Bayesian Community Detection

S.L. van der Pas and A.W. van der Vaart

Leiden University

JSM 2015

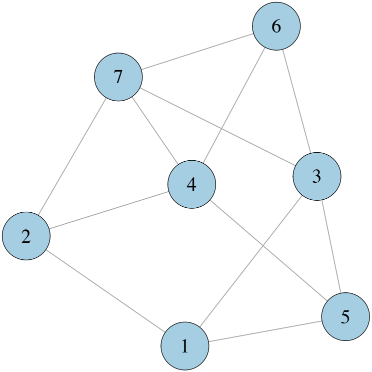
Outline

The Stochastic Block Model

Undirected graph without self-loops, of n nodes.

Observe adjacency matrix A:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$



The Stochastic Block Model

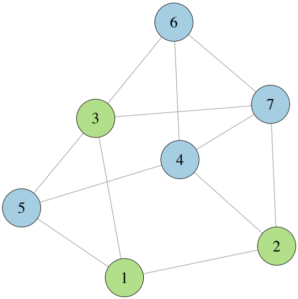
K classes.

Latent class labels: $Z = (Z_1, \ldots, Z_n)$, $Z_i \in \{1, \ldots, K\}$. For $i < j$:

$$\boxed{\mathbb{P}(A_{ij} = 1 \mid Z_i = a, Z_j = b) = P_{ab}}$$

where P is a symmetric $K \times K$-matrix of probabilities.

$$A = \left(\begin{array}{ccc|cccc} 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \end{array}\right)$$



The Stochastic Block Model

K classes.

Latent class labels: $Z = (Z_1, \ldots, Z_n)$, $Z_i \in \{1, \ldots, K\}$. For $i < j$:

$$\boxed{\mathbb{P}(A_{ij} = 1 \mid Z_i = a, Z_j = b) = P_{ab}}$$

where P is a symmetric $K \times K$-matrix of probabilities.

The Z_i are generated according to

$$\mathbb{P}(Z_i = a) = \pi_a,$$

for some $\pi \in \mathbb{R}^K$ such that $\sum_{a=1}^K \pi_a = 1$.

Goal: recovery of labels.

Other approaches

- Spectral clustering (e.g. Rohe, Chatterjee and Yu (2011), Jin (2015), Sarkar and Bickel (2015), Lei and Rinaldo (2015))
- Largest Gaps algorithm (Channarond, Daudin and Robin (2012))
- Newman-Girvan modularity (e.g. Newman and Girvan (2004))
- Likelihood modularity (Bickel and Chen (2009), Zhao, Levina and Zhu (2012))

Bayesian approach (e.g. Nowicki and Snijders (2001), McDaid et al. (2013)): theoretical results lacking.

The Bayesian modularity

The prior on the vector of class labels z:

$$\pi \sim \mathrm{Dir}\left(\frac{K+3}{2}, \dots, \frac{K+3}{2}\right)$$
$$n_1, \dots, n_K \mid \pi \sim \mathrm{Multinomial}(n, \pi)$$

Given $n_1, \dots, n_K$, the labelling z is then drawn as a random ordering of the following sequence:

$$\underbrace{1, \dots, 1}_{n_1}, \underbrace{2, \dots, 2}_{n_2}, \dots, \underbrace{K, \dots, K}_{n_K}.$$

Independently:

$$P_{ab} \sim \mathrm{Beta}\left(\tfrac{1}{2}, \tfrac{1}{2}\right), \quad 1 \leq a \leq b \leq K.$$

The Bayesian modularity

Use the posterior mode as estimator of Z.

Let

$$n_a(z) = \text{number of nodes in class } a;$$
$$n_{ab}(z) = \text{maximum possible number of edges between classes } a \text{ and } b;$$
$$O_{ab}(z) = \text{observed number of edges between classes } a \text{ and } b.$$

The Bayesian modularity

The Bayesian modularity is given by

$$Q_B(z) = \frac{1}{n^2} \sum_{a \leq b} \log B(O_{ab}(z) + \tfrac{1}{2}, n_{ab}(z) - O_{ab}(z) + \tfrac{1}{2})$$
$$+ \frac{1}{n^2} \sum_{a=1}^{K} \log \frac{\Gamma\left(n_a(z) + \frac{K+3}{2}\right)}{\Gamma(n_a(z) + 1)}.$$

The Bayesian MAP-estimator is:

$$\widehat{z} = \arg\max_z Q_B(z).$$

Weak and strong consistency

An estimator is weakly consistent if the fraction of misclassified nodes goes to zero in probability.

An estimator is strongly consistent if the number of misclassified nodes goes to zero in probability.

Strong consistency of the Bayesian modularity

Theorem [strong consistency]
If P is symmetric, every pair of rows of P is different, $0 < P < 1$, and $\pi > 0$, then the MAP classifier $\widehat{z} = \arg\max_z Q_B(z)$ is strongly consistent if the expected degree is of larger order than $(\log n)^2$.

Strong consistency of the Bayesian modularity

Theorem [strong consistency]
If P is symmetric, every pair of rows of P is different, $0 < P < 1$, and $\pi > 0$, then the MAP classifier $\widehat{z} = \arg\max_z Q_B(z)$ is strongly consistent if the expected degree is of larger order than $(\log n)^2$.

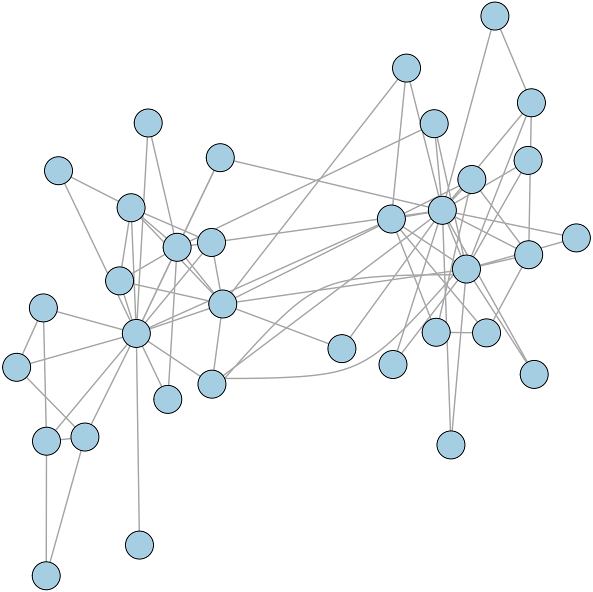
Full posterior useful for

- uncertainty quantification?
- estimating K?

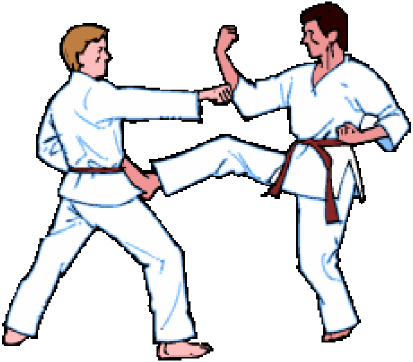
Implementation

- McDaid et al. (2013): allocation sampler, $\sim$ 10.000 nodes
- Côme and Latouche (2014): greedy inference algorithm
- tabu search (Glover, 1989)

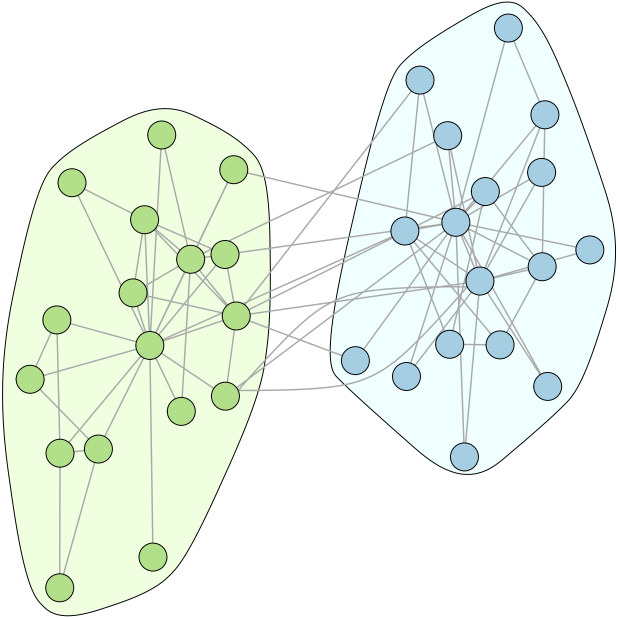
Example: Zachary's karate club



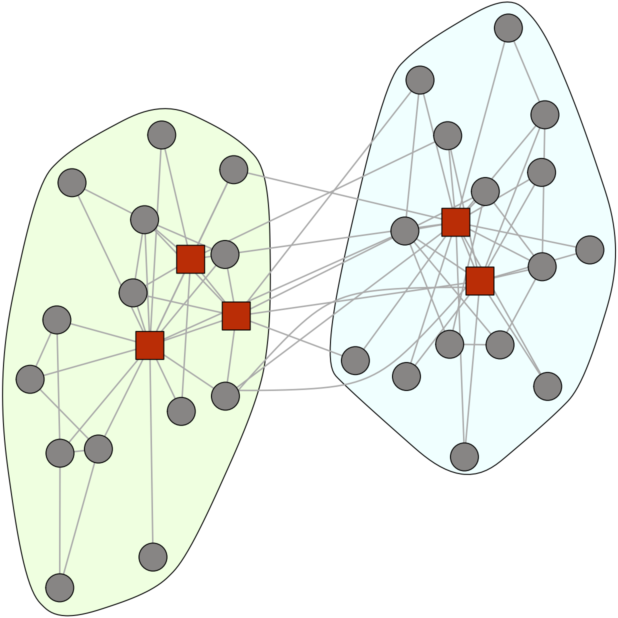
Example: Zachary's karate club



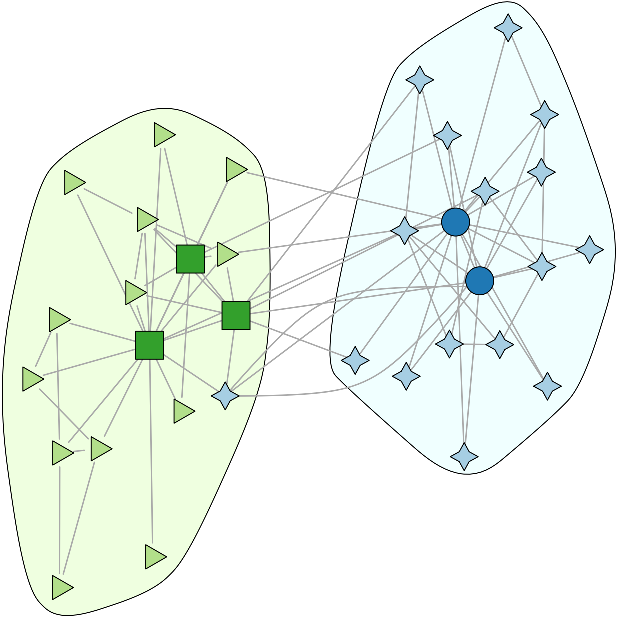
Example: Zachary's karate club



$K = 2$



$K = 4$



Conclusions

- Strongly consistent community detection with a Bayesian approach is possible, provided the expected degree is of larger order than $(\log n)^2$.
- Encourages further investigation into the full potential of the full posterior.

References

P.J. Bickel, A. Chen (2009), A Nonparametric View of Network Models and Newman-Girvan and Other Modularities, *Proceedings of the National Academy of Sciences of the United States of America* **106**, 21068-21073.

A. Channarond, J.-J. Daudin, S. Robin (2012), Classification and Estimation in the Stochastic Blockmodel Based on the Empirical Degrees, *Electronic Journal of Statistics* **6**, 2574-2601.

E. Côme, P. Latouche (2014), Model Selection and Clustering in Stochastic Block Models with the Exact Integrated Complete Data Likelihood, arXiv:1303.2962.

F. Glover (1989), Tabu Search - Part I, *ORSA Journal on Computing* **1**, 190-206.

J. Jin (2015), Fast Community Detection by SCORE, *The Annals of Statistics* **43**, 57-89.

J. Lei, A. Rinaldo (2015), Consistency of Spectral Clustering in Stochastic Block Models, *The Annals of Statistics* **43**, 215-237.

A.F. McDaid, T. Brendan Murphy, N. Friel, N.J. Hurley (2013), Improved Bayesian Inference for the Stochastic Block Model with Application to Large Networks, *Computational Statistics and Data Analysis* **60**, 12-31.

M.E.J. Newman, M. Girvan (2004), Finding and Evaluating Community Structure in Networks, *Physical Review E* **69**, 026113.

K. Nowicki, T.A.B. Snijders (2001), Estimation and Prediction for Stochastic Blockstructures, *Journal of the American Statistical association* **96**, 1077-1087.

S.L. van der Pas, A.W. van der Vaart (201?), Bayesian Community Detection. *In progress*.

K. Rohe, S. Chatterjee, B. Yu (2011), Spectral Clustering and the High-Dimensional Stochastic Blockmodel, *The Annals of Statistics* **39**, 1878-1915.

P. Sarkar and P.J. Bickel (2015), Role of Normalization in Spectral Clustering for Stochastic Blockmodels, *The Annals of Statistics* **43**, 962-990.

W. W. Zachary (1977), An Information Flow Model for Conflict and Fission in Small Groups, *Journal of Anthropological Research* **33**, 452-473.

Y. Zhao, E. Levina, J. Zhu (2012), Consistency of Community Detection in Networks under Degree-Corrected Stochastic Block Models, *The Annals of Statistics* **40**, 2266-2292.