## PhD THESIS DECLARATION

The undersigned

SURNAME Nazarov

FIRST NAME Maxim
PhD Registration Number 1465723

# Thesis title: Bayesian Modeling of Dynamic Network Data

PhD in Statistics

Cycle 25

Candidate's tutor Sonia Petrone
Year of thesis defence 2014

## DECLARES

Under his responsibility:

1) that, according to Italian Republic Presidential Decree no. 445, 28th December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30th April 1999, to keep a copy of the thesis on deposit at the "Bilioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary embargo protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

3) that the Bocconi Library will file the thesis in its "Archivio istituzionale ad accesso aperto" (institutional registry) which permits online consultation of the complete text (except in cases of a temporary embargo);

4) that, in order to file the thesis at the Bocconi Library, the University requires that the thesis be submitted online by the candidate in unalterable format to Società NORMADEC (acting on behalf of the University), and that NORMADEC will indicate in each footnote the following information:

   - thesis Bayesian Modeling of Dynamic Network Data;
   - by Nazarov Maxim;
   - defended at Università Commerciale "Luigi Bocconi" – Milano in 2014;
   - the thesis is protected by the regulations governing copyright (Italian law no. 633, 22th April 1941 and subsequent modifications). The exception is the right of Università Commerciale "Luigi Bocconi" to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis submitted online to NORMADEC is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents

of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal, and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results, and is not subject to patent or protection registrations, and therefore not subject to an embargo;

Date 14 November 2013

SURNAME Nazarov
FIRST NAME Maxim

# Contents

1 Introduction 1

2 Review of network models 3

2.1 Introduction . . . 3

2.2 Data description and presentation . . . 4

2.2.1 INFOCOM'06 dataset . . . 5

2.2.2 Basic terminology and properties . . . 5

2.2.3 Visualizing data . . . 7

2.2.4 Network summary statistics . . . 8

2.3 Main principles and goals of statistical network modeling . 9

2.4 Overview of approaches for statistical network modeling . . 10

2.4.1 Erdős and Rényi model . . . 11

2.4.2 ERGMs . . . 12

2.4.3 Latent variable models . . . 13

2.4.4 Models for dynamic network data . . . 18

2.5 Final remarks . . . 20

3 Functional representation of Markov exchangeable sequences 23

3.1 Introduction . . . 23

3.2 Functional representations . . . 24

3.3 Exchangeability . . . 25

3.4 Partial exchangeability . . . 28

3.5 Markov exchangeability and successors . . . 30

3.5.1 Representations for Markov exchangeability . . . 32

3.6 Row-column exchangeability . . . . . . . . . . . . . . . . 34
3.7 Partial Exchangeability and RCE . . . . . . . . . . . . . 37
3.8 Combining RCE and ME . . . . . . . . . . . . . . . . . . 39

**4 Dynamic latent distance model for network analysis 43**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . 43
4.2 Latent distance model for static networks . . . . . . . . . . 45
4.3 Towards dynamics: modeling network "cohesion" . . . . . 47
4.3.1 Infinite hidden Markov models . . . . . . . . . . . 49
4.4 Our proposed model . . . . . . . . . . . . . . . . . . . . 51
4.5 Estimation . . . . . . . . . . . . . . . . . . . . . . . . . 54
4.5.1 Sampling the hidden state sequence ($S_t$) . . . . . . 56
4.5.2 Sampling the iHMM parameters $\pi$ and $\eta$ . . . . . . 57
4.5.3 Sampling the "cohesion" parameters $\sigma_k^2$ . . . . . . . 58
4.5.4 Sampling the latent positions $Z$ . . . . . . . . . . 58
4.5.5 Sampling the covariate coefficients $\beta$ . . . . . . . . 59
4.6 Numerical illustrations . . . . . . . . . . . . . . . . . . . 59
4.6.1 Example 1: simulated data . . . . . . . . . . . . . 59
4.6.2 Real data . . . . . . . . . . . . . . . . . . . . . . 64
4.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . 66

**5 Final remarks and future directions 69**
5.1 Extending the static part of the dynamic latent distance model 69
5.2 Stronger individual dependence . . . . . . . . . . . . . . . 70
5.3 Dynamic eigenmodel . . . . . . . . . . . . . . . . . . . . 71
5.4 Models from the RCE-ME representation . . . . . . . . . . 72
5.5 Other directions . . . . . . . . . . . . . . . . . . . . . . . 73

# Acknowledgements

I would like to express my gratitude to the people, who made these last four years very special and exciting experience. Most importantly, to my advisor Sonia Petrone, for guiding me and continuously supporting. And for pushing me to write things up. My thanks also go to all professors in Bocconi, who introduced me to a lot of new things and ignited my interest.

Also I am grateful to my fellow PhD students, who were around during these four years, for the happy moments we spent together, whether inside the University or on our weekend trips.

I thank Isadora for her help with Chapter 4 and useful insights on how to present my work, and Daniel for the fruitful discussions of concepts presented in Chapter 3.

My brief visiting period at University of Cambridge was a useful and invigorating experience, and I would like to thank Sara and professor Ghahramani who made this possible to happen. I had a chance to discuss my work with several members of the Machine Learning lab, and it was really useful to hear other points of view on the problems discussed in the thesis.

I thank my family for always believing in me, supporting my choices, and encouraging me to never give up. Finally, my most sincere thanks go to Polina for being an exceptional friend, partner and colleague.

## Abstract

Networks are abundant in the world around us. Network and relational data arise in a variety of fields and disciplines and there is a vast choice of methods for analysis. In this work we study probabilistic and statistical approaches for analyzing such data.

Most of the literature dedicated to networks deals with the static case, when only one snapshot of network is observed. While in practice a lot of processes in life are continuously changing and evolving. So the big challenge is modeling dynamic evolution of the network data.

Idea of the thesis is to follow this aim, starting from studying underlying probabilistic notions, that act as founding stones for constructing dynamic statistical models. New representation theorems are proved for assumptions of Markov exchangeability and its combination with array exchangeability. After that, a new statistical model, that extends the latent distance model, is proposed for network dynamics, and illustrated on examples.

# Chapter 1

# Introduction

This thesis studies modeling static and dynamic network data. Network data consist of measured relations between pairs of actors, and, as such, is different from data in other forms. The implication is that there is a dependence across actors in the network, and particular assumptions are needed to model such data.

As network data are abundant around us, interest in modeling arises in many fields. The first scientific works on networks appeared in the first half of the 20th century, in the social sciences area. The mathematical basis started actively developing from 1950s, and currently the field is growing and bringing together people from different domains. In recent years contributions were made from a variety of disciplines, such as biology, computer science, statistical physics, sociology and statistics. While there are many different approaches to analyzing networks, here we discuss probabilistic and statistical models, focusing on binary data.

In many cases, the observed data represents a “snapshot” of a network at one particular moment. While in fact real-life networks are often dynamic. Literature for the static network modeling is very rich. We focus on a particular class of models that use latent variables. This class includes several proposals with different properties, but they can be put under a unifying framework, based on reasonable assumptions of invariance in distributions of the residuals in the model. Dynamic modeling is rapidly evolving, but still highly open field. The problems concern discovering changes in the structure of networks, and interpreting them to explain changes in relations between individual actors.

Our motivations include exploring theoretical foundations of statistical network models, that can shed light on better ways of defining new models. We explore the relevant probabilistic concepts, starting from notions of exchangeability for sequences and random binary matrices. New results are provided for Markov exchangeability, notion that can be applied in dynamic modeling, and its combination with array exchangeability. Functional representation theorems are developed for these cases.

From a more applied side, we study latent variable models, and possibilities to introduce dependence along time in these models. In particular the latent distance model, based on the idea of proximity in the "social space", is extended to the dynamic case by means of an infinite hidden Markov model. Our proposed model allows to capture changes in the overall cohesion of the network over time.

The thesis is organized as follows.

In Chapter 2, we introduce problems in network modeling, discuss methods of describing data and give a review of network modeling approaches, with large focus on the latent variable models and the unifying framework for them. Then, Chapter 3 is devoted to underlying probabilistic concepts. An overview of representation theorems for various notions of exchangeability is provided, after that we present our contributions on functional representations for Markov exchangeable sequences, and exchangeable arrays of such sequences. In Chapter 4 an original dynamic model is described and illustrated. Finally, we conclude in Chapter 5 with discussion of possible extensions for our proposals, and future directions of research.

# Chapter 2

# Review of network models

In this chapter we describe network data, particular properties and challenges emerging in their analysis. After that, a concise review of probabilistic approaches to network data modeling is provided.

## 2.1 Introduction

The study of networks started around the middle of the 20-th century. Nowadays network data are abundant and arise in a variety of different fields, from social networks of people or organizations to networks of protein and gene interactions. The field of network analysis developed as an interdisciplinary one and there exist many approaches from different angles. The contributions come from such disciplines as biology, computer science, statistical physics, sociology, probability and statistics. In particular, statistical analysis of such data is an active and evolving field, started from probabilistic models of 1950s.

One of the most fruitful directions is the study of “social networks”, relations between social entities, such as people, groups, communities. It comes from the social and behavioral science literature, where many topics of interest can be better described and understood in relational terms. We will not limit the discussion to necessarily “social” networks, but will speak about network and relational data in their generality. But, of course, social networks have some interesting properties that arise specifically from their nature.

The interest in networks and, in particular, the development of new methods has emerged and grew at impressive rate in the recent years, because of great amounts of such data that becomes easily accessible, and growing computational power, that allows to handle and analyze vast amounts of data in reasonable time. Almost all the interactions in the world are recorded on a continuous basis in one way or another, and the possibilities of analyzing this data are unlimited.

## 2.2 Data description and presentation

As we said, network data can arise in a variety of different fields. Some examples include friendship, working subordination and other relations between people, volume of international trade or state of conflict between countries, co-authorship of papers between scientists, protein–protein, gene–protein and gene–gene interactions in systems biology to name just a few.

What is common for all these examples is the nature of the data and their main components. Objects of interest are relations between actors, and not the actors themselves. In general, we call *network data* any data that is obtained on the set of interacting actors, and comprises of relations between them. We will also use the term “relational data”. So the main components are actors (individuals, nodes) and measurements of relations (links, ties) between the pairs of them. Actors and their relations are considered interdependent, and so network methods focus on these collections of individuals and links between them; hence some new methods are needed.

But before discussing methods, we need to introduce some notation, describe the properties of different networks and possibilities of representing the data. To make the presentation easier, we will demonstrate the ideas on an example dataset, that is later used for analysis in Chapter 4.

### 2.2.1 INFOCOM'06 dataset

The data, collected by Chaintreau et al. [2007] and downloaded from Scott et al. [2009], consist of measurements of contacts on an INFOCOM workshop in 2006. All participants were given proximity sensors, small wearable devices, to detect their interactions during the conference days. The two devices can "see" each other, if they are close enough, and then the interaction is recorded. So relations here are contacts of participants of the conference. The interactions were observed during the 4 days of the conference on 78 individuals, students and professors. This is an example of time-evolving network, where the interactions are changing over time, while the set of actors is not.

Of course, contacts between people can be described by a continuous process, but due to the way of recording, the data are discrete. To remove the noise and to simplify the dataset, the contacts were aggregated into hourly time periods. So, the relation between two individuals at hour $t$ was recorded, if there had been at least one contact between them during this hour. Also here, by design of the experiment, the contacts are supposed to be reciprocated, and the dataset was cleaned of errors in measurement, to conform. As the recorded time runs from 17:00 on the first day to 15:00 on the 4th day, that gives 94 aggregated time points. The total number of links ranges from 0 to 918 with the mean of 311 contacts per hour.

In the next subsections of this chapter, we will address this dataset as "our example".

### 2.2.2 Basic terminology and properties

As was already hinted by different examples, in the relations in question, pairs of actors can be ordered, giving a directed relation, or unordered, that gives an undirected relation. A relation can be directed or undirected by the nature of the data, or it may come from limitations of the data collection process. A special case is the possible presence of self-links, which does not make sense in some datasets (such as friendship), but can be quite common in others (such as protein interactions).

In our example, as already mentioned, the relations are undirected, meaning that each tie is reciprocated. This also holds true for such kinds of relations as people's friendship or being coauthors of a paper.

Next, the measured relation can be on a continuous or discrete scale. In simple, and the most common case, relations in question are binary, encoding the presence or absence of some property. Our example is in fact the case, and the property in question is the indication of having a contact in the particular hour. Further in the thesis, we limit our studies to the binary case, while noting that for some models, extension to the general case is immediate.

As we said, our example is time-evolving, or dynamic. If we observe the network at one point in time, or if it is produced by the aggregation over some period, it is called static. A lot of models developed in the literature are for the static case, as the additional time dimension adds to the complexity of the data.

To formalize the description of the data, often the language of graph theory is used, as graphs are a quite natural way of representing relational data. A graph $\mathcal{G} = (V, E)$ consists of a set of vertices (or nodes) $V = \{v_1, \dots, v_N\}$ and a set of edges (or links) $E = \{e_1, \dots, e_K\}$, where each $e_k$ is a pair of vertices $(v_i, v_j)$, that can be either ordered or not, and possibly include a weight; but as said, we limit our attention to binary networks. For most of the uses it will be enough to consider finite graphs, but some notions defined for infinite graphs are used in Chapter 3. Relevant definitions from the graph theory are also useful, such as degree of a vertex, diameter of a graph and so on. An extensive description of these and other notions can be found in the comprehensive book by Wasserman and Faust [1994, Chapter 4].

The alternative representation for a graph is the adjacency matrix. In the social science literature the corresponding object for social networks is called "socio-matrix". The adjacency matrix $Y = (Y_{i,j})_{1 \le i,j \le N}$ is constructed as follows: for all pairs $(i, j)$ set $Y_{i,j} = 1$ if $(v_i, v_j) \in E$ and $Y_{i,j} = 0$ otherwise. For undirected graphs, by assumption $Y_{i,j} = Y_{j,i}$, and in fact only the upper (or lower) triangle part of the matrix $Y$ is considered. As for the diagonal entries $Y_{i,i}$, often they are left undefined or considered as structural zeros.

### 2.2.3 Visualizing data

Data visualization is an important step both for preliminary exploratory analysis, and for reporting discovered structure and patterns. Clearly visualizing a large and possibly evolving dataset is a difficult task, and the methods described here are suitable, mostly, for relatively small graphs.

For the adjacency matrix, the visualization is quite straightforward. For binary data typically black squares are put when $Y_{i,j} = 1$, that is, an edge is present, and white squares when $Y_{i,j} = 0$, that is, edge is absent. Applying to our example for $t = 20$, we obtain the matrix on the left in Figure 2.1. While reordering rows and columns according to the degree of vertices gets us the matrix on the right. Other reordering schemes may be applied to emphasize highly-connected clusters that may be present in the data. So the perception can be changed a lot by a simple reordering of rows and columns. We will see in Chapter 3 that in many cases the distribution of a network can be assumed to be invariant with respect to permutations of rows and columns.

Another possible visual representation is a layout of the graph structure of nodes and links as points and lines on the plane. The chosen layout, i.e. relative positions of vertices and edges for the given graph object, can again play a big role in the perception and interpretation of the data. There are deterministic algorithms that position vertices and edges according to some pre-specified criteria, which usually include simplicity and readability of the resulting figure. Other algorithms try to highlight some desired properties of the graph, such as clustering or grouping of the nodes. Without going into much details, for the INFOCOM example four different layout algorithms are illustrated in Figure 2.2. The first is a random layout, the second one is the so called “force-directed” algorithm of Fruchterman and Reingold [1991], the third one is a layout based on singular value decomposition of the adjacency matrix, and the last one is

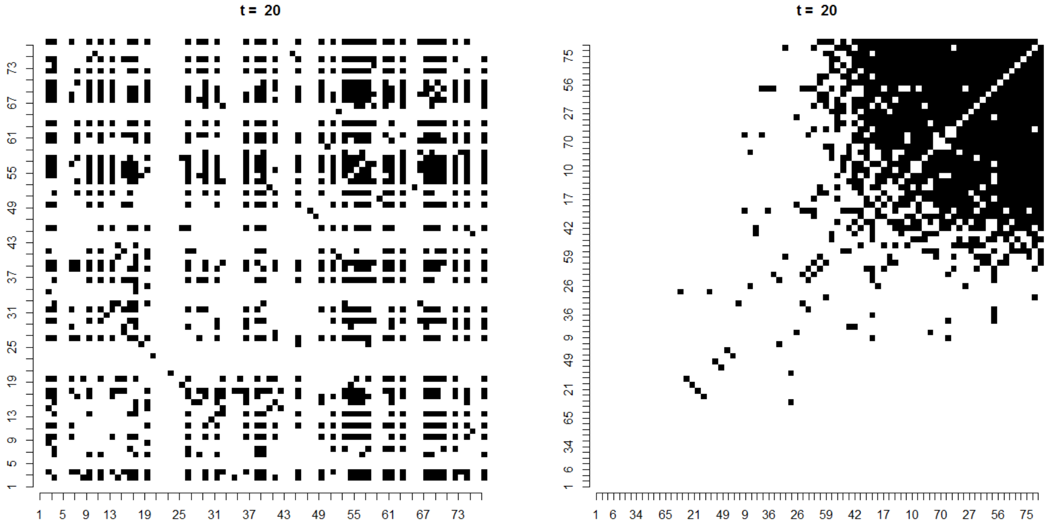


Figure 2.1: Adjacency matrix for the INFOCOM dataset at time $t = 20$, unsorted (left) and sorted by nodal degree.

based on multi-dimensional scaling of the shortest path distance matrix.

As we can see, the resulting graphs look very different. Good starting point for selection of appropriate algorithms is a recent review of the tools for graph visualization by Von Landesberger et al. [2011]. See also Salter-Townshend et al. [2012].

Some statistical models, such as the latent distance model by Hoff et al. [2002], that will be described in Section 2.4, provide model-based graphical representation of the data.

### 2.2.4 Network summary statistics

Another useful way of describing and summarizing network data is via summary statistics. They can be selected to help characterizing some local properties of the nodes of the network, or instead summarize the whole network. One simple example is the *degree* of a node. This is the number of edges to which a node belongs. If the graph is directed, then we can speak about the in-degree and out-degree for each node.

In addition to individual degrees, the total number of edges (or proportion of maximal number of edges) in a graph can give an idea of its overall density. For dynamic data, we can look at the evolution of this density (see Figure 2.3 for our data example).

Other common summaries to consider are the number of triangles and in general cliques, and stars. A $k$-clique is a subgraph, in which every node is connected with each other. So a triangle is a 3-clique. While a $k$-star is a subgraph with $k + 1$ vertices such that one of them is connected to all others with exactly one edge, i.e. there is one vertex with degree $k$ and $k$ vertices with degree of 1 each.

There are other characteristics that can be considered as summary statistics, good reference is the book by Wasserman and Faust [1994]. Several models based on the sufficient summary statistics are described in Section 2.4.

## 2.3 Main principles and goals of statistical network modeling

One of the main aims of the statistical analysis of network data is discovering the structure explaining the relations between actors. This may mean identifying groups or clusters, identification of important nodes, discovering patterns. Other goals may include prediction of missing or unobserved links. For dynamically evolving networks the task is even more difficult, as we may be interested in the evolution of these underlying network structures over time and in predicting future links in this context.

Other problems may involve studying the spread of information or, for example, decease along the network. This is conceptually a different problem than the one we consider, so we will not go into details. Our setting is that we observe a network of relations either at one point in time (static) or evolving. By evolution we imply that the set of edges may change, while the set of vertices is fixed.

When besides observing relations, there is some additional individual or pairwise information (covariates), we want to explain the links from these covariates.

An important feature of network data is that they are highly dependent, as the observations are on pairs of actors. There are some common types of dependence occurring due to this in a variety of real-life networks. These include:

Symmetry or *reciprocity* for directed data. Even if the graph is directed, meaning that $Y_{i,j}$ is not necessary equal to $Y_{j,i}$, the links tend to be reciprocated with high probability.

*Transitivity*, i.e. higher probability to have link $Y_{i,k}$ if there exist links $Y_{i,j} = 1$ and $Y_{j,k} = 1$. Informally "a friend of my friend is likely to be my friend".

*Homophily by attributes* which is tendency of actors with similar characteristics (covariates) to have higher probability of a link.

*Stochastic equivalence*, when nodes in one group have similar patterns of connectivity.

*Clustering*, i.e. existence of densely connected groups, with few connections between clusters, that may be caused by unobserved attributes.

*Degree heterogeneity*, that describes the existence of more "popular" nodes, and less popular ones.

When constructing the statistical model, the aim is to be able to capture the particular properties typical to the network in question. While modeling dependence, it is still natural to have a form of symmetry across actors. Assuming exchangeability and partial exchangeability for actors motivates a large class of models, *latent variable models*, that will be introduced in the next section.

## 2.4 Overview of approaches for statistical network modeling

Here we provide a concise overview of probabilistic models introduced for the network modeling, starting from the pioneer work of Erdős and Rényi. We do not discuss non-probabilistic approaches, such as game theoretic approach by Skyrms and Pemantle [2000] and others. A useful reference here is the book by Easley and Kleinberg [2010].

First consider modeling static network data. We remind that our focus is on binary network data, i.e. the relations are either link or no link. Let the data be encoded in the array $Y = (Y_{i,j})_{i,j=1\ldots N}$. We may also have additional covariate information for actors or pairs, that we denote by $\mathbf{X} = (X_{i,j})_{i,j\geq 1}$ where $X_{i,j} = (X_{i,j,k})_{k=1,\ldots,p}$ for each pair $(i,j)$.

For modeling of the static data, approaches provided in the literature can be broadly divided into two classes: so called "classical approach", where the presence of a link between two nodes depends on the network structure (examples include Erdős-Rényi random graph model, the so called p1 and p2 models, and exponential family random graph models) and the "latent variable approach", where the presence of a link between two nodes depends on underlying latent variables (examples are stochastic blockmodels and latent distance models).

### 2.4.1 Erdős and Rényi model

One of the earliest examples of probabilistic models for networks is the "random graph" model by Erdős and Rényi [1959]. It describes a binary graph, constructed on a set of $N$ labeled nodes with $K$ edges, by selecting uniformly at random among the set of all possible such graphs. Hence there are $\binom{\binom{N}{2}}{K}$ possibilities for undirected graphs and $\binom{N(N-1)}{K}$ for directed. This model is denoted by $G(N,K)$.

Another way to define a random graph was proposed independently by Gilbert [1959]. The graph is still defined on $N$ nodes, but instead of specifying the total number of edges, the probability of an edge between any two nodes is fixed. The presence of each link is represented by a Bernoulli random variable with parameter $\theta$, and it is independent of the presence of other links. Common notation for this model is $G(N,\theta)$

These two formulations are quite similar, and can be related by taking $\theta = \frac{K}{\binom{N}{2}}$. In this case expected probability of a link in the $G(N,K)$ model is $\theta$, while expected number of links in the $G(N,\theta)$ model is $K$. The $G(N,\theta)$ formulation is more common in the literature, as the independence assumption for edges makes analysis easier.

In terms of adjacency matrix, $P(Y_{i,j} = 1) = \theta$ independently for all pairs $(i, j)$. Then the distribution of a particular adjacency matrix $Y$ is:

$$P(Y|\theta) = \prod_{i,j} \theta^{Yi,j}(1-\theta)^{(1-Y_{i,j})},$$

where the product is taken on appropriate set of $(i, j)$ depending on whether the graph is supposed to be directed or not.

This model is quite simple, and has been studied extensively. Erdős and Rényi [1960] studied the asymptotic behavior of such graphs, depending on the relation between $N$ and $\theta$ (or $K$).

Due to this simplicity, it is not really suited for statistical analysis of the network data, as it does not assume any structure in the data and fails to capture any possible dependencies mentioned in the previous section. Furthermore, for the networks generated from the model, all vertices have approximately the same degree, which is rarely the case in real-life networks.

### 2.4.2 ERGMs

After the work of Erdős and Rényi, their model has been extended to be more suitable for statistical network modeling. Models that are still based on some network structure or statistics are sometimes termed “classical”, and include the following ones.

The $p_1$ and $p_2$ models were introduced by Holland and Leinhardt [1981] and Duijn et al. [2004], adding further parameters for the Erdős-Rényi model, to make it more expressive. These models, while being able to capture dependencies on pairs of actors, cannot capture properties referring to more than two actors, such as transitivity.

Another family of models are *Exponential family Random Graph Models* (ERGMs). These models are extension of the Markov graphs idea by Frank and Strauss [1986], and are defined in terms of joint distribution of all edges. Such network distribution is parametrized by some set of sufficient statistics and the probability distribution is assumed to be in the exponential family.

$$P(Y = y|\theta) = \exp(\theta' u(y) - \psi(\theta)),$$

where $\theta$ are parameters, $u(y)$ is a vector of sufficient statistics, that capture features of interest, and $\psi(\theta)$ is a normalizing constant. Typical choices of statistics include the number of edges, $k$-stars or $k$-cliques for different values of $k$.

The major challenge of these models is in the intractability of normalizing constant, as it involves summation over all possible networks with given sufficient statistics. There are computational methods developed to address this issue; a good review of them is provided in Hunter et al. [2012]. Also, the ERGMs may suffer from the issue of *degeneracy*, when large probability mass is concentrated on a very small proportion of the possible networks. This can be addressed by a careful choice of the sufficient statistics. The discussion of this issue is also included in Hunter et al. [2012], and recent paper by Chatterjee and Diaconis [2013].

### 2.4.3 Latent variable models

While ERGMs are useful for modeling global network characteristics, there exists another wide class of models, allowing to capture individual characteristics as well. This class of models is based on the assumption of presence of an additional (latent) layer of variables. Some additional assumptions are put on these variables, in order to capture the structure and dependencies of the networks. Most latent variable models also make the assumption that links between actors (or dyads for directed networks) are conditionally independent given the corresponding latent variables.

Many models proposed in the literature can be described as dependent on latent variables. Seemingly unrelated, they can be, in fact, put under a general framework motivated by certain symmetry considerations, namely *row-column exchangeability*. This was first explicitly described by Hoff [2008].

We explain this framework here, while row-column exchangeability and other symmetry considerations are covered in Chapter 3. Start by taking a generalized linear mixed-effects model, which is quite general model for the binary data. Keeping in mind the conditional independence of the edges it is enough to define the probability of an edge for every pair $(i, j)$:

$$P(Y_{i,j} = 1|\beta, \mathbf{X}, \Gamma) = f(\beta^T \mathbf{X}_{i,j} + \gamma_{i,j}) \tag{2.1}$$

with some link function $f$, for example inverse logit, to ensure that probability is between 0 and 1; covariate coefficients $\beta = (\beta_1, \dots, \beta_p)$ and random effects $\Gamma = (\gamma_{i,j})$. Here $\Gamma$ represents structure in the data that is not captured by covariates $\mathbf{X}$. The assumption of i.i.d. $\Gamma$ would be violated by many examples of network datasets that exhibit some of the dependence properties discussed in section 2.3. But it is reasonable to assume that its distribution is invariant under permutations of the node labels. We imply that there is no information that distinguishes actors after we have accounted for the covariates.

This invariance property of the joint distribution of $\Gamma$ is precisely *row-column exchangeability*, extension of exchangeability to the case of matrices or arrays.

Applying the representation theorem of Aldous [1981] for row-column exchangeable arrays, we get the latent variable representation for the $\Gamma$:

$$\Gamma = (\gamma_{i,j})_{1 \leq i,j \leq N} \overset{d}{=} (h(\theta, Z_i, Z_j, \varepsilon_{i,j}))_{1 \leq i,j \leq N}$$

for i.i.d. latent variables $\{Z_1, \dots, Z_N\}$, i.i.d. pair-specific effects $(\varepsilon_{i,j})_{1 \leq i,j \leq N}$ and some function $h$, that is symmetric in the second and third arguments.

This is a quite general result, telling that any statistical model for adjacency matrix with assumption of exchangeable nodes can be represented as a latent variable model. It is confirmed by the variety of models falling under this scheme. These existing models differ in the specification of particular function $h$.

So, our general framework is the following. We assume conditional independence of links given latent variables:

$$P(Y|Z, \mathbf{X}, \beta) = \prod_{i \neq j} P(Y_{i,j}|Z_i, Z_j, X_{i,j}, \beta)$$

and probabilities of each edge depend on the corresponding latent variables:

$$P(Y_{i,j} = 1 | Z_i, Z_j, X_{i,j}, \beta) = \text{logit}^{-1}(\beta_0 + \beta_1^T X_{i,j} + h(Z_i, Z_j)). \qquad (2.2)$$

Note that here the function $h$ is $h(Z_i, Z_j)$, as is common to most of the examples we consider.

Different models are constructed with different aims in mind and can be more suited to a particular problem/dataset, being able to represent the underlying properties better.

### Latent distance model

We start with the latent distance model introduced by Hoff et al. [2002]. We describe this model in further details in Section 4.2, as it acts as a building block for the novel dynamic model presented in Chapter 4, while here we briefly summarize the main idea.

The role of latent variables is played by positions $Z_i \in \mathbb{R}^d$ in unobserved "social space" which is represented by a $d$-dimensional Euclidean space. The idea is that the closer actors are in this space, the higher is the probability of a link between them. This is formalized by taking $h(Z_i, Z_j) = -\|Z_i - Z_j\|_d$ in (2.2), for all $(i, j)$; in fact, the distance is not constrained to be Euclidean as long as the triangle inequality is satisfied.

By construction, this model accounts for reciprocity and transitivity, and in presence of covariates also for homophily of the data. It also provides a model-based visual representation, which is most useful when $d$ is small (2 or 3).

The original proposal was further extended in Hoff [2005], Handcock et al. [2007] and Krivitsky et al. [2009]. The more general form of function $h$ is: $-\|Z_i - Z_j\|_d + \delta_i + \omega_j$, with $Z_i$ coming from a mixture of multivariate normal distributions, to account for clustering. Here $\delta$ and $\omega$ are sender and receiver random effects, that we expect to be different for directed data. This version additionally captures community structure (via the mixture distribution for $Z$'s) and heterogeneity in degrees.

We use the term "latent distance model", and not "latent space model" originally used in Hoff et al. [2002] to distinguish it from other models that use a latent space, and emphasize that the relations here depend on the distances. Another latent space model is also proposed in Hoff et al. [2002]. It is called "projection model", and the probability of having a link increases if the latent positions, considered as vectors, have similar directions in the space, and decreases when they have opposite directions.

### Stochastic blockmodels

Another popular example of a latent variable model is the stochastic blockmodel, formulated by Nowicki and Snijders [2001], based on the idea of blockmodels (that appeared in 1970s) with unknown block memberships. The assumption on the data here is that nodes are divided into groups and members of the same group have similar patterns of relationships.

More precisely, it is assumed that there are $K$ groups (blocks), and that probability of interaction between two actors depend only on the groups they belong to. So there is a matrix of group interaction probabilities $B = (B_{k,l})_{1 \leq k,l \leq K}$, with $B_{k,l}$ denoting the probability of interaction between members of the groups $k$ and $l$. As the group memberships are unknown, the latent variables $Z_i$ are used to indicate them. And so the function $h$ in (2.2) can be defined as $h(Z_i, Z_j) = B_{Z_i,Z_j}$, where the possible values of $Z_i$ are $\{1, \dots, K\}$.

So the assumption above is satisfied: by construction, the probability of a link between two actors depend only on their respective cluster memberships. This allows to capture such network dependencies as clustering and stochastic equivalence. Also such models have good interpretability and, therefore, are commonly used by social scientists.

An extension of this model to the case of infinite number of clusters was proposed in Kemp et al. [2006] and termed "Infinite relational model" (IRM). The idea is that the number of blocks is not required to be fixed in advance. Instead a nonparametric prior distribution is put on the block membership variables $Z$, that favors small number of groups, at the same time allowing for a potentially infinite number of groups.

Another extension of interest was proposed by Airoldi et al. [2008] to allow the cluster membership to change depending on the particular interacting pair. The latent variables $Z$ still represent cluster memberships, and $Z_{i,j}$ denotes cluster membership for actor $i$ while interacting with actor $j$, and it is given by a unit-vector of dimension $K$. The function $h$ from (2.2) is $Z_{i,j}' B Z_{j,i}$, which can be also written as before, $B_{Z_{i,j},Z_{j,i}}$ if we redefine $Z_{i,j}$ as taking values in $\{1, \dots, K\}$. The difference, is that here $Z_{i,j}$ can be different across $j$ for the same actor $i$. It is achieved by choosing an appropriate prior distribution for $Z$'s, allowing the membership of an actor to change when interacting with other actors. In particular the pair of group memberships may be different for the links $Y_{i,j}$ and $Y_{j,i}$, allowing a more flexible modeling for directed networks.

### Factor and feature models

More latent variable models assume yet another way of choosing the interaction function $h$ and the latent variables.

For example, the eigenmodel of Hoff [2008] generalizes the stochastic blockmodel and the latent distance model, but is less interpretable (patterns are in terms of eigenvectors). It is based on the eigenvalue decomposition and the function $h$ is $h(Z_i, Z_j) = Z_i' \Lambda Z_j$ with $Z_i \in \mathbb{R}^k$. The latent variables can be interpreted as $k$-dimensional vectors of unobserved factors (or features), and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_k)$ controls the feature loadings. Thus, if two actors have similar values for the same feature $m$, their probability of having a link can increase or decrease depending on the sign of $\lambda_m$. This allows the model to distinguish between structural equivalence and homophily.

Latent feature relational model by Miller et al. [2009] also assumes that individuals can be described by the set of (unobserved) features that influence the probability of having a link. Here the vector of features is binary, indicating the presence or absence of the features. This helps the interpretation of discovered features. So the latent variable $Z_i$ is a "feature vector" in $\{0, 1\}^\infty$, and the function $h$ in (2.2) is of the same kind as in the eigenmodel, $h(Z_i, Z_j) = Z_i' \Lambda Z_j$. The prior distribution for $Z$ is chosen in a way to allow potentially infinite number of features.

### Nonparametric models

The infinite relational model of Kemp et al. [2006] and the latent feature relational model of Miller et al. [2009] are examples of nonparametric models, in the sense that it is not needed to define the number of blocks or features in advance, but rather they are discovered from the data. Another example is an the infinite latent attribute model by Palla et al. [2012], that combines the ideas of feature models with cluster models, by considering partition of features into sub-clusters.

These nonparametric models were developed in the Machine Learning community and massively use Bayesian nonparametrics. The advantage of such approach is that the estimation and inference can be successfully performed even in the absence of covariates, as the highly flexible models can discover the underlying structure in data automatically. The downside, though, is that such models are usually quite complex, and hence often less interpretable.

An interesting development is presented in Lloyd et al. [2012]. The authors propose a nonparametric model, that also can be put in a general framework of (2.2), but the function $h$ is assumed to be a parameter itself. For this aim a Gaussian Process prior is used for $h$.

There is also an overview of some other models that can be framed into the general framework of latent variable models in Lloyd et al. [2012].

## 2.4.4 Models for dynamic network data

We are still under the assumption of binary network data. The focus of the thesis is to study and model dynamically evolving networks. The options to describe the evolution of networks are many, so to fix the setting, we assume that: the set of actors in the network does not change with time, but the links can evolve over time, disappear or appear; the observations are made at the discrete time-points $1, \dots, T$. In general, the underlying dynamic process may be continuous, but observed only at discrete time points.

Thus, the data can be encoded in an array $Y = (Y_{i,j,t})_{i,j=1...N,t=1...T}$, where $Y_{i,j,t}$ denotes the relation between actors $i$ and $j$ at time point $t$. We may also have additional covariate information for actors or pairs, either changing over time or not, and we denote by $\mathbf{X} = (X_{i,j,t})_{i,j,t\geq 1}$ where $X_{i,j,t} = (X_{i,j,k,t})_{k=1,...,p}$ for each triple $(i, j, t)$.

One of the first probabilistic dynamic network models was introduced by Holland and Leinhardt [1977]. Their assumption is that there is an underlying continuous Markov process, which consists of changes of one network link independently of the others. The assumption of Markov dependence is quite common for modeling dynamic evolution, whether applying it directly or via a hidden Markov models.

For many of the static models described above, extensions to dynamic case have been developed. In particular, for the latent variable setting, a straightforward idea is to assume that the underlying latent variables evolve over time in a Markovian way. Independent Markov evolution for these variables is usually assumed. Sarkar and Moore [2005] develop a dynamic version of the latent distance model of Hoff et al. [2002]. They propose independent random walk evolution for the latent positions. However, Sarkar and Moore [2005] do not use a statistical approach, but rather develop approximation algorithms for the estimation.

Foulds et al. [2011] extend the latent feature model of Miller et al. [2009] in a similar fashion, by using independent Markov evolution for each feature of each actor.

Other proposals include Rodriguez [2012], who used an infinite hidden Markov model for the block memberships variables in an extension of the infinite relational model of Kemp et al. [2006]. Westveld and Hoff [2011] extend the model with sender and receiver effects as the latent variables, by considering a first-order autoregressive structure over time for these variables. A dynamic extension of the mixed-membership stochastic blockmodel was proposed in Xing et al. [2010], by means of a dynamic state-space model for the vectors of mixed-membership.

Another interesting development in Heaukulani and Ghahramani [2013] assumes that the previously observed links influence the latent variables at the next time point. This model was introduced as an extension of the latent feature model, but the idea can be easily applied to other examples of latent variable models.

Good overviews of these and other dynamic extensions are presented in the recent review papers by Snijders [2011, Chapter 3] and Goldenberg et al. [2010, Chapter 4]. Although statistical literature for dynamic networks is rapidly evolving, many problems are still open. Motivated by that we try to construct a unifying framework for dynamic models similar to the one described in this Chapter. First steps to this are presented in Chapter 3. Also adding a dependence across latent variables corresponding to different actors in the dynamic evolution seems of interest. This direction is explored in Chapter 4.

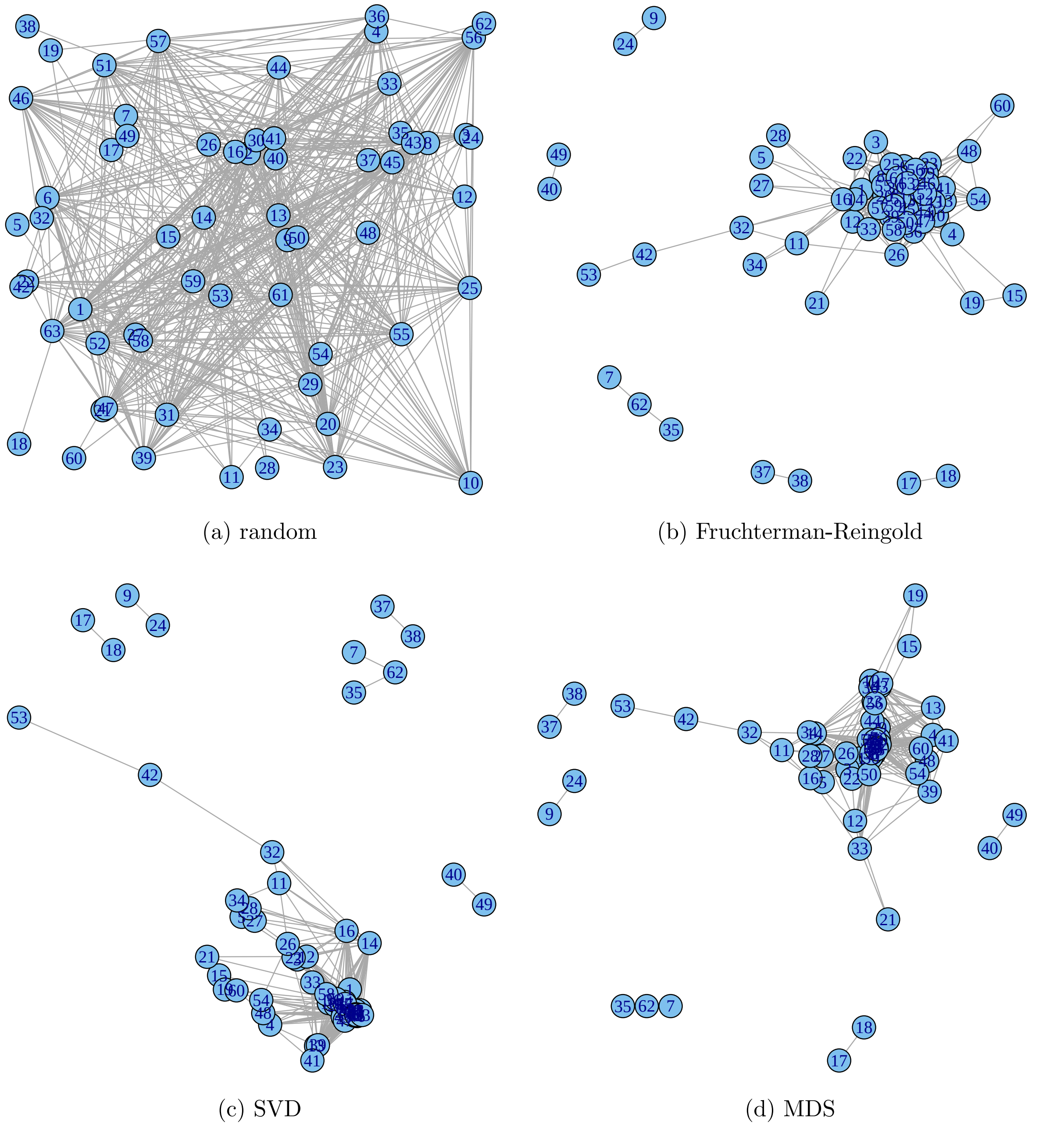
## 2.5 Final remarks

The large part of research we did not touch in this review chapter is computations.

With the rapid expansion of interest in networks, the size and complexity of currently available networks are exploding and cannot be compared with the examples used in the early models. Popular social network websites claim to include hundreds of millions of users. Protein interaction networks may include up to thousands of nodes. Thus, a big area of research is focused on developing efficient computation algorithms for the networks of such size.

For very large datasets, the development of efficient computation algorithms is a crucial challenge. Especially, for the networks that evolve constantly, and the information is updated very fast, the algorithms should keep up. Unfortunately, often it means sacrificing more expressive modeling approaches in favor of simple but fast ones.

A good review paper focusing on the challenges and solutions is Hunter et al. [2012].



(a) random

(b) Fruchterman-Reingold

(c) SVD

(d) MDS

Figure 2.2: Different graph layouts for the INFOCOM data at time $t = 20$; only nodes with non-zero degree are shown: (a) random layout, (b) layout obtained from the “force-directed” algorithm of Fruchterman and Reingold [1991], (c) layout based on the singular value decomposition of the adjacency matrix, (d) layout based on multi-dimensional scaling of the shortest path distance matrix. Created with `igraph` R-package by Csardi and Nepusz [2006]

.

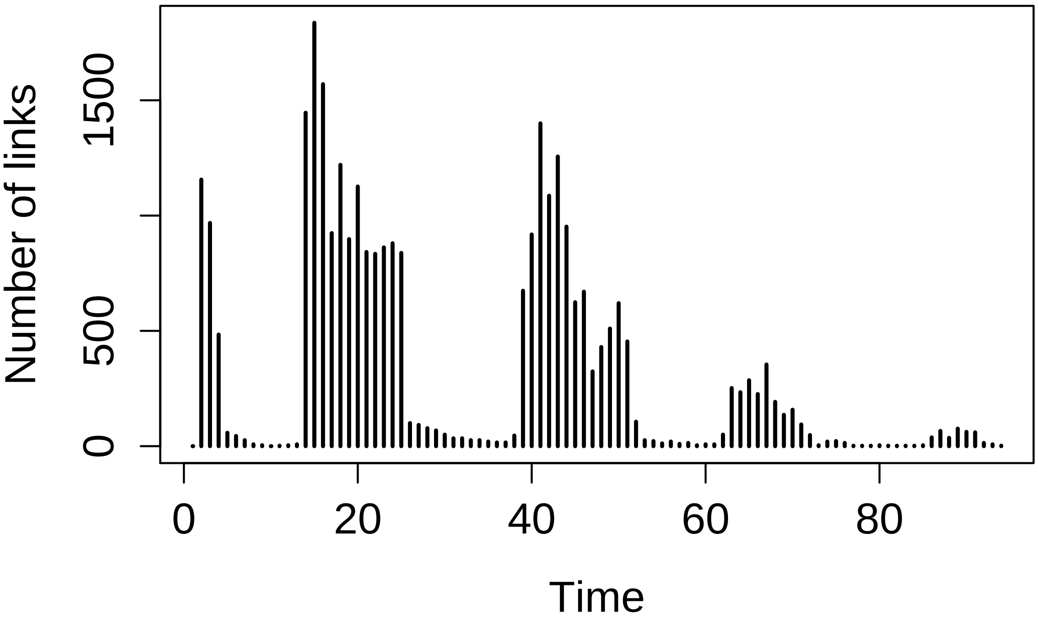


Figure 2.3: Evolution of total number of links in the INFOCOM dataset over time

# Chapter 3

# Functional representation of Markov exchangeable sequences

In this chapter we introduce some underlying concepts that are of general interest in probability and Bayesian statistics, and in particular, form basis for some approaches in modeling network data. We discuss various notions of exchangeability and different types of representation theorems, both in the spirit of de Finetti and in the functional form following Aldous [1981] and Kallenberg [2005]. This chapter provides an original contribution in giving functional representation theorem for Markov exchangeable sequences. Further a representation theorem is provided for the row-column exchangeable array of Markov exchangeable sequences. These are interesting results on their own, and can be useful for applications in dynamic network modeling.

## 3.1 Introduction

Various notions of symmetry are central for statistical modeling. The most basic type of (in)dependence is the assumption that random variables are independent and identically distributed (i.i.d.). The more general notion of *exchangeability*, discussed in the next section, is tightly connected with Bayesian approach to statistics and was advanced by Bruno de Finetti starting from the 1930-s. Further extensions of this notion, such as partial exchangeability, also emerged from work of de Finetti and others. In particular, array generalizations were thoroughly studied in the 1980-s by Aldous [1981, 1985] and Hoover [1979, 1982]. These ideas are not strictly related to Bayesian statistics, but reflect the assumptions we may have about the observations or other quantities, such as errors.

In general, various types of exchangeability assume invariance of the joint distribution of random variables under some group of permutations. A comprehensive coverage of the symmetry considerations leading to notions of exchangeability is provided by Kallenberg [2005], with Chapter 7 entirely dedicated to arrays.

## 3.2 Functional representations

For different assumptions of exchangeability described in this chapter, representation theorems have the aim of characterizing these assumptions in terms of simpler objects. The "classical" approach works for the sequences that are exchangeable (Section 3.3), partial exchangeable (Section 3.4) or Markov exchangeable (Section 3.5), while when dealing with arrays, other approach is needed. It was pioneered by Aldous [1981] and Hoover [1982] and involves functional representations, in which the joint distribution of the array can be expressed in terms of functions of collections of uniform random variables. Such an approach can also be taken to give alternative representation theorems for the former assumptions on the sequences. We specify the notation we will use, and basic definitions.

**Definition 3.2.1.** We say that a random sequence or array $X$ has a *functional representation* $f(J)$, if there exists a measurable function $f$, that acts on a family $J$ of independent uniform $U(0,1)$ random variables, and satisfies $X \overset{d}{=} (f(J))$. Here the joint distributions are equal, meaning that $f(J)$ is of the same dimension as each element of $X$ and $J$ is indexed accordingly.

This idea is explained on the examples below with particular representations.

**Definition 3.2.2.** By *coding* $\xi$ for an $S$-valued random variable $X$ we mean a representation $X \overset{d}{=} f(\xi)$ where $\xi \sim U(0,1)$, and $f$ is a measurable function $f : (0,1) \to S$.

It is easy to see that the following result holds.

**Proposition 3.2.1.** *For a Borel space $S$, each $S$-valued random variable $X$ allows a coding $f(\xi)$.*

*Proof.* First consider $S = \mathbb{R}$. Denote with $F_X$ a cumulative distribution function of $X$. Let $f(\xi) = F_X^{-1}(\xi)$, where $F_X^{-1}$ is the quantile function, i.e. $F_X^{-1}(t) = \inf\{x : F(x) \geq t\}$. Then $f(\xi) \overset{d}{=} X$. Now let $S$ be any Borel space. From the definition, there exists an isomorphism $\phi$ between $S$ and $\mathbb{R}$. Then $\bar{X} = \phi \circ X$ is a random variable on $\mathbb{R}$ with distribution function $F_{\bar{X}}$. Take a function $f = \phi^{-1} \circ F_{\bar{X}}^{-1}$. Then $f(\xi) \overset{d}{=} X$ for a $\xi \sim \mathrm{U}(0,1)$. □

Expanding the same idea for an i.i.d. sequence $X = (X_i)_{i \geq 1}$, there exists a function $f$ such that the following representation hold: $(X_1, X_2, \dots) \overset{d}{=} (f(\xi_1), f(\xi_2), \dots)$. In view of Definition 3.2.1, we can write $X \overset{d}{=} (f(\xi_i))_{i \geq 1}$ and say that $X$ has functional representation $f(\xi_i)$. Further we assume that the space $S$ is Polish, and so Borel, if not specified otherwise.

*Remark* (Notational conveniences). In the sequel we will sometimes omit lower indexing if enumeration of array goes across all dimensions. So by writing $(X_{i,j})$ we will imply $(X_{i,j})_{i,j \geq 1}$. Also we will use shortcut notation for joint equality in distribution as follows: when writing, for example, $(X_i) \overset{d}{=} (Y_i)$ we mean $(X_i)_{i \geq 1} \overset{d}{=} (Y_i)_{i \geq 1}$ that is $(X_1, X_2, \dots) \overset{d}{=} (Y_1, Y_2, \dots)$.

## 3.3 Exchangeability

Exchangeability is a probabilistic property, that generalizes the i.i.d. property. Informally, exchangeability for a sequence of random variables means that the order of its elements is not relevant. More formally

**Definition 3.3.1.** A sequence $(X_i)_{i\geq 1}$ of random variables on a measurable space $S$ is called (infinitely) exchangeable if for any $n$ and any permutation $\pi$ of $\{1, 2, \ldots, n\}$

$$(X_1, X_2, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}).$$

The well-known de Finetti representation theorem states that every exchangeable sequence can be represented as a mixture of independent and identically distributed (i.i.d.) random variables:

**Theorem 3.3.1** (de Finetti). *Let $X = (X_i)_{i\geq 1}$ be an exchangeable sequence of random variables on a measurable space $S$. Then there exists a random probability measure $\nu$ on $S$, called* directing random measure *for $X$, such that $X$ is conditionally i.i.d. given $\nu$, i.e.*

$$P(X \in \cdot|\nu) = \nu^{\infty}(\cdot) \ a.s.;$$

*moreover, $\nu$ is a.s. unique and is given by*

$$\nu(A) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I_A(X_i) \ a.s.$$

*Alternatively we can write:*

$$P(X \in \cdot) = E\nu^{\infty}(\cdot) = \int \nu^{\infty}(\cdot)\mu(d\nu)$$

*where $\mu$ (measure on probability measures on $S$) is the probability law of $\nu$, called* de Finetti measure *of the sequence $X$.*

As can be seen from the theorem, exchangeability is a natural representation of sampling at random: at first a distribution $\nu$ is chosen randomly from a measure $\mu$, and then $X$ is sampled from $\nu$.

De Finetti representation theorem plays a fundamental role in Bayesian statistics, providing motivation for the Bayesian approach: for an exchangeable sequence it motivates the use of prior distribution and conditionally independent sampling from a common distribution.

An equivalent representation theorem can be provided in a functional form, that was shown by Aldous [1981]:

**Theorem 3.3.2** (Aldous [1981, Lemma 1.5]). *An infinite sequence* $X = (X_i)_{i\geq 1}$ *is exchangeable if and only if there exists a measurable function* $f : (0,1)^2 \to S$*, such that for independent* $U(0,1)$ *random variables* $\alpha$ *and i.i.d.* $(\xi_i)_{i\geq 1}$, $X \overset{d}{=} (f(\alpha, \xi_i))_{i\geq 1}$.

*Using the Definition 3.2.1, we say that* $X$ *has a functional representation* $f(\alpha, \xi_i)$.

Comparing the two representation theorems, we can see that the role of uniform random variables is the following: $\alpha$ in Theorem 3.3.2 is used for the coding of the directing random measure $\nu$, obtained through its distribution, $\mu$. While $\xi_i$'s are used for the coding of $X_i$'s, as their distribution given $\nu$, is known.

The functional representation provides an interpretation for an exchangeable sequence: each element $X_i$ is influenced by a common effect $\alpha$ and by individual effect $\xi_i$.

*Remark.* It is interesting to look at the particular form of the function $f$ when $X$ is a binary sequence. So, suppose that each $X_i$ takes values 0 or 1 only. Then the function $f$ from Theorem 3.3.2 can be written as an indicator $f(\alpha, \xi_i) = \mathbb{I}(\xi_i < g(\alpha))$ for some function $g$. Hence, $X_i | g(\alpha) \sim \text{Bern}(g(\alpha))$. We can see clear parallels with the de Finetti theorem for $0-1$ case: denoting $g(\alpha) = \nu$, we have

$$X_i | \nu \overset{iid}{\sim} \text{Bern}(\nu),$$
$$\nu \sim \mu.$$

The theorem 3.3.2, as well as functional representations given further, can be rewritten also in terms of almost sure equality, using the so called "transfer theorem":

**Theorem 3.3.3** (Kallenberg [2002, Corollary 6.11]). *Fix two Borel spaces* $S$ *and* $T$*, a measurable mapping* $f : T \to S$*, and some random elements* $\xi$ *in* $S$ *and* $\eta$ *in* $T$ *with* $\xi \overset{d}{=} f(\eta)$*. Then there exists a random element* $\tilde{\eta} \overset{d}{=} \eta$ *in* $T$ *with* $\xi = f(\tilde{\eta})$ *a.s.*

So the alternative formulation of Theorem 3.3.2 would be:

**Theorem.** *An infinite sequence $X = (X_i)_{i\geq 1}$ is exchangeable if and only if there exists a measurable function $f : (0,1)^2 \to S$ and some $U(0,1)$ random variables $\bar{\alpha}$ and $(\bar{\xi}_i)$, such that $X = (f(\bar{\alpha}, \bar{\xi}_i))_{i\geq 1} a.s.$.*

Further in the text we will mainly consider distributional equalities for simplicity. In this case, in principle, the collections of uniform variables can be taken from one large i.i.d. collection with rich enough set of indices. For example, functional representation from the previous theorem can be expressed as $f(\xi_\emptyset, \xi_i)$, where $\xi$'s come from a collection of i.i.d. U(0,1) variables, indexed by subsets of $\mathbb{N}$. We will not pursue this generality, but such notation is used, for example, by Kallenberg [2005] and explained in the Section 7.1 of the book.

## 3.4 Partial exchangeability

The assumption of exchangeability is not always desirable. Think of samples including elements from several distinct groups, as for example in the case of a population consisting of males and females. One of the natural generalizations of exchangeability appropriate for this case is *partial exchangeability* introduced in de Finetti [1938]. We should note that the term "partial exchangeability" has been used in the literature in several different meanings (see e.g. Aldous [1981] or Diaconis and Freedman [1980]). In all cases the "partiality" means some restriction on the class of permutations, that leaves joint distribution invariant. As there is no sure agreement in the use of the term, we will use the original definition of de Finetti [1938, 1972]. But we will also see that, under certain conditions, it is in fact equivalent to the definition of Diaconis and Freedman [1980].

Informally, when we speak about partial exchangeability we do not distinguish the order of the elements inside the groups, while assuming that the elements from different groups may be distinguishable.

**Definition 3.4.1.** An array $X = (X_{i,j})_{i\in I, j\geq 1}$ is called *partially exchangeable* (by rows) if its distribution is invariant under finite permutations within rows:

$$X \overset{d}{=} (X_{i,\pi_i(j)})_{i\in I, j\geq 1}$$

for all finite permutations $\pi_1, \pi_2, \ldots$

*Remark.* The array $X$ may have a finite or countable number of rows $|I|$, while number of columns is infinite.

*Remark.* We may further assume that rows of $X$ are also exchangeable. In this case we obtain partially exchangeable RCE array. We will comment on this in section 3.6.

There exists an extension of de Finetti representation theorem 3.3.1 to this case:

**Theorem 3.4.1** (de Finetti [1938])**.** *Let $X = (X_{i,j})_{i=1,\ldots,n; j\geq 1}$ be a partially exchangeable by rows array of random variables on a measurable space $S$. Then there exists a vector of probability measures $\nu = (\nu_1, \ldots, \nu_n)$ such that rows $X_{i,\cdot}$ are conditionally independent given $\nu$, and for each $i$ the sequence $X_{i,\cdot} = (X_{i,j})_{j\geq 1}$ is i.i.d. according to $\nu_i$, i.e. for the sets $(A_i)_{i=1}^n \in S^\infty$:*

$$P(X_{1,\cdot} \in A_1, \ldots, X_{n,\cdot} \in A_n) = \int \nu_1^\infty(A_1) \ldots \nu_n^\infty(A_n) \mu(d\nu_1, \ldots, d\nu_n).$$

*The $\nu_i$ are directing random measures for $X_{i,\cdot}$, and $\mu$ (measure on vectors of probability measures) is the probability law of $\nu = (\nu_1, \ldots, \nu_n)$, called de Finetti measure of $X$.*

Using the proof of this result, it is not difficult to obtain a functional analogue of the representation theorem:

**Theorem 3.4.2.** *For an $S$-valued array $X = (X_{i,j})_{i,j\geq 1}$ of random variables the following are equivalent:*

1. *$X$ is partially exchangeable.*
2. *$X$ can be represented as $X \overset{d}{=} (f_i(\alpha, \lambda_{i,j}))_{i,j\geq 1}$ for some measurable functions $f_i : (0,1)^2 \to S$ and independent $U(0,1)$ random variables $\alpha, (\lambda_{i,j})_{i,j\geq 1}$.*

*Proof.* (1 ⇒ 2) Following the proof of de Finetti's representation theorem for partially exchangeable random variables (see Aldous [1985, Corollary 3.9]), let $\nu_i$ be the directing random measure for the row $X_{i,\cdot} = (X_{i,j})_{j\geq 1}$, and let $\mathfrak{G} = \sigma(\nu_i : i \geq 1)$. Then $\nu_i$ is the regular conditional distribution for $X_{i,j}$ given $\mathfrak{G}$, and rows $X_{i,\cdot}$ are conditionally independent given $\mathfrak{G}$.

So for all $i$, given $\mathfrak{G}$, $(X_{i,j})_{j\geq 1}$ are i.i.d. with distribution $\nu_i$ and thus have a coding $(X_{i,j})_{i,j\geq 1} \overset{d}{=} (G_{\nu_i}(\lambda_{i,j}))_{i,j\geq 1}$, where $G_{\nu_i}$ is the function obtained by Proposition 3.2.1. The measure $\nu_i$ depends on $\mathfrak{G}$ and $i$, so $(X_{i,j})_{i,j\geq 1} \overset{d}{=} (f_i(\alpha, \lambda_{i,j}))_{i,j\geq 1}$ taking additionally coding $\alpha$ for $\mathfrak{G}$.

(2 ⇒ 1) Follows immediately from the functional representation and Definition 3.4.1 of partial exchangeability. □

*Remark.* We may also write the representation in (2) as

$$(X_{i,j})_{i,j\geq 1} \overset{d}{=} (f(\alpha, i, \lambda_{i,j}))_{i,j\geq 1},$$

which suggests the interpretation of the elements of partial exchangeable array $X$ as being dependent on a general effect, a fixed row effect and an individual effect.

*Remark.* As before, it is interesting to look at the particular example for the binary case. We suppose that each $X_{i,j}$ takes values $0, 1$ only. Thus, the functions $f_i$ from theorem 3.4.2 can be rewritten as:

$$f_i(\alpha, \lambda_{i,j}) = \mathbb{I}(\lambda_{i,j} < g_i(\alpha))$$

with some other functions $g_i$. It means that $(X_{i,j})_{j\geq 1}|g_i(\alpha) \sim \text{Bern}(g_i(\alpha))$, and writing $\nu_i = g_i(\alpha)$ we can again see parallels with the de Finetti theorem for $0-1$ partially exchangeable sequences:

$$\begin{aligned} X_{i,j}|\nu_i &\overset{iid}{\sim} \text{Bern}(\nu_i), \\ (\nu_i) &\sim \mu. \end{aligned}$$

## 3.5 Markov exchangeability and successors

As we have seen, exchangeable sequences are mixtures of i.i.d. sequences. As i.i.d. is a basic assumption for sequences, in the dynamic case, Markov property can be considered as a basic type of dependence. Diaconis and Freedman [1980] studied invariance properties that can characterize mixtures of Markov chains. Following Zaman [1984] and Zabell [1995] we use the term "Markov exchangeability" to denote "partial exchangeability" as defined in Diaconis and Freedman [1980].

Markov exchangeability may be considered as a generalization of exchangeability for the case of Markov chains: if a priori we do not distinguish the order of transitions $(X_i, X_{i+1})$, then the sequences with the same transition counts and the same initial state should be assigned the same probability. More formally:

**Definition 3.5.1.** A sequence $X = (X_i)_{i\geq 0}$ of random variables on a countable state space $S$ is called *Markov exchangeable* (ME) if for any $n$ and any two sequences $\sigma = (\sigma_0, \ldots, \sigma_n) \in S^{n+1}$, $\tau = (\tau_0, \ldots, \tau_n) \in S^{n+1}$ with the same starting state and the same transition counts, $P(X_0 = \sigma_0, X_1 = \sigma_1, \ldots, X_n = \sigma_n) = P(X_0 = \tau_0, X_1 = \tau_1, \ldots, X_n = \tau_n)$. These sequences are called *equivalent* $(\sigma \sim \tau)$.

We will require two further definitions:

**Definition 3.5.2.** A sequence $X = (X_i)_{i\geq 0}$ is called *recurrent* if $P(X_i = X_0 \text{ infinitely often}) = 1$.

**Definition 3.5.3.** Consider the set $\mathbf{P}$ of random transition matrices on $S \times S$. A sequence $X = (X_i)_{i\geq 0}$ is called a mixture of Markov chains if there exists a distribution $\mu$ on the space $S \times \mathbf{P}$ such that for any $n \geq 1$:

$$P(X_0 = i_0, \ldots, X_n = i_n) = \int_{\mathbf{P}} \prod_{m=0}^{n-1} p(i_m, i_{m+1}) \mu(i_0, dp).$$

The representation theorem for Markov exchangeable sequences was given in Diaconis and Freedman [1980]:

**Theorem 3.5.1** (Diaconis and Freedman [1980, Theorem 7])**.** *If a sequence $X$ on a countable state space $S$ is recurrent, then $X$ is Markov exchangeable if and only if it is a mixture of Markov chains.*

We want to provide an equivalent functional representation for recurrent Markov exchangeable sequences. To this aim we need some auxiliary facts, connecting Markov exchangeability and partial exchangeability. Such a connection was hinted at by de Finetti [1959] and Zabell [1995], and the equivalence (under appropriate conditions) of these two notions was proved in Fortini et al. [2002].

Consider a recurrent Markov exchangeable sequence $X$ on a finite or countable state space $S$. It can be shown (see Fortini et al. [2002]) that each state $i \in S$, that is visited, recurs infinitely often (hence sequence is called *strongly recurrent*), say at times $\tau_1^i, \tau_2^i, \dots$. We can consider the sequence of *successor states* for state $i$, defined as $V_{i,j} = X_{\tau_j^i+1}$, $j \geq 1$. From the strong recurrence property, such sequences are infinite. It was shown in Zabell [1995] that for every $i$ the sequence $(V_{i,1}, V_{i,2}, \dots)$ is exchangeable.

Combining sequences of successors for each state in array, we obtain $V = (V_{i,j})_{i \in S, j \geq 1}$ with $V_{i,j}$ being the $j$-th successor of the state $i$. Fortini et al. [2002] proved that under assumption of recurrence, Markov exchangeability and partial exchangeability are equivalent, more specifically:

**Theorem 3.5.2** (Fortini et al. [2002, Theorem 2]). *The successors array $V$ of the sequence $X$ is partially exchangeable if and only if $X$ is recurrent and Markov exchangeable.*

### 3.5.1 Representations for Markov exchangeability

Now we can use the representations for partially exchangeable sequences, discussed in Section 3.4, to obtain a characterization for recurrent Markov exchangeable sequences. Starting from the functional representation from Theorem 3.4.2, we can generate array $V$ which is partially exchangeable, and construct a Markov exchangeable sequence such that $V$ is its matrix of successor states. Conversely, any recurrent Markov exchangeable sequence has a partial exchangeable successors matrix, so there exist functions $f_i$ and families of random variables such that it can be represented as in Theorem 3.4.2.

Assume that the state space $S$ is finite or countable. By relabeling the elements of the state space, we may assume, without loss of generality, that $S = \{1, \ldots, k\}$, $k \leq \infty$.

The direct consequence of the functional representation given in Theorem 3.4.2 and 3.5.2 is the following

**Theorem 3.5.3.** *Let $S = \{1, \ldots, k\}$, $k \leq \infty$ and $V = (V_{i,j})_{i \in S, j \geq 1}$. The following are equivalent:*

1. *The array $V$ has functional representation $V \overset{d}{=} (f_i(\alpha, \lambda_{i,j}))_{i \in S, j \geq 1}$, where $\alpha, (\lambda_{i,j})_{i \in S, j \geq 1}$ are i.i.d. $U(0,1)$ random variables and $(f_i)_{i \in S}$ is a collection of measurable functions $f_i : (0,1)^2 \to S$.*

2. *A sequence $X = (X_i)_{i \geq 0}$ with starting state $X_0 = 1$, generated from $V$ as successors array, is a recurrent Markov exchangeable sequence on $S$.*

*Remark.* Actually it is also possible to consider more general uncountable space as the state space $S$, see Fortini et al. [2002].

We describe in more detail the construction of a Markov exchangeable sequence from the array of successor states in Theorem 3.5.3:

Take $X_0 = 1$. Using $V$ as the array of successors, for each $\omega \in \Omega$ set for $i \geq 1$ $X_i(\omega) = V_{X_{i-1}(\omega), g(X_0(\omega), \ldots, X_{i-1}(\omega))}$, where $g(x_0, \ldots, x_{i-1})$ is the number of occurrences of the state $x_{i-1}$ among the first $i$ elements of $(x_i)_{i \geq 0}$. The constructed sequence $(X_i(\omega))_{i \geq 0}$ will be a realization of a recurrent and Markov exchangeable sequence of random variables.

So, basically, we have a recursive representation

$$X_n = f(\alpha, X_{n-1}, \lambda_{X_{n-1}, g(X_0, \ldots, X_{n-1})}).$$

Looking at this representation we see that indexing of $\lambda$ by two indexes is redundant. In fact, pair $X_{n-1}, g_n$ can appear only once (we have only one $j$-th successor of state $i$), and $(\lambda_{i,j})$ is a family of i.i.d. uniform random variables. Thus, without loss of generality, we can replace $\lambda_{X_{n-1}, g_n}$ in representation by $\lambda_n$, as in fact this element is responsible for the randomness in $X_n$ not accounted for in other elements.

So the representation changes to:

$$\forall\, n,\ X_n = f(\alpha, X_{n-1}, \lambda_n) \text{ a.s.} \tag{3.1}$$

for independent uniform$(0,1)$ random elements $\alpha$ and i.i.d. collection $(\lambda_n)_{n\geq 1}$.

Note that the recursive representation (3.1) follows from well-known recursive representation for Markov chains: $Y_n = g(Y_{n-1}, \lambda_n)$ a.s. (see, for example, Kallenberg [2002, Proposition 8.6]) by virtue of Theorem 3.5.1, as additional mixing variable can be coded as $\alpha$ as in, for example, representation for exchangeable sequences (Theorem 3.3.2).

We also note that here we use equality almost surely and not in distribution, but as we said, all the functional representation results can be put in such form using the transfer theorem 3.3.3.

Next, we would like combine the assumption of Markov exchangeability with the array exchangeability. Such assumptions give a 3-dimensional array with symmetry properties, that can be appropriate for the dynamic network modeling. We start by reminding the relevant definitions and representation theorems for exchangeable arrays, and then move to combining RCE with ME in Section 3.8.

## 3.6 Row-column exchangeability

For two-dimensional case, exchangeability may be generalized in a form of row-column exchangeability. Assume that $X = (X_{i,j})_{i,j\geq 1}$ is an infinite array of random variables. The following generalization were proposed:

**Definition 3.6.1.** $X$ is called *row-column exchangeable (RCE)* if $X \overset{d}{=} (X_{\pi(i),\sigma(j)})_{i,j\geq 1}$ for every finite permutations $\pi$ of row indices and $\sigma$ of column indices.

**Definition 3.6.2.** $X$ is called *jointly RCE* array if $X \overset{d}{=} (X_{\pi(i),\pi(j)})_{i,j\geq 1}$ for every finite permutation $\pi$. Analogously, RCE arrays are sometimes termed *separately RCE*, to highlight the differences.

Also useful is

**Definition 3.6.3.** If $(X_{i,j} : \max(i,j) \leq n)$ is independent of $(X_{i,j} : \min(i,j) > n)$ for each $n$, the array is called *dissociated.*

For RCE, jointly RCE and dissociated RCE arrays functional representation theorems were proved by Aldous [1981].

**Theorem 3.6.1** (aggregated results from Aldous [1981])**.**

1. *X is a RCE array if and only if it has a functional representation*

$$X \overset{d}{=} (f(\alpha, \xi_i, \eta_j, \lambda_{i,j}))_{i,j \geq 1}$$

2. *X is a dissociated RCE array if and only if it has a functional representation*

$$X \overset{d}{=} (f(\xi_i, \eta_j, \lambda_{i,j}))_{i,j \geq 1}$$

3. *A RCE array is a mixture of dissociated RCE arrays.*

4. *X is a jointly RCE array if and only if it has a functional representation*

$$X \overset{d}{=} (f(\alpha, \xi_i, \xi_j, \lambda_{\{i,j\}}))_{i,j \geq 1}$$

   *with $f$ symmetric in the 2nd and 3rd argument, and by writing $\lambda_{\{i,j\}}$ we mean that the collection $\lambda$ is indexed by unordered sets $\{i,j\}$, and so the same elements are used for $X_{i,j}$ and $X_{j,i}$.*

Again representations for $X$ can be interpreted in a way that $X_{i,j}$ is defined as a function of overall effect, row-effect, column-effect and individual effect. These representations are used to motivate latent variable network models, as described in Chapter 4.

A particular example that is relevant for the network modeling is the case of binary arrays $X$. In this case the functional representation can be simplified the following way:

**Proposition 3.6.2** (from Lloyd et al. [2012]). *Let $X = (X_{i,j})_{i,j\geq 1}$ be a RCE array with $X_{i,j}$ taking values in $\{0,1\}$. Then there exists a function $\Theta : (0,1)^3 \to (0,1)$ and independent collections of Uniform(0,1) random variables $\alpha, (\xi_i), (\eta_j), (\lambda_{i,j})$ such that*

$$X \overset{d}{=} (\mathbb{I}(\lambda_{i,j} < \Theta(\alpha, \xi_i, \eta_j)))$$

*where $\mathbb{I}_A(x) = 1$ when $x \in A$ and $0$ otherwise.*

*Proof.* From Theorem 3.6.1 above, $X$ has a functional representation $f(\alpha, \xi_i, \eta_j, \lambda_{i,j})$. Fixing the values of $\alpha, \xi_i, \eta_j$, the function $f(\alpha, \xi_i, \eta_j, \cdot)$ is defined on $(0,1)$ and takes values in $\{0,1\}$. Suppose it is equal to 1 on a subset $A$ of $(0,1)$ and 0 on its complement. Since $\lambda_{i,j}$ is uniform$(0,1)$, the probability of $f = 1$ is $|A|$ and $P\left(f(\alpha, \xi_i, \eta_j, \lambda_{i,j}) = 1\right) = P(\lambda_{i,j} \in A) = P(\lambda_{i,j} < |A|) = P(\mathbb{I}(\lambda_{i,j} < |A|) = 1)$. As the subset $A$ depends on $\alpha, \xi_i$ and $\eta_j$, define $\Theta(\alpha, \xi_i, \eta_j) = |A|$. This provides the desired representation.

The other side of the proof is obvious, as the function given is a particular case of functional representation that gives a RCE array, and it clearly takes values in $\{0,1\}$. □

Let us note that the representations in Theorem 3.6.1 can be also reformulated in terms of random functions as is done in Lloyd et al. [2012] and Orbanz and Roy [2013]: $X = (X_{i,j})_{i,j\geq 1}$ is RCE if and only if there is a random measurable function $F : (0,1)^3 \to S$ such that $X \overset{d}{=} (F(\xi_i, \eta_j, \lambda_{i,j}))_{i,j\geq 1}$. And similarly for jointly RCE. The $\alpha$ in representations above correspond to randomness in choosing the function $F$.

The function $\Theta$ in Proposition 3.6.2 in this terms also changes accordingly to a random $\Theta(\xi_i, \eta_j)$, or symmetric $\Theta(\xi_i, \xi_j)$ for the case of jointly RCE array, and so corresponds to the following sampling scheme, resulting in a $0-1$ array $X$:

1. sample a random function $\Theta : (0,1)^2 \to (0,1)$

2. for all $(i,j)$ sample $\xi_i, \xi_j$ from $U(0,1)$

3. compute $\Theta(\xi_i, \xi_j)$ and sample $\lambda_{i,j}$ from $U(0,1)$

4. set $X_{i,j} = \mathbb{I}(\lambda_{i,j} < \Theta(\xi_i, \xi_j))$

Alternatively, it means that $X_{i,j}|\theta_{i,j} \overset{ind}{\sim} \text{Bern}(\theta_{i,j})$ where $(\theta_{i,j})$ is RCE of particular form $\theta_{i,j} = \Theta(\xi_i, \xi_j)$.

Another remark is that functional representations for RCE arrays are closely connected with notions of graph limits. For this interesting development see Lovász [2009] and discussions in Aldous [2010], Orbanz and Roy [2013].

## 3.7 Partial Exchangeability and RCE

In section 3.4 we mentioned the possibility of assuming row exchangeability in a partially exchangeable array. Relating to the previous section, a row-column exchangeable array is a particular case of array with these properties. This is easy to see as column exchangeability is an example of partial exchangeability within rows, when all permutations are taken to be the same.

**Definition 3.7.1.** $X$ is called PE-RCE array, if it is RCE and partially exchangeable across rows, i.e. $X \overset{d}{=} (X_{\sigma(i),\pi_{\sigma(i)}(j)})$ for every finite permutations $\sigma$ and $\pi_1, \pi_2, \dots$

For such arrays, with additional property of dissociation, there is a functional representation result by Aldous [1985]:

**Theorem 3.7.1** (14.16 in Aldous [1985])**.** *$X$ is a dissociated partially exchangeable RCE array if and only if it has a functional representation $X \overset{d}{=} (f(\xi_i, \lambda_{i,j}))_{i,j \geq 1}$.*

Removing the assumption of dissociation we show that the following result holds:

**Theorem 3.7.2.** *For a RCE array $X = (X_{i,j})_{i,j\geq 1}$ the following are equivalent:*

1. *$X \overset{d}{=} (X_{i,\pi_i(j)})_{i,j\geq 1}$ for all finite permutations $\pi_1, \pi_2, \dots$* (partial exchangeability)

*2.* $X \overset{d}{=} (f(\alpha, \xi_i, \lambda_{i,j}))_{i,j\geq 1}$ *for some* $f : (0,1)^3 \to S$ *and independent uniform collections of random variables*

*3.* $X$ *is a mixture of dissociated PE-RCE arrays*

*Proof.* ($2 \Rightarrow 3$) For each $a \in [0,1]$, let $f_a(b,c) = f(a,b,c)$. By conditioning on $\alpha$ in the representation (2) we see that $X$ is a mixture (over $a$) of arrays $X^a$, where $X^a_{i,j} = f_a(\xi_i, \lambda_{i,j})$. But by Theorem 3.7.1 $X^a$ is a dissociated PE-RCE array.

($3 \Rightarrow 2$) Given the mixing distribution $\Theta = \theta$, $X$ is dissociated RCE array with property (1). So by Theorem 3.7.1 it can be represented as $f_\theta(\xi_i, \lambda_{i,j})$. Now take a coding $\alpha$ for $\Theta$. Then there exists a function $f : (0,1)^3 \to S$ such that the array $X$ has the same distribution as the array $(f(\alpha, \xi_i, \lambda_{i,j}))_{i,j\geq 1}$.

($2 \Rightarrow 1$) Same as in the case ($b \Rightarrow a$) of Theorem 3.7.1 for dissociated case, easily follows from the fact that $(\lambda_{i,j})$ are independent and identically distributed.

($1 \Rightarrow 2$) As in the proof of Theorem 3.7.1, let $\mu_i$ be the directing random measure for $(X_{i,j})_{j\geq 1}$. By de Finetti theorem for partially exchangeable sequences ([Aldous, 1985, Corollary 3.9]) for each pair $(i,j)$ we have: $\mu_i$ is a regular conditional distribution for $X_{i,j}$ given $\sigma\left(X_{i',j'} : (i',j') \neq (i,j)\right)$.

Note that we can rewrite condition (1)+RCE is the condition in Definition 3.7.1:

(1'). $X \overset{d}{=} (X_{\sigma(i),\pi_{\sigma(i)}(j)})_{i,j\geq 1}$ for all finite permutations $\sigma, \pi_1, \pi_2, \ldots$

And, in particular, the distribution of sequence of rows is exchangeable. To see this, take in (1') for all $i$, $\pi_i = id$, leaving all the elements the same, and permute only with $\sigma$.

As for each $i$, $\mu_i$ is the directing random measure for row $X_{i,\cdot} = (X_{i,j})_{j\geq 1}$, then we know from Aldous [1985, Lemma 2.15], that $\mu_i$ is the limit of the empirical distribution of sequence $X_{i,\cdot}$:

$$\mu_i(\cdot) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \delta_{X_{i,j}}(\cdot).$$

So we obtain that $\mu_i$'s are also exchangeable.

So by de Finetti theorem 3.3.1 this sequence is a mixture of i.i.d. and there exists a random measure $\nu$ such that $(\mu_1, \mu_2, \mu_3, \dots) \mid \nu = \theta \overset{iid}{\sim} \theta$. Then by usual construction, we assume coding $\alpha$ for $\nu$, i.e. $\alpha \sim U(0,1)$, $\nu \overset{d}{=} G_\nu(\alpha)$, where $G_\nu$ is the function obtained by Proposition 3.2.1 for $\nu$. Now, given $\nu = \theta$, the distribution of $\mu$ is known: $\mu_i | \nu = \theta \overset{iid}{\sim} \theta$. So we can code $\mu_i$'s by $\xi_i$ with $\mu_i \overset{d}{=} G_\mu(\xi_i)$. Then, given $\mu_i$, $(X_{i,j})_{j\geq 1}$ are i.i.d. with distribution $\mu_i$. So, finally, we code $X_{i,j} \overset{d}{=} G_{\mu_i}(\lambda_{i,j})$.

Combining the three above codings we obtain a desired functional representation $X = (X_{i,j})_{i,j\geq 1} \overset{d}{=} (f(\alpha, \xi_i, \lambda_{i,j}))_{i,j\geq 1}$ for $\alpha, (\xi_i)_{i\geq 1}, (\lambda_{i,j})_{i,j\geq 1}$ independent $U(0,1)$. □

*Remark.* Comparing Theorem 3.4.2 and 3.7.2 we can see that the removal of the assumption of exchangeability of rows replaces random $\xi_i$'s with non-random $i$'s in the representation.

## 3.8 Combining RCE and ME

Now to combine properties of row-column exchangeability and Markov exchangeability, define an infinite 3-dimensional array $\mathbf{X}$ in the following way:
Let $\mathbf{X} = (\mathbf{X}_{i,j})_{i,j\geq 1}$ be a (jointly) RCE array of sequences, i.e. each $\mathbf{X}_{i,j} = (X_{i,j,t})_{t\geq 1}$ with $X_{i,j,t} \in S = \{1, \dots, K\}, K \leq \infty$. The assumption for the time evolution is that each $\mathbf{X}_{i,j}$ is Markov exchangeable and recurrent, but these sequences are not necessary independent, as we allow dependence across $i, j$ by RCE property.

We formalize these two assumptions:

A1. $\mathbf{X}$ is a jointly RCE array of sequences, i.e. $(\mathbf{X}_{i,j}) \overset{d}{=} (\mathbf{X}_{\sigma(i),\sigma(j)})$ for all finite permutations $\sigma$ of $\mathbb{N}$, which is the same as writing $(X_{i,j,t}) \overset{d}{=} (X_{\sigma(i),\sigma(j),t})$.

A2. For every pair $(i, j)$ the sequence $(X_{i,j,t})_{t\geq 1}$ is Markov exchangeable and recurrent.

Our aim is to have a functional representation of $\mathbf{X}$. By Theorems 3.6.1 and 3.5.3, we have two functional representations for assumptions A1 and A2.

We show that they can be combined in the following result.

**Theorem 3.8.1.** *A 3-dimensional array* $\mathbf{X}$ *satisfies assumptions* $(A1)$ *and* $(A2)$ *if and only if there exist a function* $f$*, such that for independent collections of i.i.d. U(0,1) random variables the following representation holds:*

$$\mathbf{X} = (X_{i,j,t}) \overset{d}{=} \big(f(\alpha, u_i, u_j, w_{\{i,j\}}, \lambda_{\{i,j\},t}, X_{i,j,t-1})\big) \tag{3.2}$$

*Proof.* $(\Rightarrow)$:

We show a general case for (separately) RCE array, and then just change uniform random variables in the representation accordingly for jointly RCE array.

As $\mathbf{X}$ is a RCE array, we can repeat initial steps of the proof of functional representation theorem 3.6.1, taken from Aldous [1985]. Suppose that $\mathbf{X}$ is a part of the array $(X_{i,j})_{i,j\in\mathbb{Z}}$. Define $A, B, C$, the following way:

$$\begin{aligned} A &= (X_{i,j})_{i,j\le 0}, \\ B_i &= (X_{i,j})_{j\le 0}, \quad B = (B_i)_{i\ge 1}, \\ C_j &= (X_{i,j})_{i\le 0}, \quad C = (C_j)_{j\ge 1}. \end{aligned}$$

Following Aldous [1985, 14.11], we have that:

1. $B_1, B_2, \dots, C_1, C_2, \dots$ are conditionally independent and identically distributed, given $A$. Conditional distributions of $B_i$ and $C_j$ do not vary with $i$ and $j$ respectively

2. $(X_{i,j})_{i,j\ge 1}$ are conditionally independent given $(A, B, C)$. Conditional distribution of $X_{i,j}$ depends only on $(A, B_i, C_j)$.

Take a coding $\alpha$ for $A$ and condition everything on $\alpha$. Then we can choose codings $\xi = (\xi_i)$ for $B$ and $\eta = (\eta_j)$ for $C$. As, conditionally on $A$, $B$ and $C$ are independent, so are the uniform sequences $(\xi_i)$ and $(\eta_j)$. Then there exist a coding function $g_1$, such that $g_1(\alpha, \xi_i, \eta_j)$ is the conditional distribution of $X_{i,j}$, given $(A, B, C)$.

So, given $(\alpha, \xi, \eta)$, the distribution of $X_{i,j}$ is known and depends only on $(\alpha, \xi_i, \eta_j)$. As $X_{i,j} = (X_{i,j,t})_{t\geq 1}$ is a sequence, it means that the joint distribution of the whole sequence is known. And, in particular, the transition matrix $\pi^{i,j}$ of $X_{i,j}$ is

$$\pi^{i,j} \overset{d}{=} g_2(\alpha, \xi_i, \eta_j) \tag{3.3}$$

for some function $g_2$.

But, given the transition matrix, $X_{i,j}$ is a Markov chain. So, by our results from Theorem 3.5.3,

$$(X_{i,j,t})|\pi^{i,j} \overset{d}{=} \left(g_3^{\pi^{i,j}}(X_{i,j,t-1}, \lambda_t^{i,j})\right). \tag{3.4}$$

As the $X_{i,j}$ are conditionally independent, given $(\alpha, \xi, \eta)$, so the $(\lambda_t^{i,j})$ can be taken from sequences independent across $(i, j)$.

Combining (3.3) with (3.4), we obtain:

$$(X_{i,j,t})|(\alpha, \xi, \eta) \overset{d}{=} (g_4(\alpha, \xi_i, \eta_j, X_{i,j,t-1}, \lambda_{i,j,t}))$$

for some function $g_4$. And as given $(\alpha, \xi, \eta)$, the $(X_{i,j})$ are independent, we use coding $\beta_{i,j}$ to get the unconditional representation

$$(X_{i,j,t}) \overset{d}{=} (f(\alpha, \xi_i, \eta_j, \beta_{i,j}, X_{i,j,t-1}, \lambda_{i,j,t}))$$

Changing the indexing of uniform collections for jointly RCE array, we get the desired representation (3.2).

$(\Leftarrow)$:

For fixed $(i, j)$ define

$$\gamma^{(i,j)} = \begin{bmatrix} \alpha \\ \xi_i \\ \eta_j \\ \beta_{i,j} \end{bmatrix}$$

Then, because of the independence and identical distribution of the initial collections of uniform random variables, we have that $\gamma^{(i,j)}$ are independent from $\lambda_{i,j,t}$. Then

$$X_{i,j} = (X_{i,j,t})_{t\geq 1} \overset{d}{=} \left(f\left(\gamma^{(i,j)}, X_{i,j,t-1}, \lambda_{\{i,j\},t}\right)\right)_{t\geq 1}$$

and by changing function as needed we can obtain representation

$$X_{i,j} = (X_{i,j,t})_{t\geq 1} \overset{d}{=} (\bar{f}(\bar{\gamma}, X_{i,j,t-1}, \bar{\lambda}_t))$$

with $\bar{\gamma}, \bar{\lambda}_t$ being independent $U(0,1)$ random variables. And from (3.1), $X_{i,j}$ is Markov exchangeable sequence. And this holds true for every pair $(i,j)$, as needed for assumption (A2).

Now to show that (A1) also holds, let similarly

$$\tilde{\beta}_{\{i,j\}} = \begin{bmatrix} \beta_{\{i,j\}} \\ (\lambda_{\{i,j\},t}) \\ (X_{i,j,t-1}) \end{bmatrix}$$

By construction $\tilde{\beta}_{\{i,j\}})$ are i.i.d. Note that joint distribution of shifted sequences $(X_{i,j,t-1})$ have identical distributions because $\mathbf{X}$ is RCE.

Then

$$\mathbf{X} \overset{d}{=} \left(\tilde{f}(\alpha, \xi_i, \xi_j, \tilde{\beta}_{\{i,j\}})\right),$$

and by appropriate change of the function we obtain a RCE functional representation for $\mathbf{X}$ by uniform(0,1) collections. This shows, by virtue of Theorem 3.6.1, that condition (A1) is satisfied.

□

**Corollary.** *It can be shown that an alternative representation in terms of successors is the following:*

$$\mathbf{Y} = (Y_{i,j,k,l}) \overset{d}{=} \left(f_k(\alpha, \xi_i, \xi_j, \beta_{\{i,j\}}, \lambda_{\{i,j\},k,l})\right)$$

*Here $\mathbf{Y}$ is a RCE array of partially exchangeable arrays, and $Y_{i,j,k,l}$ represents l-th successor of state k in the sequence at the $(i,j)$-th element of the array $\mathbf{Y}$.*

Starting from this representation, new models can be motivated for dynamic network modeling. The arising interpretation is reasonable: the array is modeled as function of some non-changing effects on individual and pair level, by time-evolving effects, and previous values, capturing dependence.

# Chapter 4

# Dynamic latent distance model for network analysis

In this chapter we introduce a new statistical model for analyzing dynamic networks. In particular, we model the overall network cohesion over time by means of an infinite hidden Markov model. The performance of the model is illustrated on a simulated data example, and it is also applied to the INFOCOM dataset discussed in Chapter 2.

## 4.1 Introduction

One of the main aims of the statistical analysis of network data is the discovery of hidden structures underlying relations between actors. This helps to understand the nature of the network and its properties. For the dynamic networks observed over time, the underlying structures are also evolving, and capturing such evolution is of interest. Further in the text we speak about ”dynamic data” or ”dynamic networks”, assuming that the set of actors is invariant, but the set of links between them can change over time.

Many models have been proposed for static network data and most of them have extensions to the case of dynamic networks. As discussed in Chapter 2, the latent variables approach provides a rich class of models able to capture various properties of network data. Therefore it is natural to consider dynamic extensions of this class of models. The most common idea for introducing dynamics is to assume that the underlying latent variables evolve over time. Proposed dynamic extensions usually assume an independent Markov evolution for these variables, as in Foulds et al. [2011], who give a dynamic extension of the latent feature model of Miller et al. [2009]; or Sarkar and Moore [2005], who extend the latent distance model of Hoff et al. [2002]. Other proposals include Rodriguez [2012], who used an infinite hidden Markov model for class memberships variables in the infinite relational model of Kemp et al. [2006]; Westveld and Hoff [2011], where the latent variables are sender and receiver effects, and they are considered to have a first-order autoregressive structure over time.

A common limitation for most dynamic models is the assumption that the evolution of the latent variables is independent. This can be generalized by assuming an underlying global process governing the latent variable evolution.

In this chapter we present a new model for the analysis of network data evolving over time. It is a dynamic extension of the latent distance model of Hoff et al. [2002], that we describe in detail in Section 4.2. On its basis, we account for temporal evolution by means of an infinite hidden Markov model for the latent positions. Thus, our model keeps track of the general process, underlying changes in latent positions of actors, allowing to discover structural changes in the cohesion of the network over time, while keeping the advantages of the static model for each time point, i.e. being able to capture individual behavior at a fixed time. This is a simple idea, but it is worth exploring, as it introduces the assumption of a dependent evolution of the latent positions. Even with a single parameter controlling this evolution over time, some aspects, such as the implemetation of a MCMC algorithm for inference, the interpretation and the presentation of results, present challenges. Therefore, for ease of the exposition, we develop this simple one-parameter case, while the extension or generalization to a more complex and complete model is rather straightforward.

We start by reminding the latent distance model in Section 4.2, then describe the idea of capturing the cohesion of the network, and explain infinite hidden Markov models. Our dynamic model follows in Section 4.4, explaining estimation procedures in Section 4.5. Examples are shown in Section 4.6.

## 4.2 Latent distance model for static networks

Based on the idea of a social space, that is a space of unobserved latent characteristics, Hoff et al. [2002] defined the latent distance model for static network data. The idea of a social space manifests that each actor has a position, determined by his characteristics, in this space. Then for actors that are similar, their positions will be close. The latent distance model implements this assumption by letting the probability of a link to depend on the distance between actors' positions in the unobserved social space.

Additionally, the links are taken to be conditionally independent given the relative distances between actors.

The model was introduced for binary network data. Formally, let the data be encoded in the array $Y = (Y_{i,j})_{i,j=1\ldots N}$, where each $Y_{i,j}$ is 0 or 1 denoting relation between actors $i$ and $j$. Both directed or undirected relations can be modeled with this approach. Covariates may also be observed, which can be either on an individual or a pairwise level. Denote the covariate information by $\mathbf{X} = (X_{i,j})_{i,j=1\ldots N}$ where $X_{i,j} = (X_{i,j,k})_{k=1,\ldots,p}$ for each pair $(i, j)$.

Putting the model in the general framework of latent variable models, introduced in Chapter 2, here the role of latent variable for each actor $i$ is played by its position $Z_i$ in a latent space $\mathbb{R}^d$. The model itself is a logistic regression with the conditional independence assumption:

$$P(Y|Z, \mathbf{X}, \beta) = \prod_{i \neq j} P(Y_{i,j}|Z_i, Z_j, X_{i,j}, \beta)$$

$$P(Y_{i,j} = 1|Z_i, Z_j, X_{i,j}, \beta) = \text{logit}^{-1}(\beta_0 + \beta_1^T X_{i,j} - \|Z_i - Z_j\|_d)$$

The Euclidean distances, denoted by $d_{i,j} = \|Z_i - Z_j\|_d$, are used in the model specification. As discussed in Hoff et al. [2002], it is possible to substitute them with a set of other distance, as long as the triangle inequality is satisfied for $d_{i,j}$'s. The dimension of the latent space must be specified; for parsimony and better interpretability it is usually chosen to be low, i.e. 2 or 3.

The model allows to capture such properties of the dependence in the data as reciprocity, transitivity and homophily, explained in Section 2.3. Furthermore, it provides a model-based spatial representation of the data: by plotting the estimated latent positions together with the observed links, we obtain a visualization of the assumed social space. Of course, this is feasible only if the dimension of the latent space is chosen to be low.

Inference for this model has been developed both by maximum-likelihood and Bayesian approach. The fact that the likelihood is convex as a function of pairwise distances may be used for likelihood maximization and finding estimates for the distances. Then, approximated latent positions can be found by means of multidimensional scaling methods. These methods are intended for finding estimates for positions corresponding to a given set of distances. These are the estimates used in the maximum-likelihood approach. In the Bayesian approach priors are formulated for $Z$ and $\beta$ and inference on the unknown parameters is carried out by MCMC sampling. The implementation details for both approaches are provided in the original paper by Hoff et al. [2002].

One of the challenges arising in computations which is worth mentioning here, is the un-identifiability of the latent positions. Note that if we rotate, reflex or shift all the positions together, the joint likelihood will not change, as it actually depends only on the relative distances between points, which are preserved by the aforementioned transformations. These are called "Procrustean transforms". To handle this issue in the posterior computation, Hoff et al. [2002] select, within each equivalence class of configurations, the one that is closest, in mean square difference, to a reference configuration, specified beforehand. The chosen configuration is then used as the estimate of the set of actor positions in the latent space.

Extensions of the static latent distance model have been proposed, to account for further network properties, see Hoff [2005], Handcock et al. [2007], Krivitsky et al. [2009]. In particular, the prior for the $Z$'s can be taken to be a mixture of normal distributions, in order to capture a possible clustering structure amongst actors. Additionally, variables representing possible variability in sending and receiving links can be incorporated into the model.

## 4.3 Towards dynamics: modeling network "cohesion"

In the static case, as described above, the prior distributions for the latent positions $Z = (Z_i)_{i=1...N}$ are chosen such that $(Z_i)$ are a priori i.i.d. and distributed as $\mathrm{N}_d(0, \sigma^2 I_d)$. As $Z$ are latent, usually the prior knowledge is vague, so a large value of $\sigma^2$ is used. In this case, the prior information will have little effect on the posterior, which will, therefore, be driven by the data.

We are interested to study overall network "cohesion", or density of links, depending on the parameters. It is influenced by $\beta_0$ and $\sigma^2$ (and $\beta_1$, if present).

If we simulate from the model with different values of $\beta_0$ and $\sigma^2$, the resulting datasets can vary a lot in the network cohesion. In the Figure 4.1 an example is shown. The total number of links, considered as a measure of the network cohesion, varies from 87 to 12 for $\beta_0 = 1$ when $\sigma^2 = 4$ and $\sigma^2 = 16$, respectively. Smaller values of $\sigma^2$ mean that actors have closer positions, and, therefore, tend to form more links. When $\sigma^2$ grows the cohesion decreases.

At the same time, $\beta_0$ captures the maximum possible probability of a link, i.e. when the relative distance is zero. It can be seen that in the absence of covariates, if $Z_i = Z_j$ the probability of having a link is $\mathrm{logit}^{-1}(\beta_0)$ which grows quickly from $\frac{1}{2}$ when $\beta_0 = 0$ to 1 when $\beta_0$ goes to infinity.

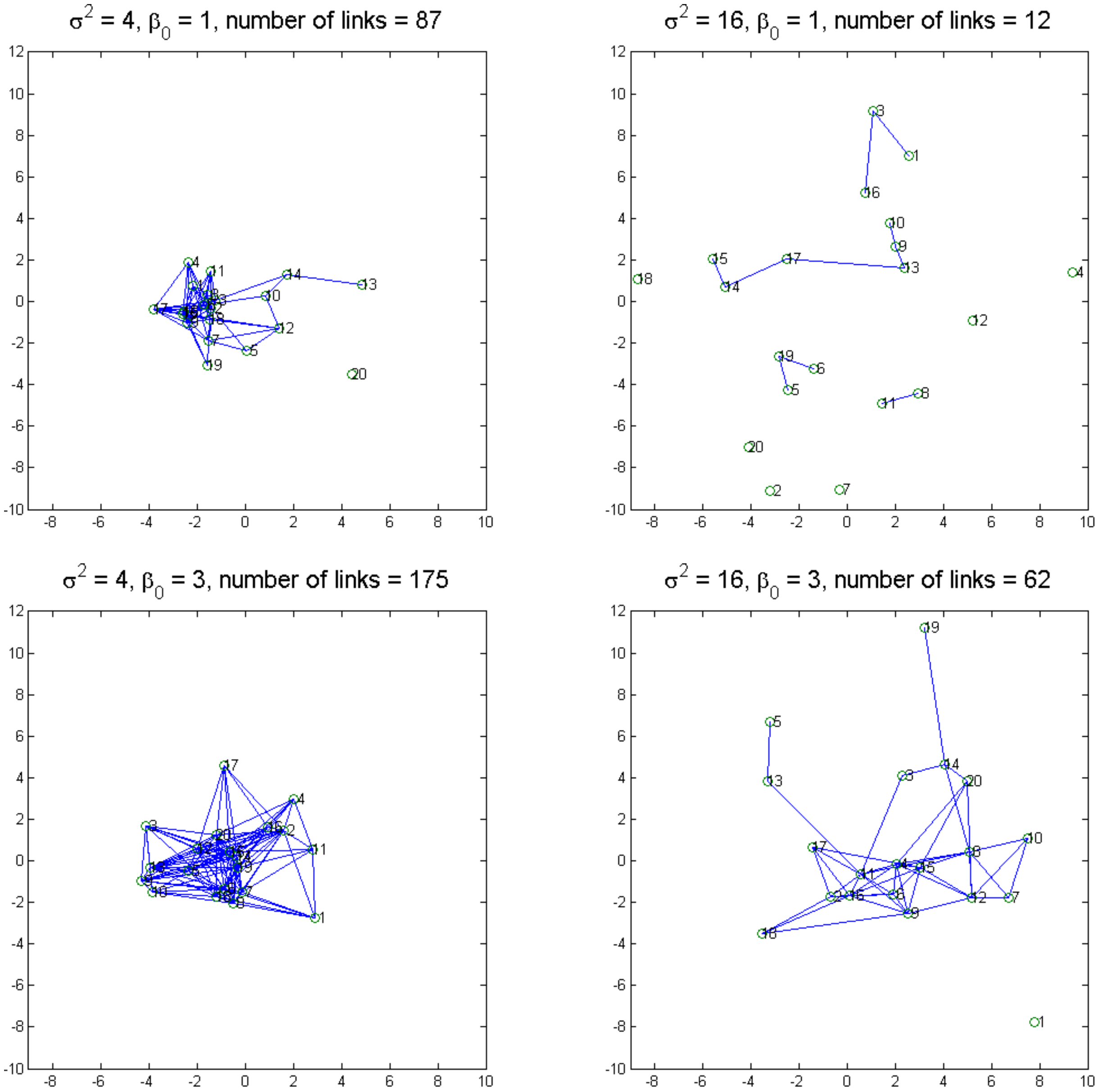


Figure 4.1: Four networks with $N = 20$, simulated from the static latent distance model with different values of parameter $\sigma^2$ in the prior distribution of the latent positions $Z$'s and different $\beta_0$. By columns $\sigma^2 = 4$ and $\sigma^2 = 16$, by rows $\beta_0 = 1$ and $\beta_0 = 3$.

Thus, the role of $\beta_0$ is to provide a baseline for the overall cohesion, which then can be changed by changing $\sigma^2$. In the second row on Figure 4.1 $\beta_0 = 3$, and the total number of links is increased to 175 and 62 correspondingly.

For dynamic data, we assume the existence of an underlying phenomenon affecting the relations between actors. This can be modeled through a process, explaining the time evolution of the latent variables. In the dynamic extension we propose for Hoff et al. [2002] model, we model changes in cohesion over time, by including $\sigma^2$ in the model, and allowing it to over time. This is a simple way of modeling time evolution, which assumes the underlying process affects all individuals equally.

As a measure of the network cohesion we choose overall number of links, that is influenced both by the static $\beta_0$ and the sequence of variance parameters $\sigma^2$. We comment on this in the examples of Section 4.6.

### 4.3.1 Infinite hidden Markov models

In order to model the time-evolution of the underlying process that governs the changes in the latent positions, we use an infinite hidden Markov model. This is a nonparametric extension of classical hidden Markov models. In this section, we briefly describe these models and some relevant computational challenges for inference.

In general, hidden Markov models (HMMs) assume that there is an observed process $Z = (Z_t)$, that depends on a latent (hidden) discrete-valued Markov process $S = (S_t)$, called the *state process.*

*Remark.* In our model $Z_t$ is the vector of all latent positions at time $t$, so it is not observed. In this section we give description of HMMs in general.

Given $S = (S_t)$, the $Z$'s are independent, with each $Z_t$ having a distribution $F(\cdot|\phi_{S_t})$, that depends on $S$ only through $S_t$. The state process $S$ is Markov, taking values in $\{1, 2, \ldots, K\}$, and its probability law is described by the initial distribution $P(S_0)$ and the transition probabilities $\pi_{i,j} = P(S_t = j|S_{t-1} = i)$, for $i, j \in \{1, \ldots, K\}$, that form a transition matrix $\pi = (\pi_{i,j})$.

A limitation of hidden Markov models is the assumption of a fixed number of possible values for the state process. This assumption may lead to inability to capture dependence structure in the data in full, as fixing a particular number of states forces the model to find these states even if the data does not justify this choice. At the same time, often one can have only vague idea on the possible number of states.

To overcome this limitation, Beal et al. [2002] introduced a nonparametric extension termed the “infinite hidden Markov model” (iHMM). The model was refined by Teh et al. [2006] in a hierarchical Bayesian way. The idea is to allow a countably infinite number of values for the hidden states, and to use a hierarchical Dirichlet process prior on the infinite-dimensional transition matrix.

For finite HMMs, a symmetric Dirichlet distribution is a common choice of prior for the transition probabilities. As a natural extension, the basic idea of iHMMs is to assign to each row $\pi_i$ of the transition matrix a Dirichlet process (DP, Ferguson [1973]) prior. However, a DP with a diffuse base distribution would not allow for a shared support for the different probabilities $\pi_k$. The solution is to use another DP as a base measure, giving the hierarchical DP construction of Teh et al. [2006]. See Section 4.4 for details.

Note that a HMM can be considered as a set of conditional finite mixture models. This can be seen by looking at the relevant part of the specification of the HMM:

$$\begin{aligned} S_t|S_{t-1}, \pi = (\pi_i)_{i=1\ldots K} &\sim \pi_{S_{t-1}} \\ Z_t|S_t, (\phi_i)_{i=1\ldots K} &\sim F(\cdot|\phi_{S_t}) \end{aligned}$$

For each value $i$ of the current state $t$ the row $\pi_i$ of the transition matrix gives the mixing proportions for the next state $t+1$. In other words, the next observation is drawn from the mixture component indexed by the value of the next state. In our notation:

$$(Z_{t+1}|S_t = i) \sim \sum_{j=1}^{K} \pi_{i,j} F(\cdot|\phi_j)$$

Thus, the infinite HMM can be defined by replacing the set of conditional finite mixture models in the finite HMM with conditional hierarchical DP mixture models.

As can be expected, inference for such models is not an easy task. We mention briefly the relevant algorithms and ideas that will surface again in the description of inference in our model. Teh et al. [2006] presented a Gibbs-sampling algorithm for the posterior estimation of the parameters of the iHMM. A drawback of this algorithm is the slow mixing due to a likely correlation in the time series data. The "beam sampler" introduced in Van Gael et al. [2008] addresses this problem by sampling the whole state sequence $S$ at once. It also exploits the idea of slice sampling by introducing auxiliary variables such that, depending on them, the number of possible trajectories of the state process is finite. The slice sampler was introduced by Neal [2003]; the idea of applying it for mixtures of Dirichlet processes first appeared in the paper by Walker [2007] and a was further extended by Kalli et al. [2011]. Our estimation algorithm is built upon the latter paper (see Section 4.5).

## 4.4 Our proposed model

We now formalize our proposed model for dynamic networks. The observed data is a discrete-time sequence $(Y_t)_{t=1\dots T}$ of binary $N \times N$ arrays $Y_t = (Y_{i,j,t})_{i,j=1\dots N}$, where $Y_{i,j,t}$ indicates the presence or absence of a link between individuals $i$ and $j$ at time point $t$. Possibly, vectors of covariates $X_{i,j,t} = (X_{i,j,t,k})_{k=1\dots p}$ are also available for each pair $(i,j)$. For the ease of notation we sometimes write $Y_{ijt}$, $Z_{it}$, $X_{ijt}$ instead of $Y_{i,j,t}$, $Z_{i,t}$, $X_{i,j,t}$.

We extend the static model presented in the section 4.2 to the case of dynamic data by allowing the latent positions $Z$ to evolve over time. As we mentioned before, our assumption is that their evolution can be described by an infinite hidden Markov model. This allows us to model a general underlying process and identify structural changes, corresponding to the evolution of the general "cohesion" of the network. The evolution of the latent positions of the actors, in this case, is not independent.

As before, each actor $i$ at time $t$ is assumed to have a (latent) position $Z_{i,t}$ in a $d$-dimensional Euclidean space, so $Z_{i,t} = (Z^l_{i,t})^d_{l=1}$. Further we assume that the latent positions evolve according to a state process $S =$

$(S_t)_{t=1,\dots T}$, which is a discrete-time Markovian process on $\{1, 2, \dots\}$. The presence (or absence) of a link between two individuals $i$ and $j$ at time $t$ is conditionally independent of all other links, given the state of the system $S_t$, and positions $Z_{i,t}$ and $Z_{j,t}$. Formally,

$$Y_t|Z_t, X_t, S_t, \beta \overset{ind}{\sim} \prod_{i \neq j} P\left(Y_{ijt}|Z_{it}, Z_{jt}, S_t, X_{ijt}, \beta\right).$$

As our interest is in binary data, for defining the distribution $P\left(Y_{ijt}|Z_{it}, Z_{jt}, S_t, X_{ijt}, \beta\right)$ it is enough to specify the probability of having a link. To this aim we use the logistic regression model:

$$P\left(Y_{ijt} = 1|Z_{it}, Z_{jt}, S_t, \beta_0, \beta_{1t}, X_{ijt}\right) \overset{ind}{\sim} \text{logit}^{-1}(\beta_0 + \beta_{1t}' X_{ijt} - \|Z_{it} - Z_{jt}\|_d),$$

where $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. This can be rewritten, for each $i, j$ and $t$, with $y_{ijt} = 0$ or 1, as:

$$\begin{aligned} &P(Y_{ijt} = y_{ijt}|Z_{it}, Z_{jt}, S_t, \beta_0, \beta_{1t}, X_{ijt}) = \\ =& \frac{\exp\left(y_{ijt}(\beta_0 + \beta_{1t}' X_{ijt} - \|Z_{it} - Z_{jt}\|_d)\right)}{1 + \exp\left(\beta_0 + \beta_{1t}' X_{ijt} - \|Z_{it} - Z_{jt}\|_d\right)}. \end{aligned} \tag{4.1}$$

Furthermore, we assume that the latent positions $(Z_t)_{t \geq 1}$ are obtained from an infinite HMM, that is:

$$Z_t|S_t, (\sigma_k^2) \overset{ind}{\sim} \prod_{i=1}^{N} \text{N}_d(Z_{it}|0, \sigma_{S_t}^2 I_d)$$

Here by $\text{N}_d$ we denote density of the $d$-variate normal distribution and $Z_t = (Z_{it})_{i \geq 1}$.

As we can see from equation (4.1), the distribution of $Y$ depends on $Z$ only via pairwise distances between actors' positions. So if we rotate, reflex or shift all the positions together, the joint distribution does not change. We will elaborate on this issue in the next section, when describing estimation procedures, but for now we note that this means that it is unnecessary to assume different means for the distributions of $Z_{it}$'s for different $t$. And for each $t$, we can always center the positions in order to have mean 0 and preserve the relative distances.

We can see that the dependence on the underlying time evolution is introduced via the variance parameter $\sigma^2_{S_t}$. As discussed in Section 4.2 this parameter controls the general cohesion of the network at time $t$. The prior on $\sigma^2_k$ is taken to be conjugate, and is given by:

$$\sigma^2_k \overset{iid}{\sim} \text{InvGam}(a_\sigma, b_\sigma),\ k = 1, 2, \dots$$

The state process $(S_t)$ takes values in $\{1, 2, \dots\}$ and, conditional on the transition matrix $\pi$, it is Markovian, i.e.

$$P(S_t = l | S_{t-1} = k, \pi) = \pi_{k,l}$$

From the specification of an infinite HMM, $\pi$ is assumed to have a hierarchical Dirichlet process prior, so that the rows $\pi_k$ of $\pi$ are conditionally independent Dirichlet processes:

$$\pi_k | \eta \overset{ind}{\sim} \text{DP}(\alpha, \eta)$$

and

$$\eta = \sum_{j=1}^{\infty} \eta_j \delta_j \text{ with } (\eta_j) \sim \text{SB}(\alpha_0)$$

where $\text{DP}(\alpha, \eta)$ is a Dirichlet Process with concentration parameter $\alpha$ and base measure $\eta$; and $\text{SB}(\alpha_0)$ represents the so called “stick breaking” construction:

$$\eta_j = v_j \prod_{k=1}^{j-1} v_k,\ (v_j) \overset{iid}{\sim} \text{Beta}(1, \alpha_0)$$

Finally, the prior distributions for $\beta = (\beta_0, \beta_1)$ are specified as follows:

$$\beta_0 \sim \text{Gamma}(a_{\beta_0}, b_{\beta_0})$$

$$\beta_{1t} \overset{iid}{\sim} \text{N}_p(\xi, \Phi)$$

Note that the prior for $\beta_0$ has a positive support. The choice is justified by the role of $\beta_0$: in the absence of covariates, if the distance between two positions is 0, the probability of having a link is $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$, so if $\beta_0 < 0$, it becomes less than $\frac{1}{2}$, which does not reflect the idea of a social space described in 4.2.

We can consider hyperparameters $\alpha$ and $\alpha_0$ as fixed, or use hyperpriors:

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha),\ \alpha_0 \sim \text{Gamma}(a_{\alpha_0}, b_{\alpha_0}).$$

## 4.5 Estimation

Summarizing our model we have:

$$P(Y = y|Z, \beta_0, \beta_1, \mathbf{X}) = \prod_{t=1}^{T} \prod_{i \neq j} \frac{\exp\left(y_{ijt}(\beta_0 + {\beta_1}_t' X_{ijt} - \|Z_{it} - Z_{jt}\|_d)\right)}{1 + \exp\left(\beta_0 + {\beta_1}_t' X_{ijt} - \|Z_{it} - Z_{jt}\|_d\right)}$$

$$Z_t|S_t, (\sigma_k^2) \overset{ind}{\sim} \prod_{i=1}^{N} \text{N}_d(Z_{it}|0, \sigma_{S_t}^2 I_d)$$

$$P(S_t = l|S_{t-1} = k, \pi) = \pi_{kl}$$

$$\pi_k|\eta \sim \text{DP}(\alpha, \eta);\ \eta = \sum_{j=1}^{\infty} \eta_j \delta_j;\ \eta_j \sim \text{SB}(\alpha_0)$$

$$\sigma_k^2 \overset{iid}{\sim} \text{InvGam}(a_\sigma, b_\sigma),\ k = 1, 2, \ldots$$

$$\beta_0 \sim \text{Gamma}(a_{\beta_0}, b_{\beta_0});\ \beta_{1t} \overset{iid}{\sim} \text{N}_p(\xi, \Phi)$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha);\ \alpha_0 \sim \text{Gamma}(a_{\alpha_0}, b_{\alpha_0}),$$

where $a_\alpha, b_\alpha, a_{\alpha_0}, b_{\alpha_0}, a_{\beta_0}, b_{\beta_0}, a_\sigma, b_\sigma, \xi, \Phi$ are hyperparameters to be specified.

We would like to estimate the parameters of the model, that is $(\sigma_k^2), \beta_0$, underlying state sequence $(S_t)$ along with transition matrix $\pi$, the positions $(Z_{i,t})$ of the actors in the latent space and, when present, covariate coefficients $\beta_1$.

We use a Bayesian approach for estimation, and, as is often the case with latent distance models, only vague information is known about the parameters, we therefore specify hyperparameters corresponding to diffuse priors on $\beta$ and $(\sigma_k^2)$.

To estimate the posterior distribution of the parameters we build a Markov chain Monte Carlo algorithm described below. The MCMC samples are used for inference on the model parameters and latent states, as well as their distributions. In addition, point estimates for the covariate coefficients and the cohesion parameters ($\sigma_k^2$) may be of interest.

For the posterior computation we build a Gibbs sampling scheme updating all the components sequentially. For the state sequence and transition matrix we implement a modification of the beam sampling algorithm by Van Gael et al. [2008], updating the whole trajectory of $(S_t)_{t=1\ldots T}$ simultaneously at each iteration. We cannot use the algorithm directly, as, in our case, the iHMM describes thee *unobserved* $(Z_{it})$, in other words we have an additional latent layer in the hierarchical representation of the model.

The parameters $\eta, \pi$ are updated following Teh et al. [2006]. The variances $(\sigma_k^2)$, for which conjugate priors were specified, are updated by sampling from the corresponding full conditional distributions. Finally, the coefficients, $\beta$, and latent positions, $Z$, are updated using random walk Metropolis-Hastings algorithm.

As we mentioned before, there is a non-identifiability issue regarding the latent positions $Z$, due to the fact that probabilities of links depend on $Z$ only through the distances between pairs of individual locations. If we rotate, reflect or move all the positions at the same time, the joint likelihood does not change. To handle this issue in the estimation procedure, we follow the approach of Hoff et al. [2002] and apply a Procrustean transformation after resampling the new positions. Such transformation chooses the closest configurations (in terms of sum of squared distances) to a predefined set of reference positions, within the class of equivalent configurations, i.e. the set of configurations with the same pair-wise distances. This is done for each time point $t$. Denote this closest configuration by $Z^*$, dropping the $t$ index, then $Z^* = \arg\min_{TZ} \operatorname{tr}(Z_0 - TZ)'(Z_0 - TZ)$, where $Z_0$ are reference positions, and transformation $T$ ranges over the set of possible rotations, reflections and shifts. If $Z$ and $Z_0$ are centered at the origin, $Z^*$ can be computed by taking $Z^* = Z_0 Z'(Z Z_0' Z_0 Z')^{-1/2} Z$. For the $Z_0$ we find approximate maximum-likelihood estimates for the latent positions. This is possible as the likelihood is convex with respect to distances, and after finding MLE for distances $d_{i,j} = \|Z_{i,t} - Z_{j,t}\|_d$, we apply multidimensional scaling to get the estimates for $(Z_{i,t})$'s. We then use these estimates also as starting point for the Markov chain in the MCMC algorithm to speed up convergence.

### 4.5.1 Sampling the hidden state sequence $(S_t)$

We build our algorithm based on the beam sampler developed by Van Gael et al. [2008]. In the beam sampler, auxiliary variables $U = (U_t)_{t\geq 1}$ are introduced, such that conditionally on $U$, the number of state trajectories with positive probability is finite. After that, dynamic programming is used to compute the conditional probabilities of each of these trajectories and thus sample whole trajectories efficiently.

For each $t$, the auxiliary variable $U_t$ is sampled from the conditional distribution

$$U_t|S_{t-1}, S_t, \pi \sim \text{Uniform}(0, S_t \pi_{S_{t-1},S_t}).$$

We have augmented the original algorithm by changing the distribution of the auxiliary variables from initial $\text{Uniform}(0, \pi_{S_{t-1},S_t})$ to expand the number of possible trajectories and speed up mixing following ideas from Kalli et al. [2011].

After sampling the $U$'s, we sample the state sequence $S_t$ given $U$ and other variables using a forward-filtering backward sampling algorithm:

First, we calculate the full conditional distributions

$$P(S_t|Z_{1:t}, U_{1:t}) = P(S_t|Z_1, \ldots, Z_t, U_1, \ldots, U_t),$$

where, for the ease of notation, we omit the dependence on $\pi$ and other variables:

$$\begin{aligned}
P(S_t|Z_{1:t},U_{1:t}) &\propto P(S_t,Z_t,U_t|Z_{1:t-1},U_{1:t-1}) \\
&= \sum_{S_{t-1}} P(Z_t|S_t)P(U_t|S_t,S_{t-1})P(S_t|S_{t-1})P(S_{t-1}|Z_{1:t-1},U_{1:t-1}) \\
&= P(Z_t|S_t)\sum_{S_{t-1}} \frac{1}{S_t}\mathbb{I}(U_t < S_t\pi_{S_{t-1},S_t})P(S_{t-1}|Z_{1:t-1},U_{1:t-1}) \\
&= \frac{P(Z_t|S_t)}{S_t} \sum_{S_{t-1}:u_t<S_t\pi_{S_{t-1},S_t}} P(S_{t-1}|Z_{1:t-1},U_{1:t-1})
\end{aligned}$$

Note that, after introducing the auxiliary variables, there are only finitely many trajectories for $S_t$ with non-zero probability. Furthermore, the sum over the possible values of $S_{t-1}$ actually involves a finite number of elements, after the truncation by $U_t$.

Then, in order to sample the complete trajectory $S$, we perform a backward pass, i.e., we first sample $S_T$ from $P(S_T|Z_{1:T},U_{1:T})$, and then, for each $t$ we sample $S_t$ given $S_{t+1}$ from the updated full conditional distribution:

$$P(S_t|S_{t+1},Z_{1:T},U_{1:T}) \propto P(S_t|Z_{1:t},U_{1:t})P(S_{t+1}|S_t,U_{t+1})$$

For a more detailed account of beam sampling see the original paper by Van Gael et al. [2008].

### 4.5.2 Sampling the iHMM parameters $\pi$ and $\eta$

Following Teh et al. [2006], the conditional distribution of $\left(\pi_{k,1},\ldots,\pi_{k,K},\sum_{l=K+1}^{\infty}\pi_{k,l}\right)$ given $S,\eta,m$ is

$$\text{Dirichlet}\left(t_{k,1}+\alpha\eta_1,\ldots,t_{k,K}+\alpha\eta_K,\alpha\sum_{i=K+1}^{\infty}\eta_i\right)$$

where $t_{k,l}$ is the number of of transitions from $k$ to $l$ in the trajectory $S=(S_t)_{t=1,\ldots T}$, and $K$ is the number of distinct states in $S$.

For the sampling of $(\eta_j)$, we again follow Teh et al. [2006], introducing further auxiliary variables $m_{i,j}$ such that

$$P(m_{i,j}=m|S,\eta,\alpha) \propto s(t_{i,j},m)(\alpha\eta_j)^m$$

for $m = 1, \ldots, N$, where $s(t,m)$ are unsigned Stirling numbers of the first kind and $m_{i,j}$ correspond to the number of tables in restaurant $i$ serving dish $j$ in the Chinese Restaurant Franchise metaphor(see Teh et al. [2006] for details).

The conditional distribution of $\left(\eta_1, \ldots, \eta_K, \sum_{l=K+1}^{\infty} \eta_l\right)$ is then

$$\text{Dirichlet}(m_{\cdot,1}, \ldots, m_{\cdot,K}, \alpha_0).$$

### 4.5.3 Sampling the “cohesion” parameters $\sigma_k^2$

As we have specified a conjugate prior for $\sigma_k^2$, the updates are obtained by sampling from the full-conditional:

$$P(\sigma_k^2|S, Z) \sim \text{InvGam}\left(a_\sigma + NT_k, b_\sigma + \frac{1}{2}\sum_{t:S_t=k}\sum_{i=1}^{N}\sum_{l=1}^{d} {Z_{i,t}^{l}}^2\right),$$

where $T_k$ is number of time points state process $S_t$ spent in the state $k$. The $\sigma_k^2$'s can be updated for all $k$ simultaneously, due to of the conditional independence assumption.

### 4.5.4 Sampling the latent positions $Z$

The full conditional distribution for $Z$ is given by

$$P(Z_{it}|S_t = k, Y, (\sigma_k^2), \beta, X) \propto P(Y|Z_{it}, \beta, X)\text{N}_d(Z_{it}|0, \sigma_k^2 I_d)$$

We use a random walk Metropolis-Hastings algorithm with a $d$-variate Normal proposal:

First, draw $Z_{it}^* \sim N_d(\cdot|Z_{it}, \tau_z^2 I_d)$ with a predefined variance parameter $\tau_Z^2$.

Accept with probability

$$\frac{P(Y|Z_{it}^*, \beta, X)\text{N}_d(Z_{it}^*|0, \sigma_{S_t}^2 I_d)}{P(Y|Z_{it}, \beta, X)\text{N}_d(Z_{it}|0, \sigma_{S_t}^2 I_d)},$$

where $\text{N}_d$ denotes the density of a $d$-variate Normal distribution.

After that, apply a Procrustean transformation to choose a configuration of positions $Z$, with minimum sum of squared distances to the reference configuration $Z_0$, as described earlier in this section.

### 4.5.5 Sampling the covariate coefficients $\beta$

For the covariate coefficients we also use random walk Metropolis Hastings step. For $\beta_0$ we use a truncated uniform proposal, as $\beta_0$ must be positive:

First, draw $\beta_0^* \sim |\mathrm{U}(\beta_0 - \delta, \beta_0 + \delta)|$ with a predefined parameter $\delta$.

Accept with probability

$$\frac{P(Y|Z, \beta_0^*, \beta_1, X)\mathrm{Ga}(\beta_0^*|a_{\beta_0}, b_{\beta_0})}{P(Y|Z, \beta_0, \beta_1, X)\mathrm{Ga}(\beta_0|a_{\beta_0}, b_{\beta_0})},$$

where Ga denotes density of a Gamma distribution.

For $\beta_{1t}$ we use a $p$-variate Normal proposal:

First, draw $\beta_{1t}^* \sim N_p(\beta_{1t}, \tau_\beta^2 I_p)$ with a predefined $\tau_\beta^2$.

Accept with probability

$$\frac{P(Y|Z, \beta_0, \beta_{1t}^*, X)\mathrm{N}_p(\cdot|\beta_{1t}^*|\xi, \Psi)}{P(Y|Z, \beta_0, \beta_1, X)\mathrm{N}_p(\beta_{1t}^*|\xi, \Psi)}.$$

## 4.6 Numerical illustrations

We apply our model to the simulated data, and to the INFOCOM dataset, described in Section 2.2.1.

### 4.6.1 Example 1: simulated data

To see how our model works, we apply it to the estimation of simulated data, generated from the model itself, and check the ability to recover a given structure and parameters.

We generate a relatively small dataset, with $N = 20$ actors and $T = 50$ time-points with $K = 3$ states, from the model with the following parameters: $\sigma^2 = [1,\ 9,\ 25]$, $\beta_0 = 2$, with $(S_t)_{t=1}^T$ generated from 3x3 transition matrix $\pi$ obtained from a uniform distribution, normalized to be a valid transition matrix. The states are chosen to represent different levels of “cohesion” along time. This is illustrated by a subset of the generated data in Figure 4.2. Note that scales are chosen for better visualization, but the higher is the value of $\sigma_k^2$, the tighter are the actors’ positions.

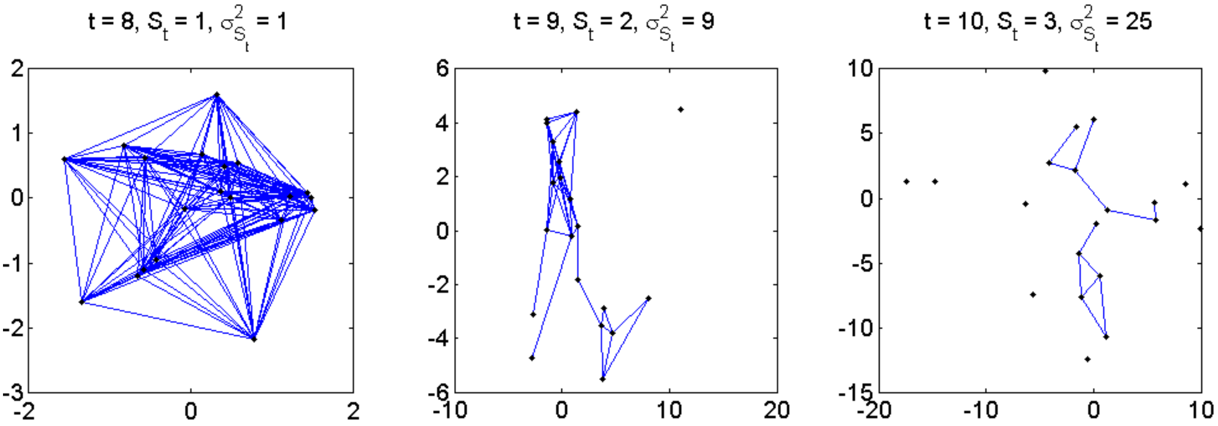


Figure 4.2: Visual representation of the subset of simulated data. Specifically, time points $8, 9, 10$ are shown, that correspond to 3 different states of the system. As variances $\sigma_2$ vary, overall cohesion of the network is clearly different.

For the estimation, we set the prior hyperparameters as follows: $a_{\beta_0} = 2, b_{\beta_0} = 1, a_\alpha = 1, b_\alpha = 1, a_{\alpha_0} = 1, b_{\alpha_0} = 1, a_\sigma = 2, b_\sigma = 1$ giving diffuse priors for parameters. Standard deviations for the proposal distributions were taken $\tau_Z = 0.5, \delta = 0.5$. These latter variables influence the efficiency of the sampling algorithm, and were chosen empirically.

We show the results of MCMC approximation after a burn-in period of 5000 iterations, taking further 10000 iterations and saving each 100th.

The estimated from posterior distribution state sequence $S_t$ is illustrated in Figure 4.3, together with posterior probabilities of belonging to each state.

The colors on the “heat-map” in Figure 4.3 represent the probability of being assigned to a particular state at a particular time point – the darker the color the higher is the probability. We have reordered the sampled state sequences in “order of appearance” to handle label switching problem. The estimated state sequence is obtained by taking, at each time point, the state with highest posterior probability. The resulting sequence (green in the figure) differs little from the true simulated sequence (differences marked with blue dots in the figure). We can observe that sampler tends to switch between 3 and 4 states, but the ambiguity is present only at time points that belonged to the third state in the true sequence.

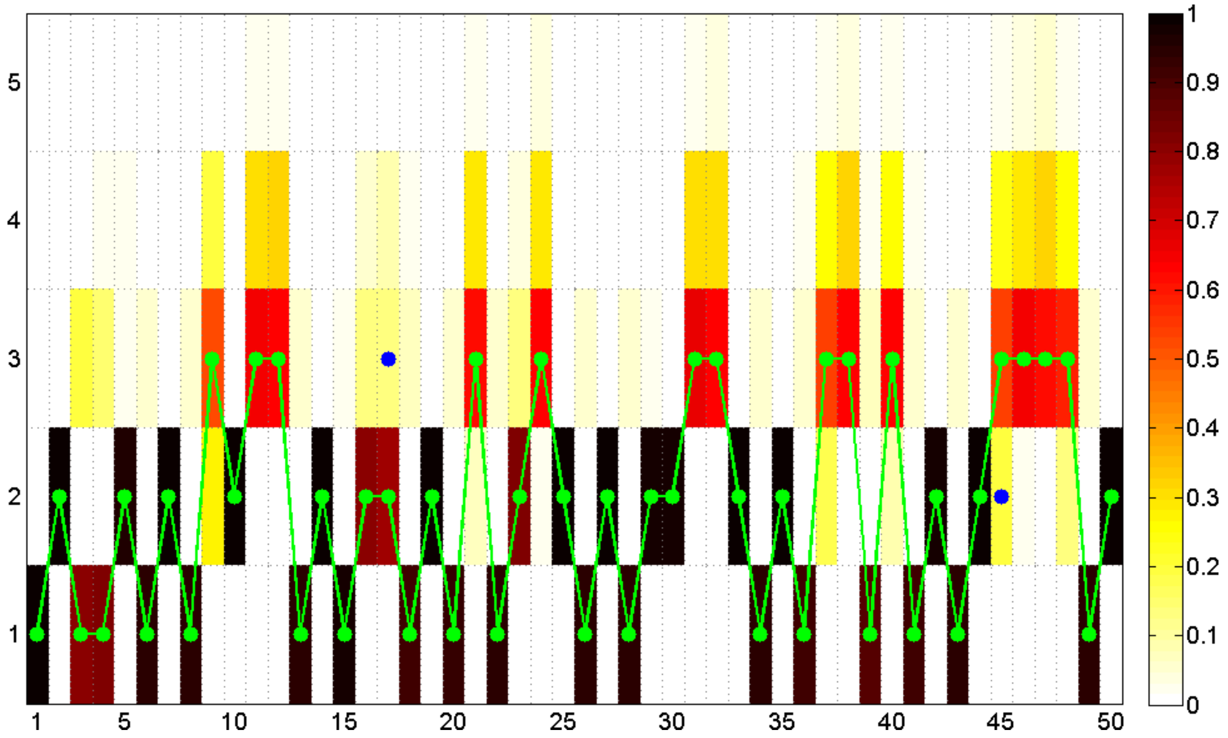


Figure 4.3: Posterior estimates of the state sequence $(S_t)$ (green) and the true states that were estimated incorrectly (blue) with the heat-map of posterior probabilities of belonging to each state.

Figure 4.4 illustrates the point estimates for MCMC approximation of the posterior expected transition probabilities, compared with the true transition matrix $\pi$ and with the matrix of normalized transition counts, computed for the true state sequence. In this example, we see that estimated $\pi$ is more close to the transition probabilities. This can be explained by the fact that the only information about $\pi$ that we have, is the generated from $\pi$ state sequence $S$.

Having checked that the model is able to discover the hidden states, we now look at the estimated values of "cohesion" parameters $\sigma^2 = (\sigma_k^2)$. We look at the variability of samples from posterior distribution of $\sigma^2$ over time in figure 4.5. The samples follow closely the estimated state sequence, apart from some inaccuracies, as, for example at $t = 9$, $\sigma^2$ has a bimodal distribution.

The figure 4.6 shows the MCMC estimates of posterior densities of $\beta_0$ and $(\sigma_k^2)$ for the estimated states $k = 1, 2, 3$.

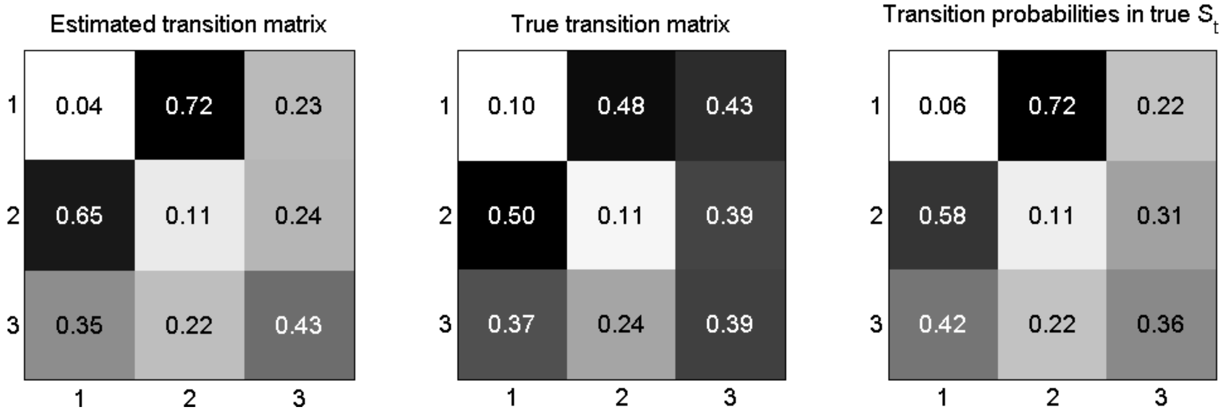


Figure 4.4: Estimated transition probability matrix (left), compared with the true transition matrix $\pi$ (center) and matrix of normalized transition counts for the true state sequence (right).

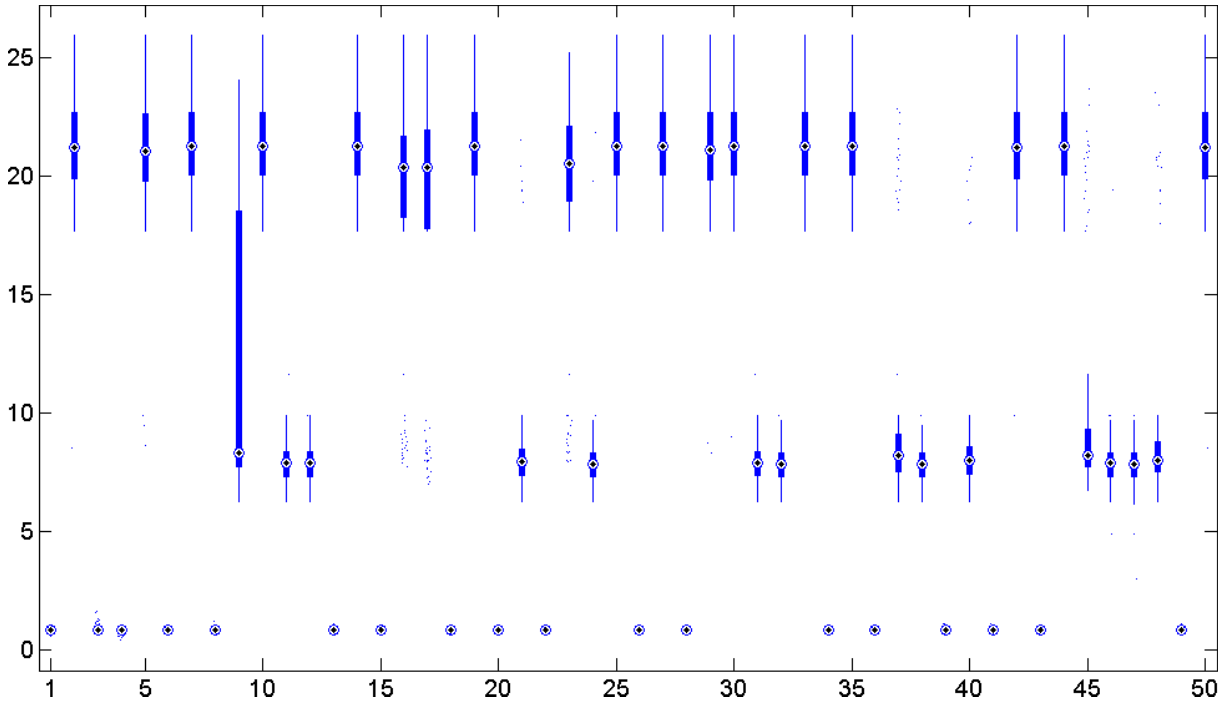


Figure 4.5: Posterior samples from the sequence of variance parameters $(\sigma^2_{S_t})$. The box-plots show summary of the posterior distributions of $\sigma^2_{S_t}$ for each $t$.

Looking at the point estimates, obtained as posterior medians (shown in the Table 4.1), and comparing with the truth, we see that we underestimate the values of all parameters. At the same time, as we have noted earlier in Section 4.3, $\sigma^2$ and $\beta_0$ are interconnected, and increase in the variances

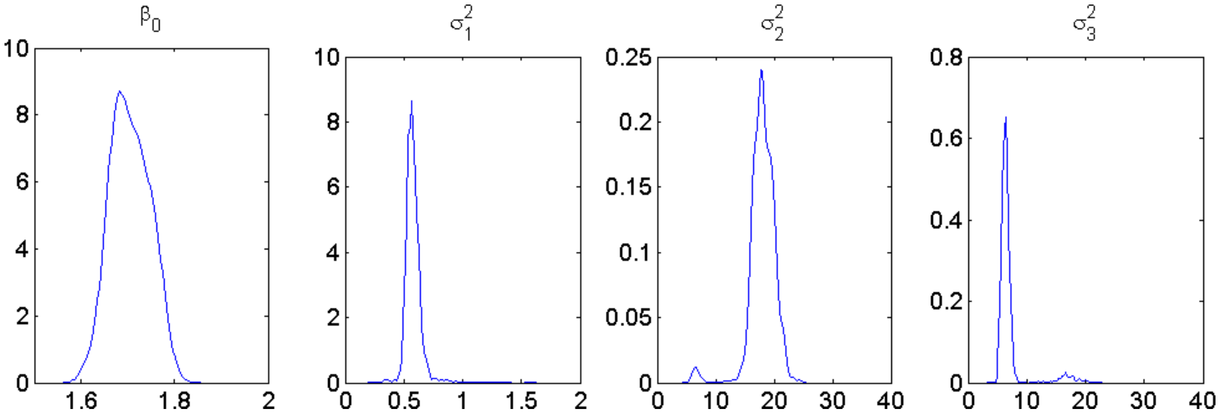


Figure 4.6: Posterior density distributions for $\beta_0$ and $\sigma_k^2$ for the 3 discovered states in the estimated state sequence $S$.

invokes the increase in the $\beta_0$ to compensate. Due to this fact, differences across variances for the states are relative.

If we fix the value of $\beta_0$ in the estimation procedure, and consider the ratios of estimated variances, we get close to the true ratios, as can be seen in Table 4.2.

| | $\beta_0$ | $\sigma_1^2$ | $\sigma_2^2$ | $\sigma_3^2$ |
|---|---|---|---|---|
| true | 2 | 1 | 25 | 9 |
| estimated | 1.70 | 0.57 | 18.02 | 6.40 |

Table 4.1: True values of the parameters and point estimates, obtained as medians of samples from posterior distributions.

| | $\sigma_1^2$ | $\sigma_2^2$ | $\sigma_3^2$ |
|---|---|---|---|
| true | 1 | 25 | 9 |
| estimated | 0.83 | 21.02 | 7.99 |
| normalized | 1 | 25.34 | 9.62 |

Table 4.2: True values of the parameters, point estimates, obtained as medians of samples from posterior distributions, and the same values normalized by the variance of state 1. $\beta_0 = 2$ fixed.

The model is able to capture changes in the “cohesion” of the network and to illustrate this, we look at the total number of links in the graph as a measure of cohesion.

In figure 4.7 we observe posterior distributions of the expected number of links for each state, and compare them to median number of links at the corresponding states in the data. We see that the true medians for all states have high posterior probabilities, as shown by percentile numbers in the figure 4.7.

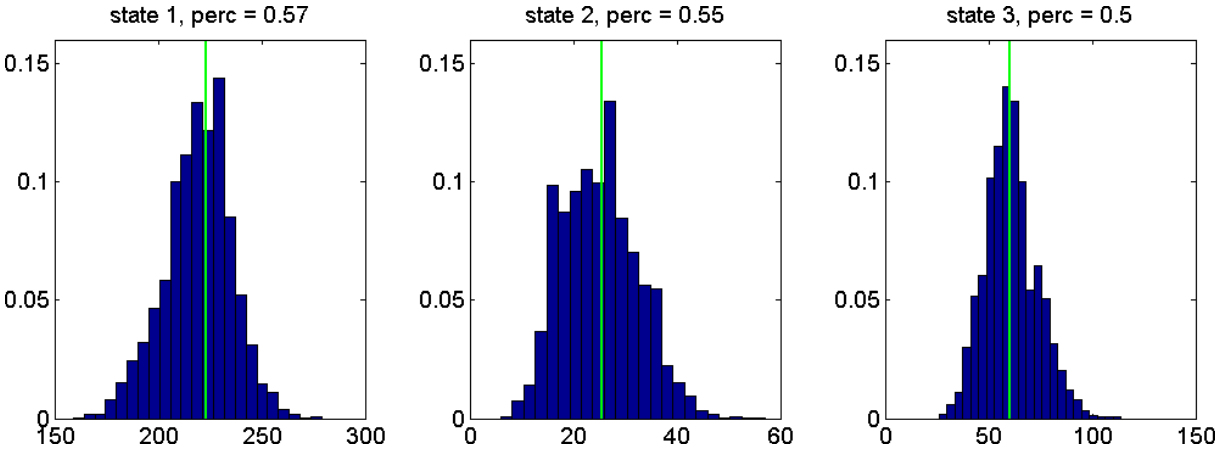


Figure 4.7: Posterior estimates for distribution of expected number of links aggregated by the 3 estimated states. The medians of number of true links at these states are shown with green lines, and corresponding percentiles are provided.

### 4.6.2 Real data

Now we use our model to analyze the INFOCOM'06 dataset, described in Section 2.2.1. To remind briefly, the data were collected on the INFOCOM conference in 2006: all participants were given a proximity sensors to detect their interactions during the conference days. The interaction was recorded if two devices "see" each other for some non-negligible amount of time.

For the analysis, data were aggregated into hourly intervals, and only the interactions, that are reciprocal, i.e. undirected, were used. Time runs from 18:00 on the first day till 16:00 on the 4th day, giving 94 time points. The total number of individuals observed is 78. The evolution of number of links was shown in the Figure 2.3.

The particularity of this dynamic network data is that there are known underlying processes for the interactions, time of day, recorded at constant discrete intervals, and events on the conference. We can expect that the hidden states discovered will have connections to these processes.

We use our model with default (as in example 1) hyperparameters, and run the MCMC chain for 30000 iterations, saving each 100-th, after the initial burn-in period of 10000. Plot of the estimated state sequence $(S_t)$ with posterior probabilities of belonging to each state over time is shown in Figure 4.8.

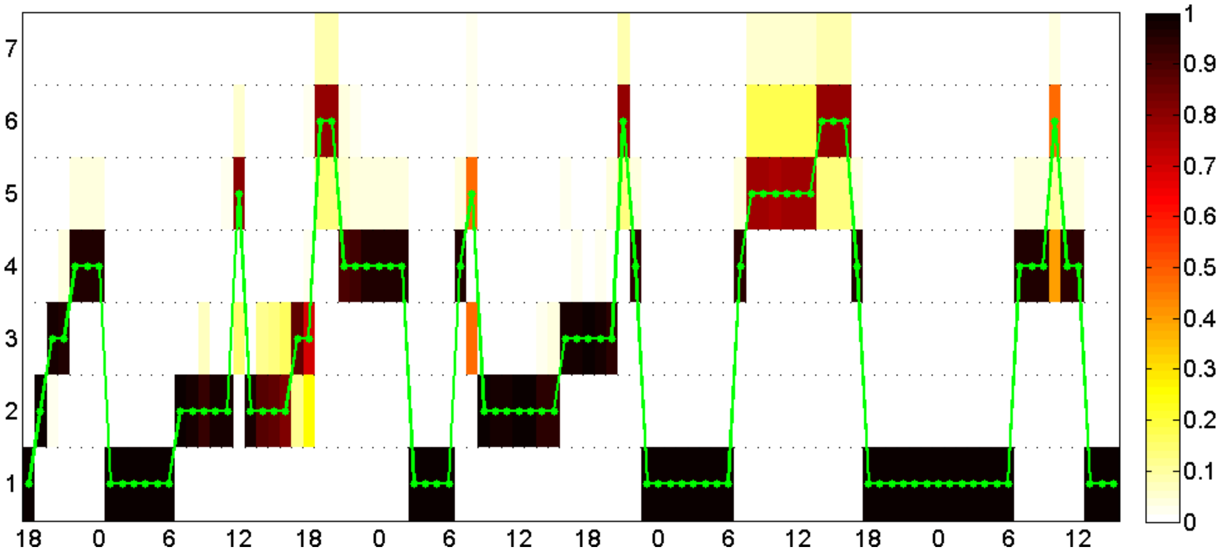


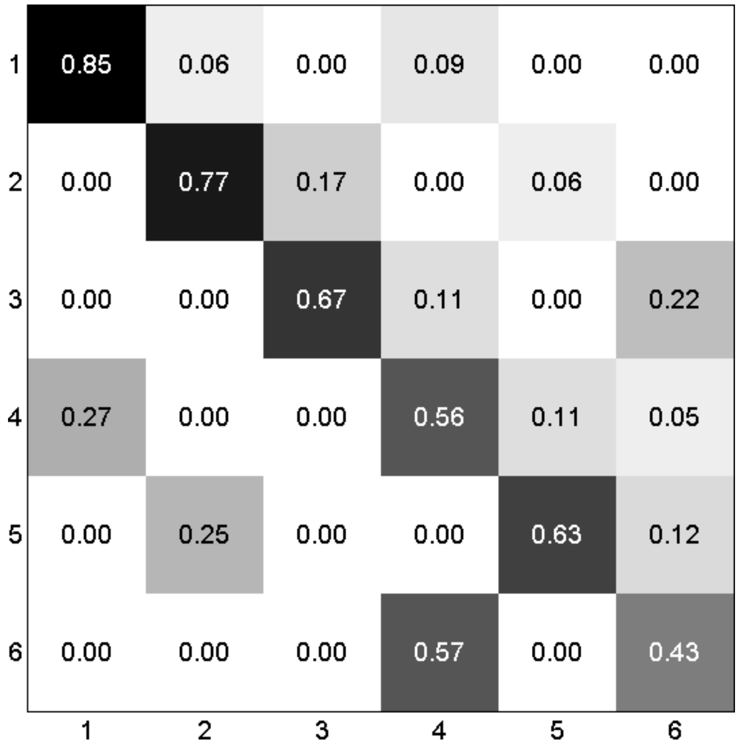
Figure 4.8: Estimated state sequence $S)$ (green), with the heat-map of posterior probabilities of belonging to each state. Time of day is on the x-axis.

Six states were discovered. By looking at the estimated state sequence, the periodicity is apparent, confirming that the discovered hidden process follows time of day. Table 4.3 presents summary of states. Some of them can be interpreted as “night time” (state 1), “active sessions” (state 2), “social events” (state 3), “late evenings” (state 4).

Looking at the estimated transition matrix (Figure 4.9), we see that with high probability system tends so stay in the same state. Also some transitions are more common, as, for example state 4 is mostly followed by state 1, that is in line with our interpretation.

| state | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| number of time points | 35 | 17 | 9 | 18 | 8 | 7 |
| average $\sigma^2$ | 424 | 14 | 21 | 141 | 32 | 63 |
| average number of links | 9 | 1101 | 612 | 47 | 389 | 109 |

Table 4.3: Summaries for discovered states of the system.



|  | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0.85 | 0.06 | 0.00 | 0.09 | 0.00 | 0.00 |
| 2 | 0.00 | 0.77 | 0.17 | 0.00 | 0.06 | 0.00 |
| 3 | 0.00 | 0.00 | 0.67 | 0.11 | 0.00 | 0.22 |
| 4 | 0.27 | 0.00 | 0.00 | 0.56 | 0.11 | 0.05 |
| 5 | 0.00 | 0.25 | 0.00 | 0.00 | 0.63 | 0.12 |
| 6 | 0.00 | 0.00 | 0.00 | 0.57 | 0.00 | 0.43 |

Figure 4.9: Estimated posterior transition probabilities between states.

## 4.7 Conclusion

We have presented an extension of a popular latent space model to the case of dynamically changing network over time. Our aim here was to capture the evolution of the general network cohesion. To this aim, we introduced an infinite hidden Markov model, that is able to capture such evolution, to describe the dynamics of latent positions.

As our proposed model is based on the static latent distance model of Hoff et al. [2002], it can also be extended as the former to accommodate more flexibility, following the ideas of proposed extensions for the original static model in Handcock et al. [2007], Krivitsky et al. [2009] or Hoff [2008].

This can allow to capture more properties of the real-life networks (such as clustering and degree heterogeneity).

For the dynamic part, the use of infinite HMM allows the discovery of states, that are characterized by changes in the overall cohesion of links in a network. The assumption of observing changes in a discrete way may not be an appropriate for some real data, if a network changes continuously. Therefore, another interesting possibility would be to explore the use of a continuous time process for the time evolution.

# Chapter 5

# Final remarks and future directions

In this short chapter we discuss possible extensions of the work presented in the thesis, and further directions of interest.

## 5.1 Extending the static part of the dynamic latent distance model

As we anticipated in Chapter 4, our proposed model is a simple idea, that can be easily extended. The static model by Hoff et al. [2002] has been extended in the recent literature by adding sender and receiver random effects (Hoff [2005]), allowing the latent positions to have a cluster structure (Handcock et al. [2007]) and, combining this two notions (Krivitsky et al. [2009]).

It would be interesting to change the static part of our model to these cases. This will allow to capture more of the typical network properties. In particular, for the models with clusters, the latent positions are assumed to come from mixture of normal distributions:

$$Z_i|\lambda, \mu, \sigma^2 \overset{ind}{\sim} \sum_{g=1}^{G} \lambda_g \mathrm{N}_d(\cdot|\mu_g, \sigma_g^2 I_d). \tag{5.1}$$

The time evolution here can be put on the group means and variance parameters $\mu_g, \sigma_g^2$ in a similar fashion to our model.

Another extension for the cluster components of the model, can be to include a Dirichlet Process mixture in place of a finite mixture in (5.1). For the dynamics we would have different DPs across $t$, and can use dependent-DP to define dependence.

Here we should note, that despite the added flexibility, and ability to discover number of clusters from the data, there are some limitations. The property of the Dirichlet Process to favor few large groups and many small groups may not be realistic for applications. At the same time extensions of the DP with different implied partition structure exist in the literature, and can be implemented.

## 5.2 Stronger individual dependence

Another line of extension of the proposed model is to add direct dependence on the previous position in the evolution of $Z_{i,t}$. In its current form our model can be applied when, for example, the covariates capture the individual temporal dependence in $Y_t$, and, therefore, the RCE residuals, that we model with latent distances, can be assumed to be fairly independent, apart from the global parameters, $\sigma^2$. Alternatively, if the network represents a tight system of actors, such as physical particles or large biological systems, that has no memory of the previous individual positions, our proposal can be used to model the global evolution, while being able to track individual interactions at the fixed time-points.

Adding stronger dependence on the previous positions can be achieved by means of a random walk:

$$Z_{i,t} \sim \mathrm{N}_d(Z_{i,t-1}, \sigma^2_{S_t})$$

Combining this idea with the existing approaches, there are four possibilities for the form of evolution of the latent positions over time:

1. *Independent latent distance models across time*

$$Z_{i,t} \sim \mathrm{N}_d(0, \sigma^2)$$

   Considers all time points to be independent, as if they are snapshots of unrelated networks. Not a very realistic assumption, but can be used as a benchmark.

2. *Our proposal, variances from a iHMM (Section 4.4)*

$$Z_{i,t} \sim \mathrm{N}_d(0, \sigma^2_{S_t})$$

   Variance changes according to a hidden state of the system. Can capture structural changes in the network, but weak individual dependence over time.

3. *Random walk evolution with fixed variance (Sarkar and Moore [2005])*

$$Z_{i,t} \sim \mathrm{N}_d(Z_{i,t-1}, \sigma^2)$$

   The positions evolve independently, with the fixed variance. The level of individual dependence is controlled by one parameter and is not changing over time.

4. *Random walk evolution with variances from a iHMM (a new proposal)*

$$Z_{i,t} \sim \mathrm{N}_d(Z_{i,t-1}, \sigma^2_{S_t})$$

   A new proposal that combines two previous ones. The individual dependence can be stronger or weaker depending on a current hidden state of the system.

## 5.3 Dynamic eigenmodel

A possible improvement of our model may also be achieved by changing the underlying static model to the “eigenmodel” by Hoff [2008]. In this model the residuals $\Gamma$ are taken to be $Z_i' \Lambda Z_j$, where $Z_i \in \mathbb{R}^d$ are vectors of latent characteristics, and $\Lambda$ is a diagonal $d \times d$ matrix, controlling how these characteristics contribute to the overall probability of link (positively or negatively depending on the sign of corresponding $\lambda_k$). The dynamic evolution can again be implemented with our idea of using a iHMM for the $Z$'s. At the same time, from the interpretation, it is reasonable to assume that characteristics should not change often, so a better time dependence structure may be preferred. One possibility can be in adding dynamics for $\Lambda$, as it is reasonable to assume that, with time, influence of different characteristics on the tendency to have a link may change.

## 5.4 Models from the RCE-ME representation

Relating Chapters 3 and 4, it would be interesting to construct models based on the representation obtained in (3.2). As the functional representation of Aldous [1981] allowed to motivate the large class of static network models and put the existing models in a common framework, we hope that our representation theorem can be of use in motivating models for dynamic networks. Looking at the general framework of Section 2.4.3 for static models, we note, that in fact, the assumption of RCE can be put in two different ways, either on the network adjacency matrix itself ($Y$ is RCE), or on the residuals in the statistical model with covariates ($\Gamma$ is RCE, in the notation of Section 2.4.3). This will also apply for our assumption of a RCE array of ME sequences. Either the sequence of adjacency matrices for each time point can be combination of RCE and ME, or the residuals in the regression model.

Also it is interesting to study different kinds of functional representations arising when additional restrictions are made, similarly to results in [Aldous, 1985, Chapter 14]. While comparing the other ways of combining RCE with ME can give additional insights on the use of representations in modeling.

## 5.5 Other directions

On a more general note about network modeling, another interesting point, is, as Orbanz and Roy [2013] write, "Exchangeable random structures are not 'sparse'... In contrast, graphs representing network data typically have a finite number of edges per vertex, and exhibit properties like power-laws and 'small-world phenomena', which can only occur in sparse graphs. Hence, even though exchangeable graph models are widely used in network analysis, they are inherently misspecified."

The absence of sparsity is a consequence of exchangeability assumption, and so, to adequately model sparse network data, other non-exchangeable assumptions should be imposed. This is an interesting open problem, and further discussion can be found in Orbanz and Roy [2013].

Also of particular interest are connections with related areas, briefly mentioned in this work, including graph limits and asymptotic properties for complex network models.

# Bibliography

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

D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

D. J. Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983*, 1117:1–198, 1985.

D. J. Aldous. More uses of exchangeability: representations of complex random structures. *arXiv preprint arXiv:0909.4339*, 2010.

M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. *Advances in Neural Information Processing Systems*, 14: 577–584, 2002.

A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, 2007.

S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *arXiv preprint arXiv:1102.2650*, 2013.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695, 2006. URL `http://igraph.sf.net`.

B. de Finetti. Sur la condition d' "équivalence partielle". *Actualités Scientifiques et Industrielles*, 739, 1938.

B. de Finetti. La probabilità e la statistica nei rapporti con l'induzione secondo i diversi punti di vistà. *Atti del Corso CIME su Induzione e Statistica*, pages 147–257, 1959.

B. de Finetti. *Probability, Induction and Statistics: The Art of Guessing.* J. Wiley & Sons, 1972.

P. Diaconis and D. Freedman. De Finetti's theorem for Markov chains. *The Annals of Probability*, 8(1):115–130, 1980.

M. A. Duijn, T. A. Snijders, and B. J. Zijlstra. p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2): 234–254, 2004.

D. Easley and J. Kleinberg. *Networks, Crowds, and Markets.* Cambridge University Press, 2010.

P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5: 17–61, 1960.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

S. Fortini, L. Ladelli, G. Petris, and E. Regazzini. On mixtures of distributions of Markov chains. *Stochastic Processes and their Applications*, 100(1):147–165, 2002.

J. Foulds, C. DuBois, A. Asuncion, C. Butts, and P. Smyth. A dynamic relational infinite feature model for longitudinal social networks. In *AI and Statistics*, volume 15, pages 287–295, 2011.

O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

T. M. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30 (4):1141–1144, 1959.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

C. Heaukulani and Z. Ghahramani. Dynamic probabilistic models for latent feature propagation in social networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 275–283, 2013.

P. Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295, 2005.

P. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 657–664. MIT Press, Cambridge, MA, 2008.

P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97 (460):1090–1098, 2002.

P. W. Holland and S. Leinhardt. A dynamic model for social networks. *Journal of Mathematical Sociology*, 5(1):5–20, 1977.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

D. N. Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 1979.

D. N. Hoover. Row-column exchangeability and a generalized model for probability. *Exchangeability in Probability and Statistics*, pages 81–291, 1982.

D. R. Hunter, P. N. Krivitsky, and M. Schweinberger. Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4):856–882, 2012.

O. Kallenberg. *Foundations of Modern Probability.* Springer Verlag, 2002.

O. Kallenberg. *Probabilistic Symmetries and Invariance Principles.* Springer Science+Business Media, 2005.

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*, pages 381–388, 2006.

P. Krivitsky, M. Handcock, A. Raftery, and P. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213, 2009. ISSN 0378-8733.

J. Lloyd, P. Orbanz, Z. Ghahramani, and D. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems 25*, pages 1007–1015, 2012.

L. Lovász. Very large graphs. *Current Developments in Mathematics*, 2008: 67–128, 2009.

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

R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96 (455):1077–1087, 2001.

P. Orbanz and D. M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. Preprint, 2013. URL `danroy.org/papers/OR-exchangeable.pdf`.

K. Palla, Z. Ghahramani, and D. A. Knowles. An infinite latent attribute model for network data. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1607–1614, 2012.

A. Rodriguez. Modeling the dynamics of social networks using bayesian hierarchical blockmodels. *Statistical Analysis and Data Mining*, 5(3): 218–234, 2012.

M. Salter-Townshend, A. White, I. Gollini, and T. B. Murphy. Review of statistical network analysis: models, algorithms, and software. *Statistical Analysis and Data Mining*, 5(4):243–264, 2012.

P. Sarkar and A. W. Moore. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):31–40, 2005.

J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD trace cambridge/haggle/imote/infocom2006 (v. 2009-05-29), May 2009. URL `http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/infocom2006`.

B. Skyrms and R. Pemantle. A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16):9340–9346, 2000.

T. A. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37:131–153, 2011.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

J. Van Gael, Y. Saatci, Y. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095. ACM, 2008.

T. Von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. In *Computer Graphics Forum*, volume 30, pages 1719–1749. Wiley Online Library, 2011.

S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation*, 36(1):45–54, 2007.

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

A. H. Westveld and P. Hoff. A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *The Annals of Applied Statistics*, 5(2A):843–872, 2011.

E. P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566, 2010.

S. L. Zabell. Characterizing Markov exchangeable sequences. *Journal of Theoretical Probability*, 8(1):175–178, 1995.

A. Zaman. Urn models for Markov exchangeability. *The Annals of Probability*, 12(1):223–229, 1984.