

Change in Connectivity in a Social Network over Time: A Bayesian Perspective

Susan Adams

Nathan Carter

Charles Hadlock

Dominique Haughton - Bentley University, Waltham, Massachusetts

George Sirbu - Thomson Medstat, Ann Arbor, Michigan

Abstract

In this paper, we propose a Bayesian methodology for examining differences between statistics of a social network at two distinct points in time. The problem has been of interest for some time in the social networks community because it is quite difficult to test whether differences over time in statistics such as overall network connectivities are significant. Several issues make this problem challenging: links in a social network tend to be dependent, and the networks at the two different points in time are likely to be dependent as well. This implies, for example, that bootstrapping a social network to address this problem may be impractical. This paper expands on a previously published Bayesian version of the p_1 model for social networks with random effects, which allows for dependence between the edges of the networks. We use the software Winbugs to obtain posterior distributions for the difference in connectivity over time and for the correlation between each actor's connectivities in the network at both points in time. We assume that this correlation is the same for all actors. We illustrate our methods with the case of a social network of collaborations (joint publications) between departments of a business university where interdisciplinary work was actively promoted. Our methods allow us to compare the tendency to make collaborative links across departments before and after the administrative initiatives.

Keywords: *Social Networks Over Time, Connectivity, p_1 models, Bayesian analysis*

Acknowledgments: We would like to express our sincere gratitude to the authors Gill and Swarz (2004) for kindly making the Winbugs code they wrote for their models available to us. It was very helpful in building the Winbugs code needed in this paper. The Winbugs code we used to simulate the posterior distributions is available on request from the authors.

Susan M. Adams (PhD Georgia Institute of Technology) is Associate Professor of Management at Bentley University where she teaches leadership and management consulting courses. Her current research focuses on how organizations affect the careers of executives and professionals, particularly in changing environments. She is a former chair of both the Careers and Management Consulting Divisions of the Academy of Management. **Nathan Carter** (PhD Indiana University) teaches mathematics at Bentley University and publishes in diverse areas, including mathematical logic, group theory visualization, and social network analysis. His book Visual Group Theory will be published by the Mathematical Association of America in 2009.

Charles Hadlock (PhD University of Illinois) is an applied mathematician with a combination of industrial and academic experience. He is currently Professor of Mathematical Sciences and Trustee Professor of Technology, Policy, and Decision Making at Bentley University. His current research focuses on complexity theory and the collapse of complex organizations and societies.

Dominique Haughton (PhD M.I.T.) is Professor of Mathematical Sciences at Bentley University. Her research interests include statistics and marketing, the analysis of living standards data, international statistics and data mining. She is editor-in-chief of the journal Case Studies in Business, Industry and Government Statistics (CSBIGS).

George Sirbu (PhD Michigan State University) is a statistician at Thomson Medstat in Ann Arbor, Michigan, where he specializes in the analysis of health care data.

For correspondence, please contact Dominique Haughton at dhaughton@bentley.edu

Introduction

The comparison of two networks at two (or more) different points in time has been approached in social network literature in the following ways. Wasserman and Iacobucci (1988) attempted to test the equality of model parameters across two or more time periods by where extensions of the p_1 model are used, but where dyads at a given point in time are still assumed to be independent. To try to overcome the difficulty of both relaxing the assumption of independent dyads (links between pairs of factors) and the observations of the same relation at two or more time points (which of course are not likely to be independent), Snijders (1996) proposed to model the longitudinal network as a simulated Markov chain, with parameters estimated via the method of moments. However, this method does not readily lend itself to testing the significance of a difference in parameter values across time periods, and is quite difficult to implement.

The paper by Faust and Skvoretz (2002) goes in another direction by introducing methods which rely on the p^* model proposed by Wasserman and Pattison (1996) and attempts to compare networks with different sizes, across different time periods and even different entities. The paper also includes a useful review of the literature to date on issues of comparing different relations on the same actors, or the same relation at several points in time, or across different groups, inclusive of the papers mentioned immediately above. However, the emphasis in the Faust and Skvoretz article is not on testing whether a particular aspect of the network has changed significantly over time.

It is this last issue – investigating shifts in a particular network parameter across two points in time – which is the focus of this

article. The problem is a challenging one, because, for instance, attempts to use procedures such the bootstrap to obtain estimates of standard errors hit against the problem that dyads are not independent, so that it is impractical to bootstrap a socio-matrix of ones and zeros which are not independent, even when only one network is under consideration (at one fixed point in time). Bootstrapping independent and identically-distributed quantities such as regression residuals is possible (and is indeed common), but, unfortunately, to bootstrap a socio-matrix, one would need to respect the dependency structure of the ones and zeros while sampling, which is very complicated. Even more intractable is bootstrapping two dependent socio-matrices made up of dependent ones and zeros.

We, therefore, propose to follow a Bayesian approach introduced in the context of social networks by Wong (1987) and extended by Gill and Swartz (2004). As outlined in the latter article, the p_1 model, with fixed effects proposed by Holland and Leinhardt (1981), was improved by Wong (1987) by incorporating random effects. The main difference is that the original p_1 model assumes independence between the links, whereas with random effects, some dependence structures can be incorporated.

The article is organized as follows: section 2 describes the data and the context that will be used to illustrate our approach; section 3 presents the Bayesian model we propose for our networks; and section 4 gives results and conclusions.

Data

In the Fall of 2001, our institution, a large, independent business school in Massachusetts, began a program specifically intended to encourage interdisciplinary research collaborations among the faculty.

The institution has been accredited by the Association to Advance Collegiate Schools of Business International (AACSB) since 1991 and has made major investments to establish a strong teaching and research capability, especially emphasizing the intersection of business and information technology.

In the period leading up to and immediately following the initial AACSB accreditation, the institution was entering a transition phase from a predominantly undergraduate teaching institution to a more comprehensive university. Buoyed by a measurable growth in reputation, rankings, and student quality, a development objective evolved into transforming the institution into a business university. It became clear that faculty could reach a new level of reputation and contribution only by focusing on major current issues facing the business world, rather than spending much of their time in isolated academic niches that might have much less impact on the practice of business. Such major current issues are almost invariably interdisciplinary in nature.

The institution decided to work proactively to facilitate the formation of faculty research teams focused on major issues that suited their interests and backgrounds, and the administration first issued a request for proposals (RFP) to the faculty near the end of 2001. Funds could be requested for a variety of purposes, such as course reductions, summer stipends, and various expense items¹; even to the point where an intensive faculty research seminar might substitute for part of the teaching load of major participants.

Considering the faculty as a social network, we let each faculty department be a node (vertex) in the network, and connections among departments be determined by faculty in a department doing

¹ Expense items include travel, data acquisition, student assistants, ongoing research seminars, etc.

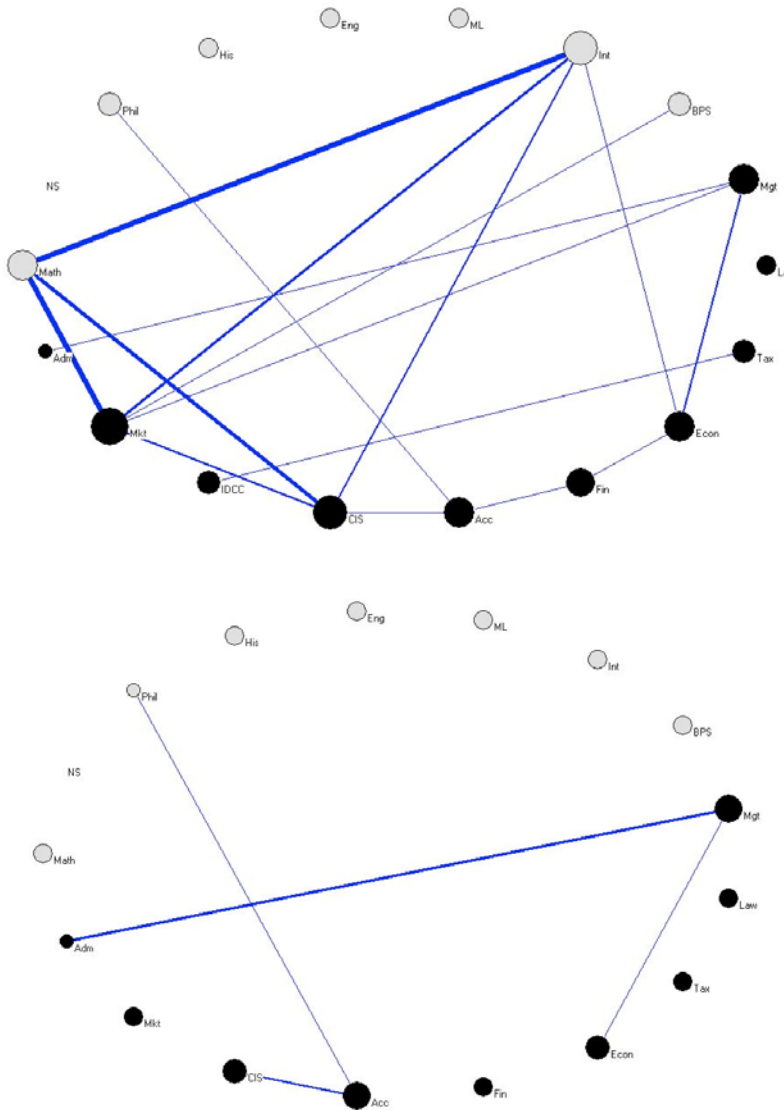
scholarly work with faculty in other departments. Therefore we consider the network of co-publication between departments (1 if at least one article was jointly authored by at least one member of each department, 0 if not) at two distinct periods of time, in '00-'01 and in '03-'04. The reason for this particular split is because of the time when some programs were launched (e.g. RFP program) and what authors believe to be a turnaround point in the university development. Note that the network is undirected.

We obtain our data from our faculty research database, which had recently been updated in connection with our re-accreditation from the AACSB. The visualization of the two networks at the two points in time is given in Figure 1. We only took into consideration journal articles, since we felt that journal articles were the best category of scholarly work to account for the quality and quantity of the faculty scholarship activity.²

Visual evidence from Figure 1 would seem to indicate a clear increase in the connectivity of the network. We will now examine this issue more formally and describe our Bayesian model in the next section.

² Since the RFP initiative had the underlying intent to encourage a better interaction among internal faculty resources, our data set does not record the departments for authors external to the institution.

Figure 1: Collaboration Network in 2000-2001 (top) and 2003-2004 (bottom)



Gray vertices are Arts and Sciences departments, black vertices are Business departments or Administration. The thickness of the lines is proportional to the number of joint publications between the two nodes it joins; the size of a vertex is proportional to the node's degree, including itself twice. Abbreviations for the departments are as follows: NS for Natural Sciences, ML for Modern Languages, Int for International Studies, BPS for Behavioral and Political Science, CIS for Computer Information Systems, IDCC for Information Design and Corporate Communication, Adm for Administration.

Model

Since in our case the ties are non-directional, the p_1 model can be written in the following way:¹

$Y_{ij11}^k = 1$ if i and j are connected, 0 otherwise

$Y_{ij00}^k = 1$ if i and j are not connected, 0 otherwise

$$\ln P(Y_{ij00}^k = 1) = \lambda_{ij}^k$$

$$\ln P(Y_{ij11}^k = 1) = \lambda_{ij}^k + \theta^k + \alpha_i^k + \alpha_j^k .$$

The index k denotes the time period and indices i and j refer to departments. Because each pair of departments (i, j) can be either linked or un-linked, the matrix Y_{ij00}^k is a simple opposite of the matrix Y_{ij11}^k in the sense that Y_{ij00}^k can be obtained from Y_{ij11}^k by replacing zeros with ones and ones with zeros. The matrix Y_{ij11}^k is often referred to as the socio-matrix, with its ones indicating where a link occurs. The probability $P(Y_{ij11}^k = 1)$ represents the probability of a link occurring between departments i and j , at time k , and $P(Y_{ij00}^k = 1)$ represents the probability that no such link exists.

We will make the convention that $k = A$ for the '00-'01 social network and $k = B$ for

¹ We adopt the Wasserman and Faust (1994) formulation.

the '03-'04 social network. The parameter θ^k is called the choice parameter and is a measure of the overall connectivity of the network, and α_i^k is the attractiveness parameter correspondent to node i for period k . We observe also that λ_{ij}^k is not a parameter, but rather a fixed constant subject to the constraint, $P(Y_{ij00}^k = 1) + P(Y_{ij11}^k = 1) = 1$.

In a Bayesian analysis, unknown parameters are considered random variables with a distribution referred to as the prior distribution, which reflects knowledge we might have about the parameters even before any data are collected. The analysis produces a posterior conditional distribution of the parameters given the data, using a MCMC (Monte Carlo Markov Chain) simulation procedure, the details of which are well documented (Congdon 2001 or Winbugs 2006). The posterior distribution is proportional to the product of the likelihood function and the prior density, and as such takes into account both the prior distribution and the data (i.e. the observed networks). Posterior densities typically *cannot be calculated in closed form*, which makes simulating from them difficult; progress in MCMC methods has made it possible to simulate from such posterior densities without fully computing their densities, and has helped give rise to the fast development of Bayesian applications. We consider the following prior distributions for the parameters of interest:

Given Σ ,

$$\begin{pmatrix} \alpha_i^A \\ \alpha_i^B \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right)$$

$$\begin{pmatrix} \theta^A \\ \theta^B \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right), \quad \text{independently of}$$

$$\begin{pmatrix} \alpha_i^A \\ \alpha_i^B \end{pmatrix},$$

with Σ given by

$$\Sigma = \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix}$$

distributed as

$$\Sigma^{-1} \sim \text{Wishart}\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, 2\right),$$

where \sim denotes “is distributed as”. The choice of our prior distributions is standard for this particular situation and implies no particular prior knowledge about where parameter values might be concentrated. For instance, the bivariate normal distribution we have chosen for the vector of attractiveness values of a node (for both periods), with a precision matrix² distributed according to a Wishart distribution as above is standard (Congdon, 2001). Note that we have assumed that the vectors with components α_i^A and α_i^B are independent a-priori of the vector with components θ^A and θ^B given Σ , because there is no particular reason for our prior belief about the overall connectivity of the network to be related to our prior belief about the attractiveness of each department. On the other hand, it seems sensible to assume that the precision

² Precision matrix is the inverse of the covariance matrix.

of our prior knowledge (represented by Σ^{-1}) is the same for the vectors with components α_i^A and α_i^B and the vector with components θ^A and θ^B , and that the a-priori correlations between α_i^A and α_i^B are the same for all i , and equal to the correlation between θ^A and θ^B .

We are interested in the difference $\Delta = \theta^B - \theta^A$ in the choice parameter θ from the former social network '00-'01 to the latter one '03-'04. We expect to find that the posterior distribution of Δ is concentrated for most of its range in the set of positive numbers. We are not arguing that this change can be entirely attributed to the programs mentioned in the introduction but merely observe that the difference Δ is a-posteriori likely to be positive. However, it is our belief that the success of the program was part of the positive change that can be seen from the analysis.

In the next section, we report the results of using an MCMC procedure, such as implemented in the software package Winbugs to generate random draws from the posterior distribution of parameters of interest. Note that Winbugs does not require that the user provide expressions for auxiliary distributions used in the procedure, only that the model which generates the data and the prior distribution be specified.

Results and Conclusions

In the table and figures below we present the results of the MCMC analysis from the model outlined in the previous section with graphs representing kernel densities for the posterior distributions of the two main parameters of interest. The statistics presented in Table 1 and the data which were used to create kernel densities for the posterior distributions of parameters of interest arose from a Winbugs analysis where we generated 200,000 iterates of the MCMC procedure. The first 4,000 iterations were used as a “burn-in”, so the summary statistics in Table 1 are in fact based on the 196,000 remaining iterates. This is necessary because the MCMC chain becomes stationary typically only after a certain number of iterations, so it is safe to compute posterior moments (means and standard deviations) from iterations arising once the chain has become stationary (further discussion of the convergence of the process is given below).

The parameters we focused on are the difference $\Delta = \theta^B - \theta^A$, the choice parameters θ^A and θ^B , their standard deviations σ_A and σ_B , and the correlation ρ between θ^A and θ^B . We also included summary posterior moments for the attractiveness of a particular department at both points in time.

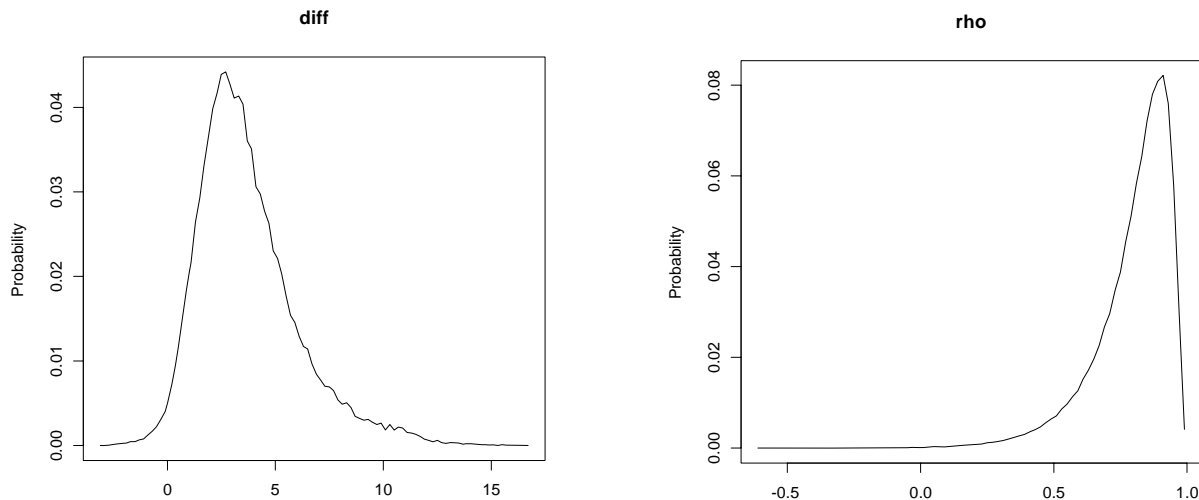
Table 1: Summary Statistics of the Posterior Distribution of Parameters of Interest, for 196,000 Iterates

Parameter	Mean	SD	MC error	2.50%	Median	97.50%
Δ	3.69	2.266	0.0824	0.282	3.313	9.346
θ^A	-7.15	2.705	0.1082	-13.64	-6.722	-2.962
θ^B	-3.46	1.286	0.0489	-6.231	-3.389	-1.096
ρ	0.798	0.1425	0.0016	0.425	0.835	0.964
σ_A	4.39	1.758	0.0471	2.1	4.022	8.838
σ_B	2.645	0.7466	0.0119	1.568	2.518	4.454
a[1,1]	-0.07125	0.7972	0.02401	-1.643	-0.07444	1.518
a[1,2]	2.257	1.486	0.05393	-0.2261	2.092	5.683

As we can see in Table 1, the 2.5% percentile (0.282) from the posterior distribution of the difference Δ is above zero so we can conclude that it is a-posteriori very likely that the choice

parameter for interdisciplinary research in the university has increased from '00-'01 to '03-'04. This is also clear from Figure 2, left panel, where the posterior density of Δ is displayed.

Figure 2. Posterior Distributions for Δ and ρ



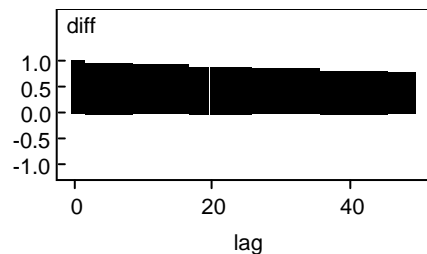
It is also interesting to notice that the correlation between choice/attractiveness parameters in '00-'01 and '03-'04 is quite

high with its posterior mean estimated at 0.798 and its left-skewed posterior distribution (Figure 2, right panel). This

makes sense given the nature of the data, where relations among the actors in the same network are likely to be preserved over time. Nevertheless due to the increased activity in the network in the latter period, the observed increase in the estimates, from 2.645 to 4.39, for the posterior mean standard deviation of the expansiveness/attractiveness parameter is quite natural; in the latter period, departments became more diverse in their propensity to engage in joint research with other departments.

When using MCMC techniques to sample from posterior distributions, an issue arises about the stationarity of the sequence of draws, which one would wish to occur after a number of “burn-in” draws. A visual examination of the history of the draws will usually suffice to conclude to stationarity; however there is always the risk that that MCMC sampler might get stuck in a sub-region of the parameter space instead of exploring the sample space, particularly if successive draws are strongly auto-correlated. For that reason, it is desirable that the auto-correlation between successive draws should decrease rapidly with the lag for each parameter of interest. It is not uncommon, and happens in the case for example of our parameter Δ on the difference between the choice parameters (referred to as *diff* on Figure 3), that this auto-correlation in fact dies down slowly. This can be seen clearly on Figure 3.

Figure 3. Auto-correlation Between Draws of the Posterior Distribution of Δ , with All 196,000 Draws Included



In that case, one can sample one out of, for example, 100 draws of the posterior, which, typically, will remove the auto-correlation problem, and then compare the posterior summaries of the smaller sample with the full sample. We are grateful to an anonymous referee for suggesting this very useful idea. In Figure 4, we can see that the auto-correlation problem has now disappeared.

Figure 4. Auto-correlation Between Draws of the Posterior Distribution of Δ , with One Out of 100 Draws (sample with 1960 draws)

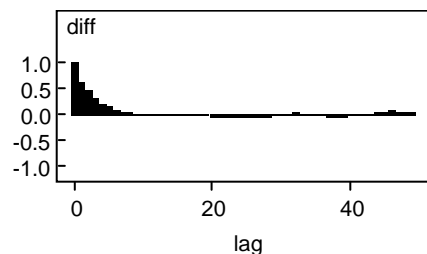


Figure 5 indicates that stationarity is not in question, particularly since the draws on Figure 5 hover about a very similar level to the level featured on a history graph for the full sample.¹ This is further supported by the large number of iterations we used. We conclude from examining this set of graphs

¹ A history graph of the full sample is not given here because it is very similar to Figure 5.

that it is quite unlikely that the MCMC chain was trapped in a sub-region of the sample space.

We have presented auto-correlation graphs for Δ and history graphs for Δ and

ρ , but graphs for the remaining parameters of interest show that stationarity holds for them as well.

Figure 5: History of draws from the posterior of Δ and ρ , on the basis of one out of 100 draws (sample with 1960 draws)

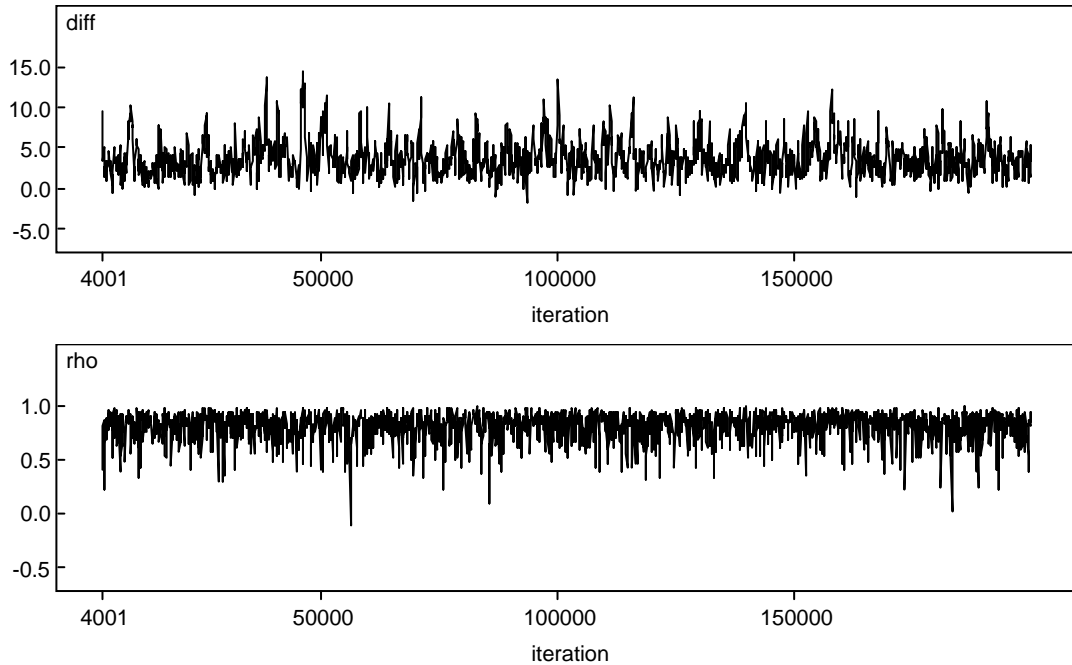


Table 2 reveals that summary statistics of the posterior distribution of the parameters of interest evaluated on the sub-sample of one out of 100 draws are close to those in Table 1 evaluated for the whole sample.

Table 2. Summary Statistics of the Posterior Distribution of Parameters of Interest, on the Basis of the One Out of 100 Sample of 1960 Draws

Parameter	Mean	SD	MC error	2.50%	Median	97.50%
Δ	3.671	2.263	0.1144	0.2885	3.329	9.289
θ^A	-7.134	2.7	0.171	-13.52	-6.714	-2.912
θ^B	-3.463	1.285	0.0804	-6.248	-3.385	-1.161
ρ	0.8005	0.1391	0.003079	0.4407	0.8372	0.9668
σ_A	4.385	1.707	0.06254	2.1	4.017	8.774
σ_B	2.649	0.7586	0.02124	1.566	2.534	4.441

To sum up, we have found that the Bayesian methodology provides a convenient and effective way of deciding whether a parameter of interest in a social network has changed over time. Of course, because of its flexibility, the methodology lends itself to other situations which might prove intractable otherwise, and a whole variety of social network models can be formulated in the Bayesian framework. For instance, Tallberg (2003) proposes a Bayesian approach to uncovering blocks within networks of actors which are similar in the sense that their probabilities of forming links with other actors are the same.

It is therefore quite likely that Bayesian methods will find further applications to problems of interest to social network researchers where no other method is readily available. However, a caveat is in order when using Bayesian methods (or even when using more classical likelihood methods). Adding too many parameters carries with it the risk of over-parameterization of a model. Over-parameterization occurs when several values of the parameters give rise to the same value of the likelihood function, leading to a situation where some parameters may not be identifiable. It is desirable to avoid over-parameterization because it can lead to some convergence problems in the MCMC procedure, even if the problems can to some extent be overcome by the choice of suitable priors. However, it is not always easy to know if a model is over-parameterized; one may be alerted to it only by unusual behavior in the MCMC iterations.¹

We also note that when attempting to compare connectivities of two networks over time, it is advisable to make sure that the networks have about the same size, as is

the case here (the number of actors differs by only one between the two time periods), since several network parameters are known to depend rather critically on network size (Anderson, Butts and Carley 1999).

¹ We refer the reader to papers by O'Neill (2005) and Rannala (2002) where the issue is discussed.

References

- Anderson, B.S., Butts, C., and Carley, K. 1999. The interaction of size and density with graph-level indices. *Social Networks*, 21: 239-267.
- Congdon, P. 2001. Bayesian Statistical Modelling. New York: Wiley.
- Faust, K. and Skvoretz, J. 2002. Comparing networks across space and time, size and species. *Sociological Methodology*, 32: 267-299.
- Gill, P. S. and Swarz, T. B. 2004. Bayesian analysis of directed graphs with applications to social networks. *Applied Statistics*, 53(2): 49-260.
- Holland, P. W. and Leinhardt, S. 1981. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76: 33-65.
- O'Neill, B. 2005. Consistency and identifiability in Bayesian analysis. Preprint School of Finance and Applied Statistics, Australian National University, available at <http://ecocomm.anu.edu.au/research/papers/pdf/05-09.pdf>.
- Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of Phylogeny. *Systematic Biology*, 51(5): 754-760.
- Snijders, T. 1996. Stochastic-oriented models for network change. *Journal of Mathematical Sociology*, 21(1-2): 149-172.
- Tallberg, C. 2003. A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1): 1 – 23.
- Wasserman, S. and Faust, K. 1994. Social network analysis: Methods and applications. Cambridge, England: Cambridge University Press.
- Wasserman, S. and Pattison, P. 1996. Logit models and logistic regressions for social networks: An introduction to Markov graphs and p^* . *Psychometrika*, 61(3): 401-425.
- Wasserman, S. and Iacobucci, D. 1988. Sequential social network data. *Psychometrika*, 53(2): 261-282.
- Winbugs. 2006. The Bugs project, Winbugs, <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>
- Wong, G. Y. 1987. Bayesian models for directed graphs. *Journal of the American Statistical Association*, 82: 140-148.