A Bayesian framework for community detection in networks

Jake M. Hofman,[1], Chris H. Wiggins[2,3]
[1]Department of Physics, Columbia University, New York, NY
[2]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY
[3]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY

Large-scale networks describing complex interactions between a multitude of objects have found application in a wide array of fields, from biology to social science to information technology [10, 3, 2]. In these applications one often wishes to *model* networks, suppressing the complexity of the full description while retaining relevant information about the structure of the interactions [12]. One such network model is a modular description in which nodes are grouped into communities, with a large number of interactions within communities and fewer interactions between communities. Many existing studies focus on defining measures of modularity and performing community detection [8, 6, 4, 7, 9, 1] but few infer model parameters, community assignments, and the number of communities within one framework. We present a Bayesian approach to network modularity which performs the above tasks and, in doing so, generalizes many previous studies on the subject.

Specfically, we consider a generative model for modular networks in which each node is randomly assigned to a module and edges between nodes are drawn from two independent Erdős-Reýni distributions. Given the data (an adjacency matrix for a particular network), we evaluate the Bayesian evidence $p(\mathcal{D}|K) = \int d\Theta \; p(\mathcal{D}|\Theta, K)p(\Theta|K)$ where $\mathcal{D}$ are the data, Θ are the parameters and latent module assignments, and K is the number of modules, by integrating over all parameter and latent variable settings. In doing so, we both infer the posterior distribution $p(\Theta|\mathcal{D}, K)$ over parameters and module assignments and use the evidence to perform complexity control, dermining the optimal number of modules.

To calculate the exact evidence, we integrate over parameters analytically and exploit implicit symmetries to compute the remaining sum over module assignments orders of magnitude faster than complete enumeration, which requires K^N integrand evaluations for N nodes and K modules. While this method can, in principle, be used to evaluate the evidence for networks of arbitrary size, runtimes scale too quickly with N to be practically applicable for large-scale networks.

To accomodate large-scale networks for which exact calculation of the evidence is computationally intractable, we use a variational Bayes [5] approach that results in an iterative algorithm which produces approximations to the posterior $p(\Theta|\mathcal{D}, K)$ and the evidence $p(\mathcal{D}|K)$. This approximation can be analytically shown to be bounded by the true evidence and the iterative algorithm converges extremely quickly. See Fig. 1 for an illustration of the iterative algorithm and experimental validation of the technique on a small network for which the exact evidence can be calculated and compared to the variational approximation. We validate the method with synthetic data and apply it to various real-world social networks. Application of this technique to one such network, Zachary's karate network [11], is shown in Fig. 1.

In explicitly considering modular network models in a generative framework we have exploited Bayesian techniques to infer posterior distributions over model parameters and module assignments from the data, while simultaneously performing complexity control to automatically determine the number of modules a given network permits. For small networks, we have shown a means for calculating the exact evidence and posterior, while for large networks we used variational techniques to arrive at suitable approximations for the quantities of interest. The developed techniques are principled, interpretable, computationally efficient, and lend themselves to future generalizations (including model selection between non-nested network models).

References

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels, 2007.

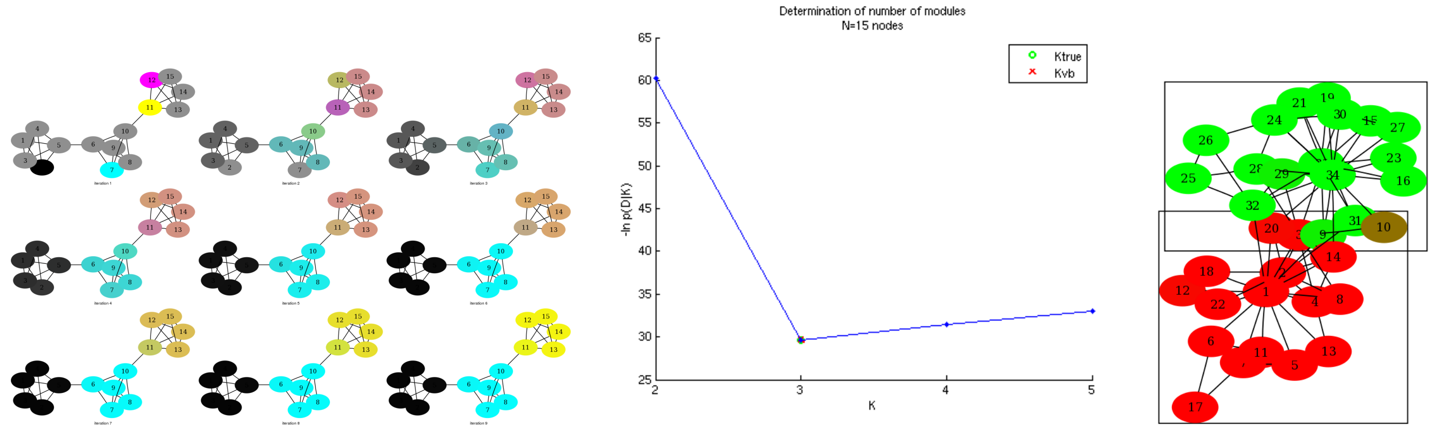

Figure 1: Left: Illustration of iterative scheme for variational inference. CMYK represent probability of being in modules 1 through 4, respectively, with probability 1; intermediate shades represent intermediate probabilities, e.g. gray represents nodes that have probability 1/K of being in each module. The algorithm is initialized with 4 possible modules, with a randomly chosen node placed in each module. Note that upon convergence only three modules have non-zero probability of occupancy. Middle: the (negative log) evidence as a function of number of modules for the network shown in the left panel. Note that the approximate evidence recovers the true number of modules. Right: The results on the karate network which recovers Zachary's empirically observed network split.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, Jan 2002.

[3] S. Brohée and J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488, Jan 2006.

[4] M. B. Hastings. Community detection as an inference problem. *arXiv*, 74(3):035102–+, 2006.

[5] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[6] M. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.

[7] M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *Proc Natl Acad Sci USA*, 104(23):9564–9, Jun 2007.

[8] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *arXiv*, 74(1):016110, 2006.

[9] M. Rosvall and C. T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proc Natl Acad Sci USA*, 104(18):7327–31, May 2007.

[10] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA*, 100(21):12123–8, Oct 2003.

[11] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[12] E. Ziv, M. Middendorf, and C. H. Wiggins. Information-theoretic approach to network modularity. *Phys Rev E*, 71(4 Pt 2):046117, Apr 2005.