Literature Review for SBMs
Bayesian SBMs and SBMs for Data Integration
Carter Allen

Stochastic block models (SBMs) have been fit in a Bayesian framework by a variety of authors. SBMs are the primary alternative to modularity-based community detection algorithms, which perform well in many settings but do not enjoy the inferential characteristics of SBMs — a generative model for network data. Generally speaking, inference on SBM models is done in either a Bayesian framework using (1) MCMC and a hierarchy of model priors, or (2) as a direct optimization problem utilizing what is known as the minimum description length (MDL) criterion. Peixoto claims and shows that inference from Bayesian SBMs with non-informative priors and MDL are equivalent. Recently, (Peixoto, 2018a) presented a nonparametric Bayesian formulation of the SBM that allows for multiple weighted edges between nodes, where edge weights can be either continuous or discrete. The concept of weighted stochastic block models was introduced by (Aicher, 2013), however their approach is slightly more restrictive in that it requires edge weights to follow an exponential family distribution.

In (Peixoto, 2018a, 2018b), "nonparametric" refers to the ability of the model to infer the number of blocks B, and thus the overall dimension of the model. In this approach, a prior is placed on the community labeling vector $\mathbf{b}$. A simple weakly informative prior for $\mathbf{b}$ is proposed by (Peixoto, 2018b) as

$$P(\mathbf{b} \mid \mathbf{n}) = \frac{\prod_r n_r!}{N!},$$

where $\mathbf{n} = (n_1, \ldots, n_B)'$ is the vector containing the number of nodes belonging to each community, and N is the total number of nodes in the network. By conditioning on the community sizes, the problem of models tending towards communities of equal sizes is avoided, however it requires an additional hyperprior on $\mathbf{n}$. The formulation in (Peixoto, 2018b) continues similarly, with priors being chosen to balance the desire to be non-informative with precaution against over or under-fitting.

A recent work (van Der Pas and van Der Vaart, 2018) builds off of (Nowicki and Snijers, 2001), which is one of the foundational papers in the SBM literature. An important distinction from Peixoto's approach is that this approach is parametric in the sense that the model dimension must be specified *a priori*. In this setting, a Dirichlet prior is placed on $\boldsymbol{\pi} = \mathbf{n}/N$, and independent and identically distributed $Beta(\beta_1, \beta_2)$ priors are placed on each P_{ab}, which is the probability of a node in community a sharing and edge with a node in community b. Finally, the authors assume the generative model for the adjacency matrix $\mathbf{A}$ to be $A_{ij} \mid \boldsymbol{\pi}, \mathbf{P} \sim Bernoulli(P_{b_i, b_j})$. For implementation, van Der Pas and van Der Vaart refer to (McDaid, 2013), which proposes an efficient algorithm for estimating the parameters in models of the type proposed by (Nowicki and Snijers, 2001). However, (Côme and Latouche, 2014) claimed this estimation routine has poor mixing properties. Also, the approach taken by (McDaid, 2013) is not easily extendable to nested block structures or multigraphs.

Summary on Bayesian SBMs: Peixoto's non-parametric approach seems to be less flexible in the choice of prior parameters, but it is quite flexible in its applicability due to its ability to (1) estimate the model dimension, (2) accommodate arbitrarily weighted edges, and (3) accommodate multiple edges between any two given nodes. Weighted multigraphs could be a promising method for implementing data integration. Peixoto has also extended this approach to nested SBMs, in which the community structure of lower level graphs are recursively modeled with higher level SBMs. Additionally, Peixoto's methods are implemented in graph-tool, which is by far the most expansive, robust, and well-documented package for fitting SBMs.

Alternatively, the parametric Bayesian approaches in the style of Nowicki and Snijers have a more traditional model hierarchy, though are restricted to simple graphs with binary edge weights. Through correspondence with McDaid, I learned that he is no longer maintaining his software, though it is available on his website as a C++ package that requires compiling. The documentation on his software is very sparse.

Summary on Data Integrated SBMs: I have not been able to find previous works describing approaches for data integration approaches using SBMs. There is literature on integrating multiple data sources to perform community detection using modularity-based algorithms, I have not found any papers that do this with the SBM. Additionally, I asked Dr. Eric Kolaczyk about this at the ICSA 2019 conference and he was not aware of any papers that conducted data integration with the SBM.

References

Peixoto, Tiago P. "Nonparametric weighted stochastic block models." *Physical Review E* 97.1 (2018): 012306.

Peixoto, Tiago P. "Nonparametric Bayesian inference of the microcanonical stochastic block model." *Physical Review E* 95.1 (2017): 012317.

Aicher, Christopher, Abigail Z. Jacobs, and Aaron Clauset. "Adapting the stochastic block model to edge-weighted networks." *arXiv preprint arXiv:1305.5782* (2013).

van der Pas, S. L., and A. W. van der Vaart. "Bayesian community detection." *Bayesian Analysis* 13.3 (2018): 767-796.

Nowicki, Krzysztof, and Tom A. B. Snijders. "Estimation and prediction for stochastic blockstructures." *Journal of the American statistical association* 96.455 (2001): 1077-1087.

McDaid, Aaron F., et al. "Improved Bayesian inference for the stochastic block model with application to large networks." *Computational Statistics & Data Analysis* 60 (2013): 12-31.

Côme, E. and Latouche, P. (2014). "Model Selection and Clustering in Stochastic Block Models with the Exact Integrated Complete Data Likelihood." ArXiv:1303.2962. MR3441229. doi: https://doi.org/10.1177/1471082X15577017.