Stochastic blockmodels

Outline

Stochastic blockmodelling
- Stochastic equivalence

Statistical model
- Estimation
- Posterior probabilities of interblock relations

Fititng stochastic blockmkodels using BLOCKS in StocNet

Selecting number of blocks

Example 1: the political network
Example 2: work frequency and importance in an organisaiton
Example 3: collaboration in a law firm

Other latent variable models for networks
- Latent space model
- Latent ultrametric model

A posteriore stochastic blockmodels (Nowicki & Snijders, 2001)

Assume that:

Actors are associated with ***unobserved*** classes, or groups (represented by different ***colours***)

the probability that a dyad has a particular relational "pattern" depends only on the (unobserved) colours of the actors

possible patterns define an ***alphabet*** of possibilities

actors with the same colour have the same pattern probabilities and hence are ***stochastically equivalent***

dyads are ***conditionally independent***, given the colours of the actors

More formally

Let Z_i be a variable denoting the ***colour*** of actor i and assume that the Z_i are iid random variables with probability $\Pr(Z_i = k) = \theta_k$

Also assume that $\Pr(\mathbf{D}_{ij} = \mathbf{a}|\mathbf{Z} = \mathbf{z}) = \eta_a(z_i,z_j)$ where:

$\mathbf{D}_{ij}$ is a vector of relational variables for the dyad (i,j)

$\mathbf{a}$ is a vector of possible values for the dyad variables (a ***pattern*** in the ***alphabet***), and

$\eta_{\mathbf{a}}(z_i,z_j)$ is the colour-dependent probability of observing the pattern $\mathbf{a}$

Also assume that the dyads are conditionally dependent given the colours $\mathbf{Z}$, so that:

the joint distribution of the $\mathbf{D}_{ij}$ given $\mathbf{Z}$ is the product of the conditional dyad probabilities

$\Pr(\mathbf{D},Z) = \Pr(\mathbf{D}|\mathbf{Z})\Pr(\mathbf{Z})$

Estimation

Nowicki and Snijders developed a Bayesian approach to estimation of θ and η using Gibbs sampling (an iterative sampling scheme for approximating posterior distribution)

They also compute posterior probabilities that:

- two actors have the same colour
- a dyad has a particular relational pattern
- interblock ties have a particular relaional pattern

Implemented in BLOCKS within Stocnet

Example 1: political network

Data file: Doreian_mat.txt

0 0 1 1 0 0 1 0 0 0 0 0 1 0
0 0 0 0 0 0 0 1 1 1 0 1 0 0
1 0 0 1 0 1 1 0 0 0 0 0 0 1
1 0 1 0 1 1 1 0 0 0 0 1 0 0
0 0 0 1 0 1 0 0 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0 0 1 0 1
1 0 1 1 0 0 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 1 1 0 1 1 0
0 1 0 0 0 0 0 1 0 1 0 0 1 0
0 1 0 0 0 0 0 1 1 0 0 0 1 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0
0 1 0 1 0 1 0 1 0 0 1 0 1 0
1 0 0 0 0 0 0 1 1 1 0 1 0 0
0 0 1 0 0 1 0 0 0 0 0 0 0 0

Some BLOCKS preliminary output

Data file is named
<C:\stocnet\temp\~Doreian_mat.txt>.
Preliminary inspection of first line of the data file:
Data file contains 14 vertices.

The following options were chosen:
Number of colors to be analyzed ranges from 2 to 4.
3 Gibbs sequences for each color.

Set of symmetric relations is
1 (0, 0) (63 dyads)
2 (1, 1) (28 dyads)
No asymmetric relations.
Total number of relations is 2.

Number of dyads with observed values is 91.
Number of dyads with missing values is 0.

Entropy of empirical dyad distribution = 0.61724

How many classes?

No of classes	*information* I_y	H_x
2	.46	.19
3	.43	.45
4	.41	.45
5	.40	.43

Values averaged over 3 runs (all within .01)

I_y is the ***information*** in the observed relations, conditional on the class structure and the inter- and intra-class probabilities (smaller values mean that observed relations are more determined by model parameters)

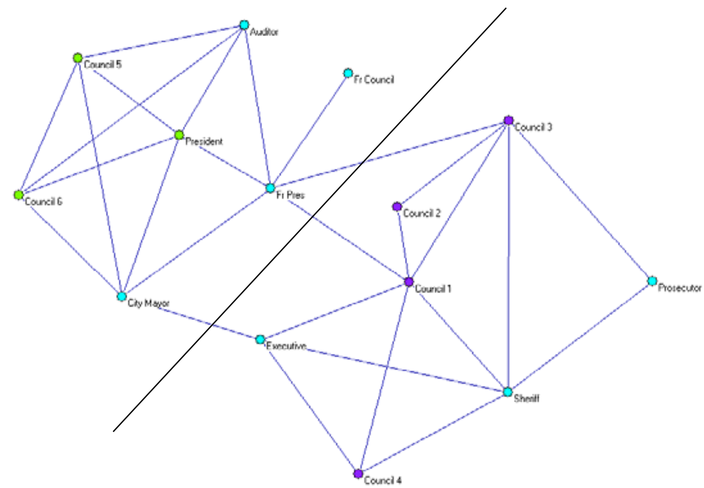
H_x measures the extent to which the class distribution defines a ***clear-cut partition*** of actors into classes (smaller values signify clearer partitions, max value of 1 corresponds to the case where all classes are equally likely for any actor)

Two-block solution: Permutation of data

```
A new ranking of the vertices was determined to bring out the block structure.
New ranking is
 1  2  3  4  5  6  7  8  9 10 11 12 13 14
 1  3  4  5  6  7 14  2  8  9 10 11 12 13

Recoded observed adjacency matrix in new ranking is as follows:
                 1       1 1 1 1
     1 3 4 5 6 7 4 2 8 9 0 1 2 3
 1     2 2 1 1 2 1 1 1 1 1 1 1 2
 3   2   2 1 2 2 2 1 1 1 1 1 1 1
 4   2 2   2 2 2 1 1 1 1 1 1 2 1
 5   1 1 2   2 1 1 1 1 1 1 1 1 1
 6   1 2 2 2   1 2 1 1 1 1 1 2 1
 7   2 2 2 1 1   1 1 1 1 1 1 1 1
14   1 2 1 1 2 1   1 1 1 1 1 1 1
 2   1 1 1 1 1 1 1   2 2 2 1 2 1
 8   1 1 1 1 1 1 1 2   2 2 1 2 2
 9   1 1 1 1 1 1 1 2 2   2 1 1 2
10   1 1 1 1 1 1 1 2 2 2   1 1 2
11   1 1 1 1 1 1 1 1 1 1 1   2 1
12   1 1 2 1 2 1 1 2 2 1 1 2   2
13   2 1 1 1 1 1 1 1 2 2 2 1 2
```

Doreian (1988) political network colour codes vote (aqua = non-voting)



Posterior probabilities that two actors have the same colour

Matrix of pairwise color equality

=================================

	1	3	4	5	6	7	14	2	8	9	10	11	12	13	
1		9	9	9	9	9	9	0	0	0	0	4	1	0	
3	9		9	9	9	9	9	0	0	0	0	4	1	0	
4	9	9		9	9	9	9	0	0	0	0	4	1	0	
5	9	9	9		9	9	9	0	0	0	0	4	1	0	
6	9	9	9	9		9	9	0	0	0	0	4	1	0	
7	9	9	9	9	9		9	0	0	0	0	4	1	0	
14	9	9	9	9	9	9		0	0	0	0	4	1	0	
2	0	0	0	0	0	0	0		9	9	9	6	9	9	
8	0	0	0	0	0	0	0	9		9	9	6	9	9	
9	0	0	0	0	0	0	0	9	9		9	6	9	9	
10	0	0	0	0	0	0	0	9	9	9		6	9	9	
11	4	4	4	4	4	4	4	6	6	6	6		7	6	**Former council member**
12	1	1	1	1	1	1	1	9	9	9	9	7		9	Former President
13	0	0	0	0	0	0	0	9	9	9	9	6	9		

Posterior probabilities of dyadic relations

Probabilities of relation 1 = (0, 0) between vertices

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

Probabilities of relation 2 = (1, 1) between vertices

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

Vertex-by-vertex analysis of fit

Overall maximum pairwise value for vertices in different classes is 0.360
Overall minimum pairwise value for vertices in the same class is 0.645

Classes now are labeled 1 to 2.

Vertices, as far as not thrown out, with their class number, their maximum pairwise value for vertices in a different class, and their minimum pairwise value for vertices in the same class;

and, if this vertex were to be deleted: the overall maximum pairwise value for vertices in different classes, and the overall minimum pairwise value for vertices in the same class:

1	1	0.357	0.989	0.360	0.645
2	2	0.010	0.646	0.360	0.645
3	1	0.354	0.991	0.360	0.645
4	1	0.355	0.991	0.360	0.645
5	1	0.360	0.985	0.357	0.645
6	1	0.351	0.991	0.360	0.645
7	1	0.353	0.990	0.360	0.645
8	2	0.009	0.646	0.360	0.645
9	2	0.009	0.645	0.360	0.646
10	2	0.009	0.646	0.360	0.645
11	2	0.360	0.645	0.113	0.893
12	2	0.113	0.747	0.360	0.645
13	2	0.010	0.646	0.360	0.645
14	1	0.351	0.985	0.360	0.645

BLOCKS appraisal of vertex-by-vertex analysis of fit

In @6:

Vertex for which throwing it out improves the fit most is 11;
this vertex now is thrown out.

…

Permuted data array, showing *post hoc* block structure

Recoded adjacency matrix with block structure

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

Posterior probabilities and frequencies for *post hoc* coloring

Posterior probabilities

Average posterior probabilities of relation 1 = (0, 0) in these blocks.

	1	2
1	0.48	0.89
2	0.89	0.42

Average posterior probabilities of relation 2 = (1, 1) in these blocks.

	1	2
1	0.52	0.11
2	0.11	0.58

Observed frequencies

Observed relative frequencies of relation 1 = (0, 0) in these blocks.

	1	2
1	0.43	0.94
2	0.94	0.38

Observed relative frequencies of relation 2 = (1, 1) in these blocks.

	1	2
1	0.57	0.06
2	0.06	0.62

Stochastic 3-block model

Matrix of pairwise color equality

```
=================================

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

Recoded adjacency matrix with block structure :

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

Example 2: Work frequency and work importance in an organisation

Crosstabulation of frequency and importance categories:

work frequency * work importance Crosstabulation

Count

		work importance					
			Critical	Important	None	Routine	Total
work frequency		2					2
	Daily		53	49		37	139
	Hourly		38	5		6	49
	Monthly		32	67	1	79	179
	None				100		100
	Weekly		48	104		58	210
Total		2	171	225	101	180	679

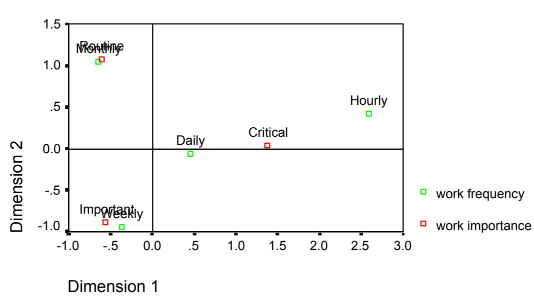
Correspondence analysis for frequency and importance of non-null work ties

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	.361	.130			.862	.862	.041	.034
2	.144	.021			.138	1.000	.043	
Total		.151	86.936	.000[a]	1.000	1.000		

a. 6 degree

Category Quantifications



A stochastic blockmodel for work ties

Code work ties as:

0 = none

1 = work interactions (at least monthly, at least routinely)

2 = critically important work interactions, occurring at least daily

Alphabet for dyadic relations:

1 = (0,0)	2 = (1,1)	3 = (0,1)	4 = (0,2)
5 = (1,2)	6 = (1,0)	7 = (2,0)	8 = (2,1)

note that (2,2) did not occur!

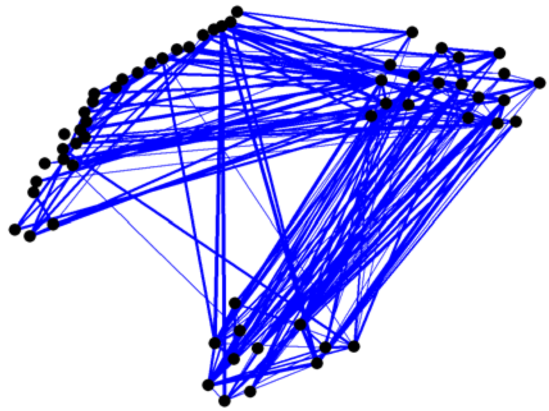
How many classes?

No of classes	*information* I_y	H_x
2	.4940	.2463
3	.4727	.4219
4	.4694	.4349
5	.4639	.4247

I_y is the information in the observed relations, conditional on the class structure and the inter- and intra-class probabilities (smaller values mean that observed relations are more determined by model parameters)

H_x measures the extent to which the class distribution defines a clear-cut partition of actors into classes (smaller values signify clearer partitions, max value of 1 corresponds to the case where all classes are equally likely for any actor)

Organisational networks



Frequency of work interactions

Relationship between 2- and 3-class solutions

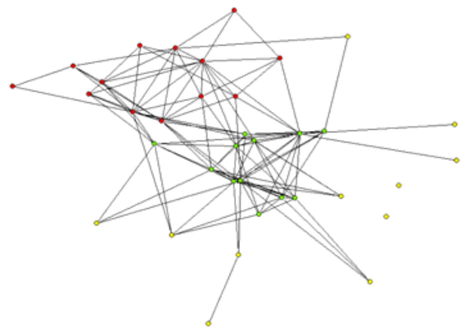
		2-class solution class 1	*2-class solution* class 2
3-class solution	class 1	1,3,4,8,11,12,16,17 19,21,26,31,34,38,39 40,43,46,47,48,54,55 55,57,58,59,60	10,24,52
	class 2		2,5,7,9,14,15,18 20,22,23,27,29 30,36,37,41,49, 50,53,56
	class 3	6,13,25,28,32,33,42,44	35,45,51

Example 3: collaboration among partners in a law firm: average posterior probabilities for blocks based on *post hoc* colouring (Lazega)

	Block 1	Block 2	Block 3
1	.05	.16	.04
2	.16	.56	.12
3	.04	.12	.42

Key:

yellow block 1
green block 2
red block 3



data with post hoc colouring

Other latent variable formulations (for positive "association" relational forms)

Latent space model (Handcock, Hoff & Raftery, 2002)

Nodes have (latent) locations in a *k*-dimensional Euclidean space
Tie probabilities are conditionally independent, given latent locations
Bayesian estimation of unobserved locations (and their uncertainty)

Latent ultramteric model (Schweinberger & Snijders, 2003)

Every pair of nodes is associated with a (latent) distance in an ultrametric space (corresponding to a hierarchy of "settings")
Tie probabilities are conditionally independent, given latent ultrametric distances
Bayesian estimation for unobserved ultrametric distances (ande hence setting hierarchy)

References

Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*, 1077-1087.

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

Schweinberger, M., & Snijders, T. A. B. (2003). Settings in social networks: a measurement model. Sociological Methodology, 33, 307-341.