Nonparametric Bayesian models of hierarchical structure in complex networks

Mikkel N. Schmidt MNSC@DTU.DK
Tue Herlau TUHE@DTU.DK
Morten Mørup MMOR@DTU.DK
Technical University of Denmark
DTU Compute
Richard Petersens Plads
2800 Kgs. Lyngby, Denmark

Editor:

Abstract

Analyzing and understanding the structure of complex relational data is important in many applications including analysis of the connectivity in the human brain. Such relational networks can have salient statistical patterns on different scales, calling for a hierarchically structured model. We propose a new non-parametric Bayesian hierarchically structured network model based on a Gibbs fragmentation tree prior, and demonstrate its ability to capture nested patterns in simulated networks. On real networks we demonstrate detection of hierarchical structure and show predictive performance on par with the state of the art. We envision that our method can be employed in exploratory analysis of large scale complex networks for example to model human brain connectivity.

1. Introduction

Complex networks are an integral part of all natural and man made systems. The complex interaction of cells in biological systems, human interaction in social relations, and the physical connections in our infrastructure can all be described in terms of complex network structures. Modeling these network structures have become an important endeavor in order to comprehend and predict the behavior of these systems.

A central organizational principle is that entities are hierarchically[1] organized and that these hierarchical structures play an important role in accounting for the patterns of connectivity in these systems (Simon, 1962; Ravasz and Barabási, 2003; Roy et al., 2007; Sales-Pardo et al., 2007; Clauset et al., 2008; Meunier et al., 2010). A common notion in modeling complex network is the idea of *communities*. Statistically, a community may be defined as a group of nodes that is more densely connected internally than externally. This notion has led to models of networks as collections of groups of nodes that determine how edges are formed (Fortunato, 2010; Mørup and Schmidt, 2012a). For instance, people in a social network may be grouped according to family, workplaces, schools, or entire countries, and it is assumed these groups determine the formation of social relations. In many complex

[1] In this work we define a hierarchy to denote a decomposition of a complex relational system into nested sets of subsystem rather than as a formal organization of successive sets of subordinates (Simon, 1962).

systems communities do not exist only at a single scale but can further be partitioned into submodules and sub-submodules, i.e., as parts within parts (Simon, 1962; Meunier et al., 2010). For instance, are the communities in a network of children best described on the level of school districts or school classes? How about social cliques within classes or year groups of classes? It seems that different answers are relevant in different contexts influencing the scale on which the network should be analyzed; but discovering the *hierarchical* structure governing such relationships in a network would tell us more than any particular choice of resolution.

Previous research on discovering hierarchical structure in networks has primarily focussed on binary trees (Ravasz et al., 2002; Ahn et al., 2010; Breiger et al., 1975; Newman and Girvan, 2004; Roy et al., 2007; Clauset et al., 2008; Roy and Teh, 2009). Given a set of nodes and a matrix of affinities between them, a commonly used tool to uncover their organization is hierarchical clustering using either agglomerative (Ravasz et al., 2002; Ahn et al., 2010) or divisive approaches (Breiger et al., 1975; Newman and Girvan, 2004). These traditional hierarchical clustering approaches, however, have the following three major drawbacks (Sales-Pardo et al., 2007):

1. They are local in their objective function and do not form a well defined global objective.

2. The number of partitions is not well defined and various heuristics are commonly invoked to determine this number.

3. The output is always a binary hierarchical tree, regardless of the underlying true organization.

Addressing the first two drawbacks, a number of non-parametric Bayesian models have been proposed: Roy et al. (2007) and Clauset et al. (2008) have studied Bayesian generative models for binary hierarchical structure in networks, assuming a uniform prior over binary trees, and Roy and Teh (2009) have proposed the Mondrian Process in which groups are formed by recursive axis-aligned bisections. Addressing the third drawback, Herlau et al. (2012) have proposed using a uniform prior over multifurcating trees with leafs terminating at groups of network nodes, and Knowles and Ghahramani (2011) have mentioned the applicability of their multifurcating Pitman-Yor Diffusion tree as a prior for learning structure in relational data.

In this work we propose a non-parametric Bayesian hierarchical network model based on multifurcating Gibbs fragmentation tree prior (McCullagh et al., 2008). We leverage Bayesian nonparametrics to devise a model that:

1. *Is generative.* This allows us to simulate networks from the model, e.g., for use in model checking, and gives a principled approach to handling missing data.

2. *Captures structure at multiple scales.* The model simultaneously learns about structures from macro scale involving the whole network to micro scale involving only a few nodes.

3. *Can infer whether or not hierarchical structure is present.* If there is no support for a hierarchy, the model can reduce to a non-hierarchical structure.

4. *Is consistent and infinitely exchangable.* The model is extendable to an infinite sequence of networks of increasing size, allowing it to increase and adapt its structure to accomodate new data.

The paper is structured as follows. In section 2 we review the Gibbs fragmentation tree process (McCullagh et al., 2008) and in section 3 we describe our model for hierarchical structure in network data using the Gibbs fragmentation tree prior. In section 4 we analyze the hierarchical structure in simulated as well as real networks; in particular, we investigate the support for hierarchical structure in structural whole brain connectivity networks derived from diffusion spectrum imaging based on the data provided by Hagmann et al. (2008). In section 5 we present our conclusions and avenues of further research.

2. Fragmentation processes and trees

Following the presentation in (McCullagh et al., 2008), we review the multifurcating Gibbs fragmentation tree process. The end result is a projective family of exchangeable distributions over rooted multifurcating trees with n leafs.

A rooted multifurcating tree can be represented by a *fragmentation* of the set of leafs. Let B be the set of leafs and $n = |B|$ the total number of leafs. Recall that a *partition* π_B of B is a set of 2 or more non-empty disjoint subsets of B, $\pi_B = \{B_1, B_2, \ldots, B_k\}$, such that the union is B. In the following we denote the size of these subsets by $n_i = |B_i|$. A *fragmentation* T_B of a set B is a collection of non-empty subsets of B defined recursively such that the set of all nodes is a member, $B \in T_B$; each member of the partition π_B is a member, $B_1 \in T_B$, $\ldots$, $B_k \in T_B$; each member of partitions of these subsets are members, and so on until we reach the singletons. Recursively we may write (McCullagh et al., 2008)

$$T_B = \begin{cases} \{B\}, & |B| = 1, \\ \{B\} \cup T_{B_1} \cup \cdots \cup T_{B_k}, & |B| \geq 2. \end{cases} \tag{1}$$

For example, the tree in Figure 3 which has leafs $B = \{1, 2, 3\}$ is represented by the fragmentation $T_B = \big\{\{1, 2, 3\}, \{1, 2\}, \{1\}, \{2\}, \{3\}\big\}$. Uniquely associated with the fragmentation is a multifurcating tree where each element in T_B above serves as a node: B is the root node, and the singletons are the leafs. To emphasize this connection, T_B is called a *fragmentation tree* (McCullagh et al., 2008). The collection of all fragmentation trees for a set B is denoted by $\mathbb{T}_B$.

Let $A \subset B$ be a nonempty proper subset of the leaf nodes. The *restriction* of T_B to A is defined as "the fragmentation tree whose root is A, whose leaves are the singleton subsets of A and whose tree structure is defined by restriction of T_B." (McCullagh et al., 2008). This is also called the *projection* of T_B onto A and denoted by $T_{B,A}$.

A *random fragmentation model* (McCullagh et al., 2008) assigns a probability to each tree $T_B \in \mathbb{T}_B$ for each finite subset B of $\mathbb{N}$. The model is said to be:

- *Exchangeable* if the distribution of T_B is invariant to permutations on B, i.e., the distribution does not depend on the labelling of the leaf nodes.
- *Markovian* if, for a given $\pi_B = \{B_1, B_2, \ldots, B_k\}$, each of the k restricted trees $T_{B_i,B}$ are independently distributed as T_{B_i}.

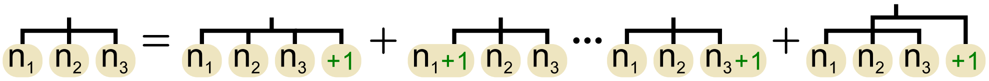

Figure 1: A Markovian consistent splitting rule satisfies the condition that the probability of a partition is equal to the sum of the probabilities of all configurations where a single extra node is added either i) as a new singleton partition on the same level as the current, ii) as a new node in one of the existing partitions, or iii) as a new singleton partition on a level above the current partition.

- *Consistent* if, for all nonempty $A \subset B$, the projection of T_B onto A is distributed like T_A.

2.1 Splitting rule

The starting point for constructing a random fragmentation model is a distribution over partitions of B. By exchangeability this distribution must be a symmetric function depending only on the size of each of the $k \geq 2$ subsets,

$$q(\pi_B) = q\big(n_1, \ldots, n_k\big), \tag{2}$$

where q is called the *splitting rule*. Abusing notation, we write the splitting rule as a function of a partition or equivalently as a function of the sizes of the subsets in the partition.

Requiring Markovian consistency places a further constraint on the splitting rule (McCullagh et al., 2008) illustrated in Figure 1,

$$\begin{aligned} q\big(n_1, \ldots, n_k\big) = q\big(n_1, \ldots, n_k, 1\big) + \\ q\big(n_1 + 1, \ldots, n_k\big) + \cdots + q\big(n_1, \ldots, n_k + 1\big) + \\ q\big(1, n_1 + \cdots + n_k\big) q\big(n_1, \ldots, n_k\big). \end{aligned} \tag{3}$$

Furthermore, McCullagh et al. (2008) impose the condition that the splitting rule is of Gibbs form,

$$q\big(n_1, \ldots, n_k\big) = \frac{a(k)}{c(n)} \prod_{i=1}^{k} w(n_i), \tag{4}$$

where $w(\cdot) \geq 0$ and $a(\cdot) \geq 0$ are some sequences of weights and $c(\cdot)$ is a normalization constant. Under these assumptions, the only admissible splitting rule is given by (McCullagh et al., 2008)

$$a(k) = \alpha^{k-2} \frac{\Gamma\big(k + \frac{\beta}{\alpha}\big)}{\Gamma\big(2 + \frac{\beta}{\alpha}\big)}, \qquad w(n_i) = \frac{\Gamma(n_i - \alpha)}{\Gamma(1 - \alpha)}. \tag{5}$$

McCullagh et al. (2008) do not give an explicit formula for the normalization constant which can be shown to be given by the following:

Proposition 1 (The normalisation constant of the splitting rule) *The normalization constant $c(n)$ in Eq. (4) is given by*

$$c(n) = \frac{1}{\alpha + \beta} \left(\frac{\Gamma(n + \beta)}{\Gamma(1 + \beta)} - \frac{\Gamma(n - \alpha)}{\Gamma(1 - \alpha)} \right). \tag{6}$$

The splitting rule probability can now be written as

$$q(n_1,\ldots,n_k) = \frac{(\alpha+\beta)\alpha^{k-2}}{\frac{\Gamma(n+\beta)}{\Gamma(1+\beta)} - \frac{\Gamma(n-\alpha)}{\Gamma(1-\alpha)}} \cdot \frac{\Gamma\big(k+\frac{\beta}{\alpha}\big)}{\Gamma\big(2+\frac{\beta}{\alpha}\big)} \prod_{i=1}^{k} \frac{\Gamma(n_i-\alpha)}{\Gamma(1-\alpha)} \tag{7}$$

$$= \frac{\big(\frac{\beta}{\alpha}\big)^{(k)} \cdot \alpha^{k-1}}{\beta^{(n)} - (-\alpha)^{(n)}} \prod_{i=1}^{k} (-\alpha)^{(n_i)}, \tag{8}$$

where $x^{(y)} = \frac{\Gamma(y+x)}{\Gamma(1+x)}$ denotes a gamma ratio. The splitting rule has two parameters, α and β: For details regarding parameter ranges, see McCullagh et al. (2008). To simplify the notation in the following we define $q(0) = q(1) = 1$.

Remark 2 (Relation to the Chinese restaurant process) *The splitting rule is closely related to the two-parameter Chinese restaurant process (CRP) (Aldous et al., 1985)*

$$p_{\text{CRP}}(n_1,\ldots,n_k) = \frac{\Gamma(\beta)\alpha^k\Gamma\big(k+\frac{\beta}{\alpha}\big)}{\Gamma(\beta+n)\Gamma\big(\frac{\beta}{\alpha}\big)} \prod_{i=1}^{k} \frac{\Gamma(n_i-\alpha)}{\Gamma(1-\alpha)}. \tag{9}$$

The CRP is a stochastic process that defines a probability distribution over partitions of a set. Its domain includes the trivial partition $k = 1$ whereas the presented splitting rule is defined on $k \geq 2$ only. Subtracting the probability which the CRP assigns to the trivial partition from the normalization constant of the CRP yields the expression for the splitting rule in Eq. (7). The CRP can be constructed sequentially: For $n = 1$ element, the CRP assigns probability one to the trivial partition. As n increases, element number $n+1$ is added to an existing set of elements in the partition with probability $\frac{n_i-\alpha}{n+\beta}$ or added to the partition as a new singleton set with probability $\frac{\beta+k\alpha}{n+\beta}$. Taking the product of n such terms yields the expression in Eq. (9) for the probability assigned to a given partition. As the number of elements goes to infinity, the CRP defines a distribution over partitions of a countably infinite set. A realization from the splitting rule can be generated by the acceptance-rejection method by sampling a partition from the CRP and rejecting if the sampled partition is the trivial partition.

2.2 The Gibbs fragmentation tree

By the Markovian property the distribution over fragmentations can then be characterized as a recursive product of the splitting rules. There will be one term for each set in the fragmentation or equivalently for each node in the tree. This gives rise to the following representation of all exchangeable, Markovian, consistent, Gibbs fragmentation processes,

$$p(T_B) = \prod_{A \in T_B} q(\pi_A), \tag{10}$$

where $q(\cdot)$ is given by Eq. (7) and as before π_A denotes the children of node A in T_B.

To illustrate how the properties of the Gibbs fragmentation tree distribution is governed by the two parameters α and β we have generated a few trees from the distribution, varying the parameters within their usual range $0 \leq \alpha < 1$, $\beta > -2\alpha$ (see Figure 2).

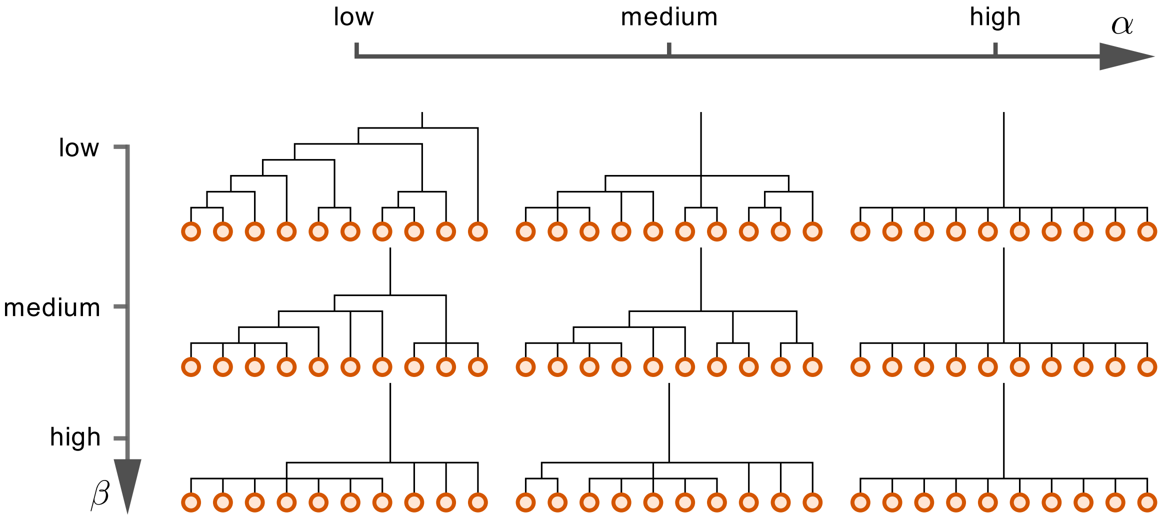

Figure 2: Multifurcating trees generated from the two-parameter Gibbs fragmentation tree process. The parameters govern the distribution of the degrees of the internal nodes in the tree. The trees shown correspond to $\alpha \in \{0.1, 0.5, 1\}$ and $\beta + \alpha \in \{0.1, 1, 10\}$.

Remark 3 (Relation to distributions over binary trees) *Clauset et al. (2008) propose a model with a uniform prior over binary trees. As a special case of the Gibbs fragmentation tree model, binary trees are obtained by the parametrization $\beta = -2\alpha$. Inserting this in Eq. (7) yields*

$$q(n_1, n_2) = \frac{1}{c(n)} \frac{\Gamma(n_1 - \alpha)\Gamma(n_2 - \alpha)}{\Gamma(1 - \alpha)^2}. \tag{11}$$

Letting $\alpha \to -\infty$ yields the uniform binary splitting rule

$$q(n_1, n_2) = \frac{2}{2^n - 2}, \tag{12}$$

which does not imply a uniform distribution over trees. A uniform distribution over binary trees can be attained by setting $\alpha = \frac{1}{2}$ yielding the splitting rule

$$q(n_1, n_2) = \frac{\Gamma(n_1 - \frac{1}{2})\Gamma(n_2 - \frac{1}{2})}{2\sqrt{\pi}\Gamma(n - \frac{1}{2})}, \tag{13}$$

which leads to (McCullagh et al., 2008)

$$p(T_{\mathrm{B}}) = 2^{n-1} \frac{\Gamma(n)}{\Gamma(2n - 1)}. \tag{14}$$

Remark 4 (Relation to the nested CRP) *Blei et al. (2007, 2003) consider a set of nested Chinese restaurant processes: First, the set B is partitioned into $\pi_B = \{B_1, \ldots, B_k\}$ according to a CRP. Next, each subset B_i is partitioned again according to a CRP with the* *same parameters, and the process is continued recursively ad infinitum (see Figure 3). This nested CRP thus defines "a probability distribution on infinitely deep, infinitely branching trees." (Blei et al., 2007) This is used as a prior distribution in a Bayesian non-parametric model of document collections by assigning parameters to each node in the tree and associating documents with paths through the tree.*

In the nested CRP, each element traces an infinite path through the tree. When a finite number of elements n is considered, they trace a tree of finite width but infinite depth. In the terminology of the random fragmentation model, the nested CRP model corresponds to a fragmentation tree using a CRP as a splitting rule. The key difference between the Gibbs fragmentation trees and the nested CRP is that the CRP splitting rule allows fragmenting into the trivial partition, i.e., it allows nodes with a single child whereas the Gibbs fragmentation tree allways has at least two children. Instead of working directly with this infinitely deep tree, we can consider the equivalence class of trees with the same branching structure by marginalizing over the internal nodes that do not branch out, yielding a tree of finite depth. The distribution for this equivalence class can be arrived at by marginalizing over the number of consecutive trivial partitions that occurs before the first "real" split. According to the CRP in Eq. (9), the trivial partition has probability

$$p_0 \equiv p\left(\pi_B = \{B\}\right) = \frac{\Gamma(\beta)\alpha\Gamma\left(\frac{\beta}{\alpha}+1\right)}{\Gamma(\beta+n)\Gamma\left(\frac{\beta}{\alpha}\right)}\frac{\Gamma(n-\alpha)}{\Gamma(1-\alpha)} = \frac{\Gamma(1+\beta)\Gamma(n-\alpha)}{\Gamma(n+\beta)\Gamma(1-\alpha)}. \tag{15}$$

We wish to marginalize over observing zero, one, two, etc. trivial partitions before the first split. To compute this marginalization, the CRP distribution must be multiplied by

$$1 + p_0 + p_0^2 + \cdots = \sum_{i=0}^{\infty} p_0^i = \frac{1}{1-p_0} = \frac{\frac{\Gamma(n+\beta)}{\Gamma(1+\beta)}}{\frac{\Gamma(n+\beta)}{\Gamma(1+\beta)} - \frac{\Gamma(n-\alpha)}{\Gamma(1-\alpha)}} \tag{16}$$

where we have used the geometric series formula to compute the infinite sum and inserted Eq. (15). Multiplying Eq. (16) by the CRP in Eq. (9) yields exactly the splitting rule of McCullagh et al. in Eq. (7) establishing the relation between the Gibbs fragmentation tree and the nested CRP.

3. Tree-structured network models

We now turn to applying the Gibbs fragmentation process as a prior in a Bayesian model of hierarchical structure in complex networks. To simplify the presentation we focus on simple graphs but note that the main ideas can be extended to more intricate relational data such as weighted and directed graphs etc. A simple graph with n nodes can be represented by a symmetric binary adjacency matrix $\boldsymbol{A}$ with element $a_{i,j} = 1$ if there is a link between node i and j.

First, consider a model in which each possible link $a_{i,j}$ is generated independently from a Bernoulli distribution (a biased coin flip) with probability $\theta_{i,j}$. Since each possible link has its own parameter, no information is shared in the model between different nodes and links and the model will not be able to generalize. Combining information between network nodes is necessary, for example by pooling the parameters for blocks of similar nodes.

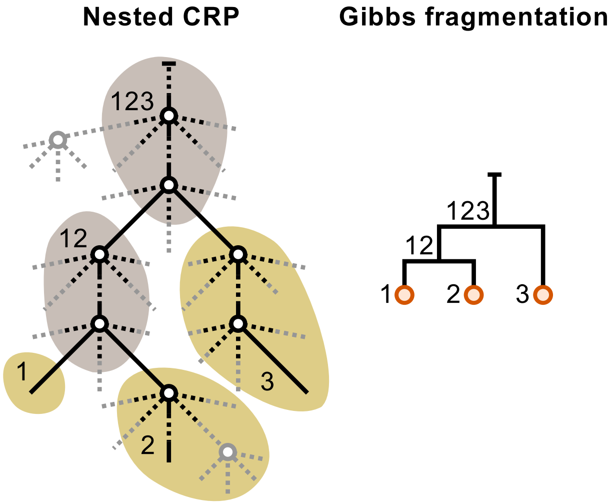

Figure 3: Illustration of the relation between the nested Chinese restaurant process (CRP) and its finite representation as a Gibbs fragmentation tree. In the nested CRP, each internal node in the tree splits into an infinite number of subtrees. Each element associated with the tree traces a infinite path starting at the root. In the illustration three elements are associated with the tree; thus, the hierarchical structure relating the observables (which is what we are ultimately interested in learning) can be represented by a Gibbs fragmentation tree of finite size. As an example, in the finite representation the root node (labeled "123") corresponds to the first common ancestral node of all observables as well as the parents and grandparents etc. of that node all the way to the root of the tree.

The particular way in which these parameters are shared is the key difference between the models we discuss here. In the stochastic blockmodel (Snijders and Nowicki, 1997; HollandKathryn Blackmond and Leinhardt, 1983) network nodes are clustered into blocks which share their probabilities of linking within and between blocks. The infinite relational model (IRM) (Kemp et al., 2006; Xu et al., 2006) is a nonparametric Bayesian extension of the stochastic blockmodel based on a CRP distribution over block structures. We consider the IRM model the state of the art in Bayesian modeling of large scale complex networks.

The link probabilities *between* the blocks in these models can either be individual for each pair of blocks (unpooled) as in the IRM model or be completely shared as a single between-block link probability (complete pooling) as in (Hofman and Wiggins, 2008). Furthermore the model can specify that blocks have more internal than external links leading to an interpretation of the blocks as communities of highly interconnected nodes (Mørup and Schmidt, 2012b).

The hierarchically structured models of complex networks proposed here correspond to nested stochastic blockmodels in which each block is recursively modelled by a stochastic blockmodel, and we use the Gibbs fragmentation tree process as a prior over the nested block structure. As in the stochastic blockmodel, links between blocks can be pooled or not, leading to models with different characteristics. Figure 4 illustrates these different approaches to pooling parameters in block structured network models, and Figure 5 illustrates a network that can be well characterized by a hierarchical block structure.

Let $\boldsymbol{A}$ denote the observed network and let T denote a fragmentation of the network nodes. The following general outline of a probabilistic generative process can be used to characterize a complex network with a hierarchical cluster structure.

1. Generate a rooted tree T where the leaf nodes corresponds to the vertices in the complex network,
$$T \sim p(T|\tau). \tag{17}$$
Each internal node in the tree corresponds to a cluster of network vertices.

2. For each internal node in the tree, generate parameters $\boldsymbol{\theta}$ that govern the probabilities of edges between vertices in each of its children,
$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|T, \rho). \tag{18}$$

3. For each pair of vertices in the network, generate an edge with probability governed by the parameters located at their common ancestral node in the tree,
$$\boldsymbol{A} \sim p(\boldsymbol{A}|\boldsymbol{\theta}, T, \xi). \tag{19}$$

Several existing hierarchical network models (Clauset et al., 2008; Herlau et al., 2012; Roy et al., 2007) are special cases of this approach with different choices for the distributions of T, θ, and $\boldsymbol{A}$. Inference in these models entails computing the posterior distribution over the latent tree structure,

$$p(T|A) = \int \frac{p(\boldsymbol{A}|\boldsymbol{\theta}, T)p(\boldsymbol{\theta}|T)p(T)}{p(\boldsymbol{A})} d\boldsymbol{\theta}. \tag{20}$$

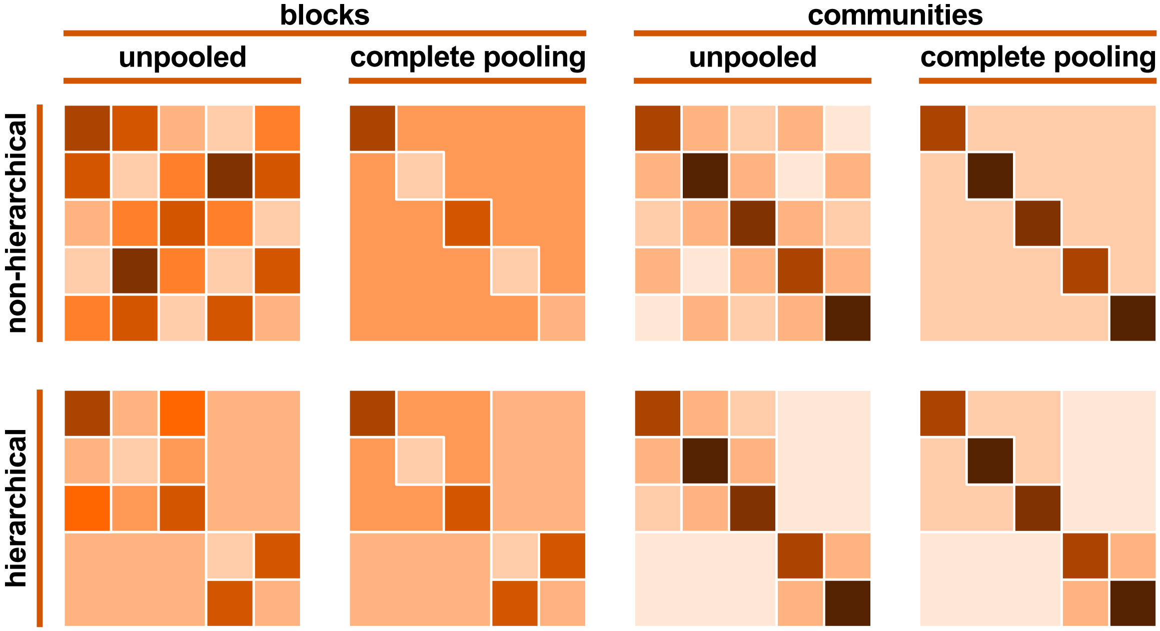

Figure 4: Illustration of different approaches to modelling (hierarchical) group structure in complex networks. The figures shows matrices of probabilities of links between groups in a network with five groups (darker color indicates higher link probability). Groups of nodes can be allowed to link to other groups with independent probabilities (denoted blocks), or restricted to have higher probability of links within than between groups (denoted communities). Furthermore, between-group link probabilities can be independent (unpooled) or shared (pooled) amongst all groups at each level of the hierarchy.

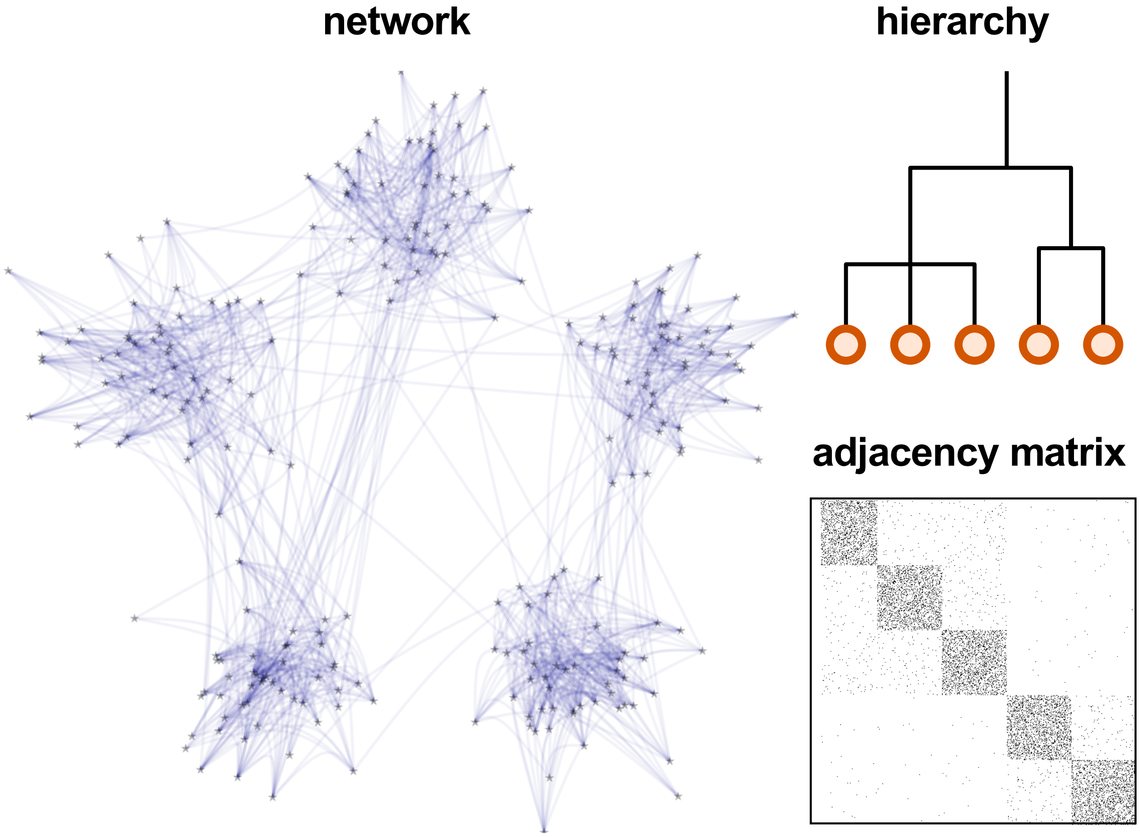

Figure 5: Simulated example of a complex network with hierarchical group structure. The network has five clusters; however, three of these (top and left) are more connected to each other than to the remaining two clusters, forming a super cluster. The goal of this work is to automatically detect such hierarchical structure and learn from data the number of clusters, their hierarchical organization, as well as the depth of the hierarchy.

In the following, we consider two hierarchical block models: An unpooled and a pooled model (see Figure 4). As a distribution over trees we use the Gibbs fragmentation model in Eq. (10). The likelihood for both models can be written as a product over all internal nodes in the tree,

$$p(\boldsymbol{A}|\boldsymbol{\theta}, T) = \prod_{\substack{B \in T \\ |B| \geq 2}} f_B(\boldsymbol{A}, \boldsymbol{\theta}_B, B). \tag{21}$$

Assume that the set B according to T fragments into the partition $\{B_1, B_2, \dots, B_k\}$ and ℓ, m denote indices of each fragment. The likelihood then has the following form,

$$f_B(\boldsymbol{A}, \boldsymbol{\theta}_B, B) = \prod_{\substack{B_\ell \in B \\ B_m \in B \\ l < m}} \prod_{\substack{i \in B_\ell \\ j \in B_m \\ i < j}} \text{Bernoulli}(\theta_{B,\ell,m}), \tag{22}$$

where the products go over each possible link between each possible pair of blocks. In the pooled model $\theta_{B,\ell,m} \equiv \theta_B$ are equal for all blocks, and in the unpooled models, $\theta_{B,\ell,m}$ are independent. We use independent Beta priors for the link probabilities,

$$p(\theta_{B,\ell,m}) = \text{Beta}(\rho^+, \rho^-). \tag{23}$$

The hyperparameters in our model are $\tau = \{\alpha, \beta\}$ and $\rho = \{\rho^+, \rho^-\}$. In all experiments these were fixed at $\alpha = \beta = \frac{1}{2}$ and $\rho^+ = \rho^- = 1$.

3.1 Implementation

Inference in the models is performed using Markov chain Monte Carlo sampling. Due to the conjugacy between the prior and likelihood for the link probabilities θ, they can be analytically marginalized allowing collapsed sampling of the tree.

We use the Metropolis-Hastings algorithm with subtree pruning and regrafting (SPR) proposals in which a subtree is removed from the tree and inserted in a new position. Assume k is a node removed from a tree T. Let T_k be the corresponding subtree rooted at k and $T_{\setminus k}$ the tree obtained by projecting out B_k. It is useful to distinguish between two types of insertion operations acting on a node h in $T_{\setminus k}$: In moves of type 1 the tree is modified by simply adding T_k as a child to node h. Notice that this require h to have at least two children, consequently, for the chain to be ergodic we must include moves of type 2 where $T_{\setminus k}$ is modified by replacing the subtree of $T_{\setminus k}$ rooted at h, $T_{\setminus k,h}$, with a new subtree with $T_{\setminus k,h}$ and T_k as its only children.

While one can simply select between all available insertion operations at random, the hierarchical organization of the network implies that it is rarely prudent to propose moves which move nodes far from where they are attached. We propose an alternative scheme where nodes are removed at random, but the set of allowed insertion moves is selected by taking the parent of the removed node, collecting all vertices in the reduced tree with a travel-distance less than or equal to two from the parent and forming the set of insertion moves of type 1 or 2 as they apply. In all simulations we choose between the two types of proposal moves with probability $\frac{1}{2}$.

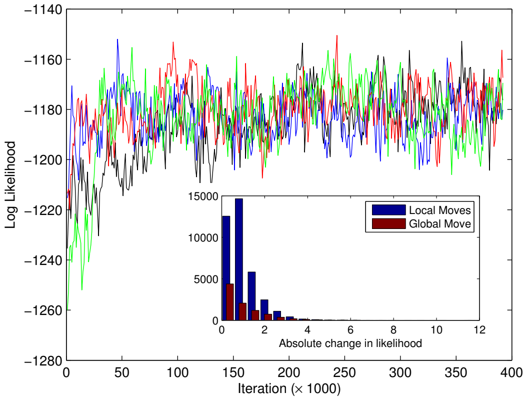

Figure 6: Log likelihood for 5 chains (shown in different color) running scan 1 of the diffusion datasets of figure 10. The frequent crossing of chains indicate reasonable mixing. Insert show histogram of (absolute) change in log likelighood for accepted moves of the two move classes, blue bars are local moves and red global. Moves to similar state not included.

By caching counts of links and non-links within each block throughout the tree, computing only the relevant updates of these counts as a move is accepted, both the pooled and unpooled model can be implemented efficiently. Removing or adding a subtree only changes cached terms associated with the path from the removed/added node to the root. While the unpooled model in the worst case contains $\frac{1}{2}n(n-1)$ θ-parameters (corresponding to the tree where the root immediately splits into n leafs), each is only associated with a single variable (the presence or absence of the single associated link) and their total contribution to the likelihood is linear in the total number of links and non-links and so one only need to keep track of these two values. In general, a large number of singleton clusters is not a computational problem, as only the sum of links and non-links between these clusters must be computed and updated.

In all experiments described in the following section, the sampler was run for 400'000 iterations. The first half of the samples were discarded for burn in, and the second half were subsampled by a facter of 1000, yielding 200 posterior samples. Figure 6 shows the result of running 5 different chains on the human brain connectivity dataset described later. As can be seen the likelihood of each chain frequently change value suggesting that the chains mix. The insert in the figure shows a histogram of the change in absolute value of log likelihood for the global and local proposal moves. The local moves are accepted more than twice as often as the global (0.093 compared to 0.036) and also contribute to significant change in log likelihood. These results were similar for all the conducted experiments.

4. Results

4.1 Simulations

The proposed models were evaluated and compared with the IRM model on four artificial networks each chosen with approximately the same overall difficulty but supporting different types of structure, see also Figure 7. The first synthetic network (*Communities*) is a network of strict community structure which can be well accounted for by all models. The second network (*IRM*) was generated according to the IRM model, the third (*Unpooled*) according to the unpooled hierarchical model and the fourth (*Pooled*) according to the pooled hierarchical model.

In Figure 7 it can be seen that the three models well infer structure in the networks they are designed for. We further see that the unpooled model is able to account for the structure of both the IRM model and the pooled model as it is closely related to the IRM model when forming a flat hierarchy while being more flexible than the pooled model when inferring hierarchical structure and is thereby able to account for the pooled hierarchical structures. Consequently, the unpooled model is able to infer the presence of hierarchical structure while reducing to a flat hierarchy corresponding to the IRM model when no such structure is supported by the data. The pooled model has a substantially reduced parameter space compared to the unpooled model. It is better able to identify structure when data indeed supports this type of hierarchical structure while it creates spurious results when the assumption of a pooled hierarchy fails: In the IRM and Unpooled networks, the pooled model clearly underfits.

In Table 1 we inspect the models ability to predict structure in multiple networks generated according to the true model parameters used to generate the networks in Figure 7. We trained each models on a single synthetic network of each type and used the trained model to predict 10 other synthetic networks generated from the same distribution. We evaluate the predictive performance in terms of the mean predictive log-likelihood and average area under curve (AUC) of the receiver operator characteristic (ROC) across the samples using link prediction on the complete network. The models' predictive performance are essentially equal except for the two cases where the pooled hierarchical model underfits.

4.2 Modeling hierarchical structure in real world networks

In order to qualitatively evaluate the proposed hierarchical models' ability to account for structure in real networks we consider the following four networks

- *NIPS:* The NIPS network is a binary graph with a total of 598 undirected links between the top 234 collaborating Neural Information Processing Systems (NIPS) authors in NIPS volumes 1 to 17 (also analyzed in (Miller et al., 2009; Mørup and Schmidt, 2012b)).
- *Football:* Network of American football games between Division IA colleges during regular season Fall 2000. The network consist of 115 colleges and 613 games (Girvan and Newman, 2002).
- *Les Miserable:* Network of the 254 co-appearances of 77 characters in the novel Les Miserables (Knuth et al., 1993).

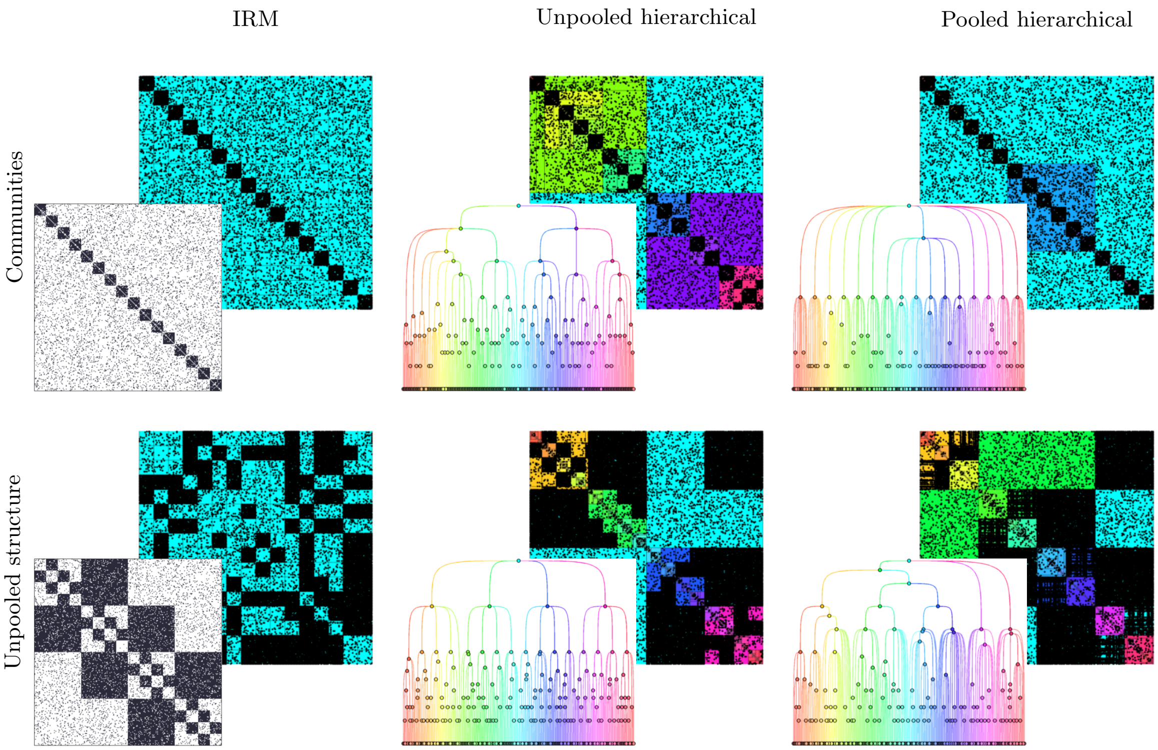

Figure 7: IRM, unpooled and pooled hierarchical modeling of four synthetic networks. The first network is generated to have strict community structure whereas the second network is generated according to the IRM model. The third and fourth networks are generated according to the unpooled and pooled hierarchical models respectively. Colors on the vertices of the inserted tree and on the corresponding adjacency matrix indicate where in the tree the parameters which explain the links reside.

Model \ Network	Communities		IRM		Unpooled		Pooled	
	$\log L$	AUC	$\log L$	AUC	$\log L$	AUC	$\log L$	AUC
IRM	-10710 ± 40	0.6769 ± 0.00138	-10700 ± 36.7	0.8964 ± 0.000572	-10720 ± 42.8	0.8989 ± 0.000751	-10680 ± 41.7	0.826 ± 0.000867
Unpooled	-10810 ± 19.4	0.675 ± 0.0008	-11010 ± 49.5	0.8947 ± 0.00077	-10780 ± 26.1	0.8995 ± 0.000435	-10790 ± 28	0.8258 ± 0.000684
Pooled	-10710 ± 23.4	0.6763 ± 0.000664	-16440 ± 23.4	0.8155 ± 0.000491	-15920 ± 42.6	0.8207 ± 0.000889	-10700 ± 25.5	0.8256 ± 0.000587

Table 1: Average predictive log likelihood ($\log L$) and AUC scores of the three models performance on the four types of networks generated. The models are trained on the networks generated in figure 7 and evaluated on 10 additional networks generated according to the true model parameters used to generate the data given in the figure.

- *Zachary:* Social network of 78 recorded friendships between 34 members of a karate club at a US university in the 1970s (Zachary, 1977).

The results of the modeling using the IRM model as well as the unpooled and pooled hierarchical models are given in Figure 8. All models appear to identify network homogeneities. In the NIPS and Les Miserable networks the IRM model extracts a large cluster that group nodes that are not well connected to each other into what more or less appear to represent a noise cluster. This has previously been reported in (Ishiguro et al., 2012) where it was proposed to extend the IRM model to explicitly account for these clusters representing noise. Rather than treating these nodes as coming from a noise cluster, both the unpooled and pooled hierarchical models are able to detect structure at a level of resolution that is substantially smaller than what the IRM model accounts for and terminate at a level where these nodes in fact form small groups of tightly connected communities. In addition both the pooled and unpooled models are able to represent structure in the graphs in terms of hierarchies and it is observed that many of the splits are multifurcating having three or more children. This well supports the notion that hierarchical structure go beyond the strict binary hierarchies considered in (Clauset et al., 2008; Roy et al., 2007; Roy and Teh, 2009). This has also been observed previously when modeling feature data by multifurcating hierarchies (Blundell et al., 2011). Thus, both the pooled and unpooled models identify prominent multifurcating hierarchical structures in the considered networks. These results support the existing literature arguing that many real world networks exhibit hierarchical structure (Simon, 1962; Ravasz and Barabási, 2003; Roy et al., 2007; Sales-Pardo et al., 2007; Clauset et al., 2008; Meunier et al., 2010).

4.3 Testing for hierarchical structure in whole brain structural connectivity

Brain networks are believed to exhibit hierarchical modularity, i.e. modules of the brain do not exist only at a single organizational scale but each module are further partitioned into submodules (Meunier et al., 2010). This type of hierarchical organization has been demonstrated to occur in both functional (Meunier et al., 2009; Ferrarini et al., 2009) and structural brain networks (Bassett et al., 2010). We presently investigate if our models indeed support the notion of hierarchical modularity in data of structural brain connectivity.

In Figure 9 we analyze the connectome of C. Elegans (Achacoso and Yamamoto, 1992; Watts and Strogatz, 1998) and the Macaque monkey right hemisphere (Hagmann et al., 2008). The C. Elegans network is the only complete connectome recorded consisting of the 306 neurons in the nematode worm Caenorhabditis Elegans (Achacoso and Yamamoto, 1992). The network forms a directed integer weighted graph having 8,799 connections (defined by synapse or gap junctions) (Watts and Strogatz, 1998). In our analysis we treat all edges as undirected and unweighted. The Macaque monkeys connectivity between 47 regions of the right hemisphere is estimated based on diffusion spectrum imaging (Hagmann et al., 2008). We consider the undirected unweighted network where a link denotes the existence of a fiber between two regions. The network has a total of 275 undirected links. From figure 9 it can be seen that both the IRM model as well as the unpooled and pooled hierarchical models are able to extract prominent structure defining network homogeneities. However, both the unpooled and pooled hierarchical models extract structures that are well

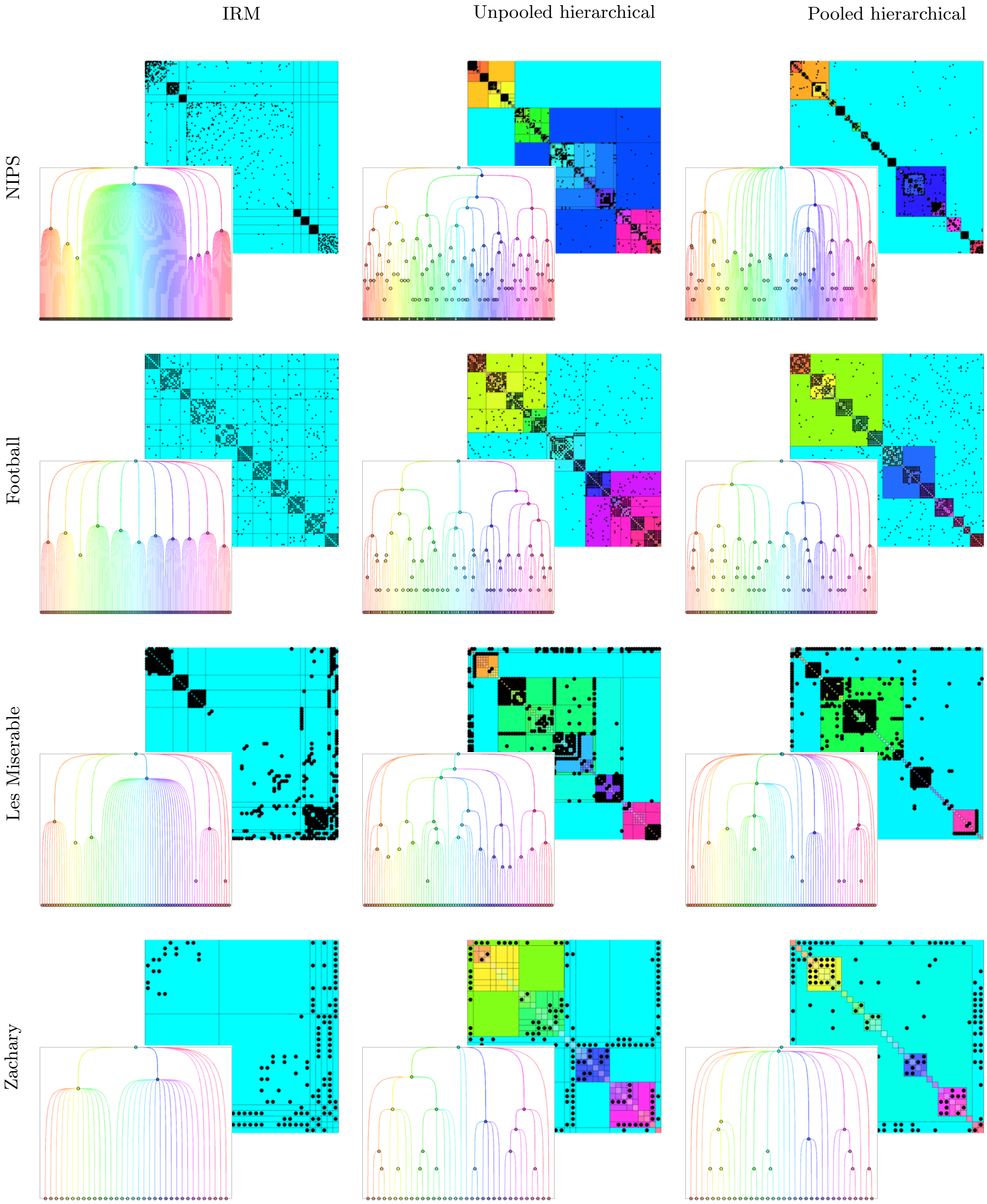

Figure 8: Analysis of the NIPS, Football, Les Miserable and Zachary networks by the IRM model as well as the unpooled and pooled hierarchical models.

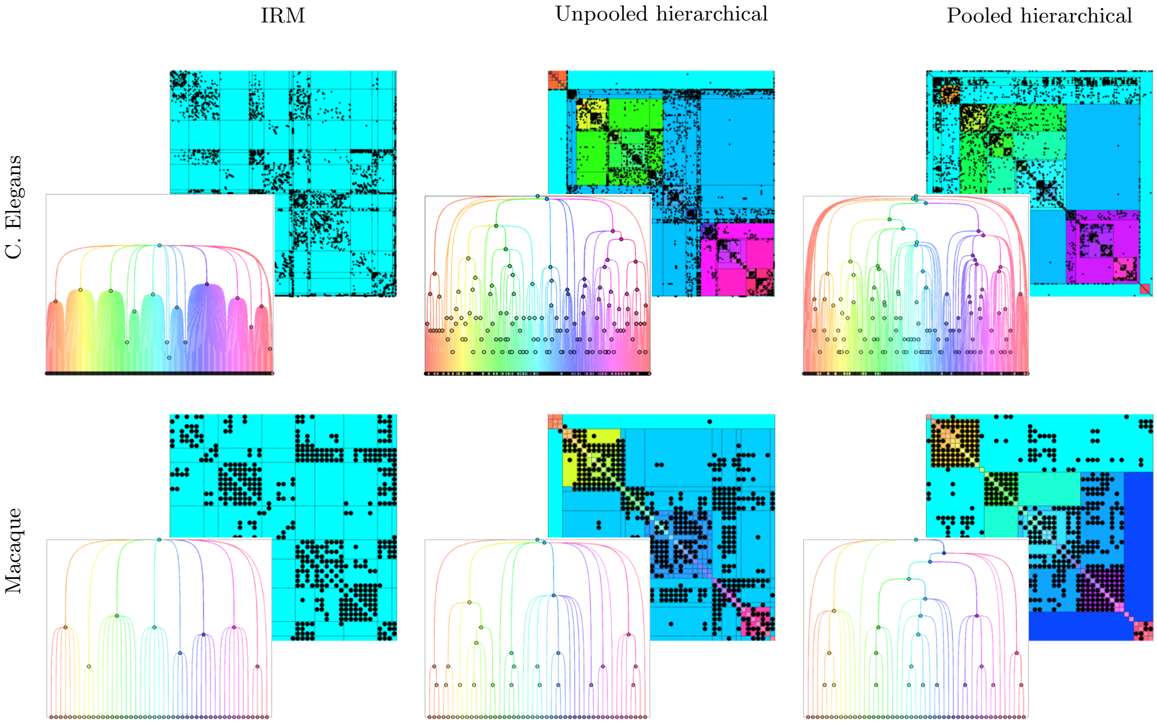

Figure 9: Analysis of the full connectome of C. Elegans (Achacoso and Yamamoto, 1992; Watts and Strogatz, 1998) as well as the Macaque monkey right hemisphere (Hagmann et al., 2008) by the IRM model as well as the unpooled and pooled hierarchical models.

in support of hierarchical modularity dividing the brain into parts and subparts of tightly connected groups of nodes.

In order to quantify if hierarchical structure is supported in structural human brain networks we consider the diffusion spectrum imaging data[2] described in (Hagmann et al., 2008) where we have access to multiple graphs defining structural connectivity across five subjects. The diffusion spectrum imaging has been used to map pathways within and across cortical hemispheres in five human participants where the first participant has been scanned twice. We consider the data at the resolution given by 66 anatomical gray matter regions (Hagmann et al., 2008). We threshold the graph such that a link exists if there is a non-zero weight in the connectivity matrix and we model the undirected binary networks where links indicate the existence of connectivity between two cortical regions. The results of the modeling of the networks by the IRM model as well as unpooled and pooled hierarchical models are given in Figure 10 where the analysis of the two separate scans of subject 1 is given. Both the IRM model as well as the pooled and unpooled hierarchical models extract prominent network homogeneities. However, from the results it is not clear if the hierarchical structure is significant compared to the IRM model.

[2] The data was downloaded from `http://connectomeviewer.org/viewer/datasets`.

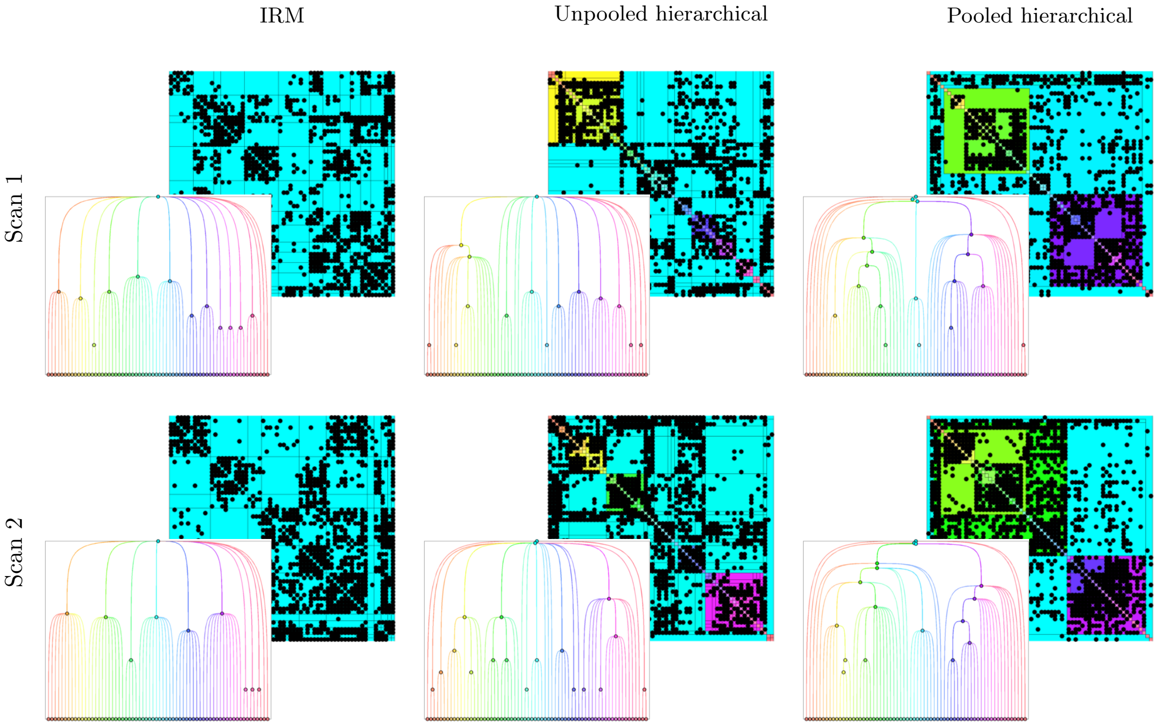

Figure 10: IRM as well as unpooled and pooled hierarchical modeling of the graphs derived from two separate diffusion imaging scans of subject 1.

To investigate this, we exploit that we have access to multiple scans, which we consider as independent samples of the “true” graph of cortical connectivity. As the first subject has been scanned twice we use these two scans to investigating how well the models are able to predict the graph derived from an independent scan of the same subject. In addition, we investigate how well the network models fitted to each subject generalize to the other subjects. We evaluate the predictive performance in terms of the mean predictive log-likelihood and average area under curve (AUC) of the receiver operator characteristic (ROC) across the samples using link prediction on the complete network of a subject (see Table 2). We included the AUC score as this is a common measure of predictive performance in networks, see also (Clauset et al., 2008; Miller et al., 2009). The IRM model performs slightly but not significantly better than the unpooled hierarchical model on all tasks. Both the IRM and the unpooled model substantially outperform the pooled hierarchical model on all tasks. Thus, the data does not support the hierarchical structure defined by the pooled hierarchy. On the other hand, as the unpooled hierarchical model is on par with the IRM model the inferred hierarchical structure has a relative benefits over the IRM model in giving an interpretable representation of how structural connectivity structure emerges at different scales.

5. Conclusion

We have proposed to use the Gibbs fragmentation tree process as a prior over multifurcating trees in two hierarchical models for relational data. An unpooled model where individual between-group interactions where independent and a pooled model where the interaction between groups at each level of the hierarchy were assumed identical. In the analysis of synthetic networks we observed that the models well identified the structure they were designed for. In real networks we found that the hierarchical models proposed were able to model structure at multiple levels and were thereby able to model structure in clusters that by the IRM model mainly resembled noise. Furthermore the hierarchical models were able to detect structure at a resolution terminating at a more detailed level than the IRM model. Thus, the two proposed hierarchical models seem to form useful frameworks for the modeling of structure emerging at multiple levels of networks and to infer from the data the number of levels of representations needed.

The analysis of brain connectivity data in C. Elegans network as well as the right hemisphere cortical connectivity of the Macaque monkey qualitatively gave some support for the notion of hierarchical modularity as proposed in (Meunier et al., 2009, 2010; Ferrarini et al., 2009; Bassett et al., 2010). However, our analysis of the human brain connectivity at the level of 66 cortical regions connectivity did not give evidence for the presence of a hierarchical structure. For predictive modeling, the unpooled model performed on par with the IRM model while providing a hierarchical account of the structure. The lack of support for the hierarchical structure might be attributed to the low resolution of this network. With only 66 cortical regions represented finer details reflecting hierarchical structure might not be visible. Thus, in future work we will analyze structural connectivity networks at a higher resolution such as the structural connectivity of 1000 regions in the data provided by (Hagmann et al., 2008).

Model \ Network	Within scan		Within subject		Between subjects	
	$\log L$	AUC	$\log L$	AUC	$\log L$	AUC
IRM	-745.271 ± 18.7209	0.912719 ± 0.00387353	-1054.27 ± 19.7743	0.822502 ± 0.0035924	-980.321 ± 10.207	0.840496 ± 0.00321595
Unpooled	-756.139 ± 21.4632	0.905174 ± 0.00702501	-1068.31 ± 36.3289	0.811378 ± 0.0131813	-1000 ± 12.8714	0.831984 ± 0.0041629
Pooled	-888.59 ± 18.2798	0.848447 ± 0.00480982	-1149.55 ± 14.0378	0.775658 ± 0.00478029	-1069.58 ± 11.9004	0.78661 ± 0.00439868

Table 2: Average log likelihood ($\log L$) and AUC scores within the same scan and the predictive log likelihood and link prediction AUC across the first subjects two scans (denoted within subject 1) and between all subjects. In parenthesis is given standard deviation across mean. (Within scan includes six samples (i.e., 5 subjects with the first subject having two independent scans), between scans include two samples and between subjects include $5 \cdot 4/2 = 10$ samples where we have used first scan of subject 1.

We believe hierarchical structure is indeed an important property of many networks including brain connectivity networks. In particular, hierarchical structure should be prominent in large scale networks where structure is likely to exist at multiple scales. Future work will focus on improving the proposed Markov chain Monte Carlo sampler for large scale inference by exploiting that the hierarchical structures inferred by the models admit sampling in parallel between the nodes belonging to separate children at given levels of the hierarchy. Furthermore, we envision the sampler at higher levels of the tree will benefit from Gibbs sampling across the multiple potential reconfigurations of the internal nodes at these higher levels.

Hierarchical modularity, i.e. systems that contain subsystems of more tightly connected nodes (that in turn may be defined by tighter connected subsystems etc.) are in the present unpooled and pooled models not explicitly accounted for. In fact, subsystems may be less densely connected than at their less detailed resolution as the density parameters are unconstrained. A benefit of keeping these parameters unconstrained is that it enable us to collapse the parameters during inference reducing the inference to sampling over tree structures. However, in future work we will aim at deriving models that explicitly accounts for hierarchical modularity using the Gibbs fragmentation tree process. Similar ideas have been proposed (Mørup and Schmidt, 2012b) for non-hierarchical models where the within-community density is constrained be higher than between community densities while admitting analytic integration of the majority of the parameters specifying between cluster interactions. We envision this can be accomplished while preserving that the model is consistent and exchangeable by exploiting ideas from (Steinhardt and Ghahramani, 2012).

The proposed framework for modeling hierarchical structure in relational data admit formal testing of hierarchical structure in networks such that the unpooled model reduce to a representation closely related to the IRM models representation when the data does not support hierarchical structure. We believe this forms a useful tool for researchers investigating and validating whether their relational systems are defined by hierarchical structures.

References

T.B. Achacoso and W.S. Yamamoto. *AY's Neuroanatomy of C. elegans for Computation*. CRC, 1992.

Y.Y. Ahn, J.P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

David Aldous, Illdar Ibragimov, and Jean Jacod. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer Berlin / Heidelberg, 1985. ISBN 978-3-540-15203-3.

Danielle S. Bassett, Daniel L. Greenfield, Andreas Meyer-Lindenberg, Daniel R. Weinberger, Simon W. Moore, and Edward T. Bullmore. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput Biol*, 6(4):e1000748, 04 2010. doi: 10.1371/journal.pcbi.1000748. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1000748.

David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Neural Information Processing Systems, Advances in*, volume 16. Bradford Book, 2003.

David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2007. URL `http://arxiv.org/abs/0710.0845`.

C. Blundell, Y. W. Teh, and K. A. Heller. Discovering non-binary hierarchical structures with Bayesian rose trees. In K. Mengersen, C. P. Robert, and M. Titterington, editors, *Mixture Estimation and Applications*. John Wiley & Sons, 2011.

Ronald L Breiger, Scott A Boorman, and Phipps Arabie. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12(3):328 – 383, 1975. ISSN 0022-2496. doi: 10.1016/0022-2496(75)90028-0. URL `http://www.sciencedirect.com/science/article/pii/0022249675900280`.

A. Clauset, C. Moore, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

L. Ferrarini, I.M. Veer, E. Baerends, M.J. van Tol, R.J. Renken, N.J.A. van der Wee, D.J. Veltman, A. Aleman, F.G. Zitman, B.W.J.H. Penninx, et al. Hierarchical functional modularity in the resting-state human brain. *Human brain mapping*, 30(7):2220–2231, 2009.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.

P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C.J. Honey, V.J. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. *PLoS biology*, 6(7):e159, 2008.

T. Herlau, M. Morup, M.N. Schmidt, and L.K. Hansen. Detecting hierarchical structure in networks. In *Cognitive Information Processing (CIP), 2012 3rd International Workshop on*, pages 1 –6, may 2012. doi: 10.1109/CIP.2012.6232913.

Jake M Hofman and Chris H Wiggins. Bayesian approach to network modularity. *Physical Review Letters*, 100(25):258701, 2008. URL `http://link.aps.org/doi/10.1103/PhysRevLett.100.258701`.

P.W. HollandKathryn Blackmond and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

K. Ishiguro, N. Ueda, and H. Sawada. Subset infinite relational models. In *AISTATS*, 2012.

Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, pages 381–388, 2006.

David A. Knowles and Zoubin Ghahramani. Pitman-Yor Diffusion Trees. In *Uncertainty in Artificial Intelligence, Proceedings of the International Conference on*, 2011.

D.E. Knuth, D.E. Knuth, and D.E. Knuth. *The Stanford GraphBase: a platform for combinatorial computing*. AcM Press, 1993.

Peter McCullagh, Jim Pitman, and Matthias Winkel. Gibbs fragmentation trees. *Bernoulli*, 14(4):988–1002, 2008.

D. Meunier, R. Lambiotte, A. Fornito, K.D. Ersche, and E.T. Bullmore. Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics*, 3, 2009.

D. Meunier, R. Lambiotte, and E.T. Bullmore. Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, 4, 2010.

K.T. Miller, T.L. Griffiths, and M.I. Jordan. Nonparametric latent feature models for link prediction. *Advances in Neural Information Processing Systems (NIPS)*, pages 1276–1284, 2009.

M. Mørup and M. N. Schmidt. Bayesian community detection. *Neural Computation*, 24(9): 2434–56, 2012a.

M. Mørup and M.N. Schmidt. Bayesian community detection. *Neural Computation*, 24(9): 2434–2456, 2012b.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+, February 2004. doi: 10.1103/PhysRevE.69.026113. URL http://dx.doi.org/10.1103/PhysRevE.69.026113.

E. Ravasz and A.L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, 2003.

E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

D.M. Roy and Y.W. Teh. The mondrian process. In *Adv. in Neural Inform. Processing Syst*, volume 21, page 27, 2009.

D.M. Roy, C. Kemp, V.K. Mansinghka, and J.B. Tenenbaum. Learning annotated hierarchies from relational data. *Advances in neural information processing systems*, 19:1185, 2007.

M. Sales-Pardo, R. Guimera, A.A. Moreira, and L.A.N. Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104 (39):15224, 2007.

H.A. Simon. The architecture of complexity. *Proceedings of the American philosophical society*, 106(6):467–482, 1962.

T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

Jacob Steinhardt and Zoubin Ghahramani. Flexible Martingale Priors for Deep Hierarchies. *International Conference on Artificial Intelligence and Statistics AISTATS*, 2012.

D.J. Watts and S.H. Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393 (6684):440–442, 1998.

Z. Xu, V. Tresp, K. Yu, and H.P. Kriegel. Infinite hidden relational models. In *Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, 2006.

W.W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.