Inferring the structure and scale of modular networks

Jake M. Hofman JMH2045@COLUMBIA.EDU
Department of Physics, Columbia University, New York, NY

Chris H. Wiggins CHRIS.WIGGINS@COLUMBIA.EDU
Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY

Keywords: Probabilistic models, variational Bayesian inference, networks, modularity, community detection

Abstract

We present an efficient, principled, and interpretable technique for inferring module assignments and for identifying the optimal number of modules in a given network, based on variational Bayesian inference for stochastic block models. We show how our method extends previous work and addresses the "resolution limit problem". We apply the technique to synthetic and real networks.

Large-scale networks describing complex interactions among a multitude of objects have found application in a wide array of fields, from biology to social science to information technology (Watts & Strogatz, 1998; Albert & Barabási, 2002). 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 (Ziv et al., 2005). One such network model groups nodes into modules, or "communities," with different densities of intra- and inter- connectivity for nodes in the same or different modules. We present here a computationally efficient Bayesian framework for inferring the number of modules, model parameters, and module assignments for such a model.

The problem of finding modules in networks (or "community detection") has recently received much attention in both the physics and machine learning literature. Most approaches in the physics literature (Newman & Girvan, 2004; Reichardt & Bornholdt, 2006; Hastings, 2006) rely on optimizing an energy-based cost function with fixed parameters over possible assignments of nodes into modules. The particular cost functions vary, but most compare a given node partitioning to an implicit null model, the two most popular being the configuration model and a limited version of the stochastic block model (SBM) (Holland & Leinhardt, 1976). It was recently shown that the "ad-hoc" choice of fixed parameters for these cost functions gives rise to the "resolution limit problem" (Fortunato & Barthélemy, 2007; Kumpula et al., 2007), wherein the parameter choice sets a lower limit on the size of detected communities as a function of the size of the network. We suggest a solution to this problem that relies on inferring the model parameters as opposed to asserting them *a priori*.

In tandem with this work, there has been progress in machine learning approaches for relational data, under both maximum likelihood (Newman & Leicht, 2007; Hugo Zanghi & Miele, 2007) and maximum evidence (Nowicki & Snijders, 2001; Kemp et al., 2004; Airoldi et al., 2007; Xu et al., 2007; Sinkkonen et al., 2007) frameworks for SBMs and related (but generally more complicated) models. While maximum likelihood methods infer model parameters, they provide only point estimates and as such are prone to overfitting; complexity control – determining the number of modules – must be handled separately, by, e.g., the Bayesian Information Criterion, an uncontrolled approximation which is only asymptotically appropriate. Maximum evidence techniques, however, infer distributions over model parameters and, as such, automatically penalize overly-complex models. Many of the models studied under a maximum evidence framework are quite complex and, as such, approximate inference is performed with sampling techniques that are computationally costly. We study a simple but effective SBM which generalizes (Hastings, 2006) and is a special case of the models used in (Nowicki & Snijders, 2001; Kemp et al., 2004; Airoldi et al., 2007). We use a variational approach for approximate maximum

Preliminary work. Under review by the International Workshop on Mining and Learning with Graphs (MLG). Do not distribute.

evidence inference that results in a computationally efficient and interpretable algorithm for module discovery.

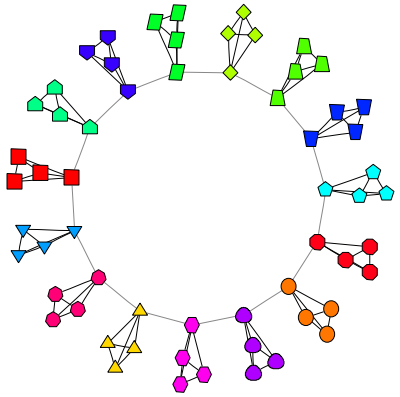

Figure 1. Results for the resolution limit test suggested in (Fortunato & Barthélemy, 2007; Kumpula et al., 2007). Our method correctly infers all 15 modules (indicated by the shape and color of nodes), whereas NG modularity optimization (Newman & Girvan, 2004) incorrectly groups pairs of neighboring cliques together.

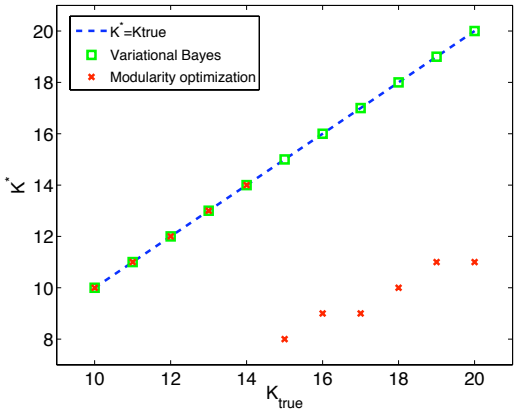

Figure 2. Results of the resolution limit test, implemented for a range of true number of modules, K_{true}, the number of 4-node cliques in the ring-like graph. Note that our method, represented by green squares, correctly infers the number of modules K^* over the entire range of K_{true}, while NG modularity initially finds the correct number of modules but undergoes a sharp transition at $K_{true} = 15$ (corresponding to Fig. 1), after which neighboring cliques are grouped together.

We specify an N-node network by its adjacency matrix $\mathbf{A}$, where $A_{ij} = 1$ if there is an edge between nodes i and j and $A_{ij} = 0$ otherwise, and define $z_i \in \{1, \ldots, K\}$ to be the unobserved module membership of the i^{th} node. We denote the set of latent variables as $Z = \{z_i\}_{i=1}^N$. We use a constrained SBM which consists of a multinomial distribution over module assignments with weights $\pi_\mu \equiv p(z_i = \mu|\vec{\pi})$ and Bernoulli distributions over edges contained within and between modules with weights $\theta_+ \equiv p(A_{ij} = 1|z_i = z_j, \vec{\theta})$ and $\theta_- \equiv p(A_{ij} = 1|z_i \neq z_j, \vec{\theta})$, respectively. In short, to generate a random undirected graph under this model we roll a K-sided die (biased by $\vec{\pi}$) N times to determine module assignments for each of the N nodes; we then flip one of two biased coins (for either intra- or inter- module connection, biased by θ_+ or θ_-, respectively) for each of the $N(N-1)/2$ pairs of nodes to determine if the pair is connected. We take beta and Dirichlet distributions as conjugate priors over $\vec{\theta}$ and $\vec{\pi}$, respectively.

Given the data (an adjacency matrix for a particular network), we evaluate the Bayesian evidence by integrating over all parameter and latent variable settings:

$$p(\mathbf{A}|K) = \sum_Z \int d\Theta \; p(\mathbf{A}, Z|\Theta, K) p(\Theta|K) \qquad (1)$$

While this method can, in principle, be used to evaluate the evidence for networks of arbitrary size, runtimes scale too quickly with N be practically applicable for real-world networks. To accommodate large-scale networks for which exact calculation of the evidence is computationally intractable, we use a variational Bayes approach (Jordan et al., 1999) that arises from the following identity:

$$\ln p(\mathbf{A}) = \left\langle \ln \frac{p(\mathbf{A}, Z, \Theta)}{q(\Theta, Z)} \right\rangle_q + D_{KL}\left(q(\Theta, Z)||p(\Theta, Z|A)\right), \quad (2)$$

where $q(\Theta, Z)$ is an arbitrary distribution and D_{KL} is the Kullback-Leibler divergence (conditional dependence on K has been suppressed for brevity). One replaces the calculation of the log-evidence by that of an approximate free energy (the expected value to the right of the equal sign), which approaches the log-evidence as $q(\Theta, Z)$ approaches the true (and unknown) posterior. To make optimization of the free energy tractable, we take a mean-field approach, assuming the form $q(\Theta, Z) = q_\Theta(\Theta) \prod_{i=1}^N q_i(z_i)$. Optimizing the free energy as a functional of q_Θ and q_i results in an iterative coordinate ascent algorithm that produces approximations to the posterior $p(\Theta, Z|\mathbf{A}, K)$ and the evidence $p(\mathbf{A}|K)$. The steps in the resulting algorithm involve sparse matrix multiplication of N-by-N and N-by-K matrices and the evaluation of digamma functions, allowing for very fast implementation – e.g. runtime for one run of variational Bayes on a synthetic network of $N = 10^5$ nodes and $K = 4$ modules is ~ 40 seconds in MATLAB on a 2GHz laptop; we note that successful inference may require several

such runs to find a global (and not local) optimum of the free energy. We add that variational inference is typically more computationally efficient than sampling approaches, often without a sizeable difference in performance, as shown in the related work by (Xu et al., 2007).

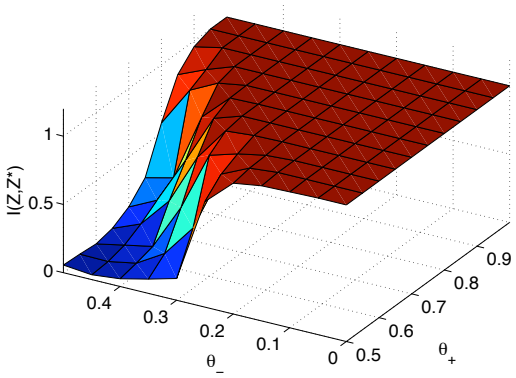

Figure 3. Test of robustness on synthetic data. For each value of θ_+ and θ_-, 5 random networks of $N = 128$ nodes and $K = 4$ modules are sampled from from the generative model, $p(\mathbf{A}, Z|\Theta, K)$, where $\pi_\mu = \frac{1}{4}$ for all samples. The algorithm is run on each network and the average mutual information between the latent module assignments, Z, and inferred module assignments, Z^*, is shown. The mutual information is computed as in (Danon et al., 2005).

We validate the method with synthetic data and apply it to an example real-world network. Fig. 1 and Fig. 2 show that our method overcomes the resolution limit (Fortunato & Barthélemy, 2007; Kumpula et al., 2007) faced by other methods (Newman & Girvan, 2004; Reichardt & Bornholdt, 2006; Hastings, 2006). Our method infers distributions over model parameters and correctly identifies the 4-node cliques as modules, whereas other methods require one to assert fixed parameter values and, as a result, incorrectly group neighboring cliques together.

Fig. 3 shows the performance of our method on synthetic networks sampled from the generative model, $p(\mathbf{A}, Z|\Theta, K)$, for a range of θ_+ and θ_- values. The algorithm successfully identifies the modular structure of the synthetic networks when such structure exists: for assortative modules in which θ_+ is sufficiently larger that θ_-, the latent and inferred module assignments are essentially identical; for $\theta_+ \approx \theta_- \approx 0.5$ the networks are approximately Erdős-Reýni random graphs, for which no modular structure is present.

Fig. 4 shows the giant component (∼7000 authors) of the American Physical Society March meeting 2008 co-authorship network. The i^{th} author is represented by the i^{th} row and i^{th} column in this matrix, and the i^{th} and j^{th} authors are connected if they co-authored a conference paper together. The rows and columns of the adjacency matrix are sorted by the results of our algorithm, revealing strong community structure which corresponds to well-defined sub-disciplines.

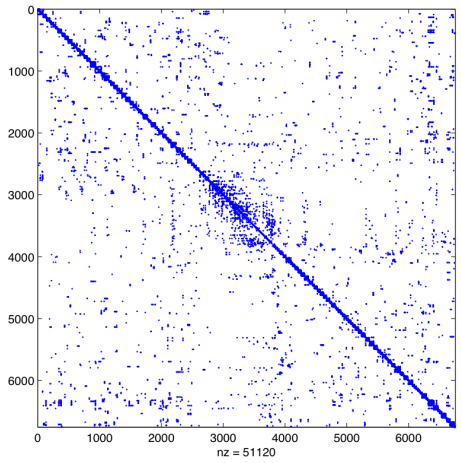

Figure 4. An analysis of the giant component of the co-authorship network compiled from the APS March meeting 2008, revealing strong community structure which corresponds to well-defined sub-disciplines, e.g. the two largest communities correspond to superconductor theorists and experimentalists.

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. We used a variational approach 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 competing network models).

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2007). Mixed membership stochastic blockmodels. *arXiv:0705.4485*.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, *74*, 47–97.

Danon, L., Díaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identifica-

tion. *Journal of Statistical Mechanics: Theory and Experiment*, P09008.

Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. *PNAS*, *104*, 36–41.

Hastings, M. B. (2006). Community detection as an inference problem. *Phys. Rev. E*, *74*, 035102(R).

Holland, P., & Leinhardt, S. (1976). Local Structure in Social Networks. *Sociological Methodology*, *7*, 1–45.

Hugo Zanghi, C. A., & Miele, V. (2007). Fast online graph clustering via Erdös-Réyni mixture. http://genome.jouy.inra.fr/ssb/preprint/SSB-RR-8.fast-onlineERMG.pdf.

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

Kemp, C., Griffiths, T., & Tenenbaum, J. (2004). *Discovering latent classes in relational data* (Technical Report). Technical Report AI Memo 2004-019, MIT.

Kumpula, J., Saramäki, J., Kaski, K., & Kertész, J. (2007). Limited resolution in complex network community detection with potts model approach. *Eur. Phys. J. B*, *56*, 41–45.

Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, *69*, 026113.

Newman, M. E. J., & Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *PNAS*, *104*, 9564–9569.

Nowicki, K., & Snijders, T. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*, 1077–1087.

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *Phys. Rev. E*, *74*, 016110.

Sinkkonen, J., Aukia, J., & Kaski, S. (2007). Inferring vertex properties from topology in large networks. In *MLG 2007*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–2.

Xu, Z., Tresp, V., Yu, S., Yu, K., & Kriegel, H. (2007). Fast inference in infinite hidden relational models. In *MLG 2007*.

Ziv, E., Middendorf, M., & Wiggins, C. H. (2005). Information-theoretic approach to network modularity. *Phys. Rev. E*, *71*, 046117.