Comment on "Model-based clustering for social networks" by Handcock et al

David R. Hunter, Penn State University

June 2, 2013

The article by Handcock et al provides an interesting and important extension of the latent space model of Hoff et al (2002). A different extension of this work — which may also be applied to the current article — allows for more explicit modeling of local network features, such as transitivity, using an exponential random graph model (ERGM).

If the matrix $\mathbf{y}$ denotes the entire network (i.e., the collection of all $y_{i,j}$), then equation (2) of Handcock et al implies that

$$P(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\beta_0^T \sum_{i,j} x_{i,j} y_{i,j}\} \exp\{\beta_1 \sum_{i,j} |z_i - z_j| y_{i,j}\}}{\kappa(\beta_0, \beta_1)}, \tag{1}$$

where $\kappa(\beta_0, \beta_1)$ is a normalizing constant.

To simplify notation in (1), let

$$g(\mathbf{y}, X) = \sum_{i,j} x_{i,j} y_{i,j} \qquad \text{and} \qquad h(\mathbf{y}, Z) = \sum_{i,j} |z_i - z_j| y_{i,j}. \tag{2}$$

Conditional on the latent positions Z, the resulting model,

$$P(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\beta_0^T g(\mathbf{y}, X) + \beta_1 h(\mathbf{y}, Z)\}}{\kappa(\beta_0, \beta_1)}, \tag{3}$$

is evidently a canonical exponential family (see, e.g., Lehmann, 1983) of distributions parameterised by (β_0, β_1) with statistics $g(\mathbf{y}, X)$ and $h(\mathbf{y}, Z)$. Therefore, conditional on Z, model (3) is an *exponential random graph model* ("graph" here is a synonym for "network"). Snijders (2002) and Robins et al (2006a) give literature reviews of these models, which are also called p-star models in the literature.

Importantly, model (3) is still an ERGM, conditional on Z, if the vector $g(\mathbf{y}, X)$ of network statistics of interest is not of the form (2) that allows the likelihood function to factor nicely as in (1). The simplest such "non-factoring" models were considered by Frank and Strauss (1986), in which $g(\mathbf{y}, X)$ contained terms such as the number of triangles in $\mathbf{y}$, $\sum_{i<j<k} y_{i,j}y_{j,k}y_{k,i}$. Much recent work in the social networks literature has focused on development of useful statistics $g(\mathbf{y}, X)$ for modeling real network data (Snijders et al, 2006; Robbins et al, 2006b), as well as explaining why some statistics, such as the number of triangles, lead to ERGMs that fail miserably at modeling these data (Handcock, 2002; Handcock, 2003).

Model (3) would give the modeler a powerful tool for exploring network structure: For instance, if the latent positions and cluster assignments of the nodes change dramatically upon the introduction of a particular network statistic into the ERGM, this suggests that the statistic captures an important aspect of network structure. Yet estimating parameters in a model such as (3) is quite difficult when $g(\mathbf{y}, X)$ is not of the form (2). In principle, the two-stage maximum likelihood estimation of Handcock et al should work, though the second stage would rely on an MCMC-based stochastic algorithm such as those described by Hunter and Handcock (2006) or Snijders (2002). The Bayesian scheme implemented here is promising, but establishing a reasonable prior for the ERGM parameter β_0 is difficult. Despite the remaining challenges, the article of Handcock et al is a real step forwards.

References

Frank, O. and D. Strauss (1986), Markov graphs, *Journal of the American Statistical Association*, **81**: 832–842.

Handcock, M. S. (2002) Statistical Models for Social Networks: Inference and Degeneracy, pp. 229 – 240 in *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, edited by Ronald Breiger, Kathleen Carley, and Philippa E. Pattison. Washington, DC: National Academy Press.

Handcock, M. S. (2003), Assessing degeneracy in statistical models of social networks, Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. Available from `http://www.csss.washington.edu/Papers/`

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), Latent space approaches to social network analysis, *Journal of the American Statistical Association*, **97**: 1090–1098.

Hunter, D. R. and M. S. Handcock (2006), Inference in curved exponential family models for networks, *Journal of Computational and Graphical Statistics*, **15**: 565–583.

Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.

Robins, G. L., P. E. Pattison, Y. Kalish, and D. Lusher (2006a), An introduction to exponential random graph (p*) models for social networks, *Social Networks*, in press.

Robins, G. L., T. A. B. Snijders, P. Weng, M. S. Handcock, and P. E. Pattison (2006b), Recent developments in exponential random graph (p*) models for social networks, *Social Networks*, in press.

Snijders, T. A. B. (2002), Markov Chain Monte Carlo estimation of exponential random graph models, *Journal of Social Structure*, **3**. Available at `www.cmu.edu/joss/content/articles/volume3/Snijders.pdf`

Snijders, T. A. B., P. E. Pattison, G. L. Robins, and M. S. Handcock (2006), New specifications for exponential random graph models, *Sociological Methodology*: **36**, in press.