Statistical modelling of the group structure of social networks

Murray Aitkin[a], Duy Vu[a] and Brian Francis[b]

[a] Department of Mathematics and Statistics,
University of Melbourne, Victoria Australia.

[b] Department of Mathematics and Statistics,
Lancaster University, Lancaster LA1 4YF United Kingdom

February 2, 2014

Abstract

This research evaluates the identification of group structure in social networks through the latent class model and a new Bayesian model comparison method for the number of latent classes. The approach is applied to a well-known network of women in Natchez Mississippi. The latent class analysis reproduces the group structure of the women identified by the original sociologists.

Keywords: social groups, latent classes, Bayesian model comparison, Natchez women.

http://dx.doi.org/10.1016/j.socnet.2014.03.002

1 Introduction

This paper investigates a statistical model for groups in a social network, which until recently has not been a major focus of analysis in the field. A recent major review of statistical modelling work in the field has been published by Goldenberg, Zheng, Fienberg and Airoldi (2009). In their summary Chapter 6 they conclude:

> Despite the many advances in network modeling over the last decade, there remains a host of unresolved issues. ... We feel that, from a statistics or machine learning perspective, the biggest breakthroughs are to be made in the areas of inference and dynamic modeling. Creating a model or perhaps fixing an existing one in such a way that provides realistic generative and inference mechanisms which can identifiably infer parameters of a large real world network would make a great contribution to the statistical network modeling community.

This paper addresses the identification of *actor* groups within the framework of a two-mode or bipartite network of *actors attending events*, through a statistical model in which the groups of actors are represented by *latent classes*, which are not directly observable, but which can be probabilistically reconstructed from the event attendance patterns of the actors.

In this process several questions are of critical importance:

- how many groups can be identified;
- the nature of the membership of the groups, for example whether individuals belong to one or to many groups;
- the nature of the event attendance patterns in the groups;
- whether other non-latent class models might give a better representation of the data.

We address these questions in a Bayesian framework, and use recent developments in Bayesian model comparisons to illuminate the choice among possible models. To show the application of the approach we use a famous data set from Davis, Gardner and Gardner (1941) analysed many times, as reported in Freeman (2003).

2 The Natchez women network

We give a detailed discussion of a simple social network which has attracted a remarkable amount of interest, and a wide variety of approaches (21 different analyses are reported and compared in Freeman 2003).

It comes from a sociological study of social interactions among women in Natchez, Mississippi in the 1930s, reported in Davis, Gardner and Gardner

(1941, hereafter DGG). The book reported a comparative study of social class in black and white society. One aspect of the study was to assess the formation or existence of "cliques", defined by the joint participation of groups of women in attending common events. The network table (Figure 1) which has caused so much interest to later analysts is reproduced from Davis et al; it gives the presence (x) or absence (...) of 18 women at 14 events. The women are named and numbered by DGG, and the events are dated from newspaper reports at the time.

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	X	X	X	X	X	X	....	X	X	....	....	....	....	....
2. Miss Laura Mandeville	X	X	X	....	X	X	X	X	....	....	....	....	....	....
3. Miss Theresa Anderson	....	X	X	X	X	X	X	X	X	....	....	....	....	....
4. Miss Brenda Rogers	X	....	X	X	X	X	X	X	....	....	....	....	....	....
5. Miss Charlotte McDowd	....	....	X	X	X	....	X	....	....	....	....	....	....	....
6. Miss Frances Anderson	....	....	X	....	X	X	....	X	....	....	....	....	....	....
7. Miss Eleanor Nye	....	....	....	....	X	X	X	X	....	....	....	....	....	....
8. Miss Pearl Oglethorpe	....	....	....	....	....	X	....	X	X	....	....	....	....	....
9. Miss Ruth DeSand	....	....	....	....	X	....	X	X	X	....	....	....	....	....
10. Miss Verne Sanderson	....	....	....	....	....	....	X	X	X	....	....	X	....	....
11. Miss Myra Liddell	....	....	....	....	....	....	....	X	X	X	....	X	....	....
12. Miss Katherine Rogers	....	....	....	....	....	....	....	X	X	X	....	X	X	X
13. Mrs. Sylvia Avondale	....	....	....	....	....	....	X	X	X	X	....	X	X	X
14. Mrs. Nora Fayette	....	....	....	....	....	X	X	....	X	X	X	X	X	X
15. Mrs. Helen Lloyd	....	....	....	....	....	....	X	X	....	X	X	X	....	....
16. Mrs. Dorothy Murchison	....	....	....	....	....	....	....	X	X	....	....	....	....	....
17. Mrs. Olivia Carleton	....	....	....	....	....	....	....	....	X	....	X	....	....	....
18. Mrs. Flora Price	....	....	....	....	....	....	....	....	X	....	X	....	....	....

Figure 1: DGG Table 1

The question of interest to analysts is how to describe the nature of the association among the women, and in particular to identify, as far as possible, subsets of the women which form coherent groups or cliques, using only their attendance at the events as data. It would help this identification if we knew more about the social events than just their dates, but no further information about them is given. We do know that these events were not the unique social events of the reported days, as DGG report that other women also attended events on the same days, but not the events in this table.

2.1 The adjacency matrix

To perform any analysis we express the table elements mathematically through the link or tie variable Y_{ij}, with the presence of woman i at event j defining $Y_{ij} = 1$, and her absence from the event defining $Y_{ij} = 0$. We use n to denote the number of rows – women – and r to denote the number of columns – events. The resulting table, shown in Table 2 is the adjacency matrix, denoted by $\mathbf{Y}$.

Marginal totals (T) have been added to the table, giving the total number of events attended by each woman, and the total number of women attending each event.

$W \backslash E$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	T
1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	8
2	1	1	1	0	1	1	1	1	0	0	0	0	0	0	7
3	0	1	1	1	1	1	1	1	1	0	0	0	0	0	8
4	1	0	1	1	1	1	1	1	0	0	0	0	0	0	7
5	0	0	1	1	1	0	1	0	0	0	0	0	0	0	4
6	0	0	1	0	1	1	0	1	0	0	0	0	0	0	4
7	0	0	0	0	1	1	1	1	0	0	0	0	0	0	4
8	0	0	0	0	0	1	0	1	1	0	0	0	0	0	3
9	0	0	0	0	1	0	1	1	1	0	0	0	0	0	4
10	0	0	0	0	0	0	1	1	1	0	0	1	0	0	4
11	0	0	0	0	0	0	0	1	1	1	0	1	0	0	4
12	0	0	0	0	0	0	0	1	1	1	0	1	1	1	6
13	0	0	0	0	0	0	1	1	1	1	0	1	1	1	7
14	0	0	0	0	0	1	1	0	1	1	1	1	1	1	8
15	0	0	0	0	0	0	1	1	0	1	1	1	0	0	5
16	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2
17	0	0	0	0	0	0	0	0	1	0	1	0	0	0	2
18	0	0	0	0	0	0	0	0	1	0	1	0	0	0	2
T	3	3	6	4	8	8	10	14	12	5	4	6	3	3	89

Table 2: Event attendance

From the adjacency matrix $\mathbf{Y}$ we can construct two other *symmetric* matrices of interest: the $n \times n$ "woman-by-woman" matrix $\mathbf{W} = \mathbf{YY}'$, and the $r \times r$ "event-by-event" matrix $\mathbf{E} = \mathbf{Y}'\mathbf{Y}$.

W	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	8	6	7	6	3	4	3	3	3	2	2	2	2	2	1	2	1	1
2		7	6	6	3	4	4	2	3	2	2	1	2	2	2	1	0	0
3			8	6	4	4	4	3	4	3	2	2	3	3	2	2	1	1
4				7	4	4	4	2	3	2	1	1	2	2	2	1	0	0
5					4	2	2	0	2	1	0	0	1	1	1	0	0	0
6						4	3	2	2	1	1	1	1	1	1	1	0	0
7							4	2	3	2	1	1	1	2	2	1	0	0
8								3	2	2	2	2	2	2	1	2	1	1
9									4	3	2	2	3	2	2	2	1	1
10										4	3	3	4	3	3	2	1	1
11											4	4	4	3	3	2	1	1
12												6	6	5	3	2	1	1
13													7	6	4	2	1	1
14														8	4	2	2	2
15															5	2	1	1
16																2	1	1
17																	2	2
18																		2

Table 3: Women connections

E	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	3	2	3	2	3	3	2	3	1	0	0	0	0	0
2		3	2	3	3	3	2	3	2	0	0	0	0	0
3			6	4	6	5	4	5	2	0	0	0	0	0
4				4	4	2	3	3	2	0	0	0	0	0
5					8	6	6	7	3	1	0	0	0	0
6						8	5	7	4	1	1	1	1	1
7							10	8	9	3	2	4	2	2
8								14	9	4	1	5	2	2
9									2	4	3	5	3	3
10										5	2	5	3	3
11											4	2	1	1
12												3	3	3
13													3	3
14														3

Table 4: Event connections

These "one-mode" networks show the connections between pairs of women through the numbers of events they have jointly attended, and the connections between pairs of events through the numbers of women jointly attending them.

The woman-by-woman matrix **W** is given in Table 3; its diagonal elements are the numbers of events attended by each woman.

The event-by-event matrix **E** is given in Table 4. The diagonal elements of **E** are the numbers of women attending each event. The upper-right off-diagonal block of zeros shows that *no* woman attended an event from the set 1-4 *and* an event from the set 10-14: these sets of events attracted different groups of women. The zeros in the **W** matrix are more scattered, but women 17 and 18 did not attend *any* event with women 2 and 4-7, and woman 5 did not attend any event with women 8, 11, 12 and 16-18.

While the one-mode "networks" of event-by-event and woman-by-woman matrices are useful, we do not use them in the modelling approach, as they both lose information from the original adjacency matrix, which we model directly. Visualisations of the two-mode and the one-mode women networks illustrate this. Figure 2, from `http://jung.sourceforge.net/applet/southern.html` gives two visualisations of the Natchez women and event connections. In the left two-mode graph, the placing of events attended (blue squares), separated by the groups of women attending them (red circles), clarifies the connections to some extent: the events in the centre are those attended by women from both groups.

In the right graph, the one-mode woman-by-woman network is shown, with the thickness of connecting lines corresponding to the number of events jointly attended. There is no clear grouping of women visible.

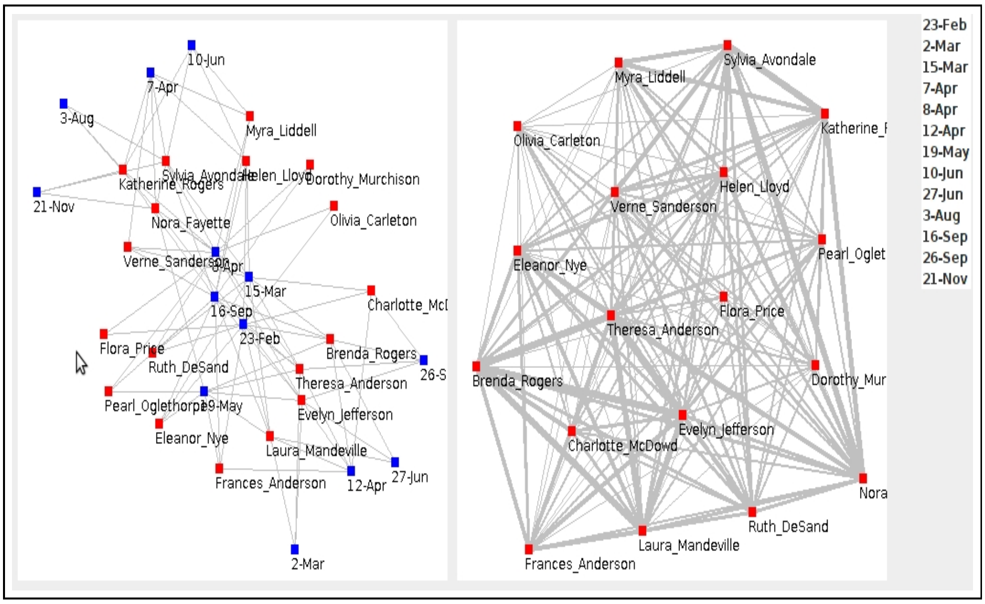

Figure 2: Southern women and events

2.2 The meaning of a group

A fundamental question which has to be addressed first is what we mean by a *group*, in this social context. We should first note than even the *word* for this subset of actors is not consistent across research fields: it is also called *community*, *clique* and *class*.

As Doreian, Batagelj and Ferligoj (2005) note, there has been an extraordinary number of definitions of the group concept. Their formulation (op. cit. p.4) is

> ... we consider a group to be *two or more individuals who are interdependent through sustained interaction.* (authors' emphasis)

Since *interdependent* is not defined, and *sustained* implies a continuing structure of the group, and since our analyses are generally of only one observation of the network, we give a weaker definition. We adopt, as a working definition of a group in the Natchez women context, *an identifiable subset of women who tend to attend the same events.* We do not yet define this *tendency*, or how it is to be used to establish the group structure. This will be considered after we examine the 21 existing analyses of this table.

2.3 The 21 analyses

Before we develop statistical models for Table 2, we should note that many analyses of the table do *not* use a statistical model, but use other combinatorial or algebraic methods for its analysis. The approaches used in these analyses are given without explanation or exposition in Table 5; full details of the approaches can be found in Freeman (2003).

An additional complication is that DGG gave *two* tables for the Natchez women event attendance, which are inconsistent. Freeman (2003) identified this inconsistency and checked the 21 analyses for their consistency with DGG's Table 1, re-doing the analyses, or asking the original authors to redo their analyses, for those which had used DGG's Table 2. We summarise his analyses in Table 5. They are ordered by time; where the same authors appear with (1), (2) or (3) these refer to different methods given in the same paper.

2.4 The gold standard

We first describe the analysis by DGG, since they carried out the research; as Freeman (2003) quoted from Davis and Warner (1939):

> they drew on "records of overt behavior and verbalizations, which cover more than five thousand pages, statistical data on both rural and urban societies, as well as newspaper records of social gatherings ..."

DGG both assigned women to groups and determined their *positions* within the groups, which they described as *core, primary* and *secondary*, though these terms were not defined. They divided the women into two overlapping groups: 1-9 in Group 1 and 9-18 in Group 2; woman 9 was assigned to both groups. Women 1-4 and 13-15 were core members, 5-7 and 11-12 were primary, and 8-10 and 16-18 were secondary.

Authors	Method
Davis, Garner and Garner (1941)	Interviews
Phillips, Conviser (1972)	Minimised entropy over groups
Breiger (1974)	Eliminated connected events from the event x event matrix
Breiger, Boorman, Arabie (1975)	Alternately clustered rows and columns
Doreian (1979)	Algebraic topology
Bonacich (1991)	Correspondence analysis
Freeman (1992)	Analysed woman x woman matrix
Freeman (1993) (1)	Genetic algorithm for woman x woman matrix
Freeman (1993) (2)	Genetic algorithm for woman x woman matrix
Freeman, White (1993) (1)	Galois lattice analysis
Freeman, White (1993) (2)	Galois lattice analysis
Everett, Borgatti (1993)	Regular graph colouring
Borgatti, Everett (1997) (1)	One-mode bipartite matrix analysed for bicliques
Borgatti, Everett (1997) (2)	One-mode bipartite matrix, tabu search algorithm
Borgatti, Everett (1997) (3)	Two-mode matrix analysed by genetic algorithm
Skvoretz, Faust (1999)	Exponential random graph (p*) model
Osbourn (1999)	All-purpose clustering and pattern-recognition algorithm
Roberts (2000)	SVD of doubly normalised adjacency matrix
Newman (2001)	Weighted proximities

Table 5: Analysis methods for Table 2

Woman 9 was assigned to both groups because in interviews she was claimed by both groups. It is clear from Table 2 that woman 9 attended events 7, 8 and 9 which were attended by women from both groups.

We may regard this analysis as a *gold standard*, in the sense that the authors had vastly more information than subsequent analysts, and so the extent to which their analysis and group structure is paralleled or replicated by other approaches is a measure of the success of the approach, of course on just this data set.

We summarise in Table 6 the group assignments reported by Freeman (2003) for all the approaches. They are re-ordered so that the approaches giving the same assignments are grouped together.

There is almost unanimous agreement that women 1-7 and 9 belong to the first group, and women 10-15 belong to the second group. Women 8 and 16-18 are less clearly identified – some analyses combine them into the two groups above, some leave them unassigned, and one has three groups, with women 8 and 16 unassigned. The Osbourn analysis is notably different from the others.

Authors	Group 1	Group 2	Group 3	Ungrouped
Davis, Garner and Garner (1941)	1-9	9-18		
Phillips, Conviser (1972)	1-9	10-18		
Bonacich (1990)	1-9	10-18		
Freeman (1993) (1)	1-9	10-18		
Borgatti, Everett (1997) (2)	1-9	10-18		
Borgatti, Everett (1997) (3)	1-9	10-18		
Roberts (2000)	1-9	10-18		
Freeman, White (1993) (1)	1-9, 16	10-18		
Breiger, Boorman, Arabie (1975)	1-7, 9	8, 10-18		
Newman (2001)	1-7, 9	8, 10-18		
Breiger (1974)	1-7, 9	10-13	14-15, 17-18	8, 16
Doreian (1979)	1-7, 9	10-15		8, 16-18
Everett, Borgatti (1993)	1-7, 9	10-15		8, 16-18
Borgatti, Everett (1997) (1)	1-7, 9	10-15		8, 16-18
Freeman (1992)	1-7, 9	10-16		8, 17-18
Freeman, White (1993) (2)	1-7, 9	10-15, 17-18		8, 16
Skvoretz, Faust (1999)	1-9	10-15, 17-18		16
Freeman (1993) (2)	1-7	8-18		
Osbourn (1999)	1-16	17-18		

Table 6: Group assignments for analysis methods in Table 5

In *none* of these analyses is woman 9 assigned to both groups; apart from the Freeman (1993) (2) analysis, she is always assigned to the first group.

The strength of membership assessments were expressed in terms of "core" or "periphery" which are qualitative and difficult to quantify. We do not discuss these, though we point out the value of the modelling approach we follow in its provision of a probability of group membership which conveys a similar idea.

3 Statistical models

3.1 Models for a random process

A striking feature of the 21 analyses is the breadth of approaches used, and their variety. Only one of these approaches – the p* analysis of Skvoretz and Faust (1999) – used a statistical model, the *exponential random graph model* described below, though the model itself did not address specifically the group structure of the ties.

We consider the presence or absence of a woman at an event as a *random process* – her attendance was determined by a possibly large number of factors unknown to us, so we represent the process outcome as a *Bernoulli random variable*, taking the value $Y_{ij} = 1$ with probability p_{ij}, and $Y_{ij} = 0$ with probability $1 - p_{ij}$. The probability of the pattern of "responses" $\{y_{ij}\}$ given the set of probabilities $\{p_{ij}\}$ over all women and all events, assuming independence of event attendance both within and among women, is

$$\Pr[\{y_{ij}\} \mid \{p_{ij}\}] = \prod_i \prod_j p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}.$$

We can bring the women and event structures into the model in several ways.

3.2 The null model – the Solomonoff-Rapoport random graph model

This is a single-parameter model, giving a constant probability $p_{ij} = p$ for every woman attending every event. The model is due originally to Solomonoff and Rapoport (1951), but is often attributed to Erdös and Rényi (1959). It has no substantive interest in general, providing only a baseline for comparison with informative models.

3.3 The "saturated" model

This model is just a re-statement of the general model, with the event attendance probabilities *completely unrelated parameters* p_{ij}, in general *different for every i and j*. We aim to improve on this model, and the null model, with parsimonious models.

3.4 The Rasch model

The Rasch model is widely used in *item response theory* (IRT) in psychology. Applied to a network, it is expressed through row and column parameters:

- Each woman i has a *propensity* θ_i to attend any event.
- Each event j has an *attractiveness* ϕ_j to any woman.
- Women attend events *independently*, and independently of each other.
- The Rasch model is a *main effect* or *additive* model, in women and events, on the logit scale:

$$\operatorname{logit} p_{ij} = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \theta_i + \phi_j.$$

The model has *no group structure* for women, and so plays the role of a *baseline* model for comparison with models with group structure.

A simpler model is the *Rasch event model* which is defined to be

$$\operatorname{logit} p_{ij} = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \theta + \phi_j.$$

This model has a separate parameter for each event but a common propensity parameter for all women. However the common propensity parameter is confounded with one of the event parameters, so it is simpler to omit it from the model.

3.5 The exponential random graph model (ERGM)

The general form of the class of (homogeneous) exponential (log-linear) random graph models is as follows (Robins, Pattison, Kalish and Lusher 2006):

$$\Pr[\{y_{ij}\}] = (1/\kappa)\exp\left[\sum_{k=1}^{K}\eta_k z_k(\mathbf{y})\right]$$

where

- the summation is over K *configuration types* – different structures of the outcomes y_{ij};
- η_k is the parameter corresponding to configurations of type k;
- $z_k(\mathbf{y}) = z_k(\{y_{ij}\})$ is the *network statistic*: the function of the outcomes y_{ij} corresponding to configuration type k;
- $\kappa = \kappa(\eta_1, ..., \eta_k)$ is a normalizing constant to ensure that the model defines a proper probability distribution.

The ERGM structures did not until recently (Handcock, Raftery and Tantrum 2007) contain any explicit *grouping* of actors.

3.6 Example – the Rasch model

For the Rasch model we have

$$\begin{aligned}
\Pr[\{y_{ij}\}] &= \prod_{i=1}^{n}\prod_{j=1}^{r} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \\
&= \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{r}\left[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})\right]\right\} \\
&= \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{r}\left[y_{ij}\log\frac{p_{ij}}{1-p_{ij}} - \log(1-p_{ij})\right]\right\} \\
&= \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{r}\left[y_{ij}(\theta_i+\phi_j) - \sum_{i=1}^{n}\sum_{j=1}^{r}\log(1+e^{\theta_i+\phi_j})\right]\right\} \\
&= \frac{1}{\kappa}\exp\left[\sum_{i=1}^{n} y_{i+}\theta_i + \sum_{j=1}^{r} y_{+j}\phi_j\right],\ \kappa = \prod_{i=1}^{n}\prod_{j=1}^{r}\left(1+e^{\theta_i+\phi_j}\right).
\end{aligned}$$

So the Rasch model is a special case of the ERGM, with two sets of network statistics y_{i+} and y_{+j} – the marginal sums across events and women. These marginal sums are reproduced by the woman and event factors in the model fitted by maximum likelihood.

3.7 The latent class model

The use of this model in social networks is recent (Snijders and Nowicki 1997, Nowicki and Snijders 2001) but it is very well-established in sociology for contingency table analysis, from Lazarsfeld and Henry (1968) and Goodman (1974) onwards.

The model specifies a *K-class latent structure* for women; the classes are distinguished by different sets of event attendance probabilities *among* classes, but identical attendance probabilities *within* classes. The class structure is not however observed; it is implied and identified by the women's different patterns of event attendance. Within each class k the model fitted is a *restricted* Rasch model, with class-and event-specific event attendance parameters and a class-specific woman propensity parameter; the overall model is a *mixed restricted Rasch model*. The formal model is

$$\begin{aligned}
\Pr[\{Y_{ij}\} \mid k, i, \{q_{jk}\}] &= \prod_{j=1}^{r} q_{jk}^{y_{ij}}(1-q_{jk})^{1-y_{ij}} \\
\Pr[\{Y_{ij}\} \mid i, \{q_{jk}\}] &= \sum_{k=1}^{K}\left[\pi_k \prod_{j=1}^{r} q_{jk}^{y_{ij}}(1-q_{jk})^{1-y_{ij}}\right] \\
\Pr[\{Y_{ij}\} \mid \{q_{jk}\}] &= \prod_{1=1}^{n}\left\{\sum_{k=1}^{K}\left[\pi_k \prod_{j=1}^{r} q_{jk}^{y_{ij}}(1-q_{jk})^{1-y_{ij}}\right]\right\} \\
\operatorname{logit} q_{jk} &= \theta_k + \phi_{jk}.
\end{aligned}$$

Since the propensity parameters θ_k and the attendance parameters ϕ_{jk} are both indexed by class k, as in the Rasch model above they are not separately identifiable without some constraints. We follow the simpler course of omitting the propensity parameters θ_k and working directly with the event attendance probabilities q_{jk}. The model is then a mixture of class- and event-specific product-Bernoullis, with attendance probabilities q_{jk} for women in class k attending event j. (For $K = 1$, the one-latent-class model is identical to the Rasch event model discussed above.) This allows much simpler computation of maximum likelihood estimates, and of posterior distributions in the Bayesian analysis. The probability that woman i is in class k may depend on woman covariates $\mathbf{x}_i$.

The model parameters π_k and q_{jk} do not have to be known or specified; they can be estimated by now-standard methods, discussed below. An important question is how to specify the number of classes K; this is discussed at length below.

3.8 The multiple membership model

In the Natchez network woman 9 was claimed by both groups – she had *multiple group membership*. The latent class model can be generalised to allow for this *a priori*, by defining a set of class membership probabilities for each actor, rather than an exclusive class membership. The resulting analysis is more complicated (see for example Airoldi, Blei, Fienberg and Xing 2008). While the formality of the latent class single membership may appear restrictive, the posterior probabilities of class membership obtained from the maximum likelihood or Bayesian analysis of the latent class model allow for uncertainty in the class membership without the requirement of a formal representation of prior uncertainty, or multiple membership, and this is adequate for our aim.

3.9 The latent space model

The latent class model has recently been generalised further into a *latent space* model (Hoff, Raftery and Handcock 2002 and Handcock, Raftery and Tantrum 2007). Actors are located in a latent Euclidean space of unknown dimension, and in this space are clustered into latent groups using a multivariate normal mixture distribution based on the positions of the actors in the latent space, which are indicated by their binary connections. This requires two levels of latency, and several difficult-to-verify assumptions. As will be seen, the much simpler latent class analysis is adequate for our aim.

4 The latent class model likelihood

Statistical models are estimated, assessed and compared through the *model likelihood*. We have a model $f(\{y_{ij}\} \mid \boldsymbol{\lambda})$ for data $\{y_{ij}\}$, depending on model parameters $\boldsymbol{\lambda}$. The likelihood is

$$L(\boldsymbol{\lambda}) = L(\boldsymbol{\lambda} \mid \{y_{ij}\}) = f(\{y_{ij}\} \mid \boldsymbol{\lambda}).$$

For the latent class model with K classes, the likelihood follows immediately from the mixture model specification above:

$$L(\boldsymbol{\lambda}) = \prod_{1=1}^{n} \left\{ \sum_{k=1}^{K} \left[\pi_k \prod_{j=1}^{r} q_{jk}^{y_{ij}} (1 - q_{jk})^{1-y_{ij}} \right] \right\},$$

with $\boldsymbol{\lambda} = (\{q_{jk}\}, \{\pi_k\}, K)$.

4.1 Maximum likelihood for mixtures

For simplicity of exposition of maximum likelihood we consider a less general mixture model with K components for data $y_1, ..., y_n$:

$$f(y \mid \theta_1, ..., \theta_K, \pi_1, ..., \pi_K, K) = \sum_{k=1}^{K} \pi_k f_k(y \mid \theta_k),$$

where the θ_k are the component-specific parameters.

The Bayesian and likelihood analyses of this model, and of latent class models, are greatly facilitated by the EM approach (Dempster, Laird and Rubin 1977) through the (counterfactual) *complete data likelihood*, in which we construct the likelihood for the "complete data set" of the response variables y_i and the *component indicator variables* Z_{ik}, with $Z_{ik} = 1$ if observation i comes from component k, and $Z_{ik} = 0$ otherwise.

The complete data likelihood and log-likelihood are

$$\begin{aligned}
CDL &= \Pr[\{y_i\}, \{Z_{ik}\}] \\
&= \Pr[\{y_i\} \mid \{Z_{ik}\}] \cdot \Pr[\{Z_{ik}\}] \\
&= \prod_{i=1}^{n} \prod_{k=1}^{K} [f_k(y_i \mid \theta_k)]^{Z_{ik}} \cdot \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{Z_{ik}} \\
CD\ell &= \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log[f_k(y_i \mid \theta_k)] + \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log(\pi_k).
\end{aligned}$$

Maximizing the log-likelihood can be achieved by *alternately* finding the conditional expectation of the complete data log-likelihood given the observed data y_{ij} and the current parameter estimates (the E step), and maximizing this conditioned complete data log-likelihood to get the next parameter estimates (the M step).

In the E step the unobserved data (the Z_{ik}) in the complete data log-likelihood are replaced by their conditional expectations given the observed data and the current parameter estimates. Since the Z_{ik} have a (marginal) multinomial distribution $M(1; \pi_1, ..., \pi_K)$, their conditional distribution for case i given the observed data will again be multinomial, $M(1; \pi_{1i}, ..., \pi_{Ki})$, with expectations ("posterior" probabilities given the observed data) π_{ki}. So the conditional expected log-likelihood and its derivatives are

$$\begin{aligned}
\mathrm{E}[CD\ell] &= \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{ki} \log[f_k(y_i \mid \theta_k)] + \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{ki} \log(\pi_k) \\
\frac{\partial \mathrm{E}[CD\ell]}{\partial \theta_k} &= \sum_{i=1}^{n} \pi_{ki} \frac{\partial \log[f_k(y_i \mid \theta_k)]}{\partial \theta_k}, \\
\frac{\partial \mathrm{E}[CD\ell]}{\partial \pi_k} &= \sum_{i=1}^{n} \left[\frac{\pi_{ki}}{\pi_k} - \frac{\pi_{Ki}}{\pi_K} \right],
\end{aligned}$$

since $\pi_K = 1 - \pi_1 - ... - \pi_{K-1}$.

So the score equations for the expected complete data log-likelihood

$$\frac{\partial \mathrm{E}[CD\ell]}{\partial \theta_k} = 0$$

are *weighted* versions of the score equations for each component parameter separately, where the weights are the posterior probabilities of component membership – the "data fractions" in each component. The EM algorithm alternates between estimating the component parameters given the current posterior probabilities, and computing the posterior probabilities given the current parameter estimates.

The same results (without the EM interpretation) are obtained by direct maximization of the observed data likelihood – the *real* likelihood. We do not give details.

4.2 Number of components

Determining the number of components in a finite mixture (or the number of latent classes in a latent class model) raises difficult inferential questions. Comparisons between models using the likelihood ratio test based on differences of the *frequentist deviance* $-2 \log L_{max}$ for each model are straightforward asymptotically for *nested* models, subject to some regularity conditions:

- the models should have their ML estimates *within* (not on the boundary of) the parameter space;
- the parameters should not be *boundary* (origin or termination) parameters for the space of the tie variables;
- the sample should be large enough for the log-likelihood in the model parameters to be approximately quadratic.

When we are comparing models with K and $K-1$ classes in a finite mixture, although the models are nested, the second regularity condition fails, as the restricted model has to have one of the π_k set to zero, a boundary value. Alternatives to the likelihood ratio test have been proposed, and a detailed comparison of some of these approaches for determining the number of classes in various latent class models was made in Nylund, Asparouhov and Muthen (2007), using sample sizes of 200, 500 and 1000.

The "naive likelihood ratio test", assuming the validity of the asymptotic χ^2 distribution, performed very poorly, with test sizes of the 5% level test of K to $K-1$ classes typically around 25%. The bootstrap likelihood ratio test, in which the critical value of the test is assessed by bootstrapping samples from the fitted $K-1$ class model, performed uniformly well. The Bayesian Information Criterion (BIC) was the best of the evaluated information criteria, though it performed badly with the smallest simulation sample size of 200, and unequal proportions in the latent classes. Since the sample size is very small in the network we are discussing, it is not clear how the bootstrap likelihood ratio test and the BIC would perform.

An important difficulty in mixture analysis is the occurence of *multiple local maxima* of the likelihood. This occurs because different membership structures of the observations may give very similar maximized likelihoods. As the number of components in the mixture increases, the number of local maxima may also increase.

The possibility of local maxima is usually allowed for by using *multiple random starting values* for the component memberships, and choosing the best of these to continue the EM iterations. This does not guarantee that the chosen "best" starting configuration gives the global maximum of the likelihood, nor does it allow the assessment of the closeness of other local maxima to the global maximum. If several local maxima are very close, then they need to be considered jointly in the interpretation of the parameter estimates and the component membership probabilities.

We investigate this issue in the analyses which follow by using 100 random starting assignments, and then iterating them *all* through 500 EM iterations to convergence, to determine the number of important local maxima in the 100 analyses. (The number of possible random assigments of the n observations to K classes is very large, of the order of K^n, so 100 random assignments is a tiny fraction of the possible number. The point is to investigate local maxima.) We do not need to investigate *all* the local maxima: a difference of 7 or more in frequentist deviances between the global and a local maximum is a ratio of maximized likelihoods of 0.03, which is sufficiently small to eliminate the local maximum from consideration. We discuss this further in the Bayesian analysis below.

5 Maximum likelihood for the latent class model

Maximum likelihood in the Rasch and latent class models is well-documented in many places, and implemented in many programs. For the latent class model, we have (with $\mathbf{y}_i = \{y_{ij}\}$),

$$\begin{aligned}
L(\boldsymbol{\lambda}) &= \prod_{1=1}^{n} \left\{ \sum_{k=1}^{K} \pi_k \prod_{j=1}^{r} q_{jk}^{y_{ij}} (1 - q_{jk})^{1-y_{ij}} \right\}, \\
\frac{\partial \log L}{\partial q_{jk}} &= \sum_{i=1}^{n} \pi_{ki} (y_{ij} - q_{jk}) \\
\pi_{ki} &= \frac{\pi_k \prod_{j=1}^{r} q_{jk}^{y_{ij}} (1 - q_{jk})^{1-y_{ij}}}{\sum_{\ell=1}^{K} \pi_\ell \prod_{j=1}^{r} q_{j\ell}^{y_{ij}} (1 - q_{j\ell})^{1-y_{ij}}} \\
\widehat{q_{jk}} &= \sum_{i=1}^{n} \pi_{ki} y_{ij} \Big/ \sum_{i=1}^{n} \pi_{ki}, \\
\widehat{\pi_k} &= \sum_{i=1}^{n} \pi_{ki}/n.
\end{aligned}$$

The maximum likelihood estimates of the q_{jk}, given the current weights π_{ki}, are weighted version of those for the product-Bernoulli model in each class. In latent class analysis for the Natchez women, the posterior probabilities of class membership can provide a formal *classification* of women to classes, based on the MAP (maximum posterior probability) – the "most probable" class, or more informatively, they provide the *degree of certainty* or *degree of membership*, of each woman in each class.

5.1 Frequentist analysis of the Natchez women

The frequentist deviances for the null model, the Rasch event model and the full Rasch women + event model are 327.29, 286.29 and 256.17. Assuming quadratic log-likelihoods in the model parameters, the models can be compared by the likelihood ratio test, giving test statistics of 41.00 with 14 degrees of freedom for null vs event and 30.12 with 18 degrees of freedom for event vs full. The full Rasch model is strongly preferred over the two simpler models; in particular, both women and events show large and significant effects. However we have no effective test of *goodness of fit* of the Rasch model against the saturated model, for the latter model with a parameter for each (i, j) pair has a frequentist deviance of *zero* by definition, because the fitted values for this model are exactly equal to the observed values.

We extend the Rasch model to the latent class models. We summarise in Table 7 the best frequentist deviances (over the random starting values) for latent class models with up to five classes. For the two-class model, there are three local minima of the deviance in the 100 random starting assignments. The global minimum is found from 95 assignments. The next-best local minimum is 16 deviance units below the best, and can be ignored.

For the three-class model, there are more than 10 local minima of the deviance in the 100 starting assignments, though only one is within 7 deviance units of the best. The global minimum of 182.87 is only 0.87 better than the second-best of 183.74. The inferior minimum moves woman 10 from class 3 to class 2.

For the four-class model, 81 of the 100 deviances are within 7 of the best. This may represent cases of very slow convergence to the global minimum (or a smaller set of local minima) but at least some of the differences are due to different assignments of women to classes which are nearly consistent in their representation of the data.

For the second-best minimum, woman 9's probability 0.92 of being in class 3 is moved to class 4, with the same probability. Woman 16 in class 3 has a small probability 0.08 of being in class 4. Table 8 gives the ML estimates of the event attendance probabilities (to 2dp) for each number of classes from the best deviances, and Table 9 gives the ML estimates of the class posterior probabilities.

A critical question is how to assess the number of classes which should be reported. Here the likelihood ratio test does not help us, for two reasons:

- while the models with increasing numbers of classes are nested, a necessary regularity condition fails, as described above;

Model	deviance	# params
Null	327.29	1
Rasch	256.17	32
Rasch event	286.29	14
$K = 2$	208.00	29
$K = 3$	182.87	44
$K = 4$	169.08	59
$K = 5$	155.98	74

Table 7: Frequentist deviances

K	1	2		3			4				5				
k	1	1	2	1	2	3	1	2	3	4	1	2	3	4	5
1	.17	.38	-	.43	-	-	.42	-	-	-	.50	.40	-	-	-
2	.17	.38	-	.43	-	-	.42	-	-	-	1.00	.20	-	-	-
3	.33	.75	-	.85	-	-	.85	.33	-	-	1.00	.80	-	-	-
4	.22	.50	-	.57	-	-	.56	.33	-	-	1.00	.40	-	-	-
5	.44	1.00	-	1.00	-	.12	1.00	1.00	-	-	1.00	1.00	-	.21	-
6	.44	.75	.20	.85	.33	.13	.85	.33	.33	.25	1.00	1.00	.24	.21	-
7	.56	.75	.40	.72	.67	.37	.72	1.00	.67	-	.50	.80	.73	.41	-
8	.78	.88	.70	.86	.67	.75	.86	.67	.67	.50	1.00	.80	.76	1.00	-
9	.67	.38	.90	.29	1.00	.87	.29	.33	1.00	1.00	1.00	-	.76	1.00	1.00
10	.28	-	.50	-	1.00	.25	-	-	1.00	-	-	-	1.00	.18	-
11	.22	-	.40	-	.33	.38	-	-	.33	-	-	-	.48	-	1.00
12	.33	-	.60	-	1.00	.38	-	-	1.00	-	-	-	1.00	.38	-
13	.17	-	.30	-	1.00	-	-	-	1.00	-	-	-	.73	-	-
14	.17	-	.30	-	1.00	-	-	-	1.00	-	-	-	.73	-	-

Table 8: Class event attendance MLE probabilities $\widehat{q_{jk}}$ for the best deviances

- the data for each class become increasingly sparse as the number of class parameters increases, since the sample size is fixed, so the log-likelihoods for the models become *increasingly non-quadratic* with increasing K.

The deviance decrements with increasing K – the "naive likelihood ratio test" statistics – do give some indication, however, as mentioned in the Nylund et al paper: the 1-2 change of 78.29 for 15 parameters is very large, that for 2-3 of 25.13 is at the 0.048 level of χ^2_{15}, and the 3-4 and 4-5 differences of 13.79 and 13.10 appear to be random noise. Nylund et al pointed out that the overstatement of significance by the naive test means that if the deviance difference is not significant at a conventional 5% level of the χ^2 test, then it will definitely *not* be significant by a better-controlled test with a correct test size. With this interpretation, we would conclude that two classes are definitely identified, three classes are questionable and more than three are not necessary.

The BIC comes to the same conclusion: penalising the deviance by $p \log n = 2.89\, p$ gives $K = 2$ as the preferred model (this is also true for AIC). If we take the two-class model as best supported by the data, the (ML estimates of the)

K	2		3			4				5				
$i\backslash k$	1	2	1	2	3	1	2	3	4	1	2	3	4	5
1	1		1			1								1
2	1		1			1				1				
3	1		1			1								1
4	1		1			1				1				
5	1		1			1				1				
6	1		1			1				1				
7	1		1			1				1				
8		1			1			1				1		
9	1		.04		.96	.08		.92				1		
10		1			1				1			1		
11		1		1					1		.14	.86		
12		1		1			1				1			
13		1		1			1				1			
14		1		1			1				1			
15		1		1					1		1			
16		1			1			1				1		
17		1			1			1					1	
18		1			1			1					1	
$\hat{\pi_k}$	.444	.556	.391	.278	.331	.393	.167	.273	.167	.278	.230	.270	.111	.111

Table 9: Probabilities $\widehat{\pi_{ki}}$ of class membership

posterior probabilities of class membership are all 1 or zero – the women appear to be classified with *certainty*, with the same assignments (given in Table 9) as given by Breiger, Boorman and Arabie (1975) and Newman (2001). Blank entries in Table 9 represent probabilities less than 0.005.

The three-class model has the same assigments as Doreian (1979), Everett and Borgatti (1993) and Borgatti and Everett (1997) (1), though their analyses leave unclassified the women 8 and 16-18, whereas the latent class analysis assigns them to a third class. This model also has estimated posterior probabilities of class membership 1 or zero except for woman 9, who was claimed by both classes in the DGG two-class analysis. In our two-class analysis she is assigned to class 1 with probability 1, but in the three-class analysis she is assigned to class 1 with probability 0.04 and class 3 with probability 0.96.

We revisit this assessment with a different approach, discussed in the next section.

6 Bayesian analysis of the latent class model

6.1 Priors and posteriors

We augment the likelihood $L(\theta)$ by a *prior distribution* $\pi(\theta \mid \phi)$ for the model parameters θ depending in general on prior ("hyper-") parameters ϕ, and use

Bayes's theorem to update the prior distribution to the *posterior distribution* $\pi(\theta \mid \mathbf{y}, \phi)$:

$$\pi(\theta \mid \mathbf{y}, \phi) = \frac{L(\theta) \cdot \pi(\theta \mid \phi)}{\int L(\theta) \cdot \pi(\theta \mid \phi)\, \mathrm{d}\theta}.$$

The denominator is a scaling term, depending on the data $\mathbf{y}$ and ϕ but not θ.

If the prior is *flat* – constant – then the posterior distribution is a simple scaled version of the likelihood. Throughout this paper, following Aitkin (2010) we use flat, *reference* or *non-informative* priors, to allow as far as possible the *data* to determine the posterior distribution through the likelihood.

Inference about these parameters is through their posterior distributions, commonly through the posterior mean, posterior median and other percentiles. Credible intervals for the parameters follow from the posterior percentiles, in a way familiar from frequentist inference but without any assumption of normality. The posterior standard deviation is often quoted, though this is useful only for normal posterior distributions.

6.2 Priors for the latent class models

We implemented the Markov chain Monte Carlo procedure in OpenBugs. For the class-specific event attendance parameters q_{jk} we used uniform priors, and for the mixture proportions π_k we used a slightly informative proper Dirichlet prior.

We specified the sequence of conditionals:

$$\begin{aligned} \text{class} &\sim M(1; \pi_1, ..., \pi_K) \\ q_{jk} \mid \text{class}\, k &\sim \text{Beta}(1, 1) \\ \Pr[y_{ij} \mid \text{class}\, k] &= \prod_{j=1}^{r} q_{jk}^{y_{ij}} (1 - q_{jk})^{1 - y_{ij}} \end{aligned}$$

6.3 Posteriors of functions of data and parameters

One of the powerful features of Bayesian analysis is its ability to provide posterior inference about complicated functions of the data and parameters. In frequentist theory we have to rely on the *delta method*, or Taylor series expansions, to obtain the asymptotic sampling distributions of non-linear functions of the model parameters, especially *ratios* of parameters.

In Bayesian theory this is unnecessary: for *posterior sampling* inference about a non-linear function $g(\theta)$ of the model parameters, we simply make M random draws $\theta^{[m]}$ of θ from $\pi(\theta \mid \mathbf{y})$, and substitute them into the function g, to give M random draws $g^{[m]} = g(\theta^{[m]})$.

6.4 Posterior distribution of class membership probabilities

A general problem with the use of posterior class membership probabilities from Bayes's theorem following the EM algorithm is familiar from the frequentist analysis of other complex models. This is the problem of *overstated precision* resulting from the substitution of ML estimates for true parameter values, without any allowance for the imprecision of the ML estimates. In frequentist analysis this is forced on us by the complexity of the exact sampling distributions, especially for non-linear functions of the parameters.

As described above, the posterior distributions of these quantities can be obtained in theory from the random draws of the parameters. Recall that the posterior probability of membership of case i in class k, in the general K-component mixture, is

$$\pi_{ki} = \frac{\pi_k f_k(y_i \mid \theta_k)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i \mid \theta_\ell)}.$$

From the posterior distributions of θ_k and π_k, we make M independent draws $\theta_k^{[m]}$ and $\pi_k^{[m]}$ and substitute them into π_{ki} to give M draws

$$\pi_{ki}^{[m]} = \frac{\pi_k^{[m]} f_k(y_i \mid \theta_k^{[m]})}{\sum_{\ell=1}^{K} \pi_\ell^{[m]} f_\ell(y_i \mid \theta_\ell^{[m]})}.$$

However, a major issue in computing posterior distributions for *any* class-specific parameter is *label-switching*. In the ML estimation of the model parameters, the class labelling of the K class parameter estimates is arbitrary, and causes no confusion. However in the MCMC iterations, the class labelling can vary during iterations and switch the class labels around, leading to class-specific posteriors that are mixtures of the posteriors from each true class. This can lead in the worst case to identical class-specific distributions for all the class parameters, despite convergence.

We follow the approach of Sperrin et al (2010), in which the labels to be attached to the M sets of posterior parameter draws are treated as *missing data* and analysed with an EM algorithm. For the two-class Natchez women model, there is almost no uncertainty about the draw labels, but there is considerably more uncertainty with the three- or more class models. We discuss this below.

6.5 Posterior distribution of the model deviance

6.5.1 General models

A particularly useful application of the posterior sampling approach is to the deviance. In Bayesian terms the deviance is $D(\theta) = -2 \log L(\theta)$. Since this is a function of both θ and the data $\mathbf{y}$, it also has a posterior distribution obtainable in this way: given M random draws $\theta^{[m]}$, we substitute them into the deviance to give M random draws $D^{[m]} = D(\theta^{[m]})$.

A full discussion of this approach, and many applications of it, are given in Aitkin (2010). Aitkin, Vu and Francis (2013) carried out a simulation study to evaluate this approach, for both normal mixtures and Bernoulli latent class models. For latent class models, identification of many classes required substantial sample sizes of actors in the simulations based on binary symptoms in psychiatric patients. Our application here is to Bayesian model comparisons of the number of latent classes. A normal mixture example can be found in Chapter 9 of Aitkin (2010).

We have K possible models, consisting of latent class models with 1,..., K classes, specified by K sets of model parameters θ_k containing the class proportion parameters π_k and the class-conditional event attendance parameters q_{jk}. For each class we obtain the posterior distribution of θ_k, and the consequent posterior distribution of the deviance D_k. The deviance is unaffected by label-switching, since it is a *symmetric function* of the class labels, and invariant under their permutation.

The comparison of the class models is equivalent to the comparison of the *distributions* of the class deviances. In large samples from regular models with MLEs internal to the parameter space, the second-order Taylor expansion of the deviance $D(\theta)$ about the MLE $\hat{\theta}$ gives the posterior distribution of the model deviance, by a Bayesian version of the derivation of the asymptotic χ^2 distribution for the likelihood ratio test statistic:

$$D_k(\theta_k) \sim D_k(\widehat{\theta}_k) + \chi^2_{p_k},$$

where p_k is the dimension of θ_k (Aitkin 2010 p. 53). It is notable that the frequentist deviance is a *location parameter* for the posterior distribution of the Bayesian deviance, which *starts from the frequentist deviance*, since this is the minimised deviance over all possible parameter values.

As will become clear in the following analyses, the asymptotic result, if it applies at all, is restricted to the simplest models: null, Rasch event and Rasch, for which the parameter dimension is small compared to the number of ties. For the mixture models, the actual posterior distributions depart from their asymptotic forms in two respects, as described in Aitkin (2010 pp. 216-220):

- the distributions *start from larger values than the frequentist deviances*, because with increasing parameter dimension it is increasingly difficult to sample by chance the MLE; and
- the distributions are *more diffuse* than the χ^2 distributions because the parameter posteriors are skewed and/or heavy-tailed.

However, the deviance distributions can in many cases be *stochastically ordered.* There are several possibilities, which are illustrated in the next section with the Natchez ladies network.

1. **Complete stochastic ordering**: the deviance cdfs are *ordered* and *do not cross* (except merge at 0 and 1); then the leftmost distribution gives the best-preferred model.

2. **Partial stochastic ordering**: the leftmost distribution does not cross the others, which may cross each other; then the leftmost distribution gives the best-preferred model.

3. **No stochastic order**: some or all of the distributions cross.

We discuss these further, and give a procedure for discriminating the models, with the analysis of the Natchez women data.

6.5.2 The saturated model

For the saturated model we can compute directly the deviance distribution. The log-likelihood is

$$\log L_{sat} = \sum_i \sum_j \left[y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij}) \right],$$

where each p_{ij} has its own posterior beta distribution, which assuming independent uniform priors for all, gives $B(2,1)$ for $y_{ij} = 1$ and $B(1,2)$ for $y_{ij} = 0$. So the deviance distribution D_{sat} for the saturated model is

$$D_{sat} = -2 \cdot \sum_i \sum_j \left[y_{ij} \log B_{ij}(2,1) + (1 - y_{ij}) \log(1 - B_{ij}(1,2)) \right].$$

The distribution of $1 - B(1,2)$ is $B(2,1)$, so the saturated deviance distribution is $D_{sat} = -2 \sum_i \sum_j \log B_{ij}(2,1)$, a sum of $2nr$ independent $-\log B(2,1)$ random variables, where n and r are the numbers of women and events.

The sum will be very accurately normal, summed over the large number nr of cells of the table, so the deviance will be normal with

$$\begin{aligned}
\mathrm{E}[D_{sat}] &= 2nr \cdot \mathrm{E}[-\log B(2,1)] = 2nr \cdot (1/2) \\
&= nr; \\
\mathrm{Var}[D_{sat}] &= 4n^2r^2 \cdot \mathrm{Var}[-\log B(2,1)] = 4n^2r^2 \cdot (1/4) \\
&= n^2r^2; \\
\mathrm{SD}[D_{sat}] &= nr.
\end{aligned}$$

So for the table $\mathrm{E}[D_{sat}] = \mathrm{SD}[D_{sat}] = 252$. The very large variance of the saturated deviance is a consequence of its very heavy parametrization and the small amount of information about the p_{ij} in each Bernoulli outcome y_{ij}. The saturated deviance distribution is so diffuse that its cdf will cross any other well-fitting *or* poorly-fitting model deviance cdf – it is *too diffuse* to serve as a baseline for well-fitting models. We do not further consider its use.

7 Bayesian analysis of the Natchez women

We initiated the Bayesian analysis with multiple searches for each K over 100 random starting assignments of women to classes, and choosing the class assignment with the highest maximized likelihood to provide starting values for the

MCMC analysis. This used Gibbs sampling for each K to alternate between draws from the conditional posterior Beta and Dirichlet distributions of the model parameters given the current latent class assignment draws, and draws from the posterior multinomial distribution of the class assignment variables given the current parameter draws.

We ran a burn-in of 10,000 runs and then made a further 10,000 runs from which we sampled every 10th run to provide 1,000 draws from the parameter posterior distribution for this K. The draws were then substituted into the K-class deviance to give the deviance posterior, and the parameter draws were also substituted, after analysis and identification of the parameter draw class labels, into the class membership probabilities to give their posterior distributions for each $k = 1, ..., K$.

We also ran a separate analysis for the Rasch model: we specified proper but very diffuse normal priors for the woman and event parameters, and thinned the 10,000 draws to 1,000 for the posterior deviance distribution. The deviance distributions for $K = 1, ..., 7$ classes and the full Rasch model are shown in Figure 3.

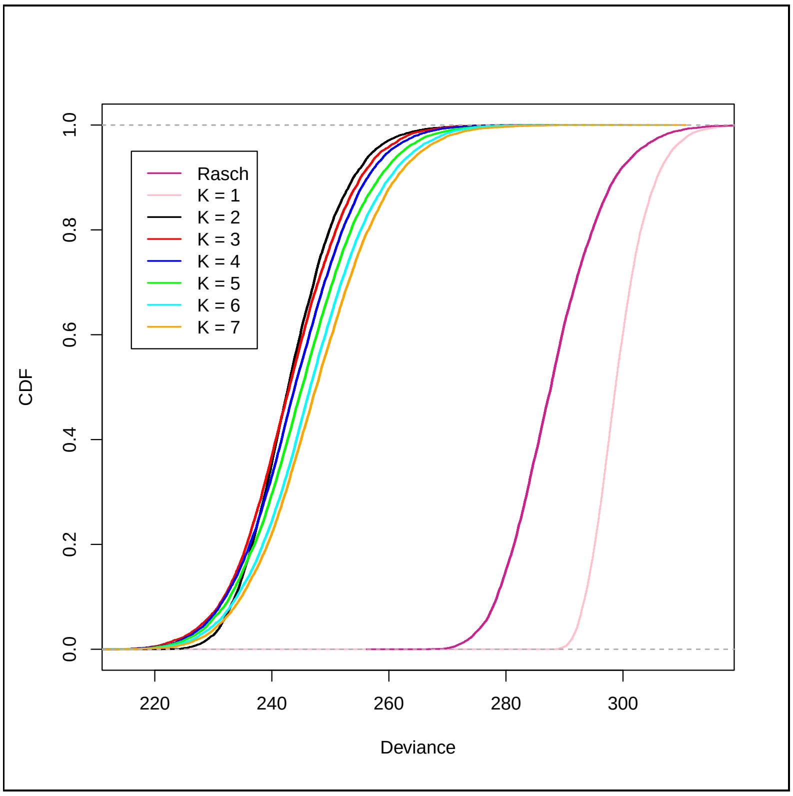

Figure 3: Deviance distributions for latent class models and the Rasch model

The one-class (event Rasch) and full Rasch models are both inferior to the class models with $K > 1$, and can be excluded. The deviance distributions for $K = 2$ and 3 improve substantially on the Rasch model, are very close and cross each other at about the 50th percentile. From $K = 4$ onwards the deviance distributions move slowly to the right, and are stochastically ordered – the cdfs do not cross. This movement is greater at higher percentiles, as expected from the increasing "leans" of the distributions with increasing numbers of parameters.

So the two- and three-class models are preferred. It is clear that there are at least two classes, but two and three classes are almost equally well-supported. The deviance distributions for the 2- and 3-class, and the null, event and full Rasch models are shown (solid curves) in Figure 4, with their χ^2 asymptotic forms (dashed curves) assuming quadratic log-likelihoods. The frequentist deviances are shown as circles on the deviance axis. The crossing of the deviance distributions for the 2- and 3-class models is much clearer. The

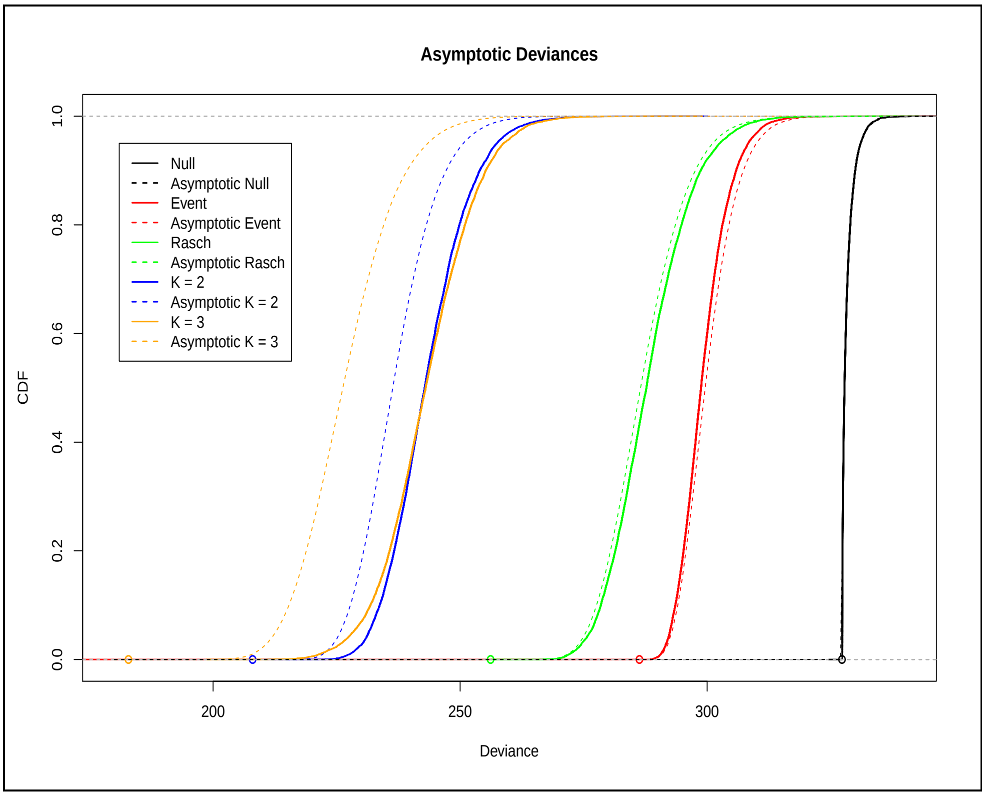

Figure 4: A subset of the Natchez women deviances with asymptotic distributions and frequentist deviances

shapes of the distributions agree well with the asymptotic distributions in all cases, but the *location* agreement deteriorates steadily, through an increasing right shift of the empirical distributions away from the asymptotic distributions with increasing numbers of model parameters.

So the Bayesian approach through comparison of posterior deviance distributions leaves us uncertain whether there are two or three classes. However we have additional information which helps.

7.1 Posterior distributions of class membership

7.1.1 The 2-class model

We noted in the frequentist analysis that for the 2-class model there were two local maxima of the likelihood in addition to the global maximum, and the best of these was 16 (frequentist) deviance units worse than the global maximum. These local maxima can therefore be ignored in considering the class membership probabilities.

We have for each draw m the parameter values $\theta^{[m]}$, from which we can compute the membership probability draws $\pi_{ki}^{[m]}$ as described in §11.4. Kernel densities of these probabilities are shown in Figure 5 for the 18 women. The distributions are extremely concentrated for almost all the women – almost spikes. The figures show the distribution of $\pi_{2i} + 1$. The horizontal scale for each woman is from 1 to 2 at the left end 1 for class 1 and zero for class 2. The vertical scale is the kernel density ordinate. It is notable that the near-spikes for women 1-6 are all at, or very near, the class 1 end, while those for women 10-18 are at, or very near, the class 2 end. However the near-spike for woman 9 is near the centre of the range, and those for women 7 and 8 diverge somewhat from the end of the range.

These results are clarified by a different form of presentation of the posterior distributions; namely those of the *component indicator variables* Z_{ik} used in the MCMC analysis. For each indicator variable we summarise the M posterior draws (0 or 1) by the *proportion of 1s* in the 1000 draws, giving a *Bayesian point estimate* of the posterior probability of class membership. The resulting proportions are given in Table 10.

It is now clear that:

- Woman 9 *belongs to both classes.*
- Woman 8 has *some affinity* with class 1, but *much more* with class 2.
- Woman 7's degree of membership in class 1 *is less* than those of women 1-6.

These results clarify the differences in the 21 analyses over where woman 8 should be placed – she has some degree of membership in both classes. The ambiguity of woman 9's membership is in accord with the "gold standard" which we described in Chapter 2.

Woman #	class 1	class 2
1	1	0
2	1	0
3	1	0
4	1	0
5	0.995	0.005
6	0.994	0.006
7	0.929	0.071
8	0.109	0.891
9	0.431	0.569
10	0.004	0.996
11	0	1
12	0	1
13	0	1
14	0	1
15	0.001	0.999
16	0.023	0.977
17	0.005	0.995
18	0.004	0.996

Table 10: Membership assignments

7.1.2 The 3-class model

For the three-class model, we noted in the frequentist analysis that there were more than 10 local minima of the deviance in the 100 starting assignments, though only one was within 7 deviance units of the best. The global minimum of 182.87 deviance units was only 0.87 units better than the second-best of 183.74. The inferior minimum moved woman 10 from class 1 to class 2.

In the Bayesian analysis of this model, we found that label-switching could not be completely eliminated by the EM "missing label" approach, partly because of the ambiguous classification of woman 10. For the best label configuration, we computed the proportions of posterior draws of the component indicator variables for classes 1, 2 and 3 for each woman. These are shown in triangular (ternary) plot form in Figure 6, with jittering of the points perpendicular to the lines.

Class 1 women are tightly packed around the class 1 apex, while class 2 women are more diffusely spread along the class 2 axes, with women 7, 8 and 9 in similar positions to the 2-class analysis, along the class 1 - class 2 axis. It is striking that *no* woman has higher probability of being in class 3 than in classes 1 or 2.

The presence of the very close local and global maxima of the likelihood, and the switching of woman 10 from class 1 to class 2 between these maxima in the frequentist analysis, make clear that the 3-class model is *unstable*, and no woman belongs to the third class with persuasive probability – this class is effectively empty.

8 Conclusions

The approach to sub-grouping of actors in social networks through latent class modelling and analysis is successful in identifying the group structure of the Natchez women, including the joint membership of both groups by woman 9; it also explains the difficulty of classifying woman 8 in some of the other approaches.

The identification of the number of groups through the posterior distributions of the competing model deviances also appears to be successful. (A detailed assessment of the performance of this approach in finite mixtures is given in Aitkin, Vu and Francis 2013.)

One important point limiting the general adoption of this approach is that it is computationally intensive for large networks, in the sense either of the number of latent classes or the dimension of the adjacency matrix. For large networks, *variational Bayes* methods, using approximations to the full posterior distribution (of the parameters and indicator variables) with simpler structure, have been found computationally effective, though the degree of agreement with the full analysis has not been clearly established. A detailed description can be found in Vu, Hunter and Schweinberger (2013).

9 Acknowledgements

In preparing this paper we have benefitted greatly from interactions with the Social Network group in the Department of Psychology, School of Psychological Sciences at the University of Melbourne. Professors Pip Pattison (a Deputy Vice-Chancellor of the University) and Garry Robins have been particularly helpful in comparing and interpreting the differences between this approach and others used in social network modelling, especially the exponential random graph model approach. We appreciate helpful comments from a referee and an editor.

We are also grateful for research support (Project DP120102902) from the Australian Research Council for Pip Pattison's participation in the project, for the support of Duy Vu for the period of this research (2012-15), and for visits by Brian Francis from the University of Lancaster.

10 References

Airolidi, E.M., Blei, D.M., Fienberg, S.E. and Xing, E.P. 2008. Mixed membership stochastic blockmodels. Journal of Machine Learning Research 9, 1981–2014.

Aitkin, M. 2010. Statistical Inference: an Integrated Likelihood/Bayesian Approach. Chapman and Hall/CRC Press, Boca Raton FL.

Aitkin, M., Vu, D. and Francis, B.J. 2013. A new Bayesian approach for determining the number of components in a finite mixture. Submitted.

Bonacich, P. 1991. Simultaneous group and individual centralities. Social Networks 13, 155–168.

Breiger, R.L. 1974. The duality of persons and groups. Social Forces 53, 181–190.

Breiger, R.L., Boorman, S.A. and Arabie, P. 1975. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. Journal of Mathematical Psychology 12, 328–383.

Borgatti, S.P. and Everett, M.G. 1997. Network analysis of 2-mode data. Social Networks 19, 243–269.

Davis, A., Gardner, B.B. and Gardner, M.R. 1941. Deep South: A Social Anthropological Study of Caste and Class. University Press, Chicago.

Davis, A. and Warner, W.L. 1939. A comparative study of American caste. In Race Relations and the Race Problem. Duke University Press, Durham NC.

Dempster, A.P., Laird, N.M and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with Discussion). Journal of the Royal Statistical Society B, 39, 1–37.

Doreian, P. 1979. On the delineation of small group structure. In H.C. Hudson (ed.) Classifying Social Data. Jossey-Bass, San Francisco.

Doreian, P., Batagelj, V. and Ferligoj, A. 2005. Generalized Blockmodeling. University Press, Cambridge.

Erdös, P. and Rényi, A. 1959. On random graphs. I. Publicationes Mathematicae Debrecen 6, 290-297.

Everett, M.G. and Borgatti, S.P. 1993. An extension of regular colouring of graphs to digraphs, networks and hypergraphs. Social Networks 15, 237–254.

Freeman, L.C. 1992. The sociological concept of "group": an empirical test of two models. American Journal of Sociology 98, 152–166.

Freeman, L.C. 1993. Finding groups with a simple genetic algorithm. Journal of Mathematical Sociology 17, 227–241.

Freeman, L.C. 2003. Finding social groups: a meta-analysis of the Southern women data. In R. Breiger, K. Carley and P. Pattison (eds.) Dynamic Social Network Modeling and Analysis. The National Academies Press, Washington, D.C.

Freeman, L.C. and White, D.R. 1993. Using Galois lattices to represent network data. In Sociological Methodology 1993, ed. P. Marsden, pp.127–146. Blackwell, Cambridge MA.

Goldenberg, A., Zheng, A.X., Fienberg, S.E. and Airoldi, E.M. 2009. A survey of statistical network models. Foundations and Trends in Machine Learning 2, 129–233.

Goodman, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61, 215–231.

Handcock, M.S., Raftery, A.E. and Tantrum, J. 2007. Model-based clustering for social networks. Journal of the Royal Statistical Society A 170, 301–354.

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.

Lazarsfeld, P.F. and Henry, N.W. 1968. Latent Structure Analysis. Houghton Mifflin, Boston.

Newman, M. 2001. The structure of scientific collaboration networks. Proceedings of the National Academy of Science 98, 404-409.

Novicki, K. and Snijders, T.A.B. 2001. Estimation and prediction for stochastic block models. Journal of the American Statistical Association 96, 1077–1087.

Nylund, K.L., Asparouhov, T. and Muthen, B.O. 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Structural Equation Modeling 14, 535-569.

Osbourn, G.C. 1999. Clustering, pattern recognition and augmentation of data visualization using VERI algorithm (reported by Freeman 2003).

Phillips, D.P. and Conviser, R.H. 1972. Measuring the structure and boundary properties of groups: some uses of information theory. Sociometry 35, 235–254.

Roberts, J.M. 2000. Correspondence analysis of two-mode networks. Social Networks 22, 65–72.

Robins, G., Pattison, P., Kalish, Y. and Lusher, D. 2007. An introduction to exponential random graph (p*) models for social networks. Social Networks 29, 173-191.

Skvoretz, J. and Faust, K. 1999. Logit models for affiliation networks. Sociological Methodology 29, 253–280.

Snijders, T.A.B. and Nowicki, K. 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of Classification, 14, 75–100.

Solomonoff, R. and Rapoport, A. 1951. Connectivity of random nets. Bulletin of Mathematical Biophysics 13, 107–117.

Sperrin, M., Jaki, T. and Wit, E. 2010. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. Statistics and Computing 20, 357-366.

Vu, D.Q., Hunter, D.R. and Schweinberger, M. 2013. Model-based clustering of large networks. The Annals of Applied Statistics 7, 1010–1039.

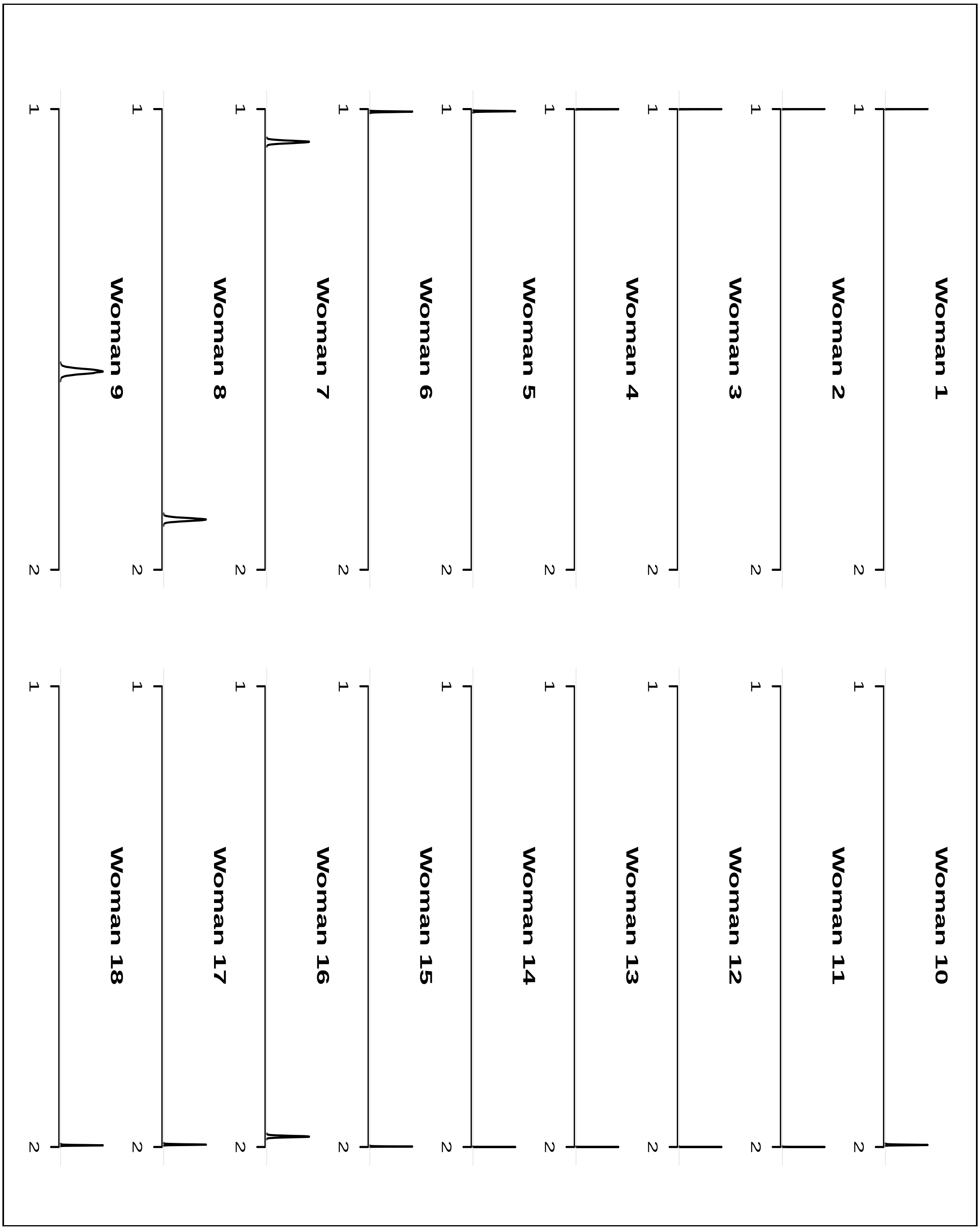

Figure 5: Natchez women 2-class membership ($\pi_{2i} + 1$)

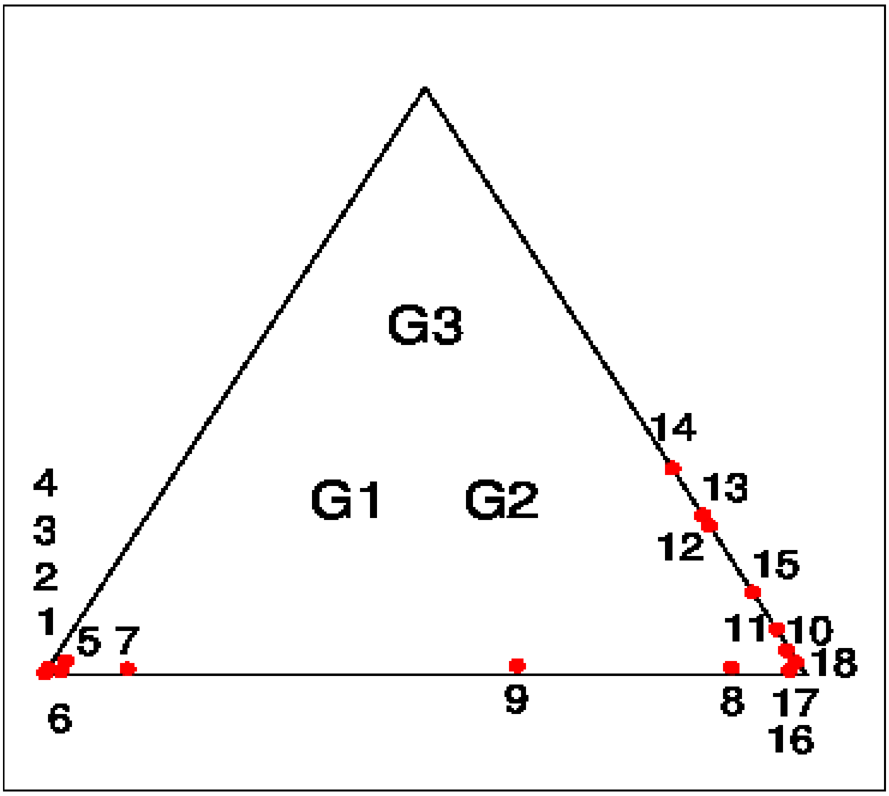

Figure 6: Natchez women: ternary plot of 3-class membership