Inferring structure in bipartite networks using the latent blockmodel and exact ICL

Jason Wyse*, Nial Friel† and Pierre Latouche‡

* School of Computer Science and Statistics, Trinity College Dublin, Ireland.

† School of Mathematical Sciences and Insight: The National Centre for Big Data Analytics, University College Dublin, Ireland

‡ Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne, 90 rue de Tolbiac, F-75634 Paris Cedex 13, France

Abstract

We consider the task of simultaneous clustering of the two node sets involved in a bipartite network. The approach we adopt is based on use of the exact integrated complete likelihood for the latent blockmodel. Using this allows one to infer the number of clusters as well as cluster memberships using a greedy search. This gives a model-based clustering of the node sets. Experiments on simulated bipartite network data show that the greedy search approach is vastly more scalable than competing Markov chain Monte Carlo based methods. Application to a number of real observed bipartite networks demonstrate the algorithms discussed.

1 Introduction

Bipartite networks are those containing two types of nodes, say types A and B. Nodes of type A may be linked to nodes of type B, and *vice versa*, but links between two nodes of the same type are not considered. There are many real life networks that can be naturally viewed in this way. Take for example relational networks where a user rates a movie. A user is a member of node type A and node type B represents the movies. One may ask a number of quantitative questions in such a situation. Can users be grouped by the types of movies they watch and rate? How many substantive genres of movies are defined by users?

Bipartite or two-mode networks have seen much attention in the social networks and machine learning literature, see for example Borgatti & Everett (1997), Doreian et al. (2004), Brusco & Steinley (2011), Brusco et al. (2013), Doreian et al. (2013). Rohe et al. (2015) discuss how their stochastic co-blockmodel may be extended to a bipartite setting. The reason for this high level of interest is due to their wide applicability and the fact that it is often very natural and fruitful to model interactions between node sets in this way. Clustering or partitioning the node sets simultaneously can reveal structure and give considerable insight into the entities in the network. Such insights may be allusive to the more classical network measures (Wasserman & Faust 1994), some of which have been adapted from the classical

literature to the bipartite or two mode situation (for example, the clustering coefficients of Opsahl (2013)). For this reason much attention is focused on clustering or grouping the node sets in tandem; this practice is referred to using many terms in the literature: bi-clustering, co-clustering, block-clustering, two-mode blockmodelling and others. One of the pioneering papers of this area was that of Hartigan (1972). Flynn & Perry (n.d.) examine asymptotic theoretical guarantees for bi-clustering.

Approaches to clustering or blockmodelling of two-mode networks fall into two classes. The first of these is often called *deterministic*, whereby the clustering of the network (or adjacency matrix) is obtained by minimizing an objective function which measures discrepancy from an ideal block structure. Examples of this are the work of Doreian et al. (2004), Brusco & Steinley (2006), Brusco & Steinley (2011), Brusco et al. (2013) and Doreian et al. (2013). The second type of approach is *stochastic*. In stochastic blockmodelling procedures, one assumes that the probability of links between the node sets in the network can be modelled by a parameterized distribution. These parameters are usually estimated (learned) and then used as a representative embodiment of the true network linking behaviour. The stochastic approach may also be referred to as model-based; that is, a statistical model is used for links in the network. Examples include the work of Govaert (1995), Govaert & Nadif (1996), Govaert & Nadif (2003), Govaert & Nadif (2005), Govaert & Nadif (2008), Rohe et al. (2015), Wyse & Friel (2012) and Keribin et al. (2013).

This paper is concerned with the stochastic approach to blockmodelling of two-mode networks. In particular, we take the latent blockmodel (LBM) developed in a series of papers by Gérard Govaert and Mohamed Nadif; Govaert (1995), Govaert & Nadif (1996), Govaert & Nadif (2003), Govaert & Nadif (2005), Govaert & Nadif (2008). We consider this model as applied in the context of bipartite networks. This is a desirable model, as it provides a model-based clustering of both node sets and has richer modelling capability than only absence/presence data for ties between nodes should this information be observed. The LBM is based around an intuitive generative structure as outlined in Section 2.2. We point out that the LBM is a different model than the stochastic blockmodel (SBM), as in, for example, Nowicki & Snijders (2001). The LBM operates on items and objects as opposed to the SBM which focusses on modelling interactions between items.

In the LBM, the posterior distribution over the latent label vectors, given the model parameters and observed data, cannot be factorized due to conditional dependency. Therefore standard optimization techniques such as the expectation maximization (EM) algorithm cannot be used directly for clustering. To tackle this issue, approximation methods like variational EM (Govaert & Nadif 2008) or stochastic EM (Keribin et al. 2010) have been proposed. Moreover, in practice, the numbers of clusters in each node set have to be estimated. This has led to treatements such as the one by Wyse & Friel (2012) who use Markov chain Monte Carlo (MCMC) to do inference for the number of clusters and the members of the nodes to the groups. We also refer to the recent work of Keribin et al. (2012) and Keribin et al. (2013) who relied on model selection criteria to estimate the number of clusters.

Unlike Govaert & Nadif (2008), Keribin et al. (2012), Keribin et al. (2013), our approach allows the number of clusters in both node sets to be estimated while simultaneously partitioning the nodes. This is based on a clustering criterion termed the exact integrated complete likelihood (ICL), and a method to search over partitions of the nodes. The main ideas of using the exact ICL come from Côme & Latouche (2013) who use this in estimation

of the SBM of Nowicki & Snijders (2001). Our work can be seen as somewhat complementary to the work in Wyse & Friel (2012), but there are many advantages to using the framework presented here, similar to the SBM in Côme & Latouche (2013). Firstly, it is more scalable than the MCMC approach and secondly we do not have to worry about the mixing rates of the MCMC algorithm in larger settings. One drawback of our new approach is that it does not provide a joint posterior distribution for the number of clusters in both node sets.

The remainder of the paper is organised as follows. Section 2 introduces ideas of blockmodels for bipartite networks (Section 2.1) and the LBM of Govaert & Nadif (2008) (Section 2.2) and discusses its relevance for this context. Section 2.3 reviews existing estimation techniques for the LBM and discusses their advantages and limitations. In Section 3 we introduce the exact ICL along with a greedy algorithm for inference purposes. We also draw strong parallels with the approach of Wyse & Friel (2012) in this section, and highlight the sensible and pragmatic nature of the algorithm. The computational complexity of the algorithm is discussed in Section 3.4. In Section 3.6 we show how large computational savings can be made over a naive implementation of this algorithm. Section 4 compares the proposed approach with the competing MCMC algorithm through a simulation study. Section 5 applies the approach to a number of datasets. We conclude with a discussion.

2 Bipartite networks and the latent blockmodel

2.1 Bipartite networks

Bipartite networks consist of possible ties between members of two different node sets. Let the node sets be A and B. There are N nodes in A, $A = \{a_1, \ldots, a_N\}$ and M in B, $B = \{b_1, \ldots, b_M\}$. We use the terms node type and set interchangeably. Node i from A may have a tie with node j from B, this tie, if it exists, is undirected. Alternatively, a tie may not exist. Either way, this information on the *linking attribute* between the two nodes is contained in the observation y_{ij}. We record all the observed information about the network in an $N \times M$ adjacency matrix

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1M} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NM} \end{pmatrix}$$

having a row for each node from A and a column for each in B. The problem of primary focus for us is the following. Can we group or partition the A nodes and the B nodes simultaneously to reveal subgroups or subsets of the A nodes that have linking attributes of similar nature to subgroups of the B nodes? We use the term "similar nature" here to highlight that we approach this problem by modelling properties of the linking attribute which could be binary in nature, categorical (> 2 categories), a count or an observation with a continuous value.

Suppose a grouping of the node sets into subsets like that mentioned above does exist.

The nodes in A are partitioned into K and those in B into G disjoint subsets

$$A = \bigcup_{k=1}^{K} A_k, \quad A_k \cap A_l = \emptyset, \forall k \neq l \quad \text{and} \quad B = \bigcup_{g=1}^{G} B_g, \quad B_g \cap B_h = \emptyset, \forall g \neq h.$$

Then the nodes in set A_k will be seen as having similar natured linking attributes to their respective nodes in set B_g. We suppose now that we model linking attributes using some parametric distribution $p(y_{ij}|\boldsymbol{\theta})$ for the y_{ij}. In order to capture differences in the nature of linking attributes between the different subsets, we allow this distribution to have a different valued parameter $\boldsymbol{\theta}_{kg}$ for each subset pairing A_k and B_g, so that if node i is in A_k and node j is in B_g, the density y_{ij} is $p(y_{ij}|\boldsymbol{\theta}_{kg})$.

2.2 Latent blockmodel

The LBM was introduced in Govaert & Nadif (2008) as a means to provide a concise summary of a large data matrix. In order to index the different possible partitionings of A and B into K and G disjoint subsets (clusters) we introduce respective label vectors $\mathbf{c}$ and $\mathbf{w}$. These are such that $c_i = k$ if node $i \in A_k$ and similarly $w_j = g$ if $j \in B_g$. Using the model for linking attributes, the probability of observing the adjacency Y can be written down, if we assume we know the partitioning of the nodes, i.e. that know $\mathbf{c}$ and $\mathbf{w}$

$$p(Y|\mathbf{c}, \mathbf{w}, \Theta, K, G) = \prod_{k=1}^{K} \prod_{g=1}^{G} \prod_{i:c_i=k} \prod_{j:w_j=g} p(y_{ij}|\boldsymbol{\theta}_{kg}).$$

Here Θ denotes the collection of $\boldsymbol{\theta}_{kg}$.

Of course, in practice $\mathbf{c}$ and $\mathbf{w}$ will not be known. Attach to each clustering of nodes $\mathbf{c}, \mathbf{w}$ a probability $p(\mathbf{c}, \mathbf{w}|\Phi, K, G)$, depending on some hyperparameters Φ. This allows the density of the adjacency matrix Y to be written down as a mixture model. Let $\mathcal{C}_K$ and $\mathcal{W}_G$ be collections containing all possible clusterings of A into K groups and B into G groups respectively. Then

$$p(Y|\Theta, \Phi, K, G) = \sum_{(\mathbf{c},\mathbf{w}) \in \mathcal{C}_K \times \mathcal{W}_G} p(\mathbf{c}, \mathbf{w}|\Phi, K, G) p(Y|\mathbf{c}, \mathbf{w}, \Theta, K, G). \tag{1}$$

Here the mixture is over all clusterings of the nodes into K and G groups. The $p(\mathbf{c}, \mathbf{w}|\Phi, K, G)$ terms represent probabilities of particular clusterings being generated parameterized by Φ.

Suppose that, *a priori*, no information exists on the joint clustering of nodes in sets A and B. Then it is reasonable to assume that

$$p(\mathbf{c}, \mathbf{w}|\Phi, K, G) = p(\mathbf{c}|\Phi, K) p(\mathbf{w}|\Phi, G).$$

Having made this assumption, LBM assumes further that there are weights associated with each subset A_k and B_g, such that $\Pr\{c_i = k|K\} = \omega_k$ and $\Pr\{w_j = g|G\} = \rho_g$. This defines a multinomial distribution for the node labels, where the weights sum to unity: $\sum_{k=1}^{K} \omega_k = 1$, $\sum_{g=1}^{G} \rho_g = 1$. The parameters Φ represent the weight vectors $\boldsymbol{\omega}, \boldsymbol{\rho}$. Thus we can write

$$p(\mathbf{c}, \mathbf{w}|\Phi, K, G) = \left(\prod_{k=1}^{K} \omega_k^{N_k} \right) \left(\prod_{g=1}^{G} \rho_g^{M_g} \right)$$

where $N_k = |A_k|$ and $M_g = |B_g|$.

Effectively the LBM defines a probability distribution over clustering of the node sets A and B into K and G disjoint subsets. Examining equation (1) one sees that the sum is over $K^N G^M$ terms, which is clearly intractable for even moderate N and M. In practice, we are interested in finding optimal or near optimal clusterings, corresponding to large values of $p(\mathbf{c}, \mathbf{w}|\Phi, K, G)p(Y|\mathbf{c}, \mathbf{w}, \Theta, K, G)$ in (1) which contribute the most to the sum. It is thus usual to rephrase the problem into one which optimises $p(\mathbf{c}, \mathbf{w}|\Phi, K, G)p(Y|\mathbf{c}, \mathbf{w}, \Theta, K, G)$ jointly over the clusterings $(\mathbf{c}, \mathbf{w})$ and parameters Φ, Θ in some way. Something to note here is that the particular cluster configurations are conditional on the value of K and G being known, something which will rarely be the case in real applications.

2.3 Estimation of the LBM

Estimation of the LBM can be performed either through hybrid variational and stochastic EM type algorithms (Govaert & Nadif 2008, Keribin et al. 2010) or through Bayesian estimation (van Dijk et al. (2009), Wyse & Friel (2012)). We refer the reader to Section 2.1.1 of Wyse & Friel (2012) for a review of the former. Here we focus on Bayesian estimation as it is most relevant for what follows. The one remark to be made about EM type estimation of the LBM however is that it conditions on the values of K and G. As of yet, there is no widely accepted information criterion for choosing the best values of K and G as outlined in Section 2.1.2 of Wyse & Friel (2012). Keribin et al. (2012), Keribin et al. (2013) relied on a ICL criterion for model selection purposes, while van Dijk et al. (2009) proposed a Gibbs sampler for Bayesian estimation of the latent block model along with an AIC-3 (Bozdogan 1994) criterion.

Wyse & Friel (2012) circumvented the problem of choosing K and G by including these as unknowns in their Bayesian formulation of LBM. They begin by taking priors on the unknowns in the model, namely, Θ, Φ, K and G and writing down the full posterior given the data

$$\begin{aligned} \pi(K, G, \mathbf{c}, \mathbf{w}, \Phi, \Theta|Y) &\propto p(\mathbf{c}, \mathbf{w}|\Phi, K, G)p(Y|\mathbf{c}, \mathbf{w}, \Theta, K, G) \\ &\quad \times \pi(\Theta|K, G)\pi(\Phi|K, G)\pi(K)\pi(G) \\ &= p(\mathbf{c}, \mathbf{w}|\Phi, K, G)\pi(\Phi|K, G) \\ &\quad \times p(Y|\mathbf{c}, \mathbf{w}, \Theta, K, G)\pi(\Theta|K, G)\pi(K)\pi(G) \end{aligned} \tag{2}$$

where the last three distributions on the right hand side are priors for Θ, K, G. Note here the assumption that, *a priori*, the number of node clusters K and G of types A and B are independent. If other prior information exists beforehand, this prior assumption can be replaced by one which represents this information. Integrating both sides of the above proportionality relation with respect to Φ and Θ returns the joint marginal distribution of

$K, G, \mathbf{c}, \mathbf{w}$ given the observed network Y:

$$\begin{aligned}\pi(K, G, \mathbf{c}, \mathbf{w}|Y) &\propto \int p(\mathbf{c}, \mathbf{w}|\Phi, K, G)\pi(\Phi|K, G)\,\mathrm{d}\Phi \\ &\quad \times \int p(Y|\mathbf{c}, \mathbf{w}, \Theta, K, G)\pi(\Theta|K, G)\,\mathrm{d}\Theta \\ &\quad \times \pi(K)\,\pi(G) \\ &= \pi(\mathbf{c}, \mathbf{w}|K, G)\pi(Y|\mathbf{c}, \mathbf{w}, K, G)\pi(K)\pi(G).\end{aligned}$$

The key observation in Wyse & Friel (2012) is that the quantities

$$\begin{aligned}\pi(\mathbf{c}, \mathbf{w}|K, G) &= \int p(\mathbf{c}, \mathbf{w}|\Phi, K, G)\pi(\Phi|K, G)\,\mathrm{d}\Phi \\ \pi(Y|\mathbf{c}, \mathbf{w}, K, G) &= \int p(Y|\mathbf{c}, \mathbf{w}, \Theta, K, G)\pi(\Theta|K, G)\,\mathrm{d}\Theta\end{aligned}$$

can be obtained exactly (analytically) using relatively standard prior assumptions. For example $\pi(\mathbf{c}, \mathbf{w}|K, G)$ can be computed exactly by assuming independent Dirichlet priors on the weights vectors

$$\pi(\Phi|K, G) = \mathrm{Dir}_K(\boldsymbol{\omega}; \alpha_0, \ldots, \alpha_0) \times \mathrm{Dir}_G(\boldsymbol{\rho}; \beta_0, \ldots, \beta_0)$$

where Dir_p represents the density of the Dirichlet distribution on the $p-1$ dimensional simplex. In this case

$$\pi(\mathbf{c}, \mathbf{w}|K, G) = \frac{\Gamma\{\alpha_0 K\}}{\Gamma\{\alpha_0\}^K} \frac{\prod_{k=1}^K \Gamma\{N_k + \alpha_0\}}{\Gamma\{N + \alpha_0 K\}} \times \frac{\Gamma\{\beta_0 G\}}{\Gamma\{\beta_0\}^G} \frac{\prod_{g=1}^G \Gamma\{M_g + \beta_0\}}{\Gamma\{M + \beta_0 G\}}$$

where $\Gamma(\cdot)$ is the gamma function. See Wyse & Friel (2012) for further details. Furthermore $\pi(Y|\mathbf{c}, \mathbf{w}, K, G)$ can be computed exactly if we make the following assumptions. Assume that the prior for Θ can be expressed as independent priors for the $\boldsymbol{\theta}_{kg}$,

$$\pi(\Theta|K, G) = \prod_{k=1}^K \prod_{g=1}^G \pi(\boldsymbol{\theta}_{kg}).$$

We can compute $\pi(Y|\mathbf{c}, \mathbf{w}, K, G)$ exactly if we can compute

$$\Lambda_{kg} = \int \pi(\boldsymbol{\theta}_{kg}) \prod_{i:c_i=k} \prod_{j:w_j=g} p(y_{ij}|\boldsymbol{\theta}_{kg})\,\mathrm{d}\boldsymbol{\theta}_{kg}$$

exactly. This happens when the prior $\pi(\boldsymbol{\theta}_{kg})$ is fully conjugate to $p(y|\theta_{kg})$. There are many widely used and standard situations where this is the case as we discuss in Section 3.2. In our notation we suppress dependence on the values of the hyperparameters, say α_0, β_0 and the hyperparameters of $\pi(\boldsymbol{\theta}_{kg})$ and similar for brevity. This is to be understood.

With these quantities available exactly, the joint marginal posterior of clusterings and number of clusters of the nodes is given by

$$\pi(K, G, \mathbf{c}, \mathbf{w}|Y) \propto \pi(\mathbf{c}, \mathbf{w}|K, G) \left(\prod_{k=1}^K \prod_{g=1}^G \Lambda_{kg} \right) \pi(K)\pi(G). \tag{3}$$

Notice that this posterior is defined over a large discrete model space. If we allow a maximum of $K_{\max}$ and $G_{\max}$ subsets for nodes in A and B, then the size of the support of the posterior is $\sum_{k=1}^{K_{\max}} \sum_{g=1}^{G_{\max}} k^N g^M$.

Wyse & Friel (2012) developed an MCMC algorithm that generates samples from (3). The algorithm moves are involved and we refer to Wyse & Friel (2012) for exact details. Node labels are sampled using a Gibbs step and there are more sophisticated moves for adding and removing clusters (that is, changing K and G). The main idea is that iteratively the chain will sample high probability configurations $(K, G, \mathbf{c}, \mathbf{w})$, so that in the sampler output one gets a joint posterior distribution for K and G with corresponding label vectors. There are upper bounds $K_{\max}$ and $G_{\max}$ placed on the number of groups the nodes may be partitioned into. Usually this will be conservatively large. A drawback of this algorithm in general is that mixing of the Markov chain may disimprove with increasing M and/or N, resulting in infrequent jumps to models with different K and G from the current state. So for high dimensional problems, one can observe most of the empirical (approximate) posterior mass centred on one or two (K, G) combinations. However, the work of Wyse & Friel (2012) is a step forward in estimation of LBMs as it is an automatic way to perform inference for K and G also, while clustering the nodes, for which no other such approach existed before.

3 Exact ICL and greedy ICL algorithm for bipartite networks

3.1 ICL and exact ICL

The integrated completed likelihood (ICL) was first introduced in Biernacki et al. (2000), as a model selection criterion, in the context of Gaussian mixture models. The rationale for using ICL is that in the finite mixture model context (similar to our situation here), analysis is often carried out using a latent label vector which it is difficult to integrate from the model. This latent vector is often termed the allocation vector (Nobile & Fearnside 2007) and it provides a clustering of the data points to component densities. For this reason Biernacki et al. (2000) argue that the evidence for a particular clustering should be taken into account when determining the number of mixture components and thus they focus on the integrated completed data likelihood. So instead of marginalising these labels from the model, ICL includes them as part of the information criterion. The number of components in the mixture (in our case values of K, G) which gives the largest ICL is the most supported by the data. For full details see Biernacki et al. (2000). Keribin et al. (2013) have used ICL in the context of the latent blockmodel to choose the number of clusters, however, it is not the same approach as we adopt here as they do not use the available form of the exact ICL. Instead, their analysis employs a penalty term in the ICL criterion, which we avoid by computing ICL analytically. Using the notation introduced earlier, the ICL can be written as

$$\mathcal{ICL} = \log \left\{ \int \int p(\mathbf{c}, \mathbf{w}|\Phi, K, G) p(Y|\mathbf{c}, \mathbf{w}, K, G) \pi(\Theta|K, G) \pi(\Phi|K, G) \, \mathrm{d}\Theta \, \mathrm{d}\Phi \right\}.$$

Typically this quantity cannot be computed exactly and so ICL is usually approximated using a high probability configuration and a penalty term. In Section 2.3 we encountered a

similar expression to $\mathcal{ICL}$ in the joint collapsed posterior of Wyse & Friel (2012), and we see that we can write

$$\begin{aligned}
\mathcal{ICL} &= \log\left\{\int\int p(\mathbf{c},\mathbf{w}|\Phi,K,G)p(Y|\mathbf{c},\mathbf{w},K,G)\pi(\Theta|K,G)\pi(\Phi|K,G)\,\mathrm{d}\Theta\,\mathrm{d}\Phi\right\} \\
&= \log\left\{\left(\int p(\mathbf{c},\mathbf{w}|\Phi,K,G)\pi(\Phi|K,G)\mathrm{d}\Phi\right)\times\left(\int p(Y|\mathbf{c},\mathbf{w},K,G)\pi(\Theta|K,G)\mathrm{d}\Theta\right)\right\} \\
&= \log\left\{\int p(\mathbf{c},\mathbf{w}|\Phi,K,G)\pi(\Phi|K,G)\mathrm{d}\Phi\right\}+\log\left\{\int p(Y|\mathbf{c},\mathbf{w},K,G)\pi(\Theta|K,G)\mathrm{d}\Theta\right\} \\
&= \log\pi(\mathbf{c},\mathbf{w}|K,G)+\log\pi(Y|\mathbf{c},\mathbf{w},K,G).
\end{aligned}$$

The conditions for the ICL to be available exactly are the same as those given in Section 2.3. If this is the case we term this quantity *exact* ICL. Now suppose we can compute the exact ICL so that

$$\mathcal{ICL}=\log\pi(\mathbf{c},\mathbf{w}|K,G)+\sum_{k=1}^{K}\sum_{g=1}^{G}\log\Lambda_{kg}.$$

Note the similarity between this expression and the log of the right hand side of (3). In fact, the only difference is that in (3) we also have two log prior terms for K and G. Thus, finding the highest ICL is almost identical to finding regions of the support of (3) with high posterior mass (bar the prior terms). This is elucidated more in Section 3.5.

3.2 Linking attribute models

We now turn attention to the types of linking attribute models for which exact ICL can be computed. The requirement for the exact ICL to be available is that

$$\Lambda_{kg}=\int\pi(\boldsymbol{\theta}_{kg})\prod_{i:c_i=k}\prod_{j:w_j=g}p(y_{ij}|\boldsymbol{\theta}_{kg})\,\mathrm{d}\boldsymbol{\theta}_{kg}$$

can be computed analytically. Λ_{kg} can be referred to as the marginal likelihood for block (k,g), as we pick up nodes from cluster k of A and cluster g of set B and average their likelihood over the prior for the parameter of the linking attributed between the two node subsets. We now give some examples of very commonly used models where the marginal likelihood can be computed exactly.

3.2.1 Absent/present linking attribute

The most commonly observed networks will be those where edges are represented by binary indicators when there is a link between a node in set A and a node in set B or not. For example, students being members of university clubs or not. For $i\in A_k$ and $j\in B_g$ the most natural way to model this is using a Bernoulli random variable with probability of a tie equal to θ_{kg}

$$p(y_{ij}|\theta_{kg})=\theta_{kg}^{y_{ij}}(1-\theta_{kg})^{1-y_{ij}}.$$

A conjugate prior for this data distribution is a beta prior

$$\theta_{kg}\sim\mathrm{Beta}(\eta,\eta)$$

with hyperparameter κ. Then it can be easily shown that

$$\Lambda_{kg} = \frac{\Gamma\{2\eta\}}{\Gamma\{\eta\}\Gamma\{\eta\}} \frac{\Gamma\{N_{kg}^1 + \eta\}\Gamma\{N_k M_g - N_{kg}^1 + \eta\}}{\Gamma\{N_k M_g + 2\eta\}}$$

where $N_{kg}^1 = \sum_{i:c_i=k} \sum_{j:w_j=g} \mathrm{I}(y_{ij} = 1)$.

3.2.2 Multinomial links with Dirichlet prior

This is a generalization of the previous model for more than two categories. Linking attributes which are categorical arising from C categories could naturally be modelled using the multinomial distribution. This could be useful say if the college years of the students in the university clubs above were also recorded e.g. junior freshman, senior freshman, junior sophister, senior sophister. The distribution is

$$p(y_{ij}|\boldsymbol{\theta}_{kg}) = \prod_{l=1}^{C} \left[\theta_{kg}^l\right]^{\mathrm{I}(y_{ij}=l)}$$

where θ_{kg}^l is the probability y_{ij} takes category l. Taking a symmetric Dirichlet prior with parameter ζ the block integrated likelihood can be computed exactly as

$$\Lambda_{kg} = \frac{\Gamma(\zeta C)}{\Gamma(\zeta)^C} \frac{\prod_{l=1}^{C} \Gamma(N_{kg}^l + \zeta)}{\Gamma(N_k M_g + \zeta C)}$$

where $N_{kg}^l = \sum\limits_{i:c_i=k} \sum\limits_{j:w_j=g} \mathrm{I}(y_{ij} = l)$ is the number taking on category l in block (k, g).

3.2.3 Poisson model for count links with Gamma prior

Count data may be modelled using a Poisson distribution with rate θ_{kg}. This could be the number of emails exchanged between students in the university clubs. Taking a Gamma(δ, γ) prior on θ_{kg} leads to

$$\Lambda_{kg} = \frac{\gamma^\delta}{\Gamma(\delta)} \frac{\Gamma(S_{kg} + \delta)}{(N_k M_g + \gamma)^{S_{kg}+\delta}} \left(\prod_{i:c_i=k} \prod_{j:w_j=g} y_{ij}! \right)^{-1}$$

where $S_{kg} = \sum\limits_{i:c_i=k} \sum\limits_{j:w_j=g} y_{ij}$ is the sum of the counts in a block.

3.2.4 Gaussian model with Gaussian-Gamma prior

For continuous data, if appropriate one can assume a Gaussian distribution with mean μ_{kg} and precision τ_{kg} for observations in a block. This may be the number of hours spent by the university students on club activities. For example, one would expect a sports club to require greater time commitment than, say, a car appreciation club. Taking a N$(\xi, \{\kappa\tau_{kg}\}^{-1})$ prior for μ_{kg} conditional on τ_{kg} and a Gamma$(\gamma/2, \delta/2)$ prior on τ_{kg}

$$\Lambda_{kg} = \frac{\pi^{-N_k M_g/2} \kappa^{1/2} \delta^{\gamma/2}}{(N_k M_g + \kappa)^{1/2}} \frac{\Gamma(\{N_k M_g + \gamma\}/2)}{\Gamma(\gamma/2)} \left(SS_{kg} + \kappa\xi^2 - \frac{(S_{kg} + \kappa\xi)^2}{N_k M_g + \kappa} + \delta \right)^{-(N_k M_g + \gamma)/2}.$$

Here $S_{kg} = \sum_{i:c_i=k} \sum_{j:w_j=g} y_{ij}$ and $SS_{kg} = \sum_{i:c_i=k} \sum_{j:w_j=g} y_{ij}^2$.

3.3 A greedy search strategy

Based on the observations about the ICL in Section 3.1 it makes sense to try to find some strategy to optimise the exact ICL. There are in essence four unknowns in our problem $K, G, \mathbf{c}, \mathbf{w}$. From Section 3.1 the largest value of the exact ICL (assuming it can be computed), gives the clustering of the nodes into subsets which is most preferable. The idea used by Côme & Latouche (2013) for the stochastic blockmodel was to optimise the ICL criterion using a greedy search over labels and the number of node clusters. We adopt a similar approach here. First define the ICL using any instance of numbers of clusters and labels

$$\text{ICL}(K, G, \mathbf{c}, \mathbf{w}) = \log \pi(\mathbf{c}, \mathbf{w}|K, G) + \sum_{k=1}^{K} \sum_{g=1}^{G} \log \Lambda_{kg}.$$

This frames the exact ICL as an objective function in parameters $(K, G, \mathbf{c}, \mathbf{w})$ which we wish to optimise. Of course, the values of K and G place constraints on the possible values of $\mathbf{c}$ and $\mathbf{w}$. The exact ICL is optimised iteratively by cycling through the following smaller optimisation techniques repeatedly until no further increases in it can be obtained.

Update node labels in set A

To update the node labels for set A, we firstly shuffle the order of nodes in the update. If node i is currently in $A_{k'}$ (cluster k'), that is $c_i = k'$, compute the change in exact ICL when moving node i to A_l for all $l \neq k'$. First of all, suppose that currently $|A_{k'}| > 1$, then the change in exact ICL is given by

$$\Delta_{k' \to l} = \text{ICL}(K, G, \mathbf{c}^*, \mathbf{w}) - \text{ICL}(K, G, \mathbf{c}, \mathbf{w})$$

where $\mathbf{c}^*$ is such that $c_r^* = c_r, r \neq i$ and $c_i^* = l$. Clearly $\Delta_{k' \to k'} = 0$. If all of the $\Delta_{k' \to l} < 0$, do not move node i from subset k'. Otherwise, move i to the node cluster A_{l^*} where $l^* = \arg\max_l \Delta_{k' \to l}$. This process is illustrated in a flow diagram in Figure 1 which summarizes the primary parts of the update.

If $|A_{k'}| = 1$, then our algorithm assumes that moving node i from $A_{k'}$ causes that subset to vanish, so that K, the number of subsets is reduced by 1. This is similar to component absorption in the case of algorithms for finite mixture models. In this case, the computation of the change in exact ICL must be modified accordingly to

$$\Delta_{k' \to l} = \text{ICL}(K - 1, G, \mathbf{c}^*, \mathbf{w}) - \text{ICL}(K, G, \mathbf{c}, \mathbf{w}).$$

This updating process is repeated for each node i in A.

Update node labels in set B

The node labels in set B are updated in an analogous way to those in set A, so we omit a description for brevity.

Algorithm initialization and termination

The algorithm is initialized randomly, with a random assignment of nodes to the subsets $A_k, B_g, k = 1\ldots, K, g = 1, \ldots, G$. Initially we assume a large number of clusters of nodes. To draw analogy with the MCMC sampler described in Section 2.3 we set $K = K_{\max}$ and $G = G_{\max}$ at the beginning. Nodes in sets A and B are then processed, and the algorithm terminates when no improvement (or further increase) in the exact ICL can be obtained. Clearly, the algorithm will cause convergence of the ICL to a local maximum, which may not necessarily be the global maximum. Thus, the greedy algorithm is run a number of times (this could be done in parallel), and the run given the highest exact ICL value used. As suggested by one reviewer, in practice, one could use another algorithm (for example the algorithm of Rohe et al. (2015) extended to the bipartite setting) to provide an initial clustering before invoking the greedy search algorithm. This may lead to faster convergence of the algorithm and quicker run times.

Using a merging move at termination

At termination of the algorithm, we can attempt to merge clusters if this increases the exact ICL further. Merging the clusters is more of a "mass node" move than the one node updates used in the greedy search. To merge two A node clusters, we compute $\text{ICL}(K-1, G, \tilde{\mathbf{c}}, \mathbf{w})$ where if attempting to merge clusters k and k', $\tilde{\mathbf{c}}$ is such that all nodes with label k' in $\mathbf{c}$ become label k nodes. This necessitates G new block marginal likelihood calculations to get

$$\text{ICL}(K-1, G, \tilde{\mathbf{c}}, \mathbf{w}) - \text{ICL}(K, G, \mathbf{c}, \mathbf{w}) = \log\left\{\frac{\pi(\tilde{\mathbf{c}}, \mathbf{w}|K-1, G)}{\pi(\mathbf{c}, \mathbf{w}|K, G)}\right\} + \sum_{g=1}^{G}\left(\log\tilde{\Lambda}_{kg} - \log(\Lambda_{kg}\Lambda_{k'g})\right)$$

where $\tilde{\Lambda}_{kg}$ is obtained by merging the sufficient statistics for blocks (k, g) and (k', g) and recomputing the marginal likelihood for these new statistics. If this difference is greater than 0 we merge the clusters k and k', otherwise no merge is performed. All pairwise merges are considered after the termination of the greedy exact ICL algorithm.

3.4 Computational complexity of the greedy algorithm

Quantifying exactly the computational complexity of the greedy algorithm is not possible, as it will be problem dependent. However, we can analyse its various components and give some guidelines and comparison with the MCMC algorithm of Wyse & Friel (2012). Assume that the average cost of computing one of the Λ_{kg} terms is Q_Λ (this clearly depends on the model chosen for the linking attributes). To update node i in cluster $A_{k'}$ we must compute $K-1$ terms of the form

$$\begin{aligned}\Delta_{k'\to l} &= \text{ICL}(K, G, \mathbf{c}^*, \mathbf{w}) - \text{ICL}(K, G, \mathbf{c}, \mathbf{w}) \\ &= \left[\log\pi(\mathbf{c}^*, \mathbf{w}|K, G) + \sum_{k=1}^{K}\sum_{g=1}^{G}\log\Lambda^*_{kg}\right] - \left[\log\pi(\mathbf{c}, \mathbf{w}|K, G) + \sum_{k=1}^{K}\sum_{g=1}^{G}\log\Lambda_{kg}\right]\end{aligned}$$

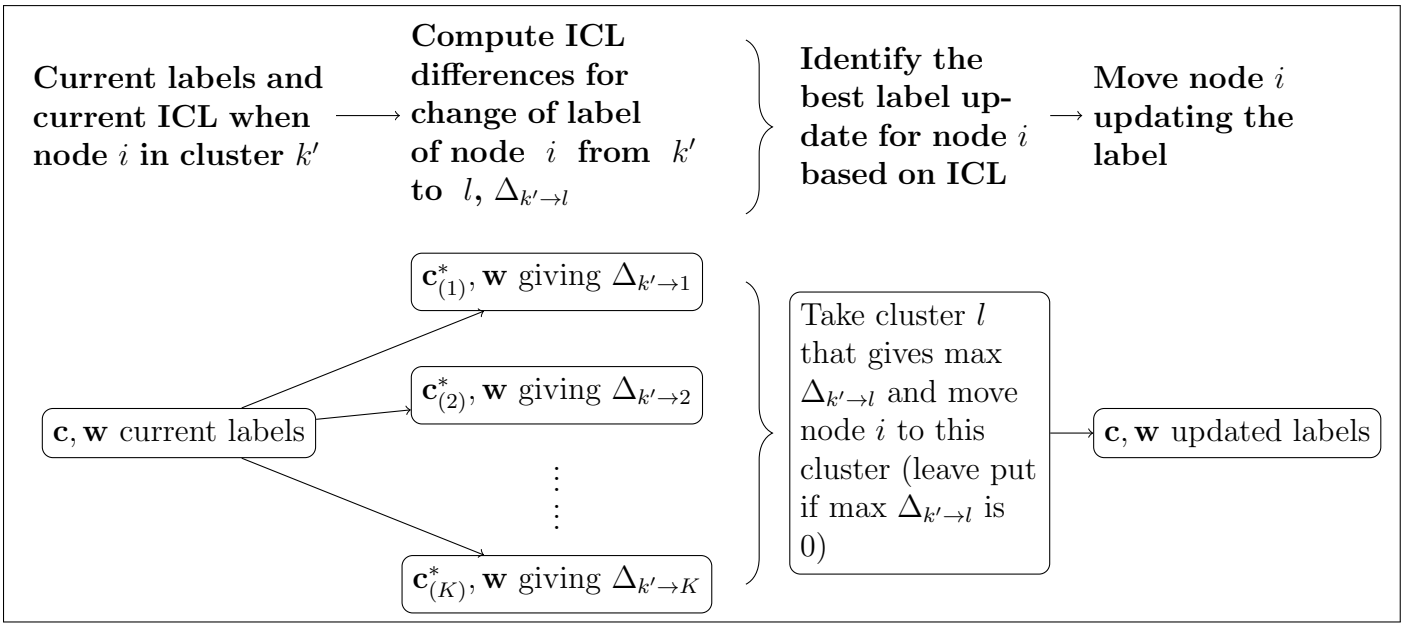

Figure 1: Flowchart showing the process for updating the label of a node i from set A. Here the $\mathbf{c}^*$ are subscripted with the candidate cluster for clarity. The top portion (boldface) shows gives an overview of each stage of the process. The bottom portion outlines the computations carried out.

assuming that $|A_{k'}| > 1$, where the Λ^*_{kg} refer to the label vector $\mathbf{c}^*$. One can see that $\Lambda^*_{kg} = \Lambda_{kg}$ for all but the affected subsets k' and l, so that cancellation results in

$$\begin{aligned} \Delta_{k' \to l} &= \left[\log \pi(\mathbf{c}^*, \mathbf{w} | K, G) + \sum_{g=1}^{G} \left(\log \Lambda^{(-i)}_{k'g} + \log \Lambda^{(+i)}_{lg} \right) \right] \\ &\quad - \left[\log \pi(\mathbf{c}, \mathbf{w} | K, G) + \sum_{g=1}^{G} \left(\log \Lambda_{k'g} + \log \Lambda_{lg} \right) \right]. \end{aligned} \tag{4}$$

Where now we have substituted the Λ^*_{kg} for more meaningful and specific notation ("$+i$" meaning add i to the block, "$-i$" meaning take i out of the block). For all of the linking attribute models mentioned in Section 3.2, one can store the sufficient statistics of each (k, g) block to compute the Λ_{kg}, meaning that Λ^*_{kg} can be computed by a slight modification of these statistics. Also we can store the Λ_{kg} from the most recent state of the algorithm. The term $\log \pi(\mathbf{c}^*, \mathbf{w} | K, G) - \log \pi(\mathbf{c}, \mathbf{w} | K, G)$ is equal to

$$\log \left(\frac{\Gamma\{N_{k'} - 1 + \alpha_0\} \Gamma\{N_l + 1 + \alpha_0\}}{\Gamma\{N_{k'} + \alpha_0\} \Gamma\{N_l + \alpha_0\}} \right).$$

We assume that the cost of computing this quantity is dominated by the computation of the Λ_{kg}. This is reasonable for all of the models in Section 3.2. This means that (4) can be computed with $O(2Q_\Lambda G)$ operations (for the case $|A_{k'}| = 1$ it can be shown that $O(Q_\Lambda G)$

operations are required for the calculation). So updating the label of node i is $O(2Q_\Lambda KG)$. One can show that updating nodes in set B has the same cost, giving $O(2NMQ_\Lambda KG)$ overall per sweep of the greedy algorithm. If we compare to Wyse & Friel (2012), the total cost of the Gibbs update for nodes for T_{it} iterations is fixed at $O(2T_{\text{it}}NMQ_\Lambda KG)$. Usually we would expect convergence of the greedy algorithm in much fewer than T_{it} iterations. Also there are other moves in Wyse & Friel (2012), which are costly but whose complexity is difficult to quantify exactly. In short, we expect our approach to be much more scalable than the MCMC based inference. Another point to note is that the performance in terms of mixing of the algorithm of Wyse & Friel (2012) usually declines as the dimensions N and/or M increase, so that although we may be expending much more computing power, we may not necessarily be gaining much more information about the clustering. The MCMC algorithm also has limitations in size, in that to run it on larger datasets may not even be feasible from a time consideration.

3.5 Why the greedy search is both sensible and pragmatic

Consider a Gibbs sampler for updating the labels of the nodes from set A for example similar to that used in Wyse & Friel (2012). Using (3) the full conditional distribution of the new label for node i with $c_i = k'$, is

$$\pi(c_i^* = l|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G) \propto \Gamma\{N_{k'} - 1 + \alpha_0\}\Gamma\{N_l + 1 + \alpha_0\} \prod_{g=1}^{G} \Lambda_{k'g}^{(-i)} \Lambda_{lg}^{(+i)}$$

for $l \neq k$ where $\mathbf{c}_{-i}$ denotes $\mathbf{c}$ without the i^{th} entry and $\Lambda_{k'g}^{(-i)}$ denotes the quantity computed by removing the statistics associated with node i. Also

$$\pi(c_i^* = k'|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G) \propto \Gamma\{N_{k'} + \alpha_0\}\Gamma\{N_l + \alpha_0\} \prod_{g=1}^{G} \Lambda_{k'g}\Lambda_{lg}$$

reflects no change in label. Notice that

$$\frac{\pi(c_i^* = l|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G)}{\pi(c_i^* = k'|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G)} \propto \frac{\Gamma\{N_{k'} - 1 + \alpha_0\}\Gamma\{N_l + 1 + \alpha_0\}}{\Gamma\{N_{k'} + \alpha_0\}\Gamma\{N_l + \alpha_0\}} \frac{\prod_{g=1}^{G} \Lambda_{k'g}^{(-i)} \Lambda_{lg}^{(+i)}}{\prod_{g=1}^{G} \Lambda_{k'g}\Lambda_{lg}} = \exp\{\Delta_{k' \to l}\}$$

from (4). Using this relation, it is straightforward to show that

$$\pi(c_i^* = l|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G) = \frac{\exp\{\Delta_{k' \to l}\}}{\sum_{k=1}^{K} \exp\{\Delta_{k' \to k}\}}. \tag{5}$$

This implies that changing the label of node i to the label that maximizes the change in exact ICL (or alternatively not changing the node label if none of the changes in exact ICL are positive) corresponds to maximizing the full conditional distribution of the label in the Gibbs sampler. So instead of stochastically sampling a label from the full conditional, we deterministically choose the label that maximizes it in our greedy exact ICL search. This is similar to the iterated conditional modes algorithm of Besag (1986). For this reason we

argue that using the exact ICL as an objective function is sensible. The pragmatic nature of the algorithm comes from the reasoning that if there is a strong cohesion to subsets (strong clustering and good separation of node subsets), then the label that will be sampled from the full conditional will usually be the one which corresponds to the largest value of the change in exact ICL. If our objective is to get the optimal partitioning of the node sets, then this appears an appealing approach.

3.6 Computational savings for the greedy ICL algorithm

There are two main ways we can reduce the computational complexity of the greedy algorithm discussed in Section 3.4. The first of these concerns the correspondence between maximizing exact ICL and the full conditional distribution of labels as shown in Section 3.5. The second is by exploiting the fact that often observed bipartite networks are sparse.

3.6.1 Greedy search pruning

One might expect that after a few sweeps of the greedy algorithm, some degree of degeneracy will appear in the full conditional distributions of the labels. In such situations one will observe that $\pi(c_i^* = l|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G)$ will be quite small compared with the full conditional for some other label, say $\pi(c_i^* = k'|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G)$. Suppose that currently $c_i = k$. Compute $\Delta_{k\to l}$ for all $l \neq k$, where $\Delta_{k\to k} = 0$. Let $k' = \arg\max_r \Delta_{k\to r}$. Consider labels l which are very unlikely for node i compared to k'

$$\frac{\pi(c_i^* = l|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G)}{\pi(c_i^* = l|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G) + \pi(c_i^* = k'|\mathbf{c}_{-i}, \mathbf{w}, Y, K, G)} < \varepsilon$$

for some small number ε. Using (5) this may be written

$$\frac{\exp\{\Delta_{k\to l}\}}{\exp\{\Delta_{k\to l}\} + \exp\{\Delta_{k\to k'}\}} < \varepsilon.$$

This can be rearranged to give

$$\Delta_{k\to k'} - \Delta_{k\to l} > T(\varepsilon) \tag{6}$$

where $T(\varepsilon)$ is a threshold depending on the value of ε. To reduce our search space in future iterations of the greedy search, we ignore computing the exact ICL for label changes to l when (6) is satisfied. In this paper we take $T(\varepsilon) = 150$. This appeared to give satisfactory results without excluding important parts of the search space and introducing error. It can be seen that removal of these poor configurations from the search space will reduce the $O(4Q_\Lambda KG)$ cost of optimizing the labels at each step. The procedure has been described here for nodes from set A but applies analogously to those in set B.

3.6.2 Sparse representations

When dealing with observed bipartite networks one often encounters much sparsity. By sparsity we mean many observed non-ties. For this reason, it can be useful to use a sparse

representation of the adjacency matrix Y, denoted Y^{s}, and use this to perform the exact ICL calculations outlined earlier. Only the non-zero valued ties of Y are stored in Y^{s}. For example, suppose we would like to compute the ICL difference given in (4) for Y^{s} stored in sparse form. We denote a non-tie generically by "0". Suppose there are m_i non-zero entries in the row corresponding to node i (from A) of Y^{s} in columns $j_1^i, \ldots, j_{m_i}^i$. Let the indexes of the non-zero columns which correspond to nodes from B and are members of B_g be stored in the set $\mathcal{J}_g^i = \{l : w_{j_l} = g\}$ and denote the nodes that have corresponding zero columns and which are members of B_g by $\tilde{\mathcal{J}}_g^i = \{l : w_l = g\} \backslash \mathcal{J}_g^i$. For each g compute

$$\upsilon_g^{i:\mathrm{nz}} = \sum_{l=1}^{m_i} \mathrm{I}(w_{j_l^i} = g)$$

which gives the number of non-zero entries in B_g which in row i of the adjacency matrix, corresponding to node i. Then the number of zero entries in columns indexed by B_g for row i is $\upsilon_g^{i:\mathrm{z}} = \upsilon_g - \upsilon_g^{i:\mathrm{nz}}$. Then

$$\begin{aligned}
\Lambda_{kg} &= \int \pi(\boldsymbol{\theta}_{kg}) \prod_{i:c_i=k} \prod_{j:w_j=g} p(y_{ij}|\boldsymbol{\theta}_{kg}) \, \mathrm{d}\boldsymbol{\theta}_{kg} \\
&= \int \pi(\boldsymbol{\theta}_{kg}) \prod_{i:c_i=k} \left[\prod_{j \in \mathcal{J}_g^i} p(y_{ij}|\boldsymbol{\theta}_{kg}) \prod_{j \in \tilde{\mathcal{J}}_g^i} p(y_{ij}|\boldsymbol{\theta}_{kg}) \right] \mathrm{d}\boldsymbol{\theta}_{kg} \\
&= \int \pi(\boldsymbol{\theta}_{kg}) \prod_{i:c_i=k} \left[p(\text{“0”}|\boldsymbol{\theta}_{kg})^{\upsilon_g^{i:\mathrm{z}}} \prod_{j \in \mathcal{J}_g^i} p(y_{ij}^{\mathrm{s}}|\boldsymbol{\theta}_{kg}) \right] \mathrm{d}\boldsymbol{\theta}_{kg}. \qquad (7)
\end{aligned}$$

Thus by storing the number of zeros, and carrying out the calculation for these terms collectively, we can save many evaluations. In the experiments below, we found that exploiting the sparsity of some problems gave a considerable speed up of the algorithm.

3.7 Algorithm types

The greedy search algorithm can be split into four types outlined in Table 1, algorithms A2 and A3 using the pruning ideas of Section 3.6.1, and A1 and A3 using the sparse representations of Section 3.6.2. Observing the algorithm types, A0 is the naive implementation and we would expect this to be the slowest. On the other end of the scale, we would expect A3 to be fastest. Between A1 and A2, we would expect A1 to be faster than A2 when the matrix is quite sparse, and A2 to be faster than A1 when there is strong separation of nodes into subsets (strong cohesion) i.e. very strong blocking. For small to moderate sizes of adjacency matrices, differences in run times may not be apparent, but for larger and sparser adjacency matrices we would expect to notice differences. When updating labels of A nodes or B nodes, these are processed in a random order. This introduces some amount of stochasticity into the ultimate destination of the greedy search. This will be investigated in due course in the examples below. For the examples we take α_0 and β_0 both equal to 1 throughout. For the pruning algorithms we allow five full sweeps of the data before pruning begins.

Algorithm	Pruning	Sparse form
A0	No	No
A1	No	Yes
A2	Yes	No
A3	Yes	Yes

Table 1: Different types of greedy search algorithm.

4 Simulation study

To investigate the performance of the greedy search approaches, we compared them with the MCMC approach of Wyse & Friel (2012) on simulated data.

4.1 Study set-up

The data was generated from the model outlined in Section 2, with $K = 5$ and $G = 5$. The label distributions were chosen as $\omega_k = 1/5, k = 1, \ldots, 5$ and $\rho_g = 1/5, g = 1, \ldots, 5$. The node sets were chosen to be of size $N = 100$ and $M = 100$, giving a 100×100 adjacency matrix. Using the absent/present linking attribute outlined in Section 3.2, the ties were generated so as to give different levels of overlap in the blocks. There are five blocks of varying clarity in the matrix following

$$\theta_{kg} = \begin{cases} 1 - q & \text{if } k = g \\ q & \text{otherwise} \end{cases}$$

where we let q vary from 0.0125 to 0.5 in steps of 0.0125, giving 40 values of q in total. As q gets closer to 0.5 the clustering task becomes more difficult, as the probability of a tie becomes constant across all blocks. Twenty datasets were generated at each value of q.

4.2 Comparing clusterings

As a means of comparing different clusterings of the data, we extend the normalized mutual information measure introduced in Vinh & Epps (2010) and used by, for example Côme & Latouche (2013). Vinh & Epps (2010) provide extensive and rigorous justification for using this such a measure for clustering comparison. We define a combined measure for the two node sets (rows and columns of the adjacency) to account for the fact that we are effectively doing two clustering tasks simultaneously. Consider first node set A. Denote the true clustering (from the simulation) by the labels $\mathbf{c}^{\mathrm{t}}$ and the estimated one (from the algorithm in question) by $\mathbf{c}^{\mathrm{e}}$. The mutual information between the two clusterings is

$$\mathcal{I}_A(\mathbf{c}^{\mathrm{e}}, \mathbf{c}^{\mathrm{t}}) = \sum_{k,l}^{K} P_{kl} \log \left(\frac{P_{kl}}{P_k^{\mathrm{e}} P_l^{\mathrm{t}}} \right)$$

where

$$P_{kl} = \frac{1}{N}\sum_{i,j}^{N} \mathrm{I}(c_i^{\mathrm{e}} = k, c_j^{\mathrm{t}} = l),\ P_k^{\mathrm{e}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}(c_i^{\mathrm{e}} = k),\ P_l^{\mathrm{t}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}(c_i^{\mathrm{t}} = l).$$

The mutual information measures how much is learned about the true clustering if the estimated one is known, and *vice versa*. We normalize this quantity when the clusterings can have a different number of clusters (as in our case)

$$\mathcal{NI}_A(\mathbf{c}^{\mathrm{e}}, \mathbf{c}^{\mathrm{t}}) = \frac{\mathcal{I}(\mathbf{c}^{\mathrm{e}}, \mathbf{c}^{\mathrm{t}})}{\max(\mathcal{H}(\mathbf{c}^{\mathrm{e}}), \mathcal{H}(\mathbf{c}^{\mathrm{t}}))}$$

with $\mathcal{H}_A(\mathbf{c}) = -\sum_{k=1}^{K} P_k \log P_k$ and $P_k = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}(c_i = k)$.

We compute the same quantity analogously for node set B, and add the two together to give an overall measure of mutual information between estimated and true clusterings as

$$\mathcal{NI}_A(\mathbf{c}^{\mathrm{e}}, \mathbf{c}^{\mathrm{t}}) + \mathcal{NI}_B(\mathbf{w}^{\mathrm{e}}, \mathbf{w}^{\mathrm{t}})$$

which has a maximum value of 2 with high agreement and minimum 0 with little or no agreement.

4.3 Different approaches

The three different approaches taken were

- a run of the collapsed MCMC algorithm of Wyse & Friel (2012) of 25,000 iterations taking the first 5,000 as a burn in
- the basic greedy search algorithm A0
- the pruned greedy search algorithm A2.

The reason the sparse forms (A1 and A3) of the greedy search algorithms are not used here is the dimension of the adjacency matrix; a 100×100 matrix runs quickly enough using algorithms A0 and A2 that exploiting any possibly sparsity is not of huge benefit here. For the MCMC algorithm, the estimated clustering was computed by first finding the MAP estimates of K and G from the approximated posterior, and then finding the MAP of the labels conditioning on these values. It should be noted here that the MCMC takes considerably longer to run than either of the greedy algorithms, which are about 2,000 times faster in this instance. The pruned algorithm is usually about 10 times faster than the basic greedy algorithm. Figure 2 shows the average value of the normalized mutual information over the twenty datasets for each value of q for each of the algorithms. The MCMC algorithm is shown in blue, the greedy algorithm in red and the pruned greedy algorithm in green. It can be seen that as the value of q approaches 0.5, the greedy algorithms slightly outperform the MCMC algorithm, with the MCMC outperforming the greedy algorithm around $q = 0.35$. However, the observation most of note here is that the greedy algorithm gives essentially the same information as the longer MCMC run in a very small fraction of the time.

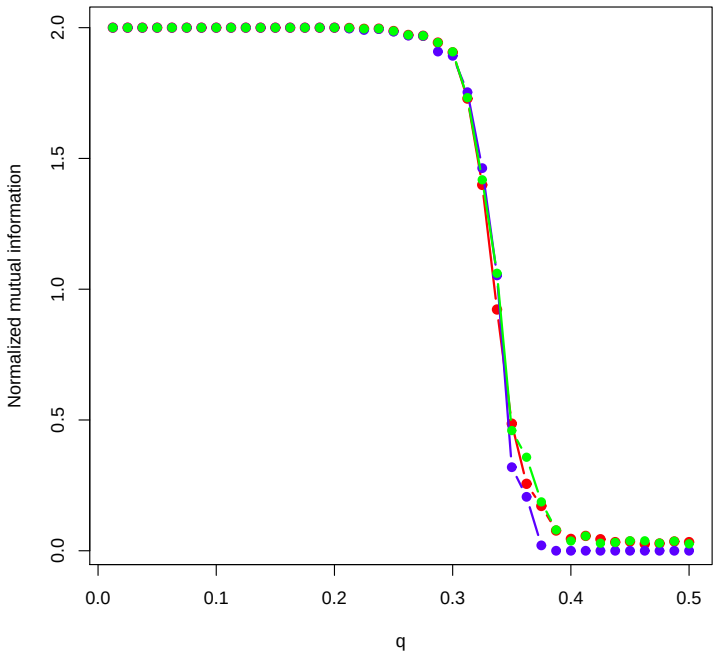

Figure 2: Comparison of the normalized mutual information for the three algorithms. MCMC is shown in blue, greedy algorithm in red and the pruned greedy algorithm in green.

Algorithm	maximum ICL	average time (sec)	(K, G)
A0	-3543.062	0.62	(6,12)
A1	-3543.062	0.56	(6,12)
A2	-3543.062	0.62	(6,12)
A3	-3543.062	0.56	(6,12)

Table 2: Results for ten runs of each of the four algorithms on Congressional voting data with same random seed.

Algorithm	maximum ICL	average time (sec)	(K, G)
A0	-3546.968	0.63	(6,11)
A1	-3543.812	0.58	(6,10)
A2	-3537.503	0.65	(6,11)
A3	-3537.550	0.64	(6,11)

Table 3: Results for ten runs of each of the four algorithms on Congressional voting data with different random seed.

5 Examples

5.1 Congressional voting data

First of all the congressional voting data was analysed by treating abstains as nay's. The data matrix is 435×16. Here the node sets to be grouped are the congressmen (435) and the issues which they vote on (16). Ten instances of the algorithms in Table 1 were run, each with two restarts and the maximum ICL recorded from the ten runs. If restarting the algorithm affects the results, one would expect this to me more pronounced in the pruned versions A2 and A3, since restarting here corresponds to resetting the search for each row/column to the set of all possible clusters as opposed to the pruned set.

To allow for direct comparison of the four algorithms, the random number generator was seeded the same- this also allows a direct time comparison for the same search progression as well as the maximum exact ICL reached. Table 2 shows the average runtime and maximum exact ICL reached for each algorithm over the ten runs. It can be seen that the sparse versions of the algorithm give a reduction in runtime with no noticeable reduction due to pruning.

The experiment was repeated, this time with different random seeds for each algorithm. Figure 3 shows the re-ordered data matrices for the four highest ICL over ten runs of the four algorithms. Here the effect of the stochasticity on the outcome of the greedy search is noted. The four algorithms all converge to different numbers of clusters, with different resulting maximum exact ICL. Visually, all four algorithms appear to give a sensible result despite the fact that there are some differences. This highlights the possibility of many local maxima in general applications and the need to run the algorithm a number of times to give it the best chance to find a close to global maximum.

To take a closer look at the effect of starting value and the randomness introduced into

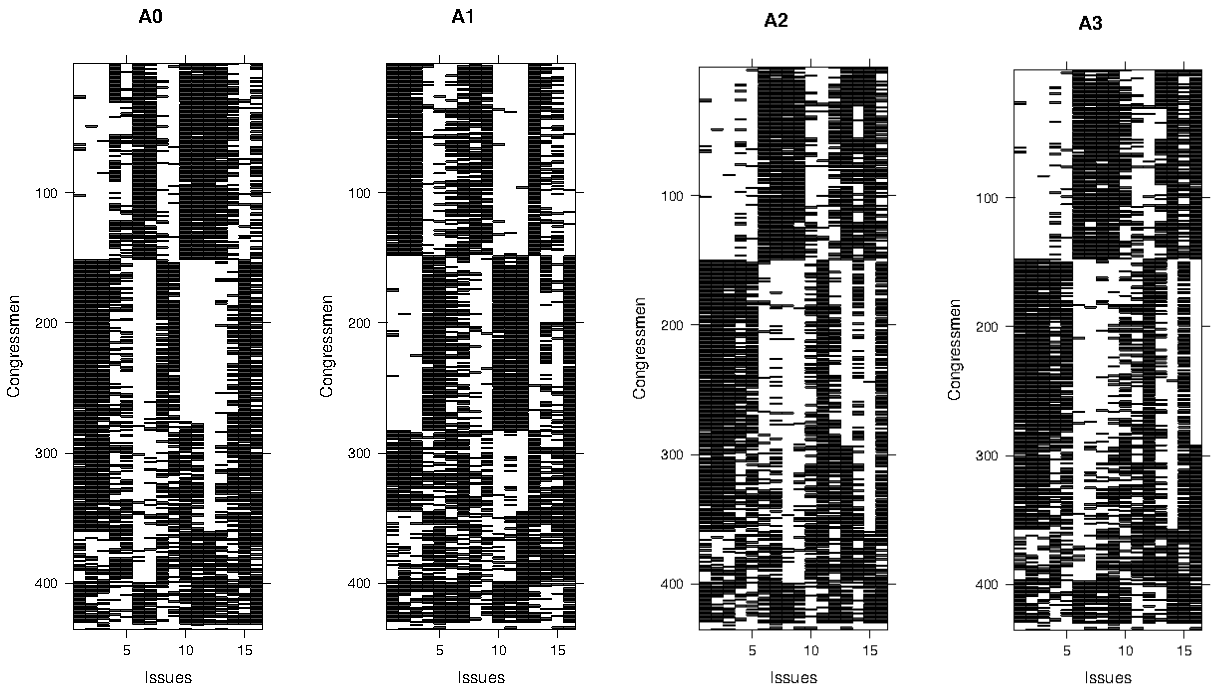

Figure 3: Re-ordered data matrix of maximum ICL configuration output over the ten runs of each algorithm.

the greedy search by processing nodes in a random fashion when updating the labels, we took 100 runs of each algorithm. Histograms of the maximum ICL obtained are shown in Figure 4. From this it can be seen that all four algorithms most often get to a maximum exact ICL around -3560.00 and less frequently obtaining values larger than this. As mentioned in Section 3.3 and suggested by one reviewer, the optimal ICL could potentially be improved by providing a good starting value for the algorithm. In the absence of such an initialization, we'd recommend running the algorithm a number of times (10 or more) and taking the output of the run that gives the maximum ICL over these. This is similar to the general approach when the EM algorithm is fitted to finite mixture models; the run which results in the maximum value of the likelihood is taken as the output run.

As a rough comparison to the MCMC based algorithm of Wyse & Friel (2012), their analysis took about thirty minutes on this network, gleaning much the same information that we have obtained here in a fraction of the time. The MCMC analysis here was coded in parallel with the implementation for computing the exact ICL and is more efficient than the original implementation in Wyse & Friel (2012), hence the difference in reported run time. The mixing of their algorithm, even for this moderately sized data was was slow for the congressman dimension with only about 1.5% of proposed cluster additions/deletions being accepted. In terms of scalability, MCMC is not an option when dealing with adjacency matrices of any large size. The fact that both the MCMC approach and the exact ICL approach are based on almost identical posteriors indicates that the greedy search is the only viable option of the two for larger applications.

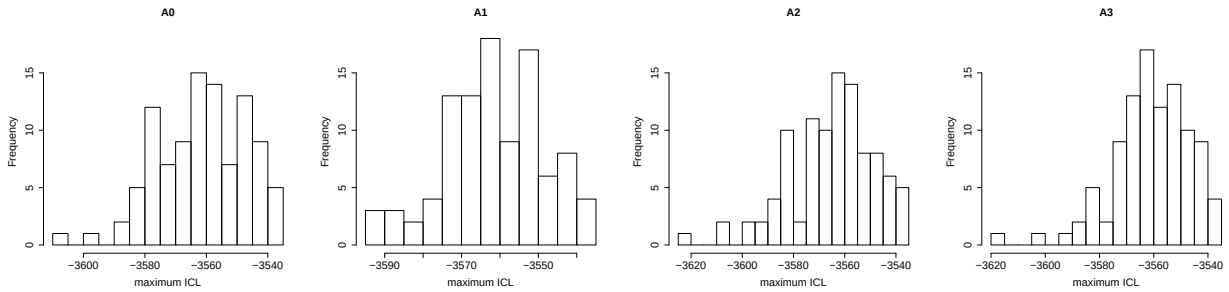

Figure 4: Histograms of maximum ICL for 100 runs of each of algorithms A0-A4.

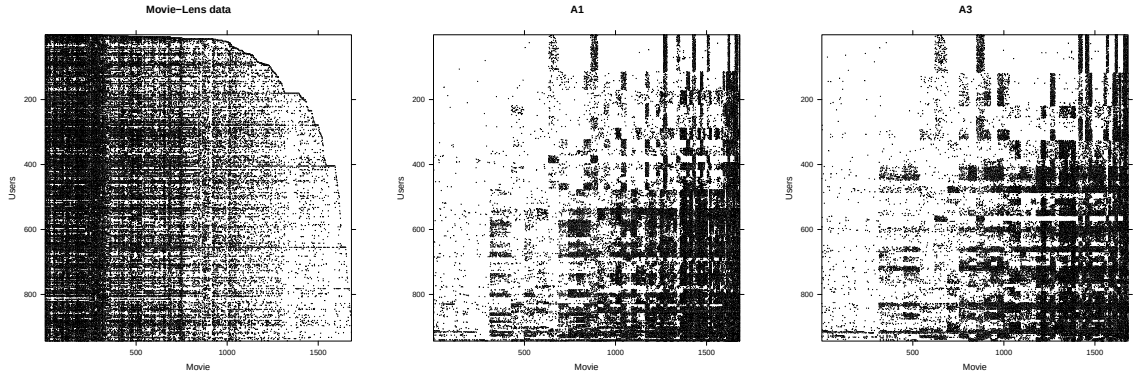

Figure 5: The Movie-lens data with 100,000 user ratings and results from the A1 and A3 algorithm.

5.2 Movie-Lens 100k

The Movie-Lens dataset gives 100,000 ratings of movies by users. In total, there are 943 users and 1,682 movies. Many users do not rate particular movies so the network is quite sparse (93%) and it is useful to use the approach mentioned in Section 3.6.2. The ratings (linking attributes) are assumed Poisson distributed, due to the ordinal nature of the ratings going from 1-5. This is a sensible model choice in this context due to both the ordinal and discrete nature of the ratings. The clustering task will cluster groups of individuals giving a similar rating to specific groups of movies. One could expect that this would glean more information than treating the 1-5 ratings as purely categorical since the Poisson model captures in some ways the ordinal features of the data. A missing rating is just given the value 0 when assuming the ratings are Poisson. The data is shown in Figure 5. One run of each of the algorithms was carried out each with two restarts with results in Table 4. The reordered data is shown for algorithms A1 and A3 in Figure 5.

A glance at these results reveals that the pruning based algorithms give different number of clusters to those that do not involve pruning. This could be due to the effect of the value of the difference in exact ICL considered for the pruning. A value of this threshold which is too high could introduce some degeneracy into the search space, with labels settling in sub-optimal groups. The result for this dataset can vary depending on the pruning threshold

Algorithm	maximum ICL	run time (sec)	(K, G)
A0	-646597.7	1098.529	(55,64)
A1	-646597.7	368.032	(55,64)
A2	-646268.2	525.818	(56,62)
A3	-646268.2	160.162	(56,62)

Table 4: Results for one run of each of the four algorithms on Movie-Lens data with same random seed.

chosen. However, pruning does give a considerable reduction in computing time which is favourable from a scalability viewpoint. This could be seen as an option when the data size is such that a run of the full algorithm is just too expensive.

6 Discussion

We have presented an approach to grouping or clustering subsets of nodes in bipartite networks which share similar linking properties. There are two main advantages to our work. The first is that we model linking patterns between nodes in the two node sets in a statistical way. The second is that we provide a principled inference technique for the number of clusters in the node sets by appealing to a known information criterion due to Biernacki et al. (2000). The fact that the exact ICL may be computed due to the tractability assumptions we make, means that it can be iteratively updated for each proposed new labelling of a node in either node set.

The approach presented here can be considered favourable to the MCMC techniques of Wyse & Friel (2012) for a few reasons. If only an optimal block structure is required our algorithm provides this in a more scalable framework than MCMC. Also, in summarizing MCMC output, often the maximum a posteriori (MAP) estimate is used as an overall summary, since it is not possible to average clusterings with a difference number of clusters. The concept of this MAP estimator is in somewhat equivalent to the configuration of labels and clusters that maximizes the ICL. Probably the most important reason is that MCMC is simply not a viable option for large bipartite networks due to the declining mixing performance observed and the long run times required in such cases. The simulation study shows that we gain as much information from the greedy algorithm as we would from a longer and more time consuming MCMC run, which is encouraging. This being said, there is always a risk that the ICL greedy algorithm we discuss could get trapped in a local maximum. Due to the fast run time, we can circumvent this somewhat in practice by running, say 10 instances of the greedy algorithm and taking the one that gives the highest ICL (indicating the best clustering). The greedy algorithm (especially the sparse versions) scale well enough to make this feasible. The risk that this will happen to the MCMC algorithm is less due to its stochastic nature. ICL does still possess many of the benefits of MCMC, the strongest of which is that it allows for automatic choice of the number of clusters. Considering the complexity of the task it performs (model selection and clustering), the ICL algorithm does appear to scale quite well, running on the Movie-Lens 100k data in less than 3 minutes,

and the congressional voting data in less than a second. For many of the applications of bi-clustering in the literature, this is competitive. In addition, we provide code (written in C with an R wrapper) which is available from the first author's webpage

`https://sites.google.com/site/jsnwyse/`

(a Makefile is provided to compile on Unix-alikes).

In Section 3.6 we showed how sparsity (a common feature of observed networks) can be exploited to provide a faster version of the greedy search exact ICL algorithm. Although not possible to provide exact quantification for the order of speed up (in terms of N and M) we observed considerable computational gain in exploiting sparsity. The applicability of these ideas may stretch well beyond implementation of the LBM. For example, it may be possible to use the same kind of representation for SBMs.

We believe that the latent blockmodel is an appealing model for bipartite networks, although to our knowledge, no authors have appeared to explicitly discuss it in the context of network modelling. One drawback of the LBM in its original form was the lack of any principled information criterion (such as Bayesian Information Criterion) for choosing the number of clusters. However, the work of Keribin et al. (2013) and the ideas presented in this paper overcome this issue. For this reason, the LBM can now be seen as a rich model for which there is capacity to model group structure in bipartite networks. The approach we take also means it is scalable for larger networks where sparsity can be exploited.

References

Besag, J. (1986), 'On the statistical analysis of dirty pictures (with discussion)', *Journal of the Royal Statistical Society, Series B* **48**, 259–302.

Biernacki, C., Celeux, G. & Govaert, G. (2000), 'Assessing a mixture model for clustering with the integrated completed likelihood', *IEEE Trans. Pattern Anal. Machine Intel* **7**, 719–725.

Borgatti, S. P. & Everett, M. G. (1997), 'Network analysis of 2-mode data', *Social Networks* **19**, 243–269.

Bozdogan, H. (1994), Mixture-model cluster analysis using model selection criteria and a new information measure of complexity, *in* H. Bozdogan, ed., 'Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach, Volume 2', Kluwer, Boston, pp. 69–113.

Brusco, M., Doreian, P., Lloyd, P. & Steinley, D. (2013), 'A variable neighbourhood search method for a two-mode blockmodelling problem in social network analysis', *Network Science* **1**, 191–212.

Brusco, M. J. & Steinley, D. (2006), 'Inducing a blockmodel structure on two-mode data using seriation procedures', *Journal of Mathematical Psychology* **50**, 468–477.

Brusco, M. & Steinley, D. (2011), 'A tabu search heuristic for deterministic two-mode blockmodelling of binary network matrices', *Psychometrika* **76**, 612–633.

Côme, E. & Latouche, P. (2013), 'Model selection and clustering in stochastic block models with the exact integrated complete data likelihood', *arXiv e-print* . arXiv:1303.2962.

Doreian, P., Batagelj, V. & Ferligoj, A. (2004), 'Generalized blockmodeling of two-mode network data', *Social Networks* **26**, 29–53.

Doreian, P., Lloyd, P. & Mrvar, A. (2013), 'Partitioning large signed two-mode networks: Problems and prospects', *Social Networks* **35**, 1–21.

Flynn, C. J. & Perry, P. O. (n.d.), 'Consistent Biclustering', *arXiv e-print* .

Govaert, G. (1995), 'Simultaneous clustering of rows and columns', *Control Cybernet* **24**(4), 437–458.

Govaert, G. & Nadif, M. (1996), 'Comparison of the mixture and the classification maximum likelihood in cluster analysis when data are binary', *Computational Statistics and Data Analysis* **23**, 65–81.

Govaert, G. & Nadif, M. (2003), 'Clustering with block mixture models', *Pattern Recognition* **36**, 463–473.

Govaert, G. & Nadif, M. (2005), 'An em algorithm for the block mixture model', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**, 643–647.

Govaert, G. & Nadif, M. (2008), 'Block clustering with Bernoulli mixture models: Comparison of different approaches', *Computational Statistics and Data Analysis* **52**, 3233–3245.

Hartigan, J. A. (1972), 'Direct Clustering of a Data Matrix', *Journal of the American Statistical Association* **67**, 123–129.

Keribin, C., Brault, V., Celeux, G. & Govaert, G. (2012), Model selection for the binary latent block model, *in* 'Proceedings of COMPSTAT'.

Keribin, C., Brault, V., Celeux, G. & Govaert, G. (2013), Estimation and Selection for the Latent Block Model on Categorical Data, Rapport de recherche RR-8264, INRIA.
URL: *http://hal.inria.fr/hal-00802764*

Keribin, C., Govaert, G. & Celeux, G. (2010), Estimation d'un modèle à blocs latents par l'algorithme em, *in* '42eme journées de Statistique'.

Nobile, A. & Fearnside, A. T. (2007), 'Bayesian finite mixtures with an unknown number of components: The allocation sampler', *Statistics and Computing* **17**, 147–162.

Nowicki, K. & Snijders, T. A. B. (2001), 'Estimation and Prediction for Stochastic Blockstructures', *Journal of the American Statistical Association* **96**, 1077–1087.

Opsahl, T. (2013), 'Triadic closure in two-mode networks: Redefining the global and local clustering coefficients', *Social Networks* **35**, 159–167.

Rohe, K., Qin, T. & Yu, B. (2015), 'Co-clustering for Directed Graphs; the Stochastic Co-Blockmodel and spectral algorithm Di-Sim', *arXiv e-print* . arXiv:1204.2296.

van Dijk, B., van Rosmalen, J. & Paap, R. (2009), 'A Bayesian Approach to Two-Mode Clustering', *Econometric Institute Report* **06**.

Vinh, N. X. & Epps, J. (2010), 'Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance', *Journal of Machine Learning Research* **11**, 2837–2854.

Wasserman, S. & Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press.

Wyse, J. & Friel, N. (2012), 'Block clustering with collapsed latent block models', *Statistics and Computing* **22**, 415–428.