stop the assessment or discard the result when (i) the number of iterations until convergence becomes large⁴⁷, as we have observed in Fig. 5b, (ii) the resulting partition does not exhibit a significant behaviour, e.g., no clear pattern in the affinity matrix, or (iii) the actual number of clusters does not increase as q increases. All of this information is available in the output of our code.

It is possible that the difficulty of selecting the number of clusters q^* arises in the model itself. Fitting with the stochastic block models is flexible; thus, the algorithm can infer not only the assortative and disassortative structures but also more complex structures. However, the flexibility also indicates that slightly different models might be able to fit the data as good as the most parsimonious model. As a result, we can obtain a gradually decreasing LOOCV curve for a broad range of q . Thus, there should be a trade-off between the flexibility of the model and the difficulty of the model assessment.

Despite the fact that model assessment based on prediction accuracy is a well-accepted principle, to the best of our knowledge, the method that is applicable to large-scale modular networks in the Bayesian framework is discussed in this study for the first time. For better accuracy and scalability, we believe that there is more to be done. One might investigate whether the LOOCV errors perform, in some sense, better than other BIC-like criteria. The relations and tendencies of the LOOCV errors compared to other model assessment criteria would certainly provide interesting future work even though it is unlikely that we can generally conclude the superiority of the criteria over the others²⁸. In practice, if resource allocation allows, it is always better to evaluate multiple criteria⁵¹.

Methods

Inference of the cluster assignments. Let us first explain the inference algorithm of the cluster assignment using the stochastic block model when the number of clusters q is given. For this part, we follow the algorithm introduced in refs 21, 34: the expectation–maximisation (EM) algorithm using BP.

An instance of the stochastic block model is generated based on selected parameters, (γ, ω) , as explained in the subsection of the stochastic block model. Because the cluster assignment of each vertex and the edge generation between each vertex pair are determined independently and randomly in the stochastic block model, its likelihood is given by

$$p(A, \sigma | \gamma, \omega, q) = \prod_{i=1}^N \gamma_{\sigma_i} \prod_{i < j} \omega_{\sigma_i \sigma_j}^{A_{ij}} (1 - \omega_{\sigma_i \sigma_j})^{1 - A_{ij}}. \quad (13)$$

The goal here is to minimise the free energy defined by $f = -N^{-1} \log \sum_{\sigma'} p(A, \sigma' | \gamma, \omega)$, which is equivalent to maximising the marginal log-likelihood. Using the identity $p(A, \sigma | \gamma, \omega) = p(\sigma | A, \gamma, \omega) p(A | \gamma, \omega)$, the free energy f can be expressed as

$$f = -\frac{1}{N} \log \frac{p(A, \sigma | \gamma, \omega)}{p(\sigma | A, \gamma, \omega)} \quad (14)$$

for an arbitrary σ . Taking the average over a probability distribution $q(\sigma)$ on both sides, we obtain the following variational expression.

$$\begin{aligned}
 f &= -\frac{1}{N} \sum_{\sigma} q(\sigma) \log \frac{p(A, \sigma | \gamma, \omega)}{q(\sigma)} \frac{q(\sigma)}{p(\sigma | A, \gamma, \omega)} \\
 &= \frac{1}{N} [-\sum_{\sigma} q(\sigma) \log p(A, \sigma | \gamma, \omega) + \sum_{\sigma} q(\sigma) \log q(\sigma) \\
 &\quad - D_{\text{KL}}(q(\sigma) \| p(\sigma | A, \gamma, \omega))].
 \end{aligned}
 \tag{15}$$

When $q(\sigma)$ is equal to the posterior distribution $p(\sigma | A, \gamma, \omega)$, the KL divergence disappears, and we should obtain the minimum. We can interpret the first and second terms as corresponding to the internal energy and negative entropy, respectively, and equation (15) as the thermodynamic relation of the free energy. For the first term, by substituting equation (13), we have

$$\begin{aligned}
 &-\frac{1}{N} \sum_{\sigma} q(\sigma) \log p(A, \sigma | \gamma, \omega) \\
 &= -\frac{1}{N} \sum_{\sigma} q(\sigma) \left[\sum_i \log \gamma_{\sigma_i} + \sum_{i < j} (A_{ij} \log \omega_{\sigma_i \sigma_j} + (1 - A_{ij}) \log (1 - \omega_{\sigma_i \sigma_j})) \right] \\
 &= -\frac{1}{N} \sum_i \sum_{\sigma} q_{\sigma}^i \log \gamma_{\sigma} - \frac{1}{N} \sum_{(i,j) \in E \sigma \sigma'} q_{\sigma \sigma'}^{ij} \log \omega_{\sigma \sigma'} + \frac{c}{2} + O(N^{-1}),
 \end{aligned}
 \tag{16}$$

where $q_{\sigma}^i = \langle \delta_{\sigma_i} \rangle_{\sigma}$ and $q_{\sigma \sigma'}^{ij} = \langle \delta_{\sigma_i} \delta_{\sigma'_j} \rangle_{\sigma}$ are the marginal probabilities, where δ_{xy} is the Kronecker delta and $\langle \dots \rangle_{\sigma}$ is the average over $q(\sigma)$. Therefore, the free energy minimisation in the variational expression requires two operations: the inference of the marginal probabilities of the cluster assignments, q_{σ}^i and $q_{\sigma \sigma'}^{ij}$, and the learning of the parameters, (γ, ω) . In the EM algorithm, they are performed iteratively.

For the E-step of the EM algorithm, we infer the marginal probabilities of the cluster assignments using the current estimates of the parameters. Whereas the exact marginalisation is computationally expensive in general, BP yields an accurate estimate quite efficiently when the network is sparse, i.e., the case in which the network is locally tree-like. Here, we do not go over the complete derivation of the BP equations, which can be found in refs 21, 52. We denote ψ_{σ}^i as the BP estimate of q_{σ}^i , which is obtained by

$$\psi_{\sigma}^i = \frac{1}{Z^i} \gamma_{\sigma} e^{-h_{\sigma}} \prod_{k \in \partial i} \left(\sum_{\sigma_k} \psi_{\sigma_k}^{k \rightarrow i} \omega_{\sigma_k \sigma} \right),
 \tag{17}$$

where Z^i is the normalisation factor with respect to the cluster assignment σ , $h_{\sigma} = \sum_{k=1}^N \sum_{\sigma_k} \psi_{\sigma_k}^k \omega_{\sigma_k \sigma}$, which is due to the effect of non-edges $(i, k) \notin E$, and ∂i indicates the set of neighbouring vertices of i . In equation (17), we also have the so-called cavity bias $\psi_{\sigma}^{i \rightarrow j}$ for an edge $(i, j) \in E$; the cavity bias is the marginal probability of vertex i without the marginalisation from vertex j . Analogously to equation (17), the update equation for the cavity bias is obtained by

$$\psi_{\sigma}^{i \rightarrow j} = \frac{1}{Z^{i \rightarrow j}} \gamma_{\sigma} e^{-h_{\sigma}} \prod_{k \in \partial i \setminus j} \left(\sum_{\sigma_k} \psi_{\sigma_k}^{k \rightarrow i} \omega_{\sigma_k \sigma} \right),
 \tag{18}$$

where $\partial i \setminus j$ indicates the set of neighbouring vertices of i except for j , and $Z^{i \rightarrow j}$ is the normalisation factor. For the BP estimate $\psi_{\sigma \sigma'}^{ij}$ of the two-point marginal $q_{\sigma \sigma'}^{ij}$ for $(i, j) \in E$, we have

$$\psi_{\sigma \sigma'}^{ij} = \frac{1}{Z^{ij}} \psi_{\sigma}^{i \rightarrow j} \omega_{\sigma \sigma'} \psi_{\sigma'}^{j \rightarrow i},
 \tag{19}$$

where Z^{ij} is the normalisation factor with respect to the assignments σ and σ' .

With these marginals in hand, in the M-step of the EM algorithm, we update the estimate of the parameters to $\hat{\gamma}$ and $\hat{\omega}$ as

$$\hat{\gamma}_{\sigma} = \frac{1}{N} \sum_{i=1}^N \psi_{\sigma}^i,
 \tag{20}$$

$$\hat{\omega}_{\sigma \sigma'} = \frac{1}{N^2 \gamma_{\sigma} \gamma_{\sigma'} \sum_{(i,j) \in E} Z^{ij}} \sum_{(i,j) \in E} \omega_{\sigma \sigma'} (\psi_{\sigma}^{i \rightarrow j} \psi_{\sigma'}^{j \rightarrow i} + \psi_{\sigma}^{j \rightarrow i} \psi_{\sigma'}^{i \rightarrow j}).
 \tag{21}$$

Here, we have used the fact that the network is undirected. These update rules can be obtained by simply taking the derivatives with respect to the parameters in equation (16) with the normalisation constraint $\sum_{\sigma} \gamma_{\sigma} = 1$.

We recursively compute equation (18) and the parameter learning, equations (20) and (21), until convergence; then, we obtain the full marginal using equation (17), which yields the estimates of the cluster assignments of the vertices.

Bethe free energy and derivation of the cross-validation errors. In the algorithm above, because we use BP to estimate the marginal probabilities, we are no longer minimising the free energy. Instead, as an approximation of the free energy, we minimise the Bethe free energy, which is expressed as

$$f_{\text{Bethe}} = -\frac{1}{N} \sum_i \log Z^i + \frac{1}{N} \sum_{(i,j) \in E} \log Z^{ij} - \frac{c}{2}, \quad (22)$$

where Z^i and Z^{ij} are the normalisation factors that appeared in equations (17) and (19). As mentioned in the Introduction section, to determine the number of clusters q^* from the Bethe free energy f_{Bethe} , we select the most parsimonious model among the models that have low f_{Bethe} . This approach corresponds to taking the maximum likelihood estimation of the parameters.

We now explain the derivations of the cross-validation errors that we used in the Results section. The LOOCV estimate of the Bayes prediction error E_{Bayes} is measured by equation (2), and its element $p(A_{ij} = 1 | A^{(i,j)})$ is given as equation (1). The first factor in the sum of equation (1) is simply $p(A_{ij} = 1 | \sigma_i, \sigma_j) = \omega_{\sigma_i \sigma_j}$, by the model definition. An important observation is that because the cavity bias $\psi_{\sigma_i}^{i \rightarrow j}$ represents the marginal probability of vertex i without the information from vertex j , this entity is exactly what we need for prediction in the LOOCV, i.e., $p(\sigma_i, \sigma_j | A^{(i,j)}) = \psi_{\sigma_i}^{i \rightarrow j} \psi_{\sigma_j}^{j \rightarrow i}$. Additionally, recall that $p(\sigma_i, \sigma_j | A_{ij} = 1, A^{(i,j)}) = \psi_{\sigma_i}^{i \rightarrow j} \omega_{\sigma_i \sigma_j} \psi_{\sigma_j}^{j \rightarrow i} / Z^{ij}$, where Z^{ij} is defined in equation (19). Thus, we have $p(A_{ij} = 1 | A^{(i,j)}) = Z^{ij}$. By using the fact that $L = O(N)$ and $p(A_{ij} = 1 | A^{(i,j)}) = O(N^{-1})$, $E_{\text{Bayes}}(q)$ can be written as

$$E_{\text{Bayes}}(q) = 1 - \frac{1}{L} \sum_{(i,j) \in E} \log Z^{ij} + O(N^{-1}). \quad (23)$$

Equation (23) indicates that the prediction with respect to the non-edges contributes only a constant; thus, $E_{\text{Bayes}}(q)$ essentially measures whether the existence of the edges is correctly predicted in a sparse network.

The Gibbs prediction error E_{Gibbs} in equation (3) can be obtained in the same manner. Using the approximation that the network is sparse, it can be written in terms of the cavity biases as

$$E_{\text{Gibbs}}(q) \simeq 1 - \frac{1}{L} \sum_{(i,j) \in E} \sum_{\sigma_i, \sigma_j} \psi_{\sigma_i}^{i \rightarrow j} \psi_{\sigma_j}^{j \rightarrow i} \log \omega_{\sigma_i \sigma_j}. \quad (24)$$

The (MAP) estimate of the Gibbs prediction error $E_{\text{MAP}}(q)$ can be obtained by replacing $\psi_{\sigma_i}^{i \rightarrow j}$ with $\delta_{\sigma_i, \text{argmax}\{\psi_{\sigma}^{i \rightarrow j}\}}$ in equation (24).

For the Gibbs training error E_{training} in equation (4), we take the average over $p(\sigma_i, \sigma_j | A)$ instead of $p(\sigma_i, \sigma_j | A^{(i,j)})$. Thus, we have

$$E_{\text{training}}(q) \simeq 1 - \frac{1}{L} \sum_{(i,j) \in E} \sum_{\sigma_i, \sigma_j} \frac{\psi_{\sigma_i}^{i \rightarrow j} \omega_{\sigma_i \sigma_j} \psi_{\sigma_j}^{j \rightarrow i}}{Z^{ij}} \log \omega_{\sigma_i \sigma_j}. \quad (25)$$

This training error can be interpreted as a part of the internal energy because it corresponds to the BP estimate of the second term in equation (16).

Note that all of the cross-validation errors, equations (23–25), are analytical expressions in terms of the parameters and cavity biases. Therefore, we can readily measure those errors by simply running the algorithm once. It should also be noted that using BP for the LOOCV itself is not totally new; this idea has been addressed in a different context in the literature, e.g., ref. 53.

Detectability of the Bayes and Gibbs prediction errors. Let us evaluate the values of the Bayes and Gibbs prediction errors, E_{Bayes} and E_{Gibbs} , for the stochastic block model with the simple community structure, i.e., $\omega_{\sigma \sigma'} = (\omega_{\text{in}} - \omega_{\text{out}}) \delta_{\sigma \sigma'} + \omega_{\text{out}}$. By substituting this affinity matrix to equations (23) and (24), we obtain

$$E_{\text{Bayes}}(q) = 1 - \log \omega_{\text{out}} - \frac{1}{L} \sum_{(i,j) \in E} \log \left(1 + \frac{\omega_{\text{in}} - \omega_{\text{out}}}{\omega_{\text{out}}} \sum_{\sigma=1}^q \psi_{\sigma}^{i \rightarrow j} \psi_{\sigma}^{j \rightarrow i} \right) \quad (26)$$

$$E_{\text{Gibbs}}(q) = 1 - \log \omega_{\text{out}} - \left(\log \frac{\omega_{\text{in}}}{\omega_{\text{out}}} \right) \frac{1}{L} \sum_{(i,j) \in E} \sum_{\sigma=1}^q \psi_{\sigma}^{i \rightarrow j} \psi_{\sigma}^{j \rightarrow i}. \quad (27)$$

As we sweep the parameter $\epsilon = \omega_{\text{out}} / \omega_{\text{in}} \rightarrow 1$ in the stochastic block model that has 2 clusters, $q=2$ is selected as long as the prediction error for $q=2$ is less than the error for $q=1$. When the number of vertices N is sufficiently large, $\omega = \omega_{\text{out}} + (\omega_{\text{in}} - \omega_{\text{out}}) \sum_{\sigma=1}^2 \gamma_{\sigma}^2$ in equation (5). Thus,

$$\begin{aligned}
 E_{\text{Bayes}}(q = 1) &= E_{\text{Gibbs}}(q = 1) \\
 &= 1 - \log \omega_{\text{out}} - \log \left(1 + \frac{\omega_{\text{in}} - \omega_{\text{out}}}{\omega_{\text{out}}} \sum_{\sigma=1}^2 \gamma_{\sigma}^2 \right),
 \end{aligned}
 \tag{28}$$

where we neglected the $O(N^{-1})$ contribution. Let us presume that the cavity bias $\psi_{\sigma}^{i \rightarrow j}$ is correlated with the planted structure up to the detectability threshold ϵ^* and $\psi_{\sigma}^{i \rightarrow j} = \gamma_{\sigma}$ for $\epsilon \geq \epsilon^*$, which is rigorously proven^{35, 36} in the case of 2 clusters of equal size. We then have

$$E_{\text{Bayes}}(q = 2) = E_{\text{Bayes}}(q = 1) \tag{29}$$

in the undetectable region. In the case of equal size clusters, $E_{\text{Bayes}}(q = 2)$ is strictly smaller than $E_{\text{Bayes}}(q = 1)$ in the detectable region because the last term in (26) is minimised when the distribution is uniform. In contrast, when $\psi_{\sigma}^{i \rightarrow j} = \gamma_{\sigma}$, the inequality $\log(1 + xy) > y \log(1 + x)$ ($x > 0, 0 < y < 1$) ensures that

$$E_{\text{Gibbs}}(q = 2) > E_{\text{Gibbs}}(q = 1). \tag{30}$$

Therefore, assuming that the parameters γ and ω are learned accurately, the Bayes prediction error E_{Bayes} suggests $q^* = 2$ all the way down to the detectability threshold, while the Gibbs prediction error E_{Gibbs} strictly underfits near the detectability threshold.

The Bethe free energy in terms of the prediction errors. We can also express the Bethe free energy f_{Bethe} in terms of the prediction errors. Ignoring the constant terms and factors, we observe that the Bayes prediction error E_{Bayes} , equation (23), is one component of the Bethe free energy, equation (22). The remaining part is the contribution of $-\sum_i \log Z^i$ that arises as follows.

Let us consider hiding a vertex and the edges and non-edges that are incident to that vertex, instead of an edge. We denote $\{A_{ij}\}_{j \in V}$ as the set of adjacency matrix elements that are incident to vertex i , and A^i as the adjacency matrix in which the information of $\{A_{ij}\}_{j \in V}$ is unobserved. For the prediction of an edge's existence, for example, because the unobserved vertex receives no information about its cluster assignment from its neighbours, we have

$$\begin{aligned}
 p(A_{ij} = 1 | A^i) &= \sum_{\sigma_i \sigma_j} p(A_{ij} = 1 | \sigma_i \sigma_j) p(\sigma_i \sigma_j | A^i) \\
 &= \sum_{\sigma_i \sigma_j} \omega_{\sigma_i \sigma_j} \gamma_{\sigma_i} \gamma_{\sigma_j} \psi_{\sigma_i}^{j \rightarrow i}.
 \end{aligned}
 \tag{31}$$

When we consider the prediction of the set of edges $\{A_{ij}\}_{j \in V}$, we find that all of the vertex pairs share the same cluster assignment for vertex i . Therefore, using the approximation that the network is sparse, we have

$$\begin{aligned}
 p(\{A_{ij}\}_{j \in V} | A^i) &= \sum_{\sigma_i} p(\sigma_i | A^i) \\
 &\quad \times \prod_{j \in \partial i} \left(\sum_{\sigma_j} p(A_{ij} = 1 | \sigma_i \sigma_j) p(\sigma_j | A^i) \right) \\
 &\quad \times \prod_{k \notin \partial i} \left(\sum_{\sigma_k} p(A_{ik} = 0 | \sigma_i \sigma_k) p(\sigma_k | A^i) \right) \\
 &\approx \sum_{\sigma_i} \gamma_{\sigma_i} e^{-h_{\sigma_i}} \prod_{j \in \partial i} \left(\sum_{\sigma_j} \psi_{\sigma_j}^{j \rightarrow i} \omega_{\sigma_i \sigma_j} \right) \\
 &= Z^i.
 \end{aligned}
 \tag{32}$$

Again, due to the sparsity, the prediction for the non-edges happened to be neglected, and it is essentially the prediction of the edges. Analogously to equation (2), we can consider a *leave-one-vertex-out* version of the Bayes prediction error, which we refer to as E_{Bayes}^v , as follows.

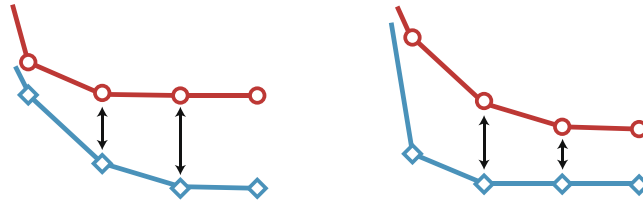


Figure 7. Schematic of the shapes of the error functions. As long as the gap between two error functions is nondecreasing, the error with the smaller value does not saturate earlier than the other.

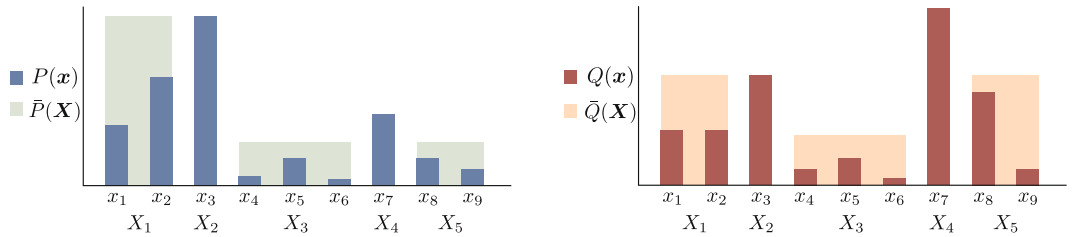


Figure 8. Example of the refinement of the probability distributions. We can regard $P(x)$ and $Q(x)$ as the refinements of $\bar{P}(X)$ and $\bar{Q}(X)$, respectively, with $\mathbf{x}^1 = \{x_1, x_2\}$, $\mathbf{x}^2 = \{x_3\}$, $\mathbf{x}^3 = \{x_4, x_5, x_6\}$, $\mathbf{x}^4 = \{x_7\}$, and $\mathbf{x}^5 = \{x_8, x_9\}$ as a possible correspondence. Note that the correspondence is not unique. If we refer only to $Q(x)$ and $\bar{Q}(X)$, the assignment of $\{x_1, x_2\}$, $\{x_3\}$, and $\{x_8, x_9\}$ is exchangeable within X_1, X_2 , and X_5 . However, $\bar{P}(X_5)$ does not coincide with $P(\{x_1, x_2\})$ or $P(\{x_3\})$; therefore, only X_1 and X_2 are exchangeable between $\{x_1, x_2\}$ and $\{x_3\}$. The same argument applies to $\{x_4, x_5, x_6\}$ and $\{x_8, x_9\}$.

$$\begin{aligned}
 E_{\text{Bayes}}^v(q) &= -\frac{1}{L} \sum_{ij} \left[A_{ij} \log p(A_{ij} = 1 | A^i) \right. \\
 &\quad \left. + (1 - A_{ij}) \log p(A_{ij} = 0 | A^i) \right] \\
 &= -\frac{1}{L} \sum_i \log \left(\prod_{j \in \partial i} p(A_{ij} = 1 | A^i) \prod_{j \notin \partial i} p(A_{ij} = 0 | A^i) \right) \\
 &= -\frac{1}{L} \sum_i \log p \left(\{A_{ij}\}_{j \in V} | A^i \right) \\
 &= -\frac{1}{L} \sum_i \log Z^i.
 \end{aligned} \tag{33}$$

Again, the edges and non-edges that are incident to the unobserved vertex i share the same cluster assignment σ_i at one end. In this definition, however, the prediction of an edge’s existence is made twice for every edge. If the Bayes prediction error E_{Bayes} is subtracted to prevent this over-counting, it is exactly the Bethe free energy f_{Bethe} up to the constant terms and the normalisation factor.

Derivation of inequality (12). Recall that the Bayes prediction error E_{Bayes} , Gibbs prediction error E_{Gibbs} , and Gibbs training error E_{training} are related via equations (9) and (10). We select the number of clusters q as the point at which the error function saturates (i.e., stops decreasing) with increasing q . For a smaller q to be selected by E_{Bayes} than by E_{Gibbs} , the gap between them, $D_{\text{KL}}(p(\sigma_i, \sigma_j | A^{(i,j)}) || p(\sigma_i, \sigma_j | A))$, must decrease (see Fig. 7). In this subsection, we explain the information monotonicity of the KL divergence and when it is applicable in the present context.

Let us consider sets of variables $\mathbf{X} = \{X_1, \dots, X_m\}$ and $\mathbf{x} = \{x_1, \dots, x_n\}$ ($n > m$). We first set a pair of probability distributions $\bar{P}(X)$ and $\bar{Q}(X)$ on X . Then we define a pair of probability distributions $P(x)$ and $Q(x)$ on x as refinements of $\bar{P}(X)$ and $\bar{Q}(X)$, respectively, if there exists a family of sets $\{\mathbf{x}^\mu\}_{\mu=1}^m$ that is a partition of x , i.e., $\mathbf{x}^\mu \cap \mathbf{x}^{\mu'} = \emptyset$ for $\mu \neq \mu'$ and $\cup_\mu \mathbf{x}^\mu = x$, that satisfies $P(\mathbf{x}^\mu) = \bar{P}(X_\mu)$ and $Q(\mathbf{x}^\mu) = \bar{Q}(X_\mu)$ for any μ . In other words, X can be regarded as the coarse graining of x . An example is given in Fig. 8. Note, however, that if X is actually constructed as the coarse graining of x , the above condition trivially holds. In general, a family that satisfies the above condition might not exist; even if it exists, it might not be unique.

The information monotonicity of the KL divergence states that for $P(x)$ and $Q(x)$, which are the refinements of $\bar{P}(X)$ and $\bar{Q}(X)$, respectively, we have

$$D_{\text{KL}}(P || Q) \geq D_{\text{KL}}(\bar{P} || \bar{Q}), \tag{34}$$

which is natural, because the difference between the distributions is more visible at a finer resolution. Equation (34) is deduced by the convexity of the KL divergence. First, we can rewrite the right-hand side of equation (34) in terms of P and Q as follows:

$$D_{\text{KL}}(\bar{P} \parallel \bar{Q}) = \sum_{\mu=1}^m P(x^\mu) \log \frac{P(x^\mu)}{Q(x^\mu)}; \quad (35)$$

Thus, if

$$\sum_{x \in x^\mu} P(x) \log \frac{P(x)}{Q(x)} \geq P(x^\mu) \log \frac{P(x^\mu)}{Q(x^\mu)} \quad (36)$$

holds for an arbitrary μ , then equation (34) holds. We split x^μ into $x_1 \in x^\mu$ and $x^\mu \setminus x_1$, and we denote the corresponding probabilities as $P_1 := P(x_1)$, $Q_1 := Q(x_1)$, $P_1^c := P(x^\mu \setminus x_1)$, and $Q_1^c := Q(x^\mu \setminus x_1)$. The right-hand side of equation (36) is then

$$\begin{aligned} (P_1 + P_1^c) \log \left(\frac{P_1 + P_1^c}{Q_1 + Q_1^c} \right) &= -(P_1 + P_1^c) \log \left(\frac{P_1}{P_1 + P_1^c} \frac{Q_1}{Q_1 + Q_1^c} + \frac{P_1^c}{P_1 + P_1^c} \frac{Q_1^c}{Q_1 + Q_1^c} \right) \\ &\leq P_1 \log \frac{P_1}{Q_1} + P_1^c \log \frac{P_1^c}{Q_1^c}, \end{aligned} \quad (37)$$

where we used the convexity of the logarithmic function. By repeating the same argument for the second term of equation (37), we obtain equation (36). Although the KL divergence is our focus, the information monotonicity holds more generally, e.g., for f -divergence⁴⁹.

We now use the information monotonicity to estimate the error functions. In the present context, the sets of variables \mathbf{X} and \mathbf{x} correspond to the cluster assignments of different q 's, (σ_i, σ_j) with q and (σ'_i, σ'_j) with q' ($q' > q$), for a vertex pair i and j . Because the labelling of the clusters is common to all of the vertices, we require that the refinement condition is satisfied with the common family of sets for every vertex pair. Under this condition, the KL divergence is nondecreasing as a function of q , which indicates that E_{Bayes} does not saturate earlier than E_{Gibbs} . Similarly, E_{training} does not saturate earlier than E_{Bayes} . Although the refinement condition that we required above is too strict to be satisfied exactly in a numerical calculation, it is what we naturally expect when the algorithm detects a hierarchical structure or the same structure with excess numbers of clusters. Moreover, the argument above is only a sufficient condition. Therefore, we naturally expect that E_{Gibbs} suggests a smaller number of clusters than E_{Bayes} and E_{training} quite commonly in practice. In addition, note that if we use a different criterion for the selection of q , e.g., the variation of the slope of the error function, the above conclusion can be violated.

Holdout method and K -fold cross-validation. Other than the LOOCV, it is also possible to measure the prediction errors by the holdout method and the K -fold cross-validation using BP. They can be conducted by randomly selecting the set of edges to be held out (i.e., the holdout set) and running BP that ignores the cavity biases sent by the held-out edges. In the holdout method, unlike the K -fold cross-validation, the test set is never used as the training set and vice versa. The results are listed in Fig. 9. We observe that the overfitting of the Bayes prediction error E_{Bayes} appears to be prevented. Although the resulting behaviours are somewhat reasonable, we have to bear in mind that they have following conceptual and computational issues.

First, we cannot expect the cross-validation assessment to work at all when the holdout set is too large, or equivalently, when the training set is too small. It is because, when a significant fraction of edges that are connected to vertices in certain clusters are held out, it is impossible to learn the underlying block structure correctly³¹. Second, when the holdout set has more than one edge, we actually need to run the whole algorithm twice for each holdout set; first to compute the posterior distribution given the training set, and then, to evaluate the predictive distribution for the holdout set based on the obtained posterior distribution. Then, we need to repeat this process as many times as the number of holdout sets. Again, none of the above processes are required for the LOOCV, because we have the analytical expression. Thus, the computational cost for the holdout method and the K -fold cross-validation are orders of magnitude larger than the LOOCV.

Besides the above issues, a more delicate treatment is required in the K -fold cross-validation. When the unobserved edges are connected, a prediction error is not given as an independent sum of the error per edge, and we need the calculation as we have done in equation (32). Therefore, the precise measurement is even more time consuming. In Fig. 9, we imposed the constraint that the edges are not connected in the holdout method. For the K -fold cross-validation, we simply ignored the fact that the unobserved edges might be connected, although this discrepancy may not be negligible.

We lose all the advantages that we had in the LOOCV when the holdout method and the K -fold cross-validation are considered. However, when the computational efficiency can be compromised, the holdout method with small holdout set can be a promising alternative.

Justification of the LOOCV estimates. It is a common mistake that one will make use of the information of the test set (the unobserved edge in the present context) for the training of the model in the process of cross-validation. Precisely speaking, however, the information of the unobserved edge does contribute to the training of the model in our procedure; thus, the cross-validation estimate here is not correct in a rigorous sense. Nevertheless, the treatment that we propose is justified because the contribution of the unobserved edge in the present setting is negligible. First, when we update the distribution of the cluster assignment $\psi_\sigma^{i \rightarrow j}$, we make use of

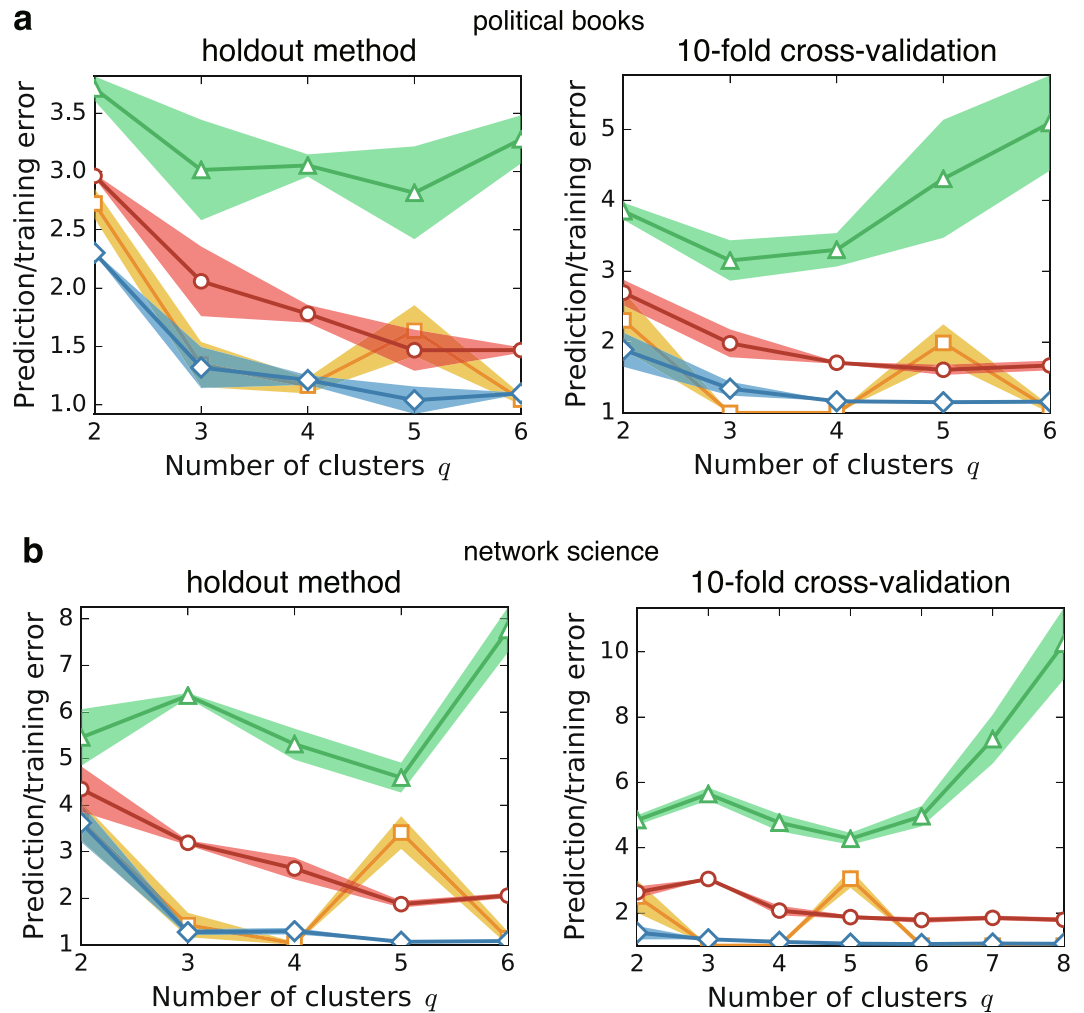


Figure 9. Cross-validation errors using the holdout method and the 10-fold cross-validation. As in Fig. 3, the four data in each plot indicate the Gibbs prediction errors E_{Gibbs} (green triangles), MAP estimates E_{MAP} of E_{Gibbs} (yellow squares), Bayes prediction errors E_{Bayes} (red circles), and Gibbs training errors E_{training} (blue diamonds). In the holdout method, we randomly selected 1% of the edges (5 edges in (a) *political books* and 10 edges in (b) *network science*) as the holdout set. The solid lines represent the average values obtained by repeating the prediction process 10 times and the shadows represent their standard errors. For the K -fold cross-validation, we set $K=10$. Although it is common to take average over different choices of K partition, we omitted this averaging process.

the information of the unobserved edge when there are loops in the network. Because we consider the sparse network, which is locally tree-like, the effect of the loops is typically negligible. Second, the update equations (20) and (21) of the parameters in the maximisation step of the EM algorithm contain the unobserved edge. However, because we consider the LOOCV in which only a single vertex pair is unobserved, this discrepancy results in a difference of only $O(N^{-1})$ to the parameter estimation. Therefore, our analytical results in terms of BP are sufficiently accurate to serve as error estimates when using the LOOCV.

Data Availability. The datasets generated during and/or analysed during the current study are available in M. E. J. Newman's repository, <http://www-personal.umich.edu/~mejn/netdata/> ref. 42.

References

1. Barabasi, A.-L. *Network Science* 1 edn. (Cambridge University Press, 2016).
2. Newman, M. E. J. *Networks: An Introduction* (Oxford university press, 2010).
3. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
4. Leger, J.-B., Vacher, C. & Daudin, J.-J. Detection of structurally homogeneous subsets in graphs. *Stat. Comput.* **24**, 675–692 (2014).
5. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821–7826 (2002).
6. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **101**, 2658–63 (2004).
7. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).

8. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**, 046110 (2008).
9. Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: First steps. *Soc. Networks* **5**, 109–137 (1983).
10. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).
11. Zhang, P. & Moore, C. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proc. Natl. Acad. Sci. USA* **111**, 18144–18149 (2014).
12. Rosvall, M. & Bergstrom, C. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **105**, 1118–1123 (2008).
13. Rosvall, M. & Bergstrom, C. T. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS One* **6**, e18209 (2011).
14. Newman, M. E. J. Spectral methods for community detection and graph partitioning. *Phys. Rev. E* **88**, 042822 (2013).
15. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 888–905 (2000).
16. Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
17. Krzakala, F. *et al.* Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci. USA* **110**, 20935–40 (2013).
18. Abbe, E. & Sandon, C. Recovering communities in the general stochastic block model without knowing the parameters. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R. (eds) *Advances in Neural Information Processing Systems 28*, 676–684 (Curran Associates, Inc., 2015).
19. Nowicki, K. & Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96**, 1077–1087 (2001).
20. Daudin, J. J., Picard, F. & Robin, S. A mixture model for random graphs. *Stat. Comput.* **18**, 173–183 (2008).
21. 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**, 066106 (2011).
22. Hayashi, K., Konishi, T. & Kawamoto, T. A tractable fully bayesian method for the stochastic block model. *arXiv preprint arXiv:1602.02256* (2016).
23. Newman, M. E. J. & Reinert, G. Estimating the number of communities in a network. *Phys. Rev. Lett.* **117**, 078301 (2016).
24. Peixoto, T. P. Parsimonious module inference in large networks. *Phys. Rev. Lett.* **110**, 148701 (2013).
25. Peixoto, T. P. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X* **4**, 011047 (2014).
26. Peixoto, T. P. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Phys. Rev. X* **5**, 011033 (2015).
27. Hastie, T. J., Tibshirani, R. J. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics (Springer, New York, 2009).
28. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection. *Statist. Surv.* **4**, 40–79 (2010).
29. Celeux, G. & Durand, J.-B. Selecting hidden markov model state number with cross-validated likelihood. *Comput. Stat.* **23**, 541–564 (2008).
30. Vehtari, A. & Ojanen, J. A survey of bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **142–228** (2012).
31. Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. Mixed membership stochastic blockmodels. In Koller, D., Schuurmans, D., Bengio, Y. & Bottou, L. (eds) *Advances in Neural Information Processing Systems 21*, 33–40 (Curran Associates, Inc., 2009).
32. Hoff, P. Modeling homophily and stochastic equivalence in symmetric relational data. In Platt, J. C., Koller, D., Singer, Y. & Roweis, S. T. (eds) *Advances in Neural Information Processing Systems 20*, 657–664 (Curran Associates, Inc., 2008).
33. Chen, K. & Lei, J. Network cross-validation for determining the number of communities in network data. *arXiv preprint arXiv:1411.1715* (2014).
34. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Inference and Phase Transitions in the Detection of Modules in Sparse Networks. *Phys. Rev. Lett.* **107**, 065701 (2011).
35. Mossel, E., Neeman, J. & Sly, A. Reconstruction and estimation in the planted partition model. *Probab. Theory Relat. Fields* **1–31** (2014).
36. Massoulié, L. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing, STOC '14*, 694–703 (ACM, New York, NY, USA, 2014).
37. Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006).
38. Levin, E., Tishby, N. & Solla, S. A. A statistical approach to learning and generalization in layered neural networks. In *Proceedings of the Second Annual Workshop on Computational Learning Theory, COLT '89*, 245–260 (1989).
39. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**, 8577–8582 (2006).
40. Peel, L., Larremore, D. B. & Clauset, A. The ground truth about metadata and community detection in networks. *arXiv:1608.05878* (2016).
41. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
42. Newman, M. E. J. <http://www-personal.umich.edu/~mejn/netdata/> (Date of access: 11/05/2015) (2006).
43. Duch, J. & Arenas, A. Community detection in complex networks using extremal optimization. *Phys. Rev. E* **72**, 027104 (2005).
44. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107 (2011).
45. Zhao, Y., Levina, E. & Zhu, J. Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Stat.* **2266–2292** (2012).
46. Yan, X. *et al.* Model selection for degree-corrected block models. *J. Stat. Mech. Theor. Exp.* **2014**, P05007 (2014).
47. Newman, M. E. J. & Clauset, A. Structure and inference in annotated networks. *Nat. Commun.* **7**, 11863 (2016).
48. Csiszár, I. Axiomatic characterizations of information measures. *Entropy* **10**, 261 (2008).
49. Amari, S.-i. & Cichocki, A. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences* **58**, 183–195 (2010).
50. Kawamoto, T. <https://github.com/tatsuro-kawamoto/graphBIX> (Date of access: 13/09/2016) (2016).
51. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87 (2012).
52. Mézard, M. & Montanari, A. *Information, Physics, and Computation* (Oxford University Press, 2009).
53. Opper, M. & Winther, O. Mean field approach to bayes learning in feed-forward neural networks. *Phys. Rev. Lett.* **76**, 1964–1967 (1996).

Acknowledgements

This work was supported by JSPS KAKENHI No. 26011023 (T.K.) and No. 25120013 (Y.K.).

Author Contributions

T.K. and Y.K. conceived the research and designed the analysis. T.K. wrote the program code, conducted analysis, and wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017