Comment on "Model-based clustering for social networks" by Handcock, Raftery and Tantrum

David Krackhardt
Professor of Organizations
The Heinz School of Public Policy and Management
Carnegie Mellon University

First, I want to emphasize the importance of the problem addressed by this paper. Cleanly identifying clusters of actors in a social system based on their social ties is an age old pursuit of generations of scholars, from sociologists and psychologists to mathematicians (e.g., Luce and Perry, 1949; Harary and Cartwright, 1956). UCINET (Borgatti, Everett and Freeman, 2002), the most commonly used package for analysis of network data, has 20 distinct methods for finding clusters or groups, each with a plethora of suboptions and parameter choices, which, depending on the data, may yield wildly differing results. This dizzying array of "solutions" begs the central question: Given the observed data, what is the right number of clusters and what is their composition? Using the Bayes factors approach to statistically answer this critical question is a major step forward out of this intellectual morass.

The paper quickly leads me to ask a couple of extending questions. First, how sensitive is this procedure to violations of the assumption on independence of dyadic observations? We know that even moderate amounts of "network autocorrelation" in the data can dramatically affect estimates of standard errors and concomitant inference tests in traditional analytic procedures (Krackhardt, 1988).

Second, should we rely on empirical demonstrations of the model to provide us with evidence that the procedure is uncovering the true, underlying group structure? The fact that the procedure recovers the same structure in the Sampson data as other prior analyses could be because the networks are so clearly clustered that it does not matter what hammer you use to pound the data, they will always reveal the same story. In the case of ties among adolescents, the fact that their method cleanly shows discrimination among grades is interesting, but does that mean it was more accurate? Suppose the result had not fallen along grade lines. Would that mean that the method was not accurately assessing real underling clusters? Or, would it mean instead that networks were clustering on some other criteria?

Both of these questions could be addressed with appropriate Monte Carlo simulations. The advantage of such simulations is that you have control over "truth", and by adding precise, known, and yet complex structures of noise, one can directly assess how well the proposed Bayesian method recaptures this underlying truth. Such simulations would help us delineate the boundary conditions within which their method is truly powerful.

Cited References

Borgatti, S. P., M. G. Everett, and L. C. Freeman (2002). UCINET for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Cartwright, D., and F. Harary (1956). Structural balance: A generalization of Heider's theory. Psychological Review, 63: 277-292.

Krackhardt, D. (1988). Predicting with networks: A multiple regression approach to analyzing dyadic data. Social Networks, 10: 359-381.

Luce, R. D. and A. Perry (1949). A method of matrix analysis of group structure. Psychometrika, 14: 95-116.