Probabilistic inference in ecological networks; graph discovery, community detection and modelling dynamic sociality

Thesis submited for the degree
Doctor of Philosophy

Ioannis Psorakis

Wolfson College

Supervisors: Prof. Stephen Roberts, Prof. Ben Sheldon

Machine Learning Research Group
Department of Engineering Science

University of Oxford

Trinity 2013

Ioannis Psorakis
Wolfson College

DPhil Thesis
Trinity Term 2013

Probabilistic inference in ecological networks; graph discovery, community detection and modelling dynamic sociality

---

Abstract

This thesis proposes a collection of analytical and computational methods for inferring an underlying social structure of a given population, observed only via timestamped occurrences of its members across a range of locations. It shows that such data streams have a modular and temporally-focused structure, neither fully ordered nor completely random, with individuals appearing in "gathering events". By exploiting such structure, the thesis proposes an appropriate mapping of those spatio-temporal data streams to a social network, based on the co-occurrences of agents across gathering events, while capturing the uncertainty over social ties via the use of probability distributions.

Given the extracted graphs mentioned above, an approach is proposed for studying their community organisation. The method considers communities as explanatory variables for the observed interactions, producing overlapping partitions and node membership scores to groups.

The aforementioned models are motivated by a large ongoing experiment at Wytham woods, Oxford, where a population of *Parus major* wild birds is tagged with RFID devices and a grid of feeding locations generates thousands of spatio-temporal records each year. The methods proposed are applied on such data set to demonstrate how they can be used to explore wild bird sociality, reveal its internal organisation across a variety of different scales and provide insights into important biological processes relating to mating pair formation.

This is my own work, except where otherwise indicated.

Candidate: Ioannis PSORAKIS

Signature: .....

Date: .....

Acknowledgements

I would like to thank Prof Stephen Roberts and Prof Ben Sheldon for being the best supervisors I could ever have, by providing me constant guidance, support and academic freedom. Steve has not only been a truly amazing mentor and source of encouragement, but also a shield from all kinds of bureaucratic hassles. Ben has been patient with my lack of zoological knowledge and always receptive to my crazy ideas with the most constructive feedback, while remaining curiously tolerant of being exposed to more Bayesian statistics than he originally intended (or wanted).

My D.Phil research has been generously supported by the Microsoft Research scholarship. I would like to thank all people in the Cambridge team, especially Dr Fabien Petitcolas and Dr Scarlet Schwiderski-Grosche, for their limitless support, coaching and vision.

All zoological data used in this work have been collected by researchers from the Edward Grey Institute. Special thanks go to Teddy Wilkin, Simon Evans, Ada Grabowska, Colin Garroway, Antica Culina, Damien Farine, Nicole Milligan, Ross Crates and Lindall Kidd.

All the research in this thesis has been conducted at the Machine Learning Research Group. I would like to thank Ash, Zach, Ed, Asif, Jan, Iead, Mike, Steve, Mark and Nauman, and even Rob, for being amazing friends and colleagues.

In the past years I have not only finished a D.Phil, but also met my significant other. Natalia has gone way beyond simply maintaining my sanity during thesis write-up, being an amazing source of encouragement and inspiration.

Finally, no words can express my gratitude towards my parents, whose support during my D.Phil has been yet another act of selfless love and kindness. This work is dedicated to them.

Published material

Papers

This work is largely based on the following peer-reviewed publications:

- **Ioannis Psorakis**, Stephen Roberts, Iead Rezek and Ben Sheldon "Inferring social network structure in ecological systems from spatio-temporal data streams", Journal of the Royal Society Interface, vol. 9 no. 76, pp 3055-3066 (2012).

  Portions of this paper appear in Chapters 6 and 8 of this thesis.

- **Ioannis Psorakis**, Iead Rezek, Zach Frankel and Stephen Roberts "Discovering latent association structure via Bayesian one-mode projection of temporal bipartite graphs ", International Conference in Machine Learning and Data Mining (MLDM), July 19-25, 2013, New York, USA (to appear in conference proceedings).

  Methods and results from this paper appear in Chapter 7. The method is extensively used for the results of Chapter 8.

- **Ioannis Psorakis**, Stephen Roberts, Mark Ebden, Ben Sheldon "Overlapping Community Detection using Nonnegative Matrix Factorization", Physical Review E 83, 066114 (2011).

  Methods and results from this paper appear in Chapter 4. The method is extensively used for the results of Chapter 8.

- Edwin Simpson, Stephen Roberts, **Ioannis Psorakis**, and Arfon Smith "Dynamic Bayesian Combination of Multiple Imperfect Classiers", book chapter of Decision

Making and Imperfection, pages 1–35, Intelligent Systems Reference Library series, Springer, 2013.

Results in this paper that are relevant to the present thesis appear in Chapter 4.

- Mark Smith, Steven Reece, Stephen Roberts, **Ioannis Psorakis** and Iead Rezek "Maritime Abnormality using Gaussian Processes", Knowledge and Information Systems, Springer (accepted – to appear).

  Results in this paper that are relevant to the present thesis appear in Chapter 4.

- Jens Krause, Stephan Krause, Robert Arlinghaus, **Ioannis Psorakis**, Stephen Roberts and Christian Rutz "Reality Mining in Ecological Systems", review paper, Trends in Ecology and Evolution (accepted – to appear).

  Elements of this work appear in Chapters 5, 6 and Chapter 8.

Data

All wild bird data, used in Chapters 5 and 8, have been collected and compiled by the members of Edward Grey Institute of Field Ornithology. Any request for data should be directed to Prof Ben Sheldon. The benchmark data sets, used in Chapter 4, are compiled by various authors (see citations in the main text) whom I warmly thank for their time and effort.

Algorithms

Algorithms used in this thesis are freely available for download from the Machine Learning Research group software page: `http://www.robots.ox.ac.uk/~parg/doku.php?id=software_page`.

Contents

Nomenclature

x	scalar variable.
$\mathbf{x}$	vector variable.
$\mathbf{X}$	matrix.
$p(x)$	Probability of x.
$H(x)$	Entropy of random variable x.
$\mathcal{N}(x; \mu, \beta^{-1})$	Gaussian probability density function over $x \in \mathbb{R}$, with mean μ and precision β^{-1}.
$\text{Pois}(x; \lambda)$	Poisson probability density function over $x \in \mathbf{N}_0$, with rate λ.
$\text{Ga}(x; \kappa, \theta)$	Gamma probability density function over $x \in \mathbb{R}_+$, with parameters k (scale) and θ (shape).
$\text{Beta}(x; \alpha, \beta)$	Beta probability density function over $x \in [0, 1]$, with shape parameters α and β.
$\text{Binom}(x; \pi, N)$	Binomial probability density function over $x \in \mathbf{N}_0$, with bias parameter π and N number of trials.
$\text{Dir}(\mathbf{x}; \mathbf{a})$	Dirichlet probability density function over $\mathbf{x} = \{x_i\}_{i=1}^{D},\ x_i \in [0, 1]$ and $D \geq 2$, with concentration parameters $\mathbf{a} = \{a_i\}_{i=1}^{D}$ for which $a_i > 0$.
$\mathcal{G}(\mathcal{V}, \mathcal{E})$	Unipartite graph with node (vertex) set $\mathcal{V}$ and edge (link) set $\mathcal{E}$.
N	Number of nodes (vertices) $\lvert\mathcal{V}\rvert$ in the network.
M	Number of edges (links) $\lvert\mathcal{E}\rvert$ in the network.

$\mathcal{G}(\mathcal{V},\mathcal{U},\mathcal{E})$	Bipartite graph with node sets $\mathcal{V},\mathcal{U}$ and edge set $\mathcal{E}$.
K	Number of nodes $\lvert\mathcal{U}\rvert$.
$\mathbf{A}$	$N \times N$ adjacency matrix of an undirected graph.
a_{ij}	Connection weight of node pair i, j given adjacency matrix $\mathbf{A}$.
$\mathbf{B}$	$N \times K$ bipartite graph incidence matrix.
b_{ik}	Connection weight of node pair i, j given incidence matrix $\mathbf{B}$.
d_i	Degree of node i.
g_{ij}	geodesic distance or shortest path length between nodes i and j.
C	Number of communities in a network.
Q	Newman-Girvan modularity.
ϵ	Nonnegative real number, for which $\epsilon < 10^{-3}$.

Chapter 1

Introduction

1.1 Motivation

Most organisms display social behaviour of some form, but the extent and duration of this behaviour varies tremendously between species and over life cycles within species. Following the explosive growth of *social network analysis* (SNA) during the past couple of decades [Buchanan and Caldarelli, 2010], there is currently great interest in employing SNA to the study of animal behaviour [Krause et al., 2009; Wey et al., 2008]. As social connections matter, for all but the most solitary asexual organisms [Whitehead, 2008], network analysis provides us a powerful conceptual framework for studying sociality at any level of group organisation (from dyads to populations) or type of interaction (sexual, cooperative, predator-prey, etc) [Krause et al., 2009; Whitehead, 2008].

The present work is motivated by a large ongoing study of a wild bird population that has been a model system in ecology and evolutionary biology [Grosler, 1993]; the great tit (GT) *Parus major* at Wytham Woods near Oxford, in which thousands of individuals are marked with transponders and a grid of recording locations generates hundreds of thousands of records each winter. Such records consist of timestamped wild bird visitations across Wytham Woods, which allow us to describe the great tit mobility patterns across each season.

Based on such a rich and complex data set, we propose a collection of statistical and computational methods for modelling and exploring an underlying social structure of the bird population, from a network analysis perspective. Through the course of our discussion, we address a series of research questions, such as "how can we infer network structure from non-relational, time-series data?", "how can we identify the overlapping social groups in a given network?" and "how can we perform our inference in sensor-generated data sets, which are to a certain extent contaminated by noise and/or missing observations?". Arguments in favour of using our proposed methods are made via comparative experiments against existing approaches, across a variety of benchmark and real-world problems.

In the next section, we present an overview of the current thesis and describe how we tackle, at each chapter, the aforementioned research questions. We conclude Section 1.2 by laying out the methodological contributions of the present thesis, while in Section 1.3 we discuss their relevance to our zoological application.

1.2 Overview

In **Chapter 2** we begin our discussion by presenting elements of modern social network analysis and introduce the notation and terminology that is extensively used throughout the present thesis. We focus on key statistical and topological properties of real-world networks, with particular emphasis on *community structure*. We introduce this concept and discuss why community detection is a challenging inference task and a major research theme in contemporary network analysis literature.

In **Chapter 3**, we provide a brief overview of the concepts and tools from probability theory that are extensively used in the present work. We place special emphasis on the *Bayesian* view of probability, along with modelling and performing inference on stochastic systems via the use of *Directed Acyclic Graphs* (DAGs - or Bayesian networks).

Following our discussion of network analysis and probability theory, we proceed in

Chapter 4 by focusing on the community structure of the inferred graphs. We present our contribution to the community detection literature, by introducing a novel approach to extracting modules that overcomes the limitations of many popular existing methods. Our method addresses the issues of describing overlapping communities and defining node membership scores, while possessing a low computational overhead. We present our model formulation, which is based on identifying the most probable community configuration that can account for our interaction data (network), given a set of prior assumptions. We then examine the module identification capabilities of our method in a range of computer-generated benchmarks and real-world applications.

In **Chapter 5**, we open our discussion by providing an overview of network analysis, as applied to animal societies. We then proceed by presenting the Wytham Woods great tit Parus Major experiment setting, discuss how wild-bird observations are being collected and describe the working format of our data. Additionally, we present various statistical and demographical information on the data set, along with surrounding zoological information that provides further insight into the problem setting.

Following the foundational material presented in the above chapters, in **Chapter 6** we address the issue of inferring social connectivity structure among individuals, based on similarities in their mobility patterns. We begin by providing a critical overview of existing methods and discuss why they might lead to erroneous topological structures. Based on our analysis of the data retrieved from the Wytham Woods sensor grid, we introduce the notion of *gathering events*, i.e. areas of observation "spikes" that may result from an underlying social trigger. We exploit such an insight, by introducing two novel methodologies that identify such gathering events in data streams and place associations between individuals based on their *co-participation* in such groups. An appropriate null model is also developed in order to account for the statistical significance of inferred ties.

Discovering associations based on bird co-participation in gathering events, as described

in Chapter 6, can be viewed as an *one-mode projection* problem from a graph theory perspective. Based on such an insight, in **Chapter 7** we introduce a novel methodology for inferring and describing links in temporal bipartite graphs, where probability distributions are placed over the presence/absence and strength of social ties.

In **Chapter 8** we apply the above methodological developments to the Wytham Woods data set and explore the wild-bird social network across a multi-year span. We examine how various graph metrics relate to zoological quantities of interest and investigate their dynamics over different parts of the great tit life circle. Additionally, we apply the above proposed models in order to understand the process of mating pair formation.

We conclude in **Chapter 9** by providing an overview of our findings, both from a methodological and zoological point of view. Based on the advantages and limitations of the present work, we provide a discussion of ongoing and future research directions. We also discuss how the methodological advances presented in this thesis can be applied to a wide range of problems beyond animal social networks. In summary, the present thesis introduces the following methodological contributions to the contemporary literature:

1. CD-NMF - Community Detection via Nonnegative Matrix Factorisation (Chapter 4); an overlapping *community detection algorithm* that produces soft-partitioning solutions and assigns node participation scores to groups, in a computationally efficient manner.

2. GEM - Gathering Events Method (Chapter 6); a Bayesian methodology for *extracting adjacency matrices* from raw sensor observations, by exploiting the statistical properties of spatio-temporal data streams.

3. BOMP - Bayesian One-Mode Projection (Chapter 7); a fully Bayesian approach for modelling uncertainty over associations in a dynamic network, which exploits past information on temporal dynamics in order to *produce a probability distribution over the presence and weight of each link*.

1.3 Findings relating to the GT ecology

The methodological contributions we enumerated in Section 1.2 have been used in Chapter 8 in order to extract and analyse the GT social networks across two seasons: the first one spanning from August 2007 to March 2008 and the second one from August 2008 to March 2009.

In particular, in Section 8.3 we show that GT networks are globally sparse but locally dense, with a strong presence of triangle formation and densely-connected communities. This particular modular organisation can imply that diffusion processes, from disease transmission to social learning, can be very fast within the wild-bird flocks, but slower throughout the entire Wytham Woods population. For the case of social learning, such hypothesis has been already confirmed, using the same GT data, by [Aplin et al., 2012].

By analysing the pairwise GT connectivity we show, in Section 8.4, that network communities consist of members with similar site-fidelity patterns. Such homophily among exploratory (and among territorial) individuals implies that there is a strong coupling between sociality and dispersion, in the sense that network communities correspond to flocks with common migration strategies. As site residence is a strong predictor of dominance in great tits [Sandell and Smith, 1991], our results show that individuals may adopt the strategy of reducing their overall competition by associating with other non-territorial migrants.

In Section 8.4 we have also examined the relationship between network connectivity and gender. In contrast to the case of site-fidelity, we found no evidence of sex-based assortativity. Instead, network communities exhibit a mixed-sex composition, with no significant temporal variability throughout each observation season. This mixed-sex flocking structure is indicative of a strong presence of dominance hierarchies [Johnsen et al., 2001], where both male-male aggression and female choice influence mating success.

By focusing our analyses on mating pairs, in Section 8.5 we found that future mating partners, or “new pairs”, already had strong participation in the same mixed-species com-

munities since winter, many months before the actual breeding season in spring. Although the exact timing of mating bond formation can not be established by flock co-membership alone [Grosler, 1993; Robertson et al., 1998], our approach provides a baseline for predicting potential mating partners, by focusing on the extracted network groups.

We have also found that pre-existing mating partners, or "old pairs", participate in the same communities throughout the observation season, without any significant temporal variability. Despite such a different co-membership profile between old and new pairs, in both cases their co-occurrences in feeding stations are more concentrated towards the breeding season.

The methodologies developed in the present thesis not only provide important insights to the great tit sociality, but also form the baseline for important future research. In Chapter 9 we discuss how our methodological contributions can be used in order to investigate genetic foundations of sociality, meme diffusion and mixed-species social networks.

Chapter 2

Graphs and Social Networks

"Understanding how parts of a system interact, is as important as understanding the parts themselves" [Service, 1999].

Networks have been an important field of study since Euler and the solution of the Königsberg bridge problem (1735). During the 20th century and before the World Wide Web era, sociologists used the network paradigm to model human interactions with the most popular studies being Granovetter's "Strength of weak ties" (1973) [Granovetter, 1973] and Milgram's "Small-world" experiments (1967) [Milgram, 1967]. In recent years, the field of network analysis has undergone an explosive growth [Buchanan and Caldarelli, 2010] as theoretical and computational advances have allowed the study of large-scale complex systems such as the World Wide Web, social media, scientific collaboration patterns, animal societies and protein interactions [Fortunato, 2010; Newman, 2010]. One of the most fascinating findings of these studies is that real-world networks exhibit significant statistical and topological similarities, allowing us to study a wide and diverse range of complex systems under a single conceptual framework [Lambiotte, 2010].

In this chapter we present elements of modern network analysis along with a critical survey of theoretical ideas and computational methods from the literature. After presenting the fundamentals of network analysis (Section 2.1) and describing the properties of real-

world networks (Section 2.2), we focus on the problem of detecting communities in networks (Section 2.3), which is one of the main themes of the present work.

2.1 Graphs, networks and complex systems

In Mathematics, a simple *graph* $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is a set $\mathcal{V}$ of N nodes or vertices, connected together by a set $\mathcal{E}$ of M edges or links. The overall connectivity profile can be described by using the *adjacency matrix* $\mathbf{A} \in \mathbb{R}^{N \times N}$, so that if $a_{ij} \neq 0$ then nodes i and j are linked together; a_{ij} can take a Boolean value (unweighted edge), a real value (weighted edge) or a signed value (directed edge). Nodes can also be connected via multi-edges or self-edges, though we do not consider such cases in the present thesis. An example graph is shown in Fig. 2.1.

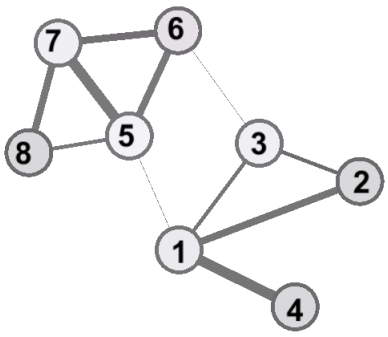


Figure 2.1: An example graph of $N = 8$ vertices and $M = 11$ edges. Edge width represents different connection weight.

Very commonly we use the term *network* to describe the simplified version of the pattern of interactions in a system (for example an Online Social Network), where nodes are individual entities and edges represent some form of association, interaction, similarity, commodity flow, or correlation between nodes. Similar to the way a map is a simplified (though useful) version of a landscape, a network describes a real-world system by focusing on the connectivity patterns of its individual components [Rosvall, 2006]. Although strictly speaking, the

term "graph" expresses the abstract mathematical structure described by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ and $\mathbf{A} \in \mathbb{R}^{N \times N}$, in some very influential works in the literature such as [Newman, 2010], the terms "graph" and "network" are used interchangeably.

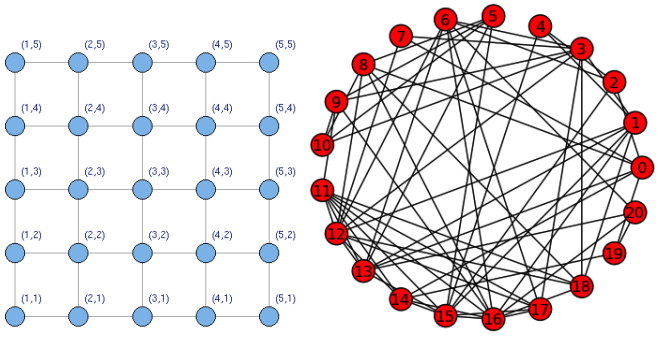
The exponentially increasing popularity of network analysis in the scientific literature [Buchanan and Caldarelli, 2010] is not only a result of the computational advances in data gathering, storage and processing technology of recent decades [Rosvall, 2006]; we have also realised that systems in nature are *complex*, i.e they are made up of a large number of entities interacting in such a way that their collective behaviour is not merely a simple combination of their individual behaviours [Newman, 2003b]. The network paradigm is therefore a very appropriate and flexible tool to describe a system at a macroscopic scale, as nodes and edges can represent any sort of entity or association. This contrasts with the traditional reductionist viewpoint of breaking down a system into its parts and focusing on each component separately [Rosvall, 2006], as networks omit the individual characteristics of each node and describe our data in a *relational* form. The key idea of this framework is that the connectivity patterns in the data have a big effect on the behaviour of the network as a system [Newman, 2010].

The fact that networks have been so popular in describing real-world systems has led Buchanan and Caldarelli to ask *"why Nature is so fond of networks?"* [Buchanan and Caldarelli, 2010]. This question is based on the relatively recent findings that networks describing complex systems possess a significant amount of similar statistical and topological properties, regardless of the application domain. We present and discuss some very important ones in the next section.

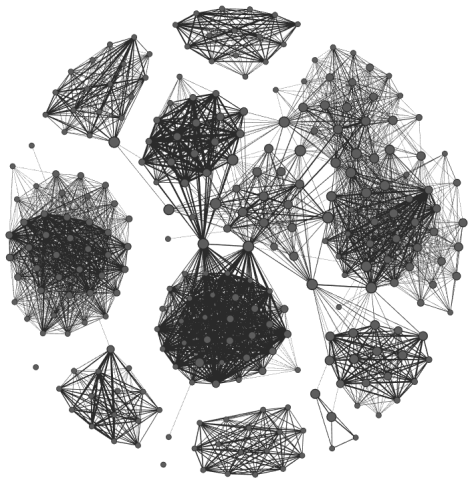
2.2 Properties of real-world networks

Consider a complex system such as the web of animal interactions in a wildlife population, or the pattern of scientific collaborations in a large research institute. We can model

such a system using a network, where each node represents an individual (a chimpanzee or a scientist) and edges reflect the presence of an association (sexual, trophic, symbiotic or cooperation). Intuitively, it is natural to conclude that not every individual is associated with everyone, nor that connections appear completely at random between any given pair in the network. Additionally, both the number of associations and the connection intensity vary between individuals, constituting a heterogeneous social structure that consists of highly connected nodes, peripheral nodes, densely connected neighbourhoods and sparse regions of loose associations.



(a) An ordered lattice (b) A random graph



(c) A real-world network

Figure 2.2: Fig. 2.2(a) shows a regular grid of 25 nodes. Fig. 2.2(b) shows an Erdős-Rényi (ER) random graph, where any pair is connected with fixed probability. Fig. 2.2(c) shows the interaction patterns in a wild bird social network (more details in Chapter 8).

If we study the connectivity patterns in a system such as the one described above, we soon realise that such networks are neither ordered lattices, as seen in Fig. 2.2(a) nor random graphs, as seen in Fig. 2.2(b). Instead, real-world networks possess a particular structure that is neither fully ordered nor completely random [Watts and Strogatz, 1998], which typically emerges from the self-organisational mechanisms of their individual components [Barabási and Albert, 1999]. This structure is often characterised by properties such as heavy-tail degree distribution, small-world effect, community structure, network resilience, degree correlations, among others [Newman, 2003b]. We particularly focus on the heavy-tail degree distribution and small-world effect properties, due to their historical significance, in Section 2.2.1 and Section 2.2.2. Community structure, due to its relevancy to the contributions of the present thesis, is discussed in more detail in Section 2.2.3.

2.2.1 Heavy-tailed degree distribution

We previously discussed that individuals have different levels of connectivity in a complex network. We define connectivity or *degree* d_i of node i as the number of edges attached to it. In the case of a *directed* network, we have a separate *in-* and *out-degree*, denoting the number ingoing $d_i^{(+)}$ and outgoing $d_i^{(-)}$ edges from a node i. By focusing our attention on undirected graphs, given the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ the degree is defined as $d_i = \sum_{j=1}^{N} a_{ij}$. One of the most fundamental properties of networks is the *degree distribution* $p(d)$, which captures the fraction of nodes in a given network possessing degree d.

The seminal work by [Barabási and Albert, 1999] showed that many real-world networks have a degree distribution that deviates significantly from the Poisson form of Erdős-Rényi random graphs [Newman et al., 2001]:

$$p(d) = \frac{\langle d \rangle^d e^{-\langle d \rangle}}{d!}, \text{ for the limit of large } N. \tag{2.1}$$

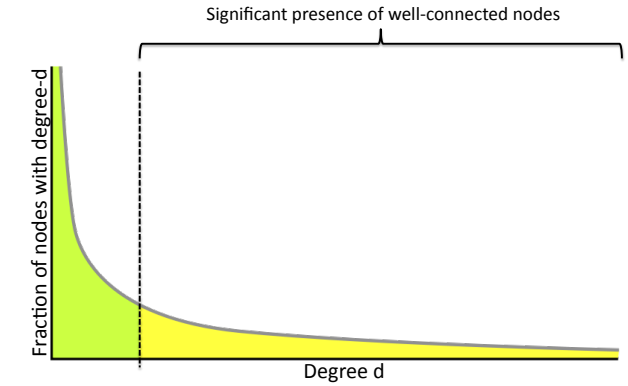
Instead, the probability $p(d)$ of picking a node with degree $d > d_{\min}$ in a real-world net-

work exhibits a *heavy tail*. In particular, the authors in [Albert and Barabási, 2002; Barabási and Albert, 1999; Newman, 2010; Newman et al., 2006] report a *power-law* distribution:

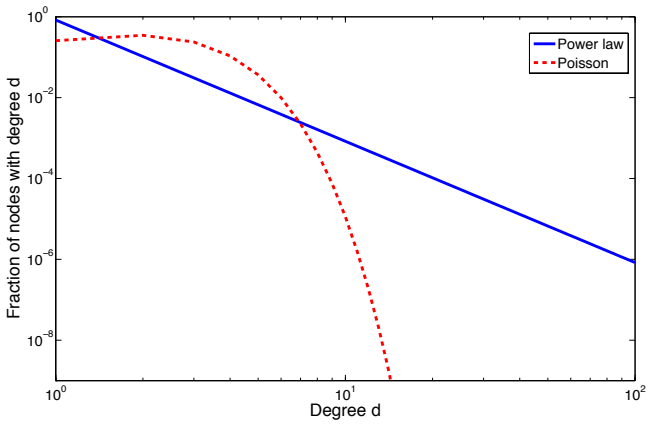
$$p(d) \propto d^{-\gamma}, \text{for } d > d_{\min}, \tag{2.2}$$

where γ is an application-dependent heuristic constant, for example $2 < \gamma < 3$, as discussed in [Newman, 2003b]. The "tail" of the power law (nodes with degree $d > d_{\min}$) decays more slowly than the tail of the Poisson, as we show in Fig. 2.3(b), implying a significant presence of highly connected individuals in the system [Newman et al., 2006]. Additionally, the average degree $\langle d \rangle$ that governs the shape of the Poisson does not play a role in the power law distribution, as probability mass is not concentrated around a mean degree value. In Fig. 2.3(a) we present the shape of such a distribution, which implies a flat pyramid of connectivity hierarchy; a large number of loosely linked individuals are attached to a much smaller number of highly connected ones, creating "star-like" structures as shown in the example of Fig. 2.2(c).

Heavy-tailed degree distributions, in particular power-laws, have been reported in many studies of complex systems, the most influential being the work of [Albert et al., 1999] on the diameter of the World Wide Web, along with the study of [Faloutsos et al., 1999] on Internet topology. Other examples include networks of Hollywood actors, protein regulatory networks and research collaborations [Barabási and Bonabeau, 2003], therefore it is very plausible to ask why heavy tails arise in such a diverse collection of real-world systems. The most popular model that explains such an emergence is the "growth by preferential attachment" by [Barabási and Albert, 1999], which states that as the network grows newly arriving nodes have a higher probability of being linked to the more connected ones that exist already, following a "rich get richer" scheme. Despite the wide presence of heavy-tails in a diverse range of systems, it is important to note that liberal use of power-law as an "one size fits all" model for network degree distributions has attracted criticism [Stumpf and Porter,



(a)



(b)

Figure 2.3: In Fig. 2.3(a) we illustrate a power-law distribution, where nodes on the left-hand (green) side have low degree (number of connections). One the right-hand (yellow) side, nodes with high degree values form "hubs" of increased network connectivity. In Fig. 2.3(b) we compare the tail decay of power law versus the Poisson degree distribution of random graphs, on a log-log scale. We can see that the "tail" of the power law decays much slower than the Poisson one, implying a significant presence of well-connected nodes in the graph.

2012], both from a data analysis (lack of sufficient statistical support) and theoretical (lack of generative mechanism) perspective.

2.2.2 Small-world effect

One of the most famous and historically significant studies in the field of network analysis was Stanley Milgram's "small-world experiment", in which letters passed from person to person in order to reach a specific individual. It was shown that if a message successfully

reached its target (as many letters were lost), it occurred in only a small number of about 6 steps[1].

In the context of graphs, a given node i can reach a node j through a, possibly large, number of distinct paths. Assuming no disconnected components in an undirected unweighted network, we define the *geodesic distance* g_{ij} to be the number of edges that comprise the shortest path between i and j. A network is said to demonstrate a small-world effect if the value of the mean geodesic distance $\langle g \rangle$ scales logarithmically or slower with network size for fixed mean degree [Newman, 2003b]. Indeed, in many empirical studies of complex networks it is shown that $\langle g \rangle$ of any pair of vertices is many levels of magnitude smaller than the number of edges. In Table II of [Newman, 2003b], the author has presented a very illustrating summary of empirical studies on various real-world systems; for example, the "biology co-authorship network" of $N = 1,520,251$ nodes and $M = 11,803,064$ edges has an average geodesic path length of $\langle g \rangle = 4.92$ and the "film actors network" of $N = 449,913$ and $M = 25,516,482$ has $\langle g \rangle = 3.48$. Especially for the case of real-world networks with power-law degree distributions, $\langle g \rangle$ scales slower than $\frac{\log n}{\log(\log n)}$, as reported in [Newman, 2003b].

Real-world networks also tend to exhibit high *transitivity* or large *clustering coefficient*, meaning that if two nodes A, B are connected to a node C, then it is high likely that A and B are also connected[2]. This results in a tendency of *triangle formation* or *densification* in the network, thus increasing the number of different ways to reach a node starting from another. Such a redundancy, in combination with the heavy-tail degree distribution that allows highly connected hubs to act as "shortcuts" between nodes, results in very low geodesic distances in the network and "ultrasmall worldness" [Cohen and Havlin, 2003]. Additionally, [Leskovec et al., 2005] showed that the increased densification (triangle formation) during network growth results in shrinkage of geodesic distances and not (sub)logarithmic increase.

[1] This is better known in popular culture as "six degrees of separation".

[2] This comes as no surprise to sociologists, as Mark Granovetter in his seminal work "The Strength of Weak Ties" [Granovetter, 1973] has already discussed transitivity in human networks.

The authors presented experiments on various data sets (for example the arXiv and patent citation networks) and presented two models for network growth that capture densification and shrinkage (along with power law degree distribution); "Community Guided Attachment" and "Forest Fire model" [Leskovec et al., 2005].

2.2.3 Community structure

Definitions

In Section 2.2.2 we discussed the formation of triangles in a complex network, which consists of dense regions of increased link connectivity. In order to examine this organisational scheme at a larger scale, let us revisit the example network of animal interactions discussed in Section 2.2. There is no need to be a seasoned field expert to realise that such a population usually consists of closely connected bands of animals (e.g. groups of primates) that interact together more than average in terms of foraging, mating, stress relief activities or communal raising of younglings. In the second example of scientific collaborations, it is intuitive to assume that scholars tend to form groups based on similar research interests or even physical proximity (same university). Additionally, with proper visualisation of a simple network, such as the one showed in Fig. 2.4, we can immediately spot the dense regions where vertices share more links inside than outside them. These latent classes of nodes that we described are called *modules* or *communities*.

Community structure is the way in which a given network is clustered into a number of latent (and possibly overlapping) classes of nodes, creating "hot-spots" of increased connectivity. Although this notion is intuitively clear [Porter et al., 2009], there is no disciplined, context-independent and universally agreed definition of what constitutes a community [Fortunato, 2010; Reichardt and Bornholdt, 2006]. Nevertheless, [Reichardt and Bornholdt, 2004] have formalised the notion of "more links inside than outside", by stating that given a network of N nodes and M edges, a community is a subset of n nodes so that:

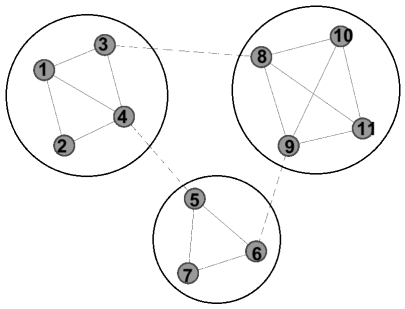


Figure 2.4: A simple example of community structure. This toy network consists of three densely connected regions with sparse communication between them.

$$\frac{2m_{in}}{n(n-1)} > \frac{2M}{N(N-1)} > \frac{m_{out}}{n(N-n)}, \tag{2.3}$$

where m_{in} is the number of intra-community edges (connecting nodes of the same community) and m_{out} the number of inter-community edges (connecting a community node to the external network). The first inequality of Eq. (2.3) suggests that network density[3] is "heavier" in node regions that form a community. The second inequality of Eq. (2.3) reflects that there is sparse communication between group members and the external network. This captures very elegantly our intuition of communities as densely connected node neighbourhoods, but suffers from a very serious drawback that the authors discuss both in [Reichardt and Bornholdt, 2004] and [Reichardt and Bornholdt, 2006]; the inequalities of Eq. (2.3) can be satisfied by chance, even in an Erdős-Rényi random graph that has no modular organisation by design. For that reason, given a module of n nodes with m_{in} intra-community and m_{out} inter-community links, we can compare the expected number of possible equivalent communities in a Erdős-Rényi random graph of the same number of nodes and edges, with pair connection

[3]Given a subset of n nodes in the graph, its density is the number of edges m that link these n nodes divided by the total number of all possible connections: $\frac{m}{\frac{1}{2}n(n-1)}$.

probability $p = \frac{2M}{N(N-1)}$:

$$\mathbb{E}(n, m_{in}, m_{out}) = \binom{N}{n}\binom{\frac{n(n-1)}{2}}{m_{in}}\binom{n(N-n)}{m_{out}} p^{m_{in}}(1-p)^{\frac{n(n-1)}{2}-m_{out}}\ p^{m_{out}}(1-p)^{n(N-n)-m_{out}}, \tag{2.4}$$

where if $\mathbb{E}(n, m_{in}, m_{out}) > 1$ then it is likely to find one such a community in a random graph of the same size, marking the border of statistical significance [Reichardt and Bornholdt, 2004].

Another definition is provided by [Palla et al., 2005] based on a k-clique[4] percolation process; modules consist of adjacent cliques, i.e k-cliques that share a $(k-1)$-clique (for example two triangles share an edge). This definition of communities as "clique percolation clusters" allows us to describe the overlapping nature of modular organisation but assumes high densification (see Section 2.2.2) inside the groups and might omit peripheral community regions of sparse connectivity. Additionally, the selection of parameter k is not obvious for any given network.

We can also define a community as a class of nodes in which the probability p_{in} of intra-community link presence is higher than the probability p_{out} of an external connection. This definition is based on the stochastic block model (SBM), where given a matrix $\mathbf{P} \in \mathbb{R}^{C\times C}$ in which C is the number of communities in the network, the element p_{ck} denotes the probability of a link between group c and k. The diagonal elements p_{cc} of $\mathbf{P}$ denote the probabilities of intra-community link presence, for which $p_{cc} > p_{ck},\ \forall\ k \neq c$. We can view this definition under the scope of a Markov model with transition matrix $\mathbf{P}$, where modules represent the system states. Thus given a random walker traversing the network who is in state c at time t, it will most likely stay in the same state at $t+1$, as the increased link density of community c will "trap" the walker inside.

[4] A k-clique is a fully connected subgraph of k nodes. For example, a triangle is a 3-clique.

Community hierarchies

We discussed that real-world networks often have some form of modular organisation, manifested as regions of increased link density. As these regions are considered relatively independent compartments of the whole system [Fortunato, 2010; Porter et al., 2009], it is fair to ask if communities exist within communities. There are many examples of real-world networks such as scientific collaborations, where groups of similar research area can be broken down to different labs, different cliques of scholars, to even simple pairs of long time associaties. Michelle Girvan and Mark Newman discussed the issue of *hierarchical community structure* in [Girvan and Newman, 2002], showing that starting from the whole network we can iteratively break it down to groups, eventually ending up to individual nodes. This hierarchy can be represented by a dendogram, an example of which is shown in Fig. 2.5.

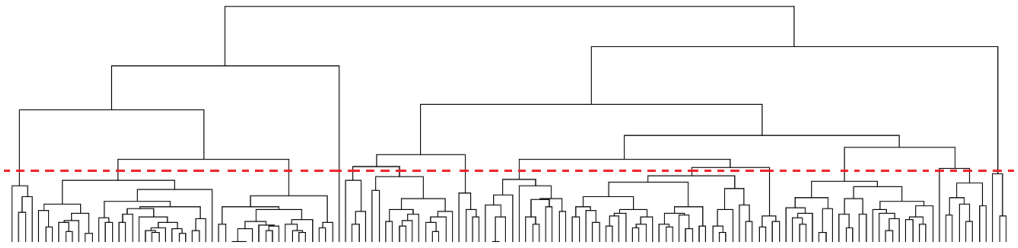


Figure 2.5: A dendogram showing the different resolutions of community structure in a wild-bird social network, as extracted using the Girvan-Newman algorithm [Girvan and Newman, 2002], ran until every node is a single partition and using recalculation of edge betweeness at each step. The horizontal dashed line represents a specific community division.

Such dendograms appear in many applications of community detection, which we discuss in more detail in Section 2.3.3.

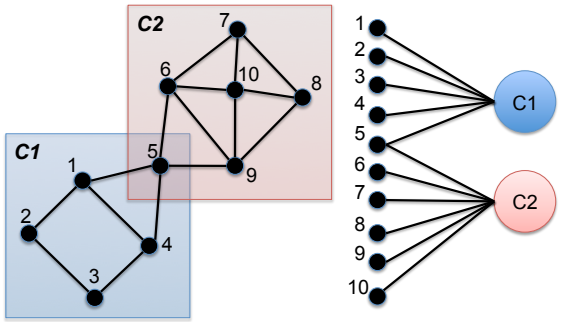
Following the above discussion and various definitions of community structure, in Section 2.3 we focus on the problem of detecting these classes of nodes and evaluating the quality of our partition. Further discussion on community structure itself and its origins in natural systems is presented in the stimulating analysis “The Road to Modularity” by [Wagner et al., 2007].

2.3 Community detection in networks

Community structure is a significant property of real-world networks as it is often considered to account for the functional characteristics of the system under study [Fortunato, 2010; Newman, 2010; Porter et al., 2009; Reichardt and Bornholdt, 2006]. In this section we present the problem of *community detection* or *module identification*, that consists of extracting these latent classes of nodes from a given network. We also discuss issues such as evaluating the quality of our partitions, the computational complexity of the algorithms, along with a brief summary of the most important methods from the literature.

2.3.1 Definitions

Consider the simple toy graph of Fig. 2.6(a), described by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{10\times 10}$ where we can immediately distinguish the two densely connected compartments C1 and C2. As we are not sure of the membership of node 5, it is fairly reasonable to consider it as an *overlap* of C1 and C2. We can describe our partition with a bipartite graph, as seen in Fig. 2.6(b), with *incidence matrix* $\mathbf{B} \in \mathbb{R}^{10\times 2}$ so that $b_{ic} = 1$ expresses that node-i belongs to a group-c and zero otherwise.



(a) Our toy network (b) Decomposed into two communities

Figure 2.6: An example community partition, on a simple toy graph of $N = 10$ nodes.

As the "ground truth" is not available to us in such settings, the problem of community

detection can be viewed as finding the appropriate $\mathbf{B} \in \mathbb{R}^{N \times C}$ that maximises a certain fitness criterion. The first challenge is obvious; we have to define an appropriate quality function, in a setting where there is no universally agreed definition of what constitutes a community (see Section 2.2.3). The second challenge is that we do not know a priori the number C of modules, placing our problem in a different framework than that of classic *graph partitioning* [Fortunato, 2010; Porter et al., 2009]. Finally, even if we define an appropriate quality function, the incidence matrix $\mathbf{B}$ defines a large solution space, where brute-force explorations are unfeasible even for small networks [Fortunato, 2010; Holmstrom et al., 2009; Porter et al., 2009]. The problem becomes worse if instead of having a Boolean incidence matrix, we require the elements b_{ic} to encode some form of "participation strength" of node i to community c.

The first challenge is discussed in the next section. The other two issues are addressed differently by various approaches, which we shall cover in Section 2.3.3.

2.3.2 Quality of network partitions

Given a network with adjacency matrix $\mathbf{A} \in \mathbb{R}_{(+)}^{N \times N}$, along with a candidate division $\mathbf{B}$ of size $N \times C$, how can we quantify the quality of the solution described by $\mathbf{B}$? In the next sections we describe such criteria for evaluating community structure.

Newman-Girvan Modularity

Consider a partition described by an incidence matrix $\mathbf{B} \in \mathbb{R}^{N \times C}$, in a network $\mathcal{G}$ with adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. Let us restrict our attention to the simple case where each node belongs only in one community (no overlaps between modules) and each b_{ic} is a Boolean variable (no measure of participation strength of nodes to groups). Based on our discussion in Section 2.2.3, we would have a "good" solution if the communities described by $\mathbf{B}$ have a *statistically surprising* link density. Thus we define an appropriate null network, with same

number of nodes and degree sequence as $\mathcal{G}$, but with random wiring of links so that they connect vertices with no regard to modular organisation. Thus we require that our partition of $\mathcal{G}$ has a higher fraction of intra-community links than the null model.

To quantify this, we use Mark Newman's "assortative mixing matrix" $\mathbf{E} \in \mathbb{R}^{C \times C}$ [Newman, 2002], expressed as:

$$\mathbf{E} = \left(\sum_{ij} a_{ij} \right)^{-1} \mathbf{B}^{\mathsf{T}} \mathbf{A} \mathbf{B}. \tag{2.5}$$

Each element e_{ij} denotes the fraction of edges that connect members of community i to community j. The sum E_c of row (or column) c of $\mathbf{E}$ encodes the fraction of edges in the network that have their one end lying on a node belonging to community c. In the case where edges connect nodes regardless of any modular organisation (null model), the fraction of links connecting community i to j would be $e_{ij}^{(null)} = E_i E_j$. The diagonal elements of $\mathbf{E}$ and $\mathbf{E}^{(null)}$ express the fraction of intra-community links in each network, which we seek to compare. Therefore, we define *modularity* Q [Newman and Girvan, 2004] as:

$$\begin{aligned} Q &= \sum_{c=1}^{C} (e_{kk} - E_k^2) \\ &\Leftrightarrow \frac{1}{2M} \sum_{i=1}^{N} \sum_{j=1}^{N} (a_{ij} - \frac{d_i d_j}{2M}) \delta(\sigma_i, \sigma_j), \end{aligned} \tag{2.6}$$

where δ is the Kronecker delta and $\delta(\sigma_i, \sigma_j) = 1$ expresses that nodes i and j belong to the same community. The theoretical range of modularity values is $(-1 + \frac{1}{C}, 1 - \frac{1}{C})$ [Holmstrom et al., 2009] and larger Q imply better community division. Newman and Girvan suggest that most real-world networks have modularity values ranging from 0.3 to 0.7 [Newman and Girvan, 2004]. Eq. (2.6) can be also applied to weighted networks [Newman, 2004], where the null model is now defined based on the strength sequence (instead of degree sequence) of our given graph.

The introduction of Newman-Girvan modularity has stimulated extensive interest in the field of community detection, mainly in terms of developing methods for extracting modules via direct optimisation of Q. We can think of Eq. (2.6) as a function over partitions, so that $f : \mathbf{B} \rightarrow Q$, which we seek to maximise. Unfortunately, such a function defines a rugged and complicated surface [Good et al., 2010; Holmstrom et al., 2009] and the total number of possible partitions grows exponentially [Fortunato, 2010; Holmstrom et al., 2009] at increasing network sizes, thus making the task of global maximisation of Q computationally unfeasible. Various approximation schemes are employed for *modularity optimisation*, which we shall discuss in more detail in Section 2.3.3.

Modularity, in its original formulation presented above and in [Newman and Girvan, 2004], cannot be directly applied to the case of networks with overlapping community structure, neither in the case where nodes have different participation strength to different modules. Additionally, modularity allows us to compare different divisions only for the same network, as Q tends to achieve high values for larger graphs [Fortunato, 2010]. It is also possible for partitions of random graphs to have high modularity values, without intrinsically possessing any community organisation [Fortunato, 2010]. Finally, modularity has a *resolution limit* problem, which favours divisions with small number of communities as discussed in [Fortunato and Barthélémy, 2007].

Hamiltonian

[Reichardt and Bornholdt, 2004, 2006] proposed an approach to evaluating community structure using the popular Potts model from Statistical Mechanics. Consider a system of N spins (nodes) that can be in C states (communities) and their topology is described by a nearest-neighbour interaction (adjacency) matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. Neighbouring spins (connected pairs) that belong to the same state (community) are considered ferromagnetic interactions that lower the total energy of the system. Inversely, neighbours of different state have repulsive

force that increases the total energy. We can describe the effect on the total energy $\mathcal{H}$ of the system of these two competitive terms of "cohesion versus anti-cohesion" using the *Hamiltonian* [Reichardt and Bornholdt, 2006]:

$$\mathcal{H} = -\sum_{i<j} \phi\ a_{ij}\delta(\sigma_i, \sigma_j) + \sum_{i<j} \psi(1 - a_{ij})\delta(\sigma_i, \sigma_j). \tag{2.7}$$

The first term of $\mathcal{H}$ "rewards" linked pairs i, j that belong to the same group (when $\delta(\sigma_i, \sigma_j) = 1$) by some quantity ϕ, while the second term "penalises" the allocation of disconnected pairs to the same community by a factor of ψ. The reward ϕ and penalty ψ weights can be controlled by a single parameter γ so that $\phi = 1 - \gamma\frac{d_i d_j}{2M}$ and $\psi = \gamma\frac{d_i d_j}{2M}$. Applying this to Eq. (2.7) following [Reichardt and Bornholdt, 2006], we have:

$$\mathcal{H} = -\sum_{i<j}(a_{ij} - \zeta\frac{d_i d_j}{2M})\delta(\sigma_i, \sigma_j), \tag{2.8}$$

which is the negative Newman-Girvan modularity with a *resolution parameter* ζ. Similar to Q, global minimisation of $\mathcal{H}$ is computationally intractable and approximation schemes such as simulated annealing or Monte Carlo optimisation are employed [Reichardt and Bornholdt, 2004]. Additionally, Eq. (2.7) and (2.8) cannot be directly applied for overlapping communities with different node belonging coefficients.

Extracting modules via approximation of $\mathcal{H}$ for various values of ζ (typically $0.01 < \zeta < 100$) allows us to look at different community resolutions in the network and larger ζ lead to denser communities with low node population [Reichardt and Bornholdt, 2006]. This may address the resolution limit of modularity we discussed in Section 2.3.2 and presented in [Fortunato and Barthélémy, 2007], in the sense that we can perform exploratory search at different levels granularity ζ. Based on such an idea, authors in [Onnela et al., 2012] have exploited the Hamiltonian parameter ζ in order to define a similarity metric between networks, based their structural similarity profile across a range of community resolutions.

Modularity in overlapping community structure

As we discussed in Section 2.3.2, the derivation of modularity Q in Eq. (2.6) cannot be directly applied to networks with overlapping community structure. [Nicosia et al., 2009] have extended the definition of Q, in order to take into account the case where $\mathbf{B} \in \mathbb{R}^{N \times C}$ describes an overlapping community division and each element B_{ic} denotes the participation strength or "belonging coefficient" [Nicosia et al., 2009] of node i to community k.

Based on Eq. (2.6), we can write the Newman-Girvan modularity as:

$$Q = \frac{1}{2M} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(a_{ij} \delta(\sigma_i, \sigma_j) - \frac{d_i d_j}{2M} \delta(\sigma_i, \sigma_j) \right), \tag{2.9}$$

which is a quantity that runs over all edges and calculates their contribution to Q. Notice that the role of Kronecker delta $\delta(\sigma_i, \sigma_j)$ is to "switch off" the effect of any link that does not have both ends (nodes) lying on the same group. In the case of overlapping community structure, we seek to weight the contribution of each edge based on "how strongly" its connected pair belongs to the same module. For that reason, Nicosia *et al.* proposed a weight coefficient β_{ijc} for each link connecting nodes i and j given a community k, which is a function of the participation strengths of b_{ic} and b_{jc} of i and j to c [Nicosia et al., 2009]; $\mathcal{F} : b_{ik}, b_{jc} \rightarrow \beta_{ijc}$. A similar quantity $\beta_{ijc}^{(\text{null})}$ is defined for the second term of Eq. (2.9) that defines the null model. Thus we rewrite Eq. (2.9) as [Nicosia et al., 2009]:

$$Q = \frac{1}{2M} \sum_{c=1}^{C} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(a_{ij} \beta_{ijc} - \frac{d_i d_j}{2M} \beta_{ijc}^{(\text{null})} \right). \tag{2.10}$$

The authors acknowledge that the choice of $\beta_{ijc} = \mathcal{F}(b_{ic}, b_{jc})$ is somewhat arbitrary [Nicosia et al., 2009]. Additionally, the formulation of $\beta_{ijc}^{(\text{null})}$ as the mean of all β_{ijc} is problematic, because it implicitly involves the same node participation strengths b_{ic}, b_{jc} to communities of the original division.

2.3.3 Community detection algorithms

In this section, we present an overview of popular and theoretically significant module identification methodologies from the literature. A comprehensive overview is presented in the survey of [Fortunato, 2010], along with a discussion on computational complexity isses.

Topology based methods

The Girvan-Newman algorithm (GN): We start with one of the most well-known and historically important community detection algorithms, presented in [Girvan and Newman, 2002], which is based on the notion of *edge betweeness*. The core assumption of the method is that inter-community edges are high flow paths in the network that act as "communication highways" between nodes. By defining edge betweeness as the fraction of shortest paths between any pair of nodes that pass through an edge, the idea is that by iterative removal of such links we isolate groups of nodes that form a community. The algorithm is divisive, i.e. it produces a dendogram (see Section 2.2.3) describing the modular organisation at different resolutions, where we select the best tree layer based on modularity Q. The computational complexity of GN is governed by constant calculation of shortest paths, yielding $\mathcal{O}(M^2N)$. Such computational load is deteriorated for densely connected graphs [Fortunato, 2010]. Additionally, the algorithm cannot describe overlapping communities or quantify how strongly each node belongs to a group.

Clique Percolation Method (CPM): In Section 2.2.3 we discussed that communities can be viewed as agglomerations of fully connected subgraphs such as triangles. Palla *et al.* have proposed an algorithm that explores the network in order to find adjacent k-cliques, i.e cliques that share a $(k-1)$-clique [Palla et al., 2005]. CPM is a popular approach for finding overlapping communities in networks and it has been also extended to the case of weighted graphs in [Farkas et al., 2007]. The drawbacks of the method is that it does not quantify the par-

ticipation strength of each node across different communities. Additionally, the problem of finding maximal subgraphs given a network scales exponentially and the method has a severe computational overhead $\mathcal{O}(e^n)$ [Danon et al., 2005], although it has been shown that in real-world networks the algorithm terminates in a reasonable time window [Fortunato, 2010; Palla et al., 2005].

Spectral methods

Donetti-Mũnoz method (DM): In this method we utilise the *Laplacian* matrix $\mathbf{L} \in \mathbb{R}^{N \times N}$, defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{2.11}$$

where $\mathbf{D}$ a matrix containing the node degrees d_i on its diagonal and 0 elsewhere, while $\mathbf{A}$ an adjacency matrix. The idea is to use the values of K eigenvector components of $\mathbf{L}$ and project each individual to a K-dimensional space [Donetti and Muñoz, 2004]. Then, by defining a distance-based similarity measure (such as the ones described in the previous section), we apply hierarchical clustering to produce a dendogram of possible network partitions. The main drawback of this method is that the number K of eigenvectors is not known a priori and the method performance heavily relies on choosing a proper value. Finally, no overlapping communities or node participation strengths are produced.

Spectral Partitioning (SP): SP is a divisive method by [Newman, 2006] that builds a community hierarchy dendogram by performing bisections of each module, by utilising the *modularity matrix* $\mathbf{M} \in \mathbb{R}^{N \times N}$, defined as $m_{ij} = a_{ij} - \frac{d_i d_j}{2M}$. Thus given the full graph, each bisection is performed by computing the leading eigenvector of $\mathbf{M}$ and by dividing the vertices into two groups according to the signs of the elements of this vector. For each subgroup $\mathbf{g}$ we perform the same division scheme but with an updated modularity matrix

$m_{ij}^{(c)} = m_{ij} - \delta_{ij} \sum_{c \in \mathbf{g}} M_{ic}$. The method has the excellent property of identifying indivisible groups where a community cannot have further divisions with positive modularity, if the all the eigenvalues of the modularity matrix are non-positive. It also provides good results in most real and artificial problems but cannot account for the overlapping nature of community structure, or node multi-membership with different weights.

The computational performance of these spectral methods is governed by the process of eigenvector extraction. In fast implementations (such as using the Microsoft "Lightspeed toolbox" in MATLAB) these methods can be reasonably scalable for large data sets.

Optimisation methods

Extremal Optimisation (EO): [Duch and Arenas, 2005] proposed a method for direct approximation of the Newman-Girvan modularity Q. The core idea is that the overall Q can be expressed as the sum of modularity contributions from each node. Thus given an individual i assigned to group r, its contribution is:

$$\lambda_i = \frac{d_{r(i)}}{d_i} - E_{r(i)}, \tag{2.12}$$

where $d_{r(i)}$ is the number of neighbours of i that belong to the same community as i (intra-community degree) and $E_{r(i)}$ is the r-th element of the "E_k" vector we described in Eq. (2.6) of Section 2.3.2. The quantity λ_i is normalised in the interval $[-1, 1]$ and the overall Q can be easily recovered by $Q = \frac{1}{2M} \sum_{i=1}^{N} k_i \lambda_i$.

The algorithm starts by creating a random bisection of the original graph. Then it calculates each λ_i and moves the least contributing nodes to another group. Due to the random initialisation scheme and to avoid local maxima we usually select randomly one of the n least contributing nodes to change its membership. After a number of moves, if the modularity does not improve we proceed recursively by applying the algorithm to each partition. From the resulting dendogram we select the partition with the highest modularity. Although EO

is initialisation dependent [Duch and Arenas, 2005] and slow $\mathcal{O}(N^2 \log N)$ [Danon et al., 2005], it provides good modularity scores across a variety of problems, as we will show in Chapter 4.

Louvain Method: The Louvain method, proposed by [Blondel et al., 2008], is an agglomerative method of iteratively merging individual nodes and groups, based on each move's contribution to the Newman-Girvan modularity Q. Starting from the bottom layer of the dendogram, as the one shown in Fig. 2.5, where every individual is a community by itself, we proceed by grouping together nodes based on modularity gain, in an approach similar to Extremal Optimisation. During the next steps, we build a new network where is group is a node, named "meta-community" and proceed in a similar fashion. Because in every layer we are dealing with a smaller graph (network of communities) the bulk of computational load is concentrated on the first iterations of the algorithm. The authors state that Louvain has a "near-linear" computational complexity, based on various experiments on "large ad hoc modular networks" [Blondel et al., 2008]. From our experiments, the method produced state of the art results in finding the maximum of Q, similar or better than Extremal Optimisation (more details in Section 4.3.3 of Chapter 4). Although Louvain is one of the best methods available of optimising modularity, it is still a "hard partitioning" method, giving solutions with no community overlaps, or node participation rates.

Other optimisation methods include the classic *simulated annealing* [Kirkpatrick et al., 1983], which can be applied either to Newman modularity Q or the Potts Hamiltonian $\mathcal{H}$ [Guimerà and Amaral, 2005]. Various experiments on artificial networks presented in [Lancichinetti and Fortunato, 2009] have shown that simulated annealing provides top performance in approximating such fitness functions. Unfortunately the method is slow, making its application prohibitive for large networks.

Link communities

The work of Evans and Lambiotte [Evans and Lambiotte, 2009] detects communities of links — in contrast to node communities, which occupy the vast body of the literature [Fortunato, 2010; Porter et al., 2009] — after performing a lossless transformation of the adjacency matrix to a "line graph". By assigning links, rather than nodes, among communities, the method allows a node to participate naturally in more than one group, as determined by the labels assigned to its adjacent links. The advantages of this approach have also been presented in [Ahn et al., 2010].

Ball-Karrer-Newman (BKN) method: Of particular interest is the approach of [Ball et al., 2011], where each link is assigned one out of C "colours", while nodes are associated with a "propensity" θ_{ic} to have links of color c. Nodes belonging to a community, correspond to graph regions that are densely connected with edges of the same color.

Assuming a multi-graph structure with self-edges, the authors consider $\theta_{ic}\theta_{jc}$ to be the expected number of links of color c, under a Poisson noise model, connecting the node pair i, j. The probability of generating the observed undirected graph $\mathcal{G}$ with adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ is thus given by:

$$\begin{aligned} p(\mathcal{G}|\theta) &= \prod_{i<j} \frac{\left(\sum_C \theta_{ic}\theta_{jc}\right)^{a_{ij}}}{a_{ij}!} \exp\left(-\sum_c \theta_{ic}\theta_{jc}\right) \\ &\quad \times \prod_i \frac{\left(\frac{1}{2}\sum_c \theta_{ic}\theta_{ic}\right)^{a_{ii}/2}}{(a_{ii}/2)!} \exp\left(-\frac{1}{2}\sum_c \theta_{ic}\theta_{ic}\right), \end{aligned} \tag{2.13}$$

where the product in the first factor runs over the upper diagonal part of the adjacency matrix $\mathbf{A}$ and the second concerns the diagonal elements (self-edges) a_{ii}. Detecting the overlapping modules corresponds to finding the appropriate propensity θ_{ic} values that maximise Eq. (2.13). This is accomplished via a bespoke Expectation-Maximisation (EM) scheme,

presented in [Ball et al., 2011], along with a detailed discussion on implementation details. The method provides excellent module identification capabilities in LFR graphs and real-world networks, with a key drawback the lack of determining the total number of "colours" (communities) C.

Other methods

Infomap: Modular organisation can be seen as a compressed version of the network, transmitted via a communication channel between sender Y and receiver X. Based on this idea, [Rosvall and Bergstrom, 2007, 2008] proposed that the best community division corresponds to the signal Y that contains the most information about X. Therefore, we seek to minimise the conditional information $H(X|Y)$ between sender and receiver:

$$H(X|Y) = \log \left[\prod_{i=1}^{K} \binom{n_i(n_i-1)/2}{l_{ii}} \prod_{i>j} \binom{n_i n_j}{l_{ij}} \right], \tag{2.14}$$

where n_i is the number of nodes in community i, l_{ij} the links connecting modules i and j and K is the number of communities. The trivial solution of $X = Y$ that gives $H(X|Y) = 0$ is avoided by placing some additional constraints regarding the quantities n, m, K based on the Minimum Description Length principle [Grunwald, 2007]. The constrained target function is optimised via simulated annealing giving state of the art results in terms of modularity Q, comparable to the Louvain [Blondel et al., 2008] and BKN [Ball et al., 2011] methods described previously. Although the optimisation scheme makes Infomap slower than Louvain, it is considered one of the most attractive methods for community detection [Fortunato, 2010] in settings where we are not interested in describing community overlaps.

Label propagation: [Raghavan et al., 2007] have proposed an algorithm with very attractive computational scalability properties, which considers communities as node labels that propagate through the network via a copying process; at each iteration, each node takes the

label shared by the majority of its immediate neighbours. In cases where there is no dominant label (as in the initialisation phase, where each node has its own label) one of the majority ones are randomly picked. In that way, a particular label propagates through the network and communities (i.e. nodes with the same label) arise naturally, due to the sparse-connectivity "bottlenecks" between modules. The solutions produced by label propagation are sensitive to initial conditions and random label selection at no-majority cases, so multiple runs may need to be performed in order to compare solutions. An interesting exploitation of this stochasticity is proposed in [Leung et al., 2009] in order to produce overlapping partitions, based on node multi-memberships across algorithm runs. Though such an approach has its merits given the computational efficiency of the algorithm, overlapping solutions are viewed as residuals of the algorithm stochasticity, rather than an inherent property of the graph community organisation. Therefore, for large graphs we may need to run Label propagation an computationally prohibitive number of times, in order to perform appropriate sampling of different partitions.

Other approaches include the algorithm of [Lancichinetti et al., 2009], which seeks a local maximum of the community "fitness" function (based on internal link density) by modifying nodes' community "appropriateness" scores through a series of inclusion-exclusion moves. The final method we present is from [Nepusz et al., 2008], which propose that communities should be comprised of "similar" nodes, assuming that a distance (inverse similarity) between nodes is defined. When a partition matrix, representing a reasonable partition, is multiplied by itself it would then be expected to approximate the similarity matrix; this leads to a nonlinear constrained optimisation problem. The number of communities of the proposed incidence matrix is selected by performing multiple runs and selecting the one with the highest fitness score based on a Newman modularity-like function.

2.4 Discussion

In this chapter we have presented an overview of networks as tools to describe complex systems in nature, society and technology. We discussed that real-world networks have a structure that is neither fully ordered nor completely random, thus raising the need for a different toolset than the one already available from classic Graph Theory. We also presented the concept of community structure, which is a prominent feature in the vast majority of complex networks, affecting their form and function. Detecting such latent groups of individuals at any given network is a challenging task, both as an inference and as a computational efficiency problem. For that reason, we presented a brief overview of popular community detection algorithms in the literature, discussing their strengths and weaknesses in both of these aspects. In Chapter 4 we present our contribution to the community detection literature, by proposing a novel method that addresses the issues of overlapping communities and weighted node participation to modules in a computationally efficient manner.

Another important issue of the network approach is that it is not always clear how to map a real life complex system to a graph. Although systems such as the UK telephone network or the U.S. power grid have an obvious web-like structure, this is not apparent in applications such as describing wild animal populations or stock price correlations. Different ways of defining an association (and its intensity) among different individuals may lead to different graph topologies, as we discuss in Chapters 6 and 7.

Finally, we need to underline that regardless of how accurate or efficient a module identification algorithm is, its output defines *structural* communities, which might be different from *functional* ones. Again, we have to take into account our modelling assumptions and incorporate application domain knowledge to our result interpretations. Further discussion is provided in Chapter 8, where we seek to model a wild bird population in a social network context, via an appropriate mapping of spatio-temporal sensor data.

Chapter 3

Elements of Probability Theory

3.1 Introduction

All analysis presented in this thesis is based on a probabilistic approach to modelling complex systems, where we employ the tools from statistical inference to reason over quantities of interest under *conditions of uncertainty*. In this chapter we introduce the methodologies, nomenclature and terminology used throughout the thesis, along with a discussion of their applications to social networks in general. For further reading, a deep theoretical analysis of probability theory is presented in [Jaynes, 2003] while a more practical discussion, from a machine learning perspective, is presented in [Bishop, 2007].

Uncertainty is inherent in the study of all natural and technological systems, from global climate prediction models to quantifying investment risk on Japanese sovereign bonds. It stems from a variety of factors; our inability to perform exact modelling of the multiplicity of parameters that govern a particular system, along with the noise-inducing imperfections of our observational media and the finite size of our available data sets.

Consider the example of a die roll, where although there is nothing inherently "random" about the outcome, our inability to model all the variables affecting it (momentum, angle, velocity, acceleration) forces us to associate each result with a *plausibility score* rather than

a function that deterministically gives “1” or “6” given all initial conditions.

Another example of uncertainty comes from the inherent limitations of our observational equipment; tracking a vehicle via GPS, measuring the pH of a solution or estimating the age of a fossil always involves some measure of error induced by inaccuracies of our sensors. Therefore the value of the variable of interest is not a single number, but a range of possible values where some are more plausible than others.

Finally, uncertainty arises from the finite nature of the data sets we are able work with. Consider the example of the US National Election, where forecasts are based not on the entirety of the voting body, but on a subset or sample of the registered voters. Appropriate sample selection and understanding of its limitations, allows political analysts to make very accurate predictions of election outcomes.

Performing the above *inference tasks* of quantifying the plausibility of an outcome, filtering-out noise and producing general conclusions from limited data, requires:

1. A model - that is, a conceptual representation of the system’s function, based on a collection of *variables* and their *associations*. Some of the variables in our model are “hidden” or *latent* and constitute our inference target, as they usually reflect the underlying factors that give rise to the observed data.

2. Prior knowledge - that captures the *context* within which we perform our inferences, resulting from past repetitions of an experiment, or expert domain knowledge that we seek to incorporate into our inference scheme.

3. An (efficient) inference engine - that applies the rules of Probability Theory in order to quantify our belief over the values of the latent variables in our model. Such solutions might be *exact* or *approximate*, given the formulation of our model and the available computational resources.

In the next sections we present the elements of Probability Theory that allow us to ad-

dress the inference tasks discussed in the present work. Our presentation is accompanied by examples within a social network analysis framework.

3.2 Definitions

Consider a *stochastic variable* a that quantifies the strength of social tie (edge weight) between Alice and Bob, say in terms of the number of hours they spend together per week. As in deterministic variables, a is associated with a domain that defines its value range, in our case $a \in [0, 168]$. Figuring out an exact value for a is infeasible, not only due to the technical (and possibly ethical) obstacles of monitoring their whereabouts 24/7, but also due to the fact that the number of hours they spend together varies from week to week.

As we cannot define an exact value for a, the next best thing is to capture the *plausibility* of each value $[0, 168]$ using an appropriate function $p(.)$, so that $p : [0, 168] \rightarrow [0, 1]$. We can thus express our belief over the interaction weight between Alice and Bob as $p(a)$, where $p(.)$ a *probability density function*, for which we have:

$$p(a) \in [0, 1], \text{ and } \sum_a p(a) = 1. \tag{3.1}$$

The formulation of Eq. (3.1) can be extended to the case where a is a continuous stochastic variable:

$$p(a) \geq 0, \text{and } \int_a p(a)\ da = 1. \tag{3.2}$$

Our belief function $p(.)$ may have a functional form (such as Gaussian, Poisson, Gamma, etc), or it can have an ad hoc structure, produced by an empirical histogram from repeated observations. The only requirement is that it has to conform with the requirements of non-negativity and sum to 1 over its domain.

Given the stochastic variable a and the associated belief function $p(.)$, we can define the

mean, or *expected value* of a as:

$$\mathbb{E}[a] = \sum_a ap(a), \quad \text{or} \quad \mathbb{E}[a] = \int_a ap(a)\ da, \tag{3.3}$$

where in the classical probability framework, the expected value is interpreted as the statistical mean $\frac{1}{N}\sum_i a_i$ obtained by observing a N times, for $N \to \infty$.

The variation of a around its mean value $\mathbb{E}[a]$ is given by the *variance* var$[a]$, defined by:

$$\begin{aligned} \text{var}[a] &= \mathbb{E}[(a - \mathbb{E}[a])^2], \\ &= \mathbb{E}[a^2] - \mathbb{E}[a]^2. \end{aligned} \tag{3.4}$$

It is worth noting that both the mean and variance defined in Eq. (3.3) and Eq. (3.4) can be generalised to the case of a function $f(a)$ over a, as:

$$\mathbb{E}[f(a)] = \sum_a f(a)p(a), \quad \text{or} \quad \mathbb{E}[f(a)] = \int_a f(a)p(a)da, \tag{3.5}$$

$$\text{var}[a] = \sum_a p(a)(f(a) - \mathbb{E}[f(a)])^2, \quad \text{or} \quad \text{var}[f(a)] = \int_a p(a)(f(a) - \mathbb{E}[f(a)])^2 da. \tag{3.6}$$

3.3 Graphical Models

Let us revisit the social tie example, where we seek to model our belief over the edge weight between two individuals. If we consider the modelling assumption that the number of hours a Alice and Bob spend together per week results from the number c of college communities they both participate in, how would our belief on a change given c? We express such probabilistic dependency as $p(a|c)$, where c denotes the *context* within which we quantify our beliefs. It encodes our modelling assumption that some underlying factor c affects the value of a, thus

our belief on a depends, or *is conditioned*, on our belief on c.

The formulation of $p(a|c)$ expresses both a *belief* and *dependency* structure in our model. We graphically illustrate such expressions in the form of *Directed Acyclic Graphs* (DAGs), an example shown in Fig. 3.1(a), where stochastic variables (such as a) are drawn as circles and their dependencies as directed edges. Each variable is governed by a probability distribution while the arrow expresses conditional dependency, so that any knowledge about c alters our belief about a.

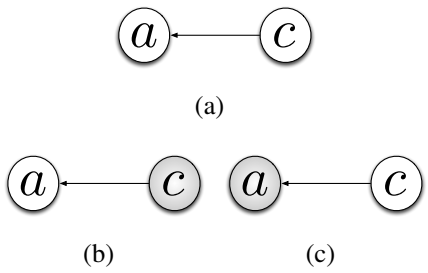


Figure 3.1: We illustrate three examples of representing the conditional structure between edge weight a and community co-membership c via a Directed Acyclic Graph (DAG). Fig. 3.1(a) represents the general case of conditional dependency, while Fig. 3.1(b) and 3.1(c) illustrate cases where one of the two variables is observed (dark coloured circle).

In Figs. 3.1(a) and 3.1(c), we also consider the cases where one of the two variables is *observed*, i.e. its value is known. Given our social tie example and Fig. 3.1(b), we can reformulate our inference question as "how does our belief on the number of hours Alice and Bob spend together change, given that we know the number of college communities they both participate in?", while Fig. 3.1(c) reads "given that we have observed how many hours Alice and Bob have spent together, what does it tell us about the number of college communities they might both be members of?".

Directed Graphical Models represent our conceptual representations of the system under study, laying out our assumptions on how its internal machinery (set of variables) is interconnected. For the case shown in Fig. 3.1(a), we can express the overall probability of our model as the *joint distribution* $p(a, b)$, which factorises as:

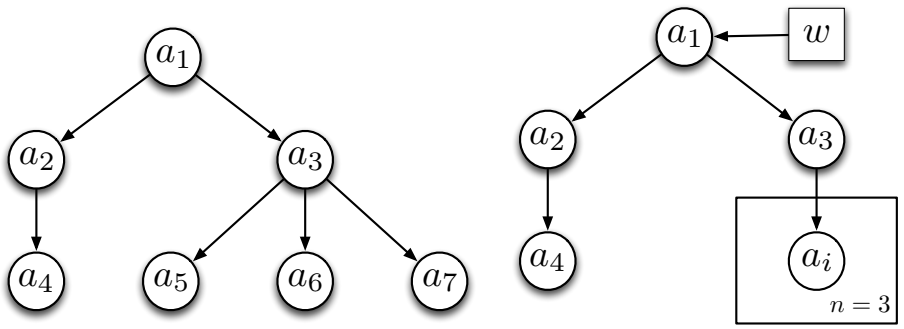
$$p(a, c) = p(a|c)p(c), \tag{3.7}$$

based on the *product rule* of probability. In the case where a and c are independent, we have $p(a|c) = p(a)$ thus the joint factorises as $p(a, c) = p(a)p(c)$.

Given the dependency (arrow from c to a) stated in Fig. 3.1(a), the *marginal distribution* $p(a)$ of a is given by:

$$\begin{aligned} p(a) &= \sum_c p(a, c) \\ &= \sum_c p(a|c)p(c), \end{aligned} \tag{3.8}$$

which is termed the *sum rule* of probability. It expresses that, given the graphical structure of our model, in order to express our belief about a without any explicit knowledge about the value of c, we must marginalise its *parent node* c, thus taking into account all of its possible values, weighted by their likelihood $p(c)$. This process is also known as "integrating-out".



(a) (b)

Figure 3.2: Two examples of DAGs, where in Fig. 3.2(b) we have explicitly stated the existence of a deterministic variable and compacted the common ancestors of a_3 to a single group a_i.

The above rules can be used to write down the equations for more complicated graphi-

cal models. Consider the general case pictured in Fig. 3.2(a). Given the interdependencies expressed by the arrows, we can write the joint distribution $p(a_1, ..., a_7)$ over the model as:

$$p(a_1, ..., a_7) = p(a_1)p(a_2|a_1)p(a_4|a_2)p(a_3|a_1)p(a_5|a_3)p(a_6|a_3)p(a_7|a_3), \tag{3.9}$$

where the probability of a given stochastic variable a_i, given our model, is dependent only on its direct predecessors or *parents* pa(a_i). Given any arbitrary DAG topology, the following factorisation of the joint distribution always holds:

$$p(a_1, ..., a_N) = p(a_1) \prod_{i=2}^{N} p(a_i|\text{pa}(a_i)), \tag{3.10}$$

where we have considered N the total number of nodes, a_1 to be the root node of the DAG (as no cycles are allowed) and pa(a_i) the set of parent nodes of a_i.

Further extensions to the notation include the use of squares to express *deterministic variables* in our model, shown for w in Fig. 3.2(b), for which there is no associated uncertainty. Dark coloured nodes, shown as a_i in Fig. 3.2(b), express stochastic variables with observed value. Additionally, multiple variables with identical dependencies can be replaced with a single node nested in a plate, shown as c_k and b_i in Fig. 3.2(b), along with a label that denotes the cardinality of such variable set.

Finally, we make the important distinction between *observations*, *latent variables* and *parameters* in a graphical model. Consider the DAG of Fig. 3.3, where there are N observations (dark coloured circle) and for each a_i there is a corresponding parent b_i. The N stochastic variables b_i are conditioned on a collection of K other variables c, which are in turn dependent on a fixed deterministic component w.

By adopting the canonical notation from [Bishop, 2007], we consider b_i as *latent variables* because they grow with the number N of observations. The variables c_k and w are considered as *parameters* in our model, as they are independent of the size of observed data. Finally, we may want to explicitly state a variable's position in the depth of the DAG, by

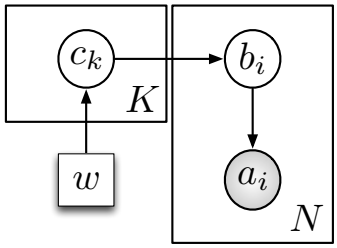


Figure 3.3: An example graphical model with N observed variables, where each a_i is conditioned on a latent variable b_i. Additionally, there are K parameters, controlled by a single hyper-parameter w.

using the "hyper-" prefix. In the example of Fig. 3.3, we call c_k a parameter and w a hyper-parameter.

3.4 Bayesian inference

Let us recall the example of modelling the edge weight between two individuals in the form of a stochastic variable a. In Fig. 3.1(a) we have formulated a probabilistic dependency between the number of hours a Alice and Bob spend each week and the number c of college communities they are both members of, where c controls the value of a.

Consider the case presented in Fig. 3.1(c), where given a known value of the social tie strength (a is observed) between the two individuals, we seek to make an inference about their co-membership score c. Using the sum and product rules presented in the previous section, we can express our belief on c given a as:

$$p(c|a) = \frac{p(a|c)p(c)}{p(a)} \tag{3.11}$$

$$= \frac{p(a|c)p(c)}{\sum_c p(a|c)p(c)}, \tag{3.12}$$

which is known as *Bayes' Rule*. If we employ a "cynical" interpretation of Eq. (3.11) and

(3.12), we would say that Bayes' rule is nothing more that a direct application of the sum and product rules of probability, along with an appropriate re-arrangement of terms.

The above statement is only partially true; Bayes' rule indeed obeys the fundamentals of Probability Theory and can be easily recovered from any factorisation of a joint distribution. The real insight lies on how the involved terms are being interpreted. Let us take a closer look at Eq. (3.11), while recalling our social tie example, which can be seen as rule via which we move from $p(c)$ to $p(c|a)$:

$$\begin{aligned} p(c|a) &= \frac{p(a|c)}{p(a)}p(c), \text{or} \\ p(c|a) &\leftarrow p(c) \end{aligned} \tag{3.13}$$

Recall that c denotes the number of common social communities Alice and Bob are members of, which given our modelling assumption in Fig. 3.1(a) directly affects the number of hours a they spend together. What we get from Eq. (3.11) and (3.13), is a rule on how to *update our prior belief* $p(c)$ based on an observation about a. Our initial belief state $p(c)$ is called *the prior* while the updated distribution $p(c|a)$ on c is called *the posterior*. In the term $p(a|c)/p(a)$ involved in the update, $p(a|c)$ is called the *likelihood* as it expresses the probability of the observation conditioned on the latent variable, while $p(a)$ is termed the *marginal likelihood* or *evidence* as it defines our belief over a with the effect of c integrated-out. Thus we can express Bayes' Rule in the general form:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \tag{3.14}$$

Bayes' Rule provides a rigorous, systematic and mathematically sound means to quantify the shift in our belief on a stochastic quantity of interest, every time we receive some data. The ramifications of that extend beyond the ivory tower of mathematical elegance; the fact

that a Bayesian posterior can be used as the prior once we receive new data and so on, forms the basis of most modern machine learning and computational intelligence systems. In the present work applications of Bayes' Rule, along with Directed Acyclic Graphs, are used throughout the methodological development of models in the next chapters.

3.5 Information Theory

Consider a discrete stochastic variable a, associated with an appropriate belief function $p(.)$. If we observe a particular outcome a_i from a, we define the *Shannon information* $h(a_i)$ as:

$$h(a_i) = -\log_m p(a_i), \tag{3.15}$$

given a logarithm base m, which defines which units of measurement are used (e.g. $m = 2$ for bits or $m = e$ for nats) while the negative sign ensures non-negativity as $p(a_i) \in [0, 1]$. The expression $h(a_i)$ quantifies our "degree of surprise" in observing a_i, where less plausible outcomes yield more information, due to $-\log_m p(a_i) \to \infty$ as $p(a_i) \to 0$.

Based on the definition Eq. (3.15), we can define the expected information content conveyed by a stochastic variable a with $p(a)$ as the *entropy* $H(a)$:

$$H(a) = -\sum_a p(a) \log_m p(a). \tag{3.16}$$

By assuming that each value of a is a symbol emitted via a channel with an encoder $p(a)$, the entropy $H(a)$ expresses the average number of transmitted information units. It also expresses the "uneven-ness" of the distribution $p(a)$, so that uniform distributions yield the richest information content (thus requiring more, say, bits to encode) in contrast to distributions where a single outcome is dominant. As an example, consider a plain text file where the domain of a is the ASCII character set and $p(a)$ is the associated empirical histogram from parsing the file. A text file that contains only the letter "A", thus $p(a = \text{"A"}) = 1$ and

0 otherwise, yields $H(a) = 0$ entropy[1]. IN contrast, a text file containing a chapter from George R.R. Martin's "Song of Ice and Fire" will contain a very diverse character histogram, thus yielding a very high $H(a)$. The entropy in both cases defines not only the information content as a statistical quantity, but also the level of *compression* that can be achieved without corrupting the original content.

Consider two discrete probability distributions $p(.)$ and $q(.)$ over the same stochastic variable a. We define the *cross entropy* $H_{p,q}(a)$ or $H_{q,p}(a)$ as:

$$H_{p,q}(a) = -\sum_a p(a) \log_m q(a), \text{ or} \tag{3.17}$$

$$H_{q,p}(a) = -\sum_a q(a) \log_m p(a), \tag{3.18}$$

where from Gibb's inequality, we have $H_p(a) \geq H_{p,q}(a)$ and $H_q(a) \geq H_{q,p}(a)$, due to the error associated with using a different encoder $q(.)$ for transmitting the information $\log_m p(.)$. We can use the difference $H_{p,q}(a) - H_p(a)$ as a dissimilarity measure between distributions in the same domain of a, which is the Kullback-Leibler (KL) divergence $\text{KL}(p||q)$:

$$\begin{aligned} \text{KL}(p||q) &= H_{p,q}(a) - H_p(a) = \\ &= -\sum_a p(a) \log_m q(a) - \left(-\sum_a p(a) \log_m p(a) \right) = \\ &= -\sum_a p(a) \log_m \frac{q(a)}{p(a)}. \end{aligned} \tag{3.19}$$

Note that although the similar equation holds for $\text{KL}(q||p)$, the KL-divergence is not symmetrical $\text{KL}(p||q) \neq \text{KL}(q||p)$.

Finally, we can use the expressions above to quantify the *degree of coupling* between two

[1] Given that $0 \log_m 0 = 0$.

stochastic variables a and b, as the KL-divergence between the joint $p(a, b)$ and the product of their marginals $p(a)p(b)$. Such a quantity, called the *mutual information* $I(a, b)$ is defined as:

$$\begin{aligned} I(a,b) &= \mathrm{KL}(p(a,b)||p(a)p(b)) = \\ &= -\sum_{a,b} p(a,b) \log_m \frac{p(a)p(b)}{p(a,b)}. \end{aligned} \tag{3.20}$$

Note that all of the equations presented in this section can be appropriately extended to the continuous case:

$$H(a) = -\int_a p(a) \log_m p(a) \mathrm{d}a, \tag{3.21}$$

$$H_{q,p}(a) = -\int_a q(a) \log_m p(a) \mathrm{d}a, \tag{3.22}$$

$$\mathrm{KL}(p||q) = -\int_a p(a) \log_m \frac{q(a)}{p(a)} \mathrm{d}a, \tag{3.23}$$

$$I(a,b) = -\int_{a,b} p(a,b) \log_m \frac{p(a)p(b)}{p(a,b)} \mathrm{d}a \; \mathrm{d}b. \tag{3.24}$$

3.6 Closing remarks

In this chapter we summarise the key elements of Probability Theory that we use extensively in the thesis, particularly the methodological contributions presented in Chapters 6, 4 and 7. Our mode of work is generally based on i) formulating our models and hypotheses by laying out the system variables and their interdependencies as presented in Section 3.3 ii) manipulating the model structure using the tools from Section 3.4 in order to express the posterior distributions over our inference targets and iii) utilising, when required, the tools of Information Theory from Section 3.5 in the formulation of objective functions, for efficient inference.

Chapter 4

Overlapping Community Detection via Bayesian Nonnegative Matrix Factorisation

4.1 Introduction

Community structure is a significant property of real-world networks as it is often considered to account for the functional characteristics of the system under study [Fortunato, 2010; Newman, 2010; Porter et al., 2009; Reichardt and Bornholdt, 2006]. Although the notion of "community" appears intuitive [Fortunato, 2010; Porter et al., 2009] (for example people form cliques in social networks and web pages of similar content have links to one another) there is no disciplined, context-independent definition of what is a community [Fortunato, 2010; Reichardt and Bornholdt, 2006]. Instead, we tend to adopt the loose definition that communities are sets of nodes with dense connections internally and sparser connections between groups [Fortunato, 2010; Newman and Girvan, 2004; Porter et al., 2009]. The task of identifying such subgraphs in a given network can be challenging, both from an inferential and a computational complexity perspective [Fortunato, 2010; Newman, 2010].

One of the key issues in community detection is describing the overlapping nature of network modules. Traditional "hard-partitioning" algorithms [Blondel et al., 2008; Duch and Arenas, 2005; Reichardt and Bornholdt, 2004; Rosvall and Bergstrom, 2007] may yield excellent identification results, but omit the important characteristic of real-world networks where a node may participate in more than one group (for example, individuals belong to various social circles and scientists may participate in more than one research group). A popular approaches to tackling this problem, such as Clique Percolation Method (CPM) [Palla et al., 2005] and the BNK algorithm [Ball et al., 2011], may identify overlapping modules but have their own shortcomings, either in terms of computational complexity or model order selection as we discussed in Chapter 2. For further reading, more comprehensive reviews of community detection algorithms are presented in [Fortunato, 2010] and [Porter et al., 2009].

In this work we propose a novel approach to community detection based on Bayesian nonnegative matrix factorisation (NMF) [Tan and Févotte, 2009]. The advantages of this methodology are: i) overlapping or soft-partitioning solutions, where communities are allowed to share members; ii) soft-membership distributions, which quantify how strongly each individual participates in each group; iii) excellent module identification capabilities; and iv) the method does not suffer from the drawbacks of modularity optimisation methods, such as the resolution limit. On the downside, our approach does not possess the excellent computational scalability of methods such as Label Propagation [Raghavan et al., 2007], thus it cannot be directly applied to graphs of millions of nodes without a distributed implementation.

This chapter is organised as follows; in Section 4.2.1 we present the intuition behind our community detection via NMF (CD-NMF) approach, while in Section 4.2.2 we formally present the model, by providing its mathematical foundations and implementation details. In Section 4.3 we run our community detection scheme across a series of tests: an illustrative example graph in Section 4.3.1, a series of artificially generated graphs with observed com-

munity structure in Section 4.3.2, a collection of widely adopted benchmark graphs in Section 4.3.3, random graphs with no community structure in Section 4.3.4 and finally a collection of real-world problems in Section 4.3.5. Application of CD-NMF on a real-world ecological network, which describes the social structure of a wild-bird population, is presented in Chapter 8.

4.2 Model formulation

4.2.1 Background

Let us begin by viewing community structure as an underlying mechanism that drives social tie formation in a network. Consider the University of Oxford, which is a large institution consisting of many divisions, academic departments, colleges and societies. University members participate across such groups to a variable extent and such participation has a strong effect on their social circle; it defines the subset of Oxford members with whom they spend time with, collaborate, do sports, organise events, etc. Our key modelling assumption is that co-participation of two people across a given range of communities, i.e. homophily in group membership profile, gives rise to a higher interaction rate or, in the probabilistic context presented in Chapter 3, increases our belief that those individuals are interacting.

To encode the above ideas in a mathematical framework, consider the C-dimensional vectors $\mathbf{w}_i$ and $\mathbf{w}_j$, where each element w_{ic} denotes the *participation strength* of individual i in a community c. The expected association strength $\hat{a}_{ij}$ between the individuals i and j based on their membership profile is thus given by:

$$\hat{a}_{ij} = \phi(\mathbf{w}_i, \mathbf{w}_j), \tag{4.1}$$

where ϕ is a predefined *association function*. Such functions define how the participation strength w_{ic} of individual i to community c contributes to the overall interaction rate a_{ij}

between i and j. In the present work, we consider ϕ to be a simple dot product between the membership vectors[1]:

$$\hat{a}_{ij} = \mathbf{w}_i^{\mathsf{T}} \mathbf{w}_j, \tag{4.2}$$

while for a pool of N individuals participating in C communities based on $\mathbf{W} \in \mathbb{R}_{(+)}^{N \times C}$, the overall connectivity pattern is given by the adjacency matrix $\hat{\mathbf{A}} \in \mathbb{R}_{(+)}^{N \times N}$:

$$\hat{\mathbf{A}} = \mathbf{W}\mathbf{W}^{\mathsf{T}}, \tag{4.3}$$

which describes an undirected network where all link weights are strictly nonnegative[2].

The above discussion can be generalised to the case of directed graphs, by allowing asymmetric interaction rates between pairs of individual nodes. Let us consider an example network of N web pages connected to each other via hyperlinks, where edge weights $\hat{a}_{ij}$ denote user traffic from page i to j through click rates. We assume that this pool of web sites can be organised into C thematic groups that correspond to overlapping, topic-focused user communities on politics, fashion, sports, etc. Given a topic c (e.g. Japanese poetry), a certain page i can either "serve" c with intensity h_{ci} (e.g. a blog with haiku poems), or "seek" c with intensity w_{ic} (e.g. a haiku forum where users post/share links to original material). Based on those two different community membership perspectives, the expected intensity by which page i directs users to page j is based on their thematic compatibility is:

$$\hat{a}_{ij} = \varphi(\mathbf{w}_i, \mathbf{h}_j), \tag{4.4}$$

where φ is a given *compatibility function* and $\mathbf{w}_i, \mathbf{h}_j$ are C-dimensional vectors, where each element w_{ic}, h_{cj} encodes the "seek/serve" community membership intensities described above. As in the undirected case, we consider φ to be a linear function:

[1] Further association functions ϕ can be considered, as discussed in Section 4.5.

[2] In the present work we have assumed that individuals cannot have negative participation in a community.

$$\hat{a}_{ij} = \mathbf{w}_i^{\mathsf{T}} \mathbf{h}_j, \tag{4.5}$$

from which we express the network adjacency matrix via the following factorisation:

$$\hat{\mathbf{A}} = \mathbf{W}\mathbf{H}, \tag{4.6}$$

where $\hat{\mathbf{A}} \in \mathbb{R}_{(+)}^{N \times N}, \mathbf{W} \in \mathbb{R}_{(+)}^{N \times C}$ and $\mathbf{H} \in \mathbb{R}_{(+)}^{C \times N}$, with the undirected example of Eq. (4.3) being the special case $\mathbf{H} = \mathbf{W}^{\mathsf{T}}$ of the equation described above.

Based on the above discussion, we pose the community detection problem in the following manner: if connection weight among individuals in a network results from the degree of homophily in their community membership profiles, which membership structure is the most plausible for explaining the observed connections? In other words, given an adjacency matrix $\mathbf{A}$, which factorisation $\hat{\mathbf{A}} = \mathbf{W}\mathbf{H}$ approximates $\mathbf{A}$ as well as possible given a particular distance metric?

In the following section, we present a probabilistic model in which community membership is an explanatory latent variable for the observed link weights and propose an inference scheme for discovering the appropriate values w_{ic}, h_{ci} for the factors $\mathbf{W}, \mathbf{H}$ of Eq. (4.6).

4.2.2 Probabilistic Model

Consider the graphical model of Fig. 4.1. The observed variable a_{ij} denotes the nonnegative count of interactions between two individuals i, j in a weighted network with adjacency matrix $\mathbf{A} \in \mathbb{R}_{(+)}^{N \times N}$. Based on the discussion in Section 4.2.1, there are C "hidden" classes of nodes in the network, which directly affect the interaction rates a_{ij} in the following sense: the more two individuals interact, the more likely they are to belong to the same communities.

We assume that the observed adjacency matrix $\mathbf{A} \in \mathbb{R}_{(+)}^{N \times N}$ is a noise-corrupted instance of a latent matrix $\hat{\mathbf{A}} \in \mathbb{R}_{(+)}^{N \times N}$, in which the elements $\hat{a}_{ij}$ denote the expected number of in-

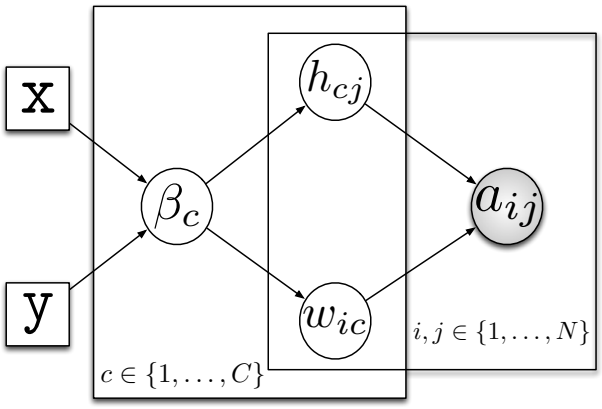


Figure 4.1: Graphical model showing the generation of count processes $\mathbf{A}$ from the latent structure $\mathbf{W}$ and $\mathbf{H}$, the components of which have scale hyper-parameters β_c. The hyper-hyper-parameters a, b are fixed.

teractions between nodes i and j under a Poisson model $a_{ij} \sim \text{Pois}(\hat{a}_{ij})$. Such a noise model is based on the fact that the Poisson is the maximum entropy (most "non-committal") distribution [Bishop, 2007; Jaynes, 2003] for random variables that represent counts (in our case, number of interactions). This modelling choice has also been employed in previous NMF models, with the ones from [Tan and Févotte, 2009] and [Cemgil, 2009] being of particular interest. Following Eq. (4.6) and based on the discussion in Section 4.2.1, we consider the expectation matrix $\hat{\mathbf{A}} \in \mathbb{R}_{(+)}^{N \times N}$ to be composed of two nonnegative matrices $\mathbf{W} \in \mathbb{R}_{(+)}^{N \times C}$ and $\mathbf{H} \in \mathbb{R}_{(+)}^{C \times N}$ so that $\hat{\mathbf{A}} = \mathbf{WH}$. From this decomposition and the model described above, each link weight a_{ij} is a draw from a Poisson distribution with rates:

$$\hat{a}_{ij} = \sum_{c=1}^{C} w_{ic} h_{cj}, \tag{4.7}$$

and the overall likelihood of the observed interactions $\mathbf{A}$ under the Poisson model is given from:

$$
\begin{aligned}
p(\mathbf{A}|\hat{\mathbf{A}}) &= \prod_{i=1}^{N}\prod_{j=1}^{N} \text{Pois}(a_{ij};\hat{a}_{ij}) \\
&= \prod_{i=1}^{N}\prod_{j=1}^{N} \frac{\hat{a}_{ij}^{a_{ij}} e^{-\hat{a}_{ij}}}{a_{ij}!} \\
&= \prod_{i=1}^{N}\prod_{j=1}^{N} \frac{\left(\sum_{c=1}^{C} w_{ic}h_{cj}\right)^{a_{ij}} e^{-\sum_{c=1}^{C} w_{ic}h_{cj}}}{a_{ij}}.
\end{aligned} \tag{4.8}
$$

The inner rank C expresses the unknown number of communities and each element w_{ic}, h_{cj} in row i of $\mathbf{W}$ and column j of $\mathbf{H}$ represents the contribution of a single latent community c to $\hat{a}_{ij}$. In other words, the expected number of times $\hat{a}_{ij}$ two individuals i, j interact is a result of their *mutual participation* in the same communities. Based on our community extraction objective and the probabilistic model of Fig. 4.1, our task is to infer the appropriate values of the latent variables $w_{ic}, h_{cj}, \forall i, j \in \{1, ..., N\}$ and $c \in \{1, ..., C\}$, given our available observations $\mathbf{A} \in \mathbb{R}_{(+)}^{N\times N}$.

Finding the appropriate values of w_{ic}, h_{cj} by direct maximisation of Eq. (4.8) is infeasible, because in the typical community-detection setting the value of C, which corresponds to *complexity* or *model order*, is initially unknown. In previous work [Wang et al., 2008; Zhang et al., 2007], the issue of inferring the appropriate number of communities has been addressed by performing multiple optimisation schemes across various C and selecting one that yields the highest Newman-Girvan modularity Q [Newman and Girvan, 2004]. In our setting, we assume an upper bound to the number of communities C and employ a probabilistic approach by considering w_{ic}, h_{cj} as random variables, governed by a Half-Normal distributions with zero mean and a given precision (inverse variance) β:

$$p(w_{ic}|\beta_c) = \mathcal{HN}(w_{ic}; 0, \beta_c^{-1}) = \sqrt{\frac{2\beta_c}{\pi}} \exp\left\{-\frac{1}{2}\beta_c w_{ic}^2\right\}, \tag{4.9}$$

$$p(h_{cj}|\beta_c) = \mathcal{HN}(h_{cj}; 0, \beta_c^{-1}) = \sqrt{\frac{2\beta_c}{\pi}} \exp\left\{-\frac{1}{2}\beta_c h_{cj}^2\right\}. \tag{4.10}$$

Such a zero-mean prior on the membership scores $w_{ic}, h_{cj} \in \mathbb{R}_{(+)}$, allows us to express our belief that individuals have a sparse participation profile across communities, so that the overall network itself is globally sparse but locally dense (due to the fact that interactions result from co-participation in communities via Eq. (4.7)).

From the perspective of our probabilistic model, the elements w_{ic}, h_{cj} of the c-th column of $\mathbf{W}$ and c-th row of $\mathbf{H}$, which refer to the same community c, are governed by a common precision term β_c. The effect of this *shrinkage* or *automatic relevance determination* [MacKay, 1995] prior structure is to "switch off" to zero irrelevant communities c, which do not contribute into explaining the observed interactions $\mathbf{A}$. This approach, originally presented in [Tan and Févotte, 2009] and which we explain in more detail in the following section, avoids the computational load of multiple runs and is free of the resolution bias problems [Fortunato and Barthélémy, 2007] of modularity.

Based on the above, our inference scheme for the probabilistic model of Fig. 4.1 consists of finding the appropriate community membership scores $w_{ic}, h_{cj} \forall i, j \in \{1, ..., N\}$, so that:

1. each $\hat{a}_{ij} = \sum_{c=1}^{C} w_{ic} h_{cj}$ is as close as possible to the observed link weight a_{ij}, given a Poisson model of Eq. (4.8).

2. assuming an initialisation of $C = N$, "irrelevant" communities c_0 yield zero membership scores $w_{ic_0}, h_{c_0 j} = 0 \ \forall i, j \in \{1, ..., N\}$ and can be removed, thus reducing C to the effective number of communities $C_\star$.

Both steps of the above process are presented in the next section, via the formulation of an appropriate objective function.

4.2.3 Posterior-based cost function

Based on the factorisation implied by the graphical model of Fig. 4.1, we can express the joint distribution over all variables as:

$$p(\mathbf{A}, \mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) = p(\mathbf{A}|\mathbf{W}, \mathbf{H})p(\mathbf{W}|\boldsymbol{\beta})p(\mathbf{H}|\boldsymbol{\beta})p(\boldsymbol{\beta}), \tag{4.11}$$

where a_{ij} is the observed variable, w_{ic}, h_{cj} are our latent variables and β_c are the model parameters. The hyper-parameters $\mathtt{x}, \mathtt{y}$ are fixed deterministic variables, therefore not expressed in the above equation. Using Bayes' rule on Eq. (4.11), the posterior over our latent space $\mathcal{X} = \{\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}\}$ is:

$$\begin{aligned} p(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}|\mathbf{A}) &= \frac{p(\mathbf{A}|\mathbf{W}, \mathbf{H})p(\mathbf{W}|\boldsymbol{\beta})p(\mathbf{H}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{A})}, & (4.12)\\ &\propto p(\mathbf{A}|\mathbf{W}, \mathbf{H})p(\mathbf{W}|\boldsymbol{\beta})p(\mathbf{H}|\boldsymbol{\beta})p(\boldsymbol{\beta}). & (4.13) \end{aligned}$$

Our aim is to maximise the model posterior given the observations, or equivalently minimise the negative log posterior, which can be regarded as an energy (or error) function $\mathcal{U}$. Noting that $p(\mathbf{A})$ is constant with respect to the inference over $\mathcal{X} = \{\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}\}$, the maximum of Eq. (4.12) with respect to $\mathcal{X}$ is the same as the one for Eq. (4.13). Thus by taking the negative logarithm of Eq. (4.13) we define the following minimisation objective:

$$\mathcal{U} = -\log p(\mathbf{A}|\mathbf{W}, \mathbf{H}) - \log p(\mathbf{W}|\boldsymbol{\beta}) - \log p(\mathbf{H}|\boldsymbol{\beta}) - \log p(\boldsymbol{\beta}). \tag{4.14}$$

The first term of Eq. (4.14) is the log-likelihood of our data, derived from the probability $p(\mathbf{A}|\mathbf{W}, \mathbf{H}) = p(\mathbf{A}|\hat{\mathbf{A}})$ of observing every interaction a_{ij} given a Poisson rate $\hat{a}_{ij}$. Therefore we express the negative log-likelihood of a single observation a as:

$$-\log p(a|\hat{a}) = -a\log\hat{a} + \hat{a} + \log a!. \tag{4.15}$$

Using the Stirling approximation to second order:

$$\log a! \approx a\log a - a + \frac{1}{2}\log(2\pi a), \tag{4.16}$$

we can write Eq. (4.15) as:

$$-\log p(a|\hat{a}) \approx a\log\left(\frac{a}{\hat{a}}\right) + \hat{a} - a + \frac{1}{2}\log(2\pi a). \tag{4.17}$$

Based on the above, the negative log-likelihood for every link a_{ij} can be expressed as:

$$\begin{aligned} -\log p(\mathbf{A}|\hat{\mathbf{A}}) &= -\sum_{i=1}^{N}\sum_{j=1}^{N}\log p(a_{ij}|\hat{a}_{ij}) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(a_{ij}\log\frac{a_{ij}}{\hat{a}_{ij}} + \hat{a}_{ij} - a_{ij} + \frac{1}{2}\log(2\pi a_{ij})\right) + \text{const.} \end{aligned} \tag{4.18}$$

Following [Tan and Févotte, 2009], we place independent half-normal priors over the columns of $\mathbf{W}$ and rows of $\mathbf{H}$ with precision (inverse variance) parameters $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_C]^{\mathsf{T}} \in \mathbb{R}^C$. The negative log priors over $\mathbf{W}$ and $\mathbf{H}$ are then given by:

$$\begin{aligned} -\log p(\mathbf{W}|\boldsymbol{\beta}) &= -\sum_{i=1}^{N}\sum_{c=1}^{C}\log\mathcal{HN}(w_{ic}; 0, \beta_c^{-1}) \\ &= \sum_{i=1}^{N}\sum_{c=1}^{K}\left(\frac{1}{2}\beta_c w_{ic}^2\right) - \frac{N}{2}\log\beta_c + \text{const}, \end{aligned} \tag{4.19}$$

$$\begin{aligned} -\log p(\mathbf{H}|\boldsymbol{\beta}) &= -\sum_{c=1}^{K}\sum_{j=1}^{N}\log\mathcal{HN}(h_{cj}; 0, \beta_c^{-1}) \\ &= \sum_{c=1}^{K}\sum_{j=1}^{N}\left(\frac{1}{2}\beta_c h_{cj}^2\right) - \frac{N}{2}\log\beta_c + \text{const}. \end{aligned} \tag{4.20}$$

The net effect of $\boldsymbol{\beta}$ on the latent variables $\mathbf{W}$ and $\mathbf{H}$ can be considered as follows; given our error function implied by Eq. (4.14), our goal is to descend its surface to a point of minimum. By taking the negative derivative with respect to a single element of $\mathbf{W}$ or $\mathbf{H}$, we have:

$$-\frac{\partial\left(-\log p(\mathbf{W}|\boldsymbol{\beta})\right)}{\partial w_{ic}} = -\beta_c w_{ic}, \tag{4.21}$$

which expresses that incremental changes in w_{ic}, as we iterate towards a solution, are proportional to the negative gradient of the energy function. Thus the effect of the prior is to promote a shrinkage to zero of w_{ic}, with a rate proportional to β_c. A large β_c represents a belief that the half-normal distribution over w_{ic} has small variance, and hence w_{ic} is expected to lie close to zero. In Eq. (4.14), the priors and the likelihood function (which quantifies how well we explain the data) are combined, with the net effect that columns of $\mathbf{W}$ (and rows of $\mathbf{H}$) which have little effect in explaining the observed data, are shrinked to zero. This generic approach is known in the statistics literature as shrinkage or ridge regression [Bernardo and Smith, 1994] and in the machine-learning community as automatic relevance determination [Bishop, 2007].

Based on the above, we now have to define a prior structure for the parameters $\boldsymbol{\beta}$. By assuming all β_c are independent[3], we follow the approach of [Tan and Févotte, 2009] and [Bernardo and Smith, 1994] by placing a Gamma distribution over them with fixed hyper-parameters $\mathtt{x}, \mathtt{y}$. This choice of prior distribution satisfies the non-negativity constraint of β_c and allows efficient model inference, as we show in the next section. The negative log hyper-priors are thus:

[3]This corresponds to the belief that the existence of one community is not dependent upon others. Clearly, there will be situations in which this can be extended to allow a full inter-dependency between communities. We do not consider this here, however. Allowing dependency is similar to the notion of *structure priors* discussed in [Penny and Roberts, 2002].

$$
\begin{aligned}
-\log p(\boldsymbol{\beta}) &= -\sum_{c=1}^{K} \log \mathcal{G}(\beta_c; \mathtt{x}, \mathtt{y}) \\
&= \sum_{c=1}^{K} (\beta_c \mathtt{y} - (\mathtt{x}-1)\log \beta_c) + \text{const.}
\end{aligned} \tag{4.22}
$$

The objective function $\mathcal{U}$ of Eq. (4.14) can be expressed as the sum of Eq. (4.18), (4.19), (4.20) and Eq. (4.22):

$$
\begin{aligned}
\mathcal{U} &= \sum_i \sum_j \left[a_{ij} \log \left(\frac{a_{ij}}{\sum_{c=1}^{C} w_{ic} h_{cj}} \right) + \sum_{c=1}^{C} w_{ic} h_{cj} \right] + \\
&+ \frac{1}{2} \sum_k \left[\left(\sum_i \beta_c w_{ic}^2 \right) + \left(\sum_j \beta_c h_{cj}^2 \right) - 2N \log \beta_c \right] \\
&+ \sum_c (\beta_c \mathtt{y} - (\mathtt{x}-1)\log \beta_c) + \text{const},
\end{aligned} \tag{4.23}
$$

thus defining our minimisation target, which consists of two competitive terms:

- the log-likelihood (first line in the above equation) that seeks to achieve maximum descriptive power of the observed data, by minimising the divergence between $\mathbf{A}$ and $\hat{\mathbf{A}}$.

- the log-priors that act as regularisation terms, by forcing the redundant community membership scores w_{ic}, h_{cj} (which do not contribute to explaining the observed interactions) to zero. Columns of $\mathbf{W}$ and rows $\mathbf{H}$ with zero elements can be removed, thus leading to the effective number of communities $C_\star$.

In the next section we propose an appropriate methodology for efficient minimisation of the optimisation target in Eq. (4.23).

4.2.4 Parameter inference

To optimise Eq. (4.23) with respect to $\mathbf{W}, \mathbf{A}$ and $\boldsymbol{\beta}$ we follow [Berry et al., 2007; Lee and Seung, 1999, 2000; Tan and Févotte, 2009] by adopting the fast fixed-point algorithm presented in [Tan and Févotte, 2009] that involves consecutive updates of $\mathbf{W}, \mathbf{H}$, and $\boldsymbol{\beta}$ until a convergence measure has been satisfied (a maximum number of iterations, or a tolerance `eps` on the cost function $\mathcal{U}$). The pseudocode is presented in Algorithm 1, while implementation details and complexity are discussed in Section 4.2.5. The solution consists of $\mathbf{W}_\star \in \mathbb{R}_{(+)}^{N \times C_\star}$ and $\mathbf{H}_\star \in \mathbb{R}_{(+)}^{C_\star \times N}$ for which $\hat{\mathbf{A}} = \mathbf{W}_\star \mathbf{H}_\star$ represents the expectation network given our observation data $\mathbf{A}$ and latent variable structure. The inner rank $C_\star$ denotes the effective number of latent modules in the network.

Algorithm 1 Community Detection using NMF

Require: Adjacency matrix $\mathbf{A} \in \mathbb{R}_{+}^{N \times N}$, initial C_0, fixed Gamma hyper-parameters a, b.
Ensure: Matrix operation $\frac{\mathbf{X}}{\mathbf{Y}}$ as *element-by-element* division.
Ensure: Matrix operation $\mathbf{X} \cdot \mathbf{Y}$ as *element-by-element* multiplication.
Ensure: $\mathbf{B} \in \mathbb{R}_{(+)}^{C \times C}$ as a matrix with elements β_c in the diagonal and zero elsewhere.

Auxiliary inputs $\mathbf{W}_0, \mathbf{H}_0$ from previous runs. If not present, initialise to random values.
while $\Delta\, \mathcal{U} >$ `eps` **do**
  $\mathbf{H} := \left(\frac{\mathbf{H}}{\mathbf{W}^\intercal \mathbf{1} + \mathbf{B}\mathbf{H}} \right) \cdot \left[\mathbf{W}^\intercal \left(\frac{\mathbf{A}}{\mathbf{W}\mathbf{H}} \right) \right]$
  $\mathbf{W} := \left(\frac{\mathbf{W}}{\mathbf{1}\mathbf{H}^\intercal + \mathbf{W}\mathbf{B}} \right) \cdot \left[\left(\frac{\mathbf{A}}{\mathbf{W}\mathbf{H}} \right) \mathbf{H}^\intercal \right]$
  $\beta_c := \frac{N + \mathtt{x} - 1}{\frac{1}{2}\left(\sum_i w_{ic}^2 + \sum_j h_{cj}^2 \right) + \mathtt{y}}$
end while
$C_\star :=$ # of non-zero columns of $\mathbf{W}$ or rows of $\mathbf{H}$
$\mathbf{W}_\star := \mathbf{W}$ with zero columns removed
$\mathbf{H}_\star := \mathbf{H}$ with zero rows removed
return $\mathbf{W}_\star \in \mathbb{R}_{(+)}^{N \times C_*}, \mathbf{H}_\star \in \mathbb{R}_{(+)}^{C_* \times N}$

In the case of undirected graphs, $\mathbf{W}_\star = \mathbf{H}_\star^\intercal$ (as $\mathbf{A}$ is symmetric) and represents the $N \times C_\star$ incidence matrix of a bipartite graph of N nodes and $C_\star$ communities. Each element $w_{ic}^\star$ (or $h_{ki}^\star$) expresses the degree of participation of individual i in community c while each normalised row of $\mathbf{W}_\star$ (or column of $\mathbf{H}_\star$) expresses a soft-membership distribution over communities given a certain node. Therefore this bipartite graph describes the overlap-

ping community structure of our network, where nodes are allocated to multiple groups with varying participation scores.

The interaction matrix $\mathbf{A}$ is approximated by a sum $\hat{\mathbf{A}} = \sum_c \mathbf{w}_c^\star \mathbf{h}_c^\star$, where $\mathbf{w}_c^\star$ is the c-th column and $\mathbf{h}_c^\star$ is the c-th row vector of the community matrices $\mathbf{W}_\star$ and $\mathbf{H}_\star$ respectively. Therefore, $\hat{\mathbf{A}}$ is a summation of $C_\star$ rank-1 matrices $\hat{\mathbf{A}}^{(c)} = \mathbf{w}_c^\star \mathbf{h}_c^\star$ and each $\hat{\mathbf{A}}^{(c)}$ expresses the expected number of pairwise interactions in the context of community c. Thus if two nodes i, j have non-zero participation rates $w_{ic}^\star, h_{cj}^\star$ to community c, then the expected link weight for this dyad would also be non-zero, due to $\hat{\mathbf{A}}_{ij}^{(c)} = w_{ic}^\star h_{cj}^\star$.

Based on the above, each inferred community c corresponds to a fully-connected subgraph in the expectation network described by $\hat{\mathbf{A}}$. This implies that even if we have not observed a direct link between members of the same community, the fact that they are positioned in the same node neighbourhood increases the probability (our "posterior belief" in a Bayesian context) that either i) they are indeed connected, but we have missed the link due to incomplete data or ii) they are not directly linked yet, but there exists a pressure or bias from their common social circle to be connected (in a similar fashion to transitivity [Granovetter, 1973] and graph densification [Leskovec et al., 2005]).

4.2.5 Implementation details and complexity

As discussed in Section 4.2.4, parameter inference is performed by a series of update equations for the latent variables in the model. The computational load is governed chiefly by the matrix multiplication $\mathbf{WH}$, appearing in the denominator of the element-by-element division in steps 3 and 4 of Algorithm 1, which is of order $\mathcal{O}(N^2C)$. In practice, such a cost can be significantly reduced if we exploit the fact that real-world networks (and their corresponding adjacency matrices) are often sparse [Clauset et al., 2004]: the dot products $\sum_c w_{ic}h_{cj}$ within $\mathbf{WH}$ need not be calculated when $a_{ij} = 0$, thus reducing significantly the effect of the quadratic term N^2 in our theoretical complexity expression. For the case of undirected net-

works, in which $\mathbf{A} = \mathbf{A}^{\mathsf{T}}$, the dot product operations are halved because $\mathbf{WH}$ is symmetric, and halved again because step 4 of Algorithm 1 is redundant ($\mathbf{W} = \mathbf{H}^{\mathsf{T}}$).

Community detection methods such as CD-NMF, which operate upon the full adjacency matrix $\mathbf{A}$, can be memory inefficient when implemented naively. The quadratic complexity, $\mathcal{O}(N^2)$, can be mitigated by loading into memory only certain columns/rows of $\mathbf{A}$ when needed, as no holistic operations (such as inversion or multiplication) are required by Algorithm 1 for $\mathbf{A}$ or $\hat{\mathbf{A}}$. In addition, all element-by-element division and multiplication operations can be parallelised, as there are no data dependencies among the threads.

In the next section, we present an illustrative example of this community extraction scheme, followed by experimental results from various artificial and real-world networks.

4.2.6 Related work

The CD-NMF method we present in this chapter bears many similarities with the BKN community detection algorithm introduced by [Ball et al., 2011], which we briefly discussed in Section 2.3.3. In both CD-NMF and BKN each observed weight a_{ij} is considered a draw from a Poisson distribution with a latent rate $\hat{a}_{ij}$, which quantifies the degree of similarity between the community membership profiles of nodes i and j. From the perspective of CD-NMF and based on the discussion of Section 4.2.1, such a degree of similarity results from the compatibility function $\hat{a}_{ij} = \sum_c w_{ic} h_{cj}$, with w_{ic}, h_{cj} being the "seek/serve" membership scores of i and j in community c. The BKN approach considers a generative model structure in which each node i has a "propensity" θ_{ic} for edges of "colour" c, where colours correspond to communities. In this setting there is a common type of membership score (no distinction between the "seek/serve" intensities w_{ic}, h_{cj} we discussed in Section 4.2.1) thus the latent Poisson rate is given by $\hat{a}_{ij} = \sum_c \theta_{ic}\theta_{jc}$. Nevertheless, BKN allows nodes to participate in more than one community, with a varying membership score.

For the BKN model, community memberships are inferred by finding the values of θ_{ic}

that maximise the probability of observing each link a_{ij} under a Poisson model with rate $\hat{a}_{ij} = \sum_c \theta_{ic}\theta_{jc}$. A similar maximisation objective has been used for CD-NMF with the addition of certain regularisation terms, as seen in Eq. (4.23), which allow the model to automatically determine the effective number of communities. We consider this as the main advantage of CD-NMF over BKN, as the latter requires a-priori knowledge of the value of C in order to infer the appropriate node membership scores.

The two community detection algorithms also differ in the way upon which they perform parameter inference: the CD-NMF approach uses a coordinate descent method that involves a series of multiplicative update steps (see details in Section 4.2.5), while BKN employs a computationally efficient Expectation-Maximisation scheme. In [Ball et al., 2011] the authors report a computational cost of $\mathcal{O}(MC)$, where M the total number of edges and C the number of communities, making the algorithm much more appropriate for large graphs.

In summary, both CD-NMF and BKN employ a generative model, where overlapping community structure is viewed as a collection of latent node labelings that account for the observed number of pairwise interactions. Inference on such labelings is performed via finding the most likely community configuration that can generate the observed weights a_{ij}, under a Poisson noise model. The key difference between the methods lies on the fact that CD-NMF employs an additional set of model parameters, which allow the automatic determination of the effective number of communities. Although BKN requires a-priori knowledge of the community number C, its key advantage lies on the more attractive computational cost, making the method appropriate for large-scale problems.

4.3 Applications

4.3.1 An illustrative example

Consider the small toy graph of Fig. 6.10 with $N = 16$ nodes and $M = 25$ edges of varying weights. We use such a network for illustrating the behaviour of our method, as it contains a variety of different cases of communities: a densely-connected clique (nodes 1 to 5), a "ring" community (nodes 9 to 12), a clique that lies in between two larger communities (nodes 6 to 8), a loosely connected and relatively isolated one (nodes 14–16), along with a singular node (13) with no strong presence towards any groups.

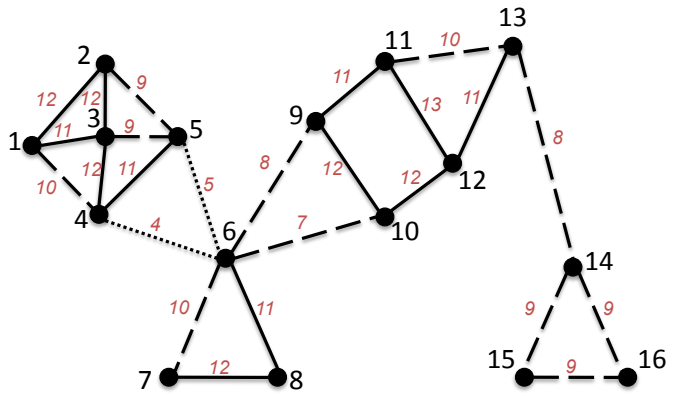


Figure 4.2: An undirected weighted toy graph with 16 nodes. The three different line styles express the differing strengths of interaction between each pair of nodes. Solid line expresses strong connection weight, while dashed lines correspond to weaker ties. The actual edge weight is shown in italicised red font.

We extract the community structure of this network using CD-NMF, along with:

- Extremal Optimisation (EO) [Duch and Arenas, 2005],
- Spectral Partitioning (SP)[4] [Newman, 2006],
- Weighted Clique Percolation Method (wCPM) [Farkas et al., 2007],

[4]The leading eigenvector at each step is calculated via the power iteration method [Lanczos, 1950] and the final partition was fine-tuned via the re-arrangement scheme proposed in [Newman, 2006].

and illustrate the results from the latter three methods in Fig. 4.3.

Though a small example graph, each community detection method we applied yielded different modules and node allocations. Hard-partitioning methods such as EO and SP produce such inconsistencies mainly due to the "broker" nature of nodes such as $6, 9$ or 10 that lie on high-flow paths in the network, making them difficult to assign on one module or the other [Fortunato, 2010]. Although this issue is addressed by wCPM, which allows node membership to multiple modules, it does not provide some measure of "participation strength" or "degree of belief" in membership.

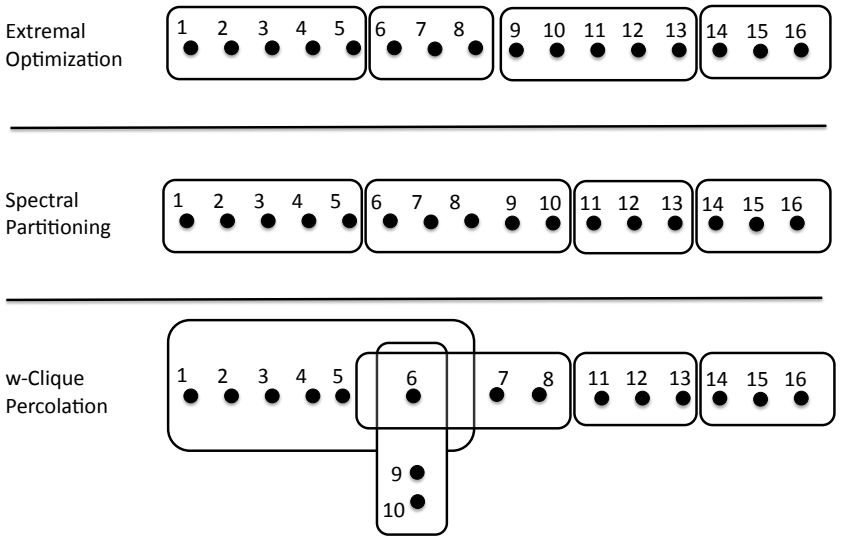


Figure 4.3: Node allocations to communities for three different community detection methodologies.

By applying CD-NMF we extracted $C_\star = 4$ overlapping groups as shown in Fig. 4.4. We can see that our method does not force node allocations to a single group, but instead allows the "broker" individuals described above to participate in more than one community. This soft-partitioning solution allows us to describe the different aspects of an individual's sociality as a collection of (possibly intersecting) sets of nodes, where each set may play a different role or function in the whole network [Fortunato, 2010].

Allocating nodes to multiple modules, as in Fig. 4.4, is only one part of the solution. We

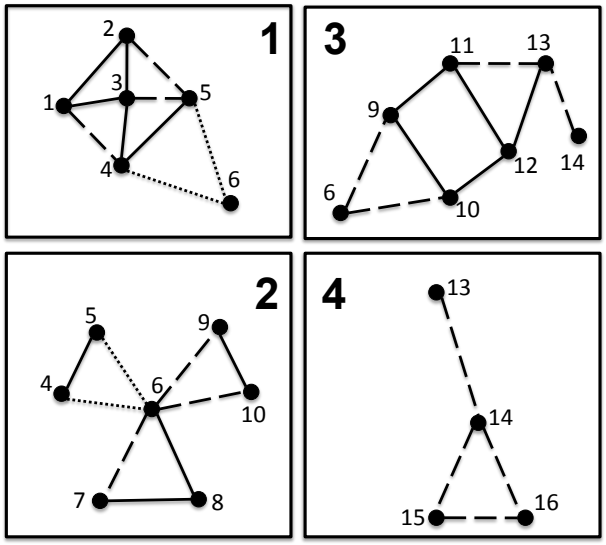


Figure 4.4: The toy graph of Fig. 4.2 decomposed into $C_\star = 4$ overlapping communities using CD-NMF.

also capture the degree of participation of nodes in each community by using the incidence matrix $\mathbf{W}_\star$ described in the previous section. Fig. 4.5(a) shows $\mathbf{W}_\star \in \mathbb{R}_{(+)}^{16\times 4}$ where different colours indicate various levels of participation of nodes in communities. We can see that the matrix is not of a clear block diagonal form, as an individual can have some form of membership in multiple groups.

In our framework, community allocation is not a Boolean decision but a belief; each node is assigned a membership distributed over communities, as seen in Fig. 4.5(b). We can see that mediator nodes of high "betweenness", such as $i = 6$, have a more entropic distribution (similar to the concept of "bridgeness" [Nepusz et al., 2008]) while for nodes such as $i = 4$ or $i = 14$ we have much more confident allocations.

Having soft-membership distributions not only allows us to describe our confidence in assigning node i to community c, but also to quantify the degree of "fuzziness" in the network. In Fig. 4.5(b), nodes such as $i = 6$ that lie on community boundaries have a membership distribution that is closer to uniform. Based on such a property, we can use the average entropy of node membership distributions as a "modularity-like" measure, in order to quantify

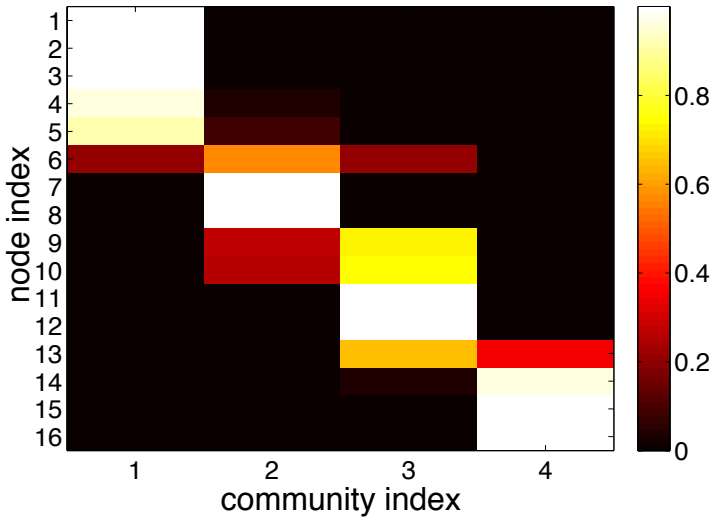
the "fuzziness of community structure in the graph.

From a classic nonnegative matrix optimisation perspective, as originally applied to image decomposition [Lee and Seung, 1999], communities can be viewed as *basis structures*, captured in each column and row of the factors $\mathbf{W}$ and $\mathbf{H}$. We consider each basis structure as a community c, which has a "total binding energy" that is allocated to the atoms (nodes) based on $\mathbf{w}_c^\star$ or $\mathbf{h}_c^\star$. For example in Fig. 4.6, we take the second column of $\mathbf{W}_\star$ and draw a color map (left frame) based on the intensity of $w_{i2}, i \in \{1, ..., N\}$. Elements of $\mathbf{W}_\star$ with non-zero energy, correspond to nodes that participate in such basis structure and form a subset of the whole network (right frame). We can see that this basis community in Fig. 4.6 is dominated by nodes 6, 7 and 8 that contribute most of the binding energy, while the peripheral nodes 4, 5, 9, 10 have some minor participation.

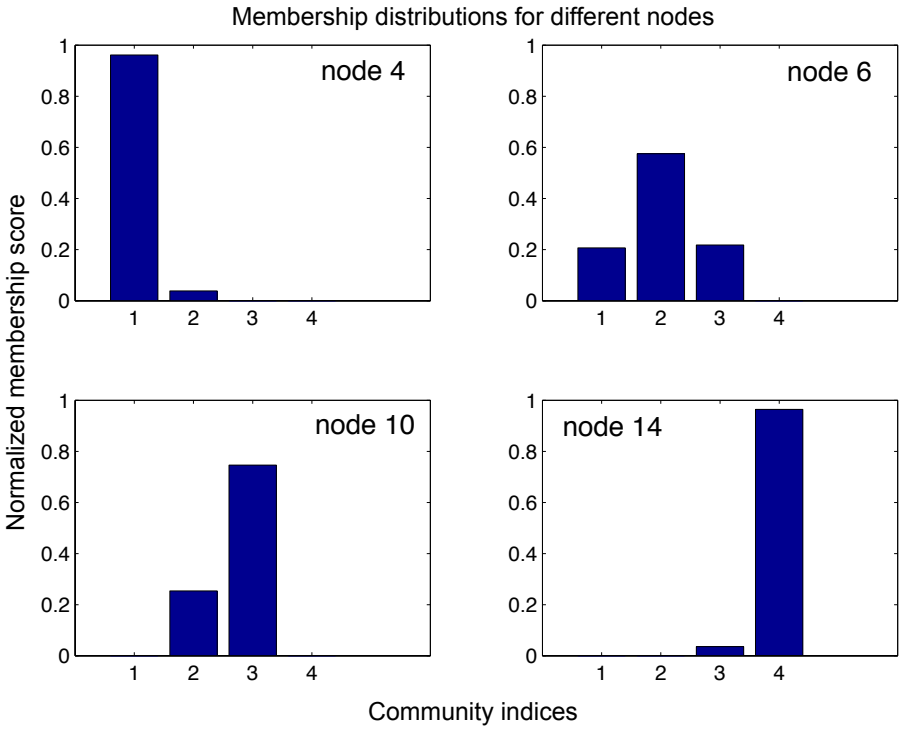
We can view real-world social networks using the framework we described above; groups of individuals are structures bound together with a given energy, by spending time together, communicating, cooperating, etc. Every individual contributes to a range of communities a certain amount of energy, which can be seen as his/her degree of membership. High-energy members can be regarded as *focal individuals* in a group, while social structures with members of uniform contribution can be regarded as *teams* that are held together because of equal participation of their members. Finally, under this framework we can identify highly social individuals, who belong to many groups with high amount of participation.

4.3.2 Tests on artificial graphs with observed community structure

In this section we test the module identification capabilities of NMF, against realisations of the Girvan-Newman (GN) random graph [Girvan and Newman, 2002]. Such a graph consists of $N = 128$ nodes and has a fully-observed community structure of of $C = 4$ modules (with $n = 32$ nodes each), average degree of $\langle d \rangle = 16$ and a variable inter-community degree $\langle d_{out} \rangle$ that controls the module cohesiveness of the network.



(a) Color map of the incidence matrix $\mathbf{W}_\star \in \mathbb{R}^{16\times 4}_{(+)}$.



(b) Soft membership distributions for various nodes in our toy network.

Figure 4.5: Fig. 4.5(a) shows the node allocations proposed by our algorithm. Colours close to white indicate strong participation of node i (vertical axis) to community c (horizontal axis). Fig. 4.5(b) shows example (normalised) rows of $\mathbf{W}_\star$ that correspond to the membership distribution of different nodes.

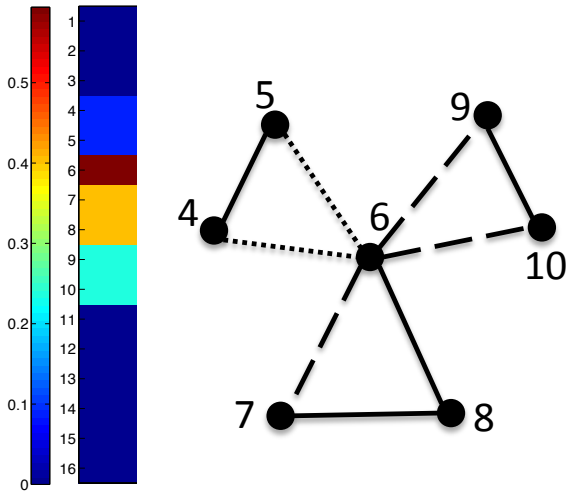
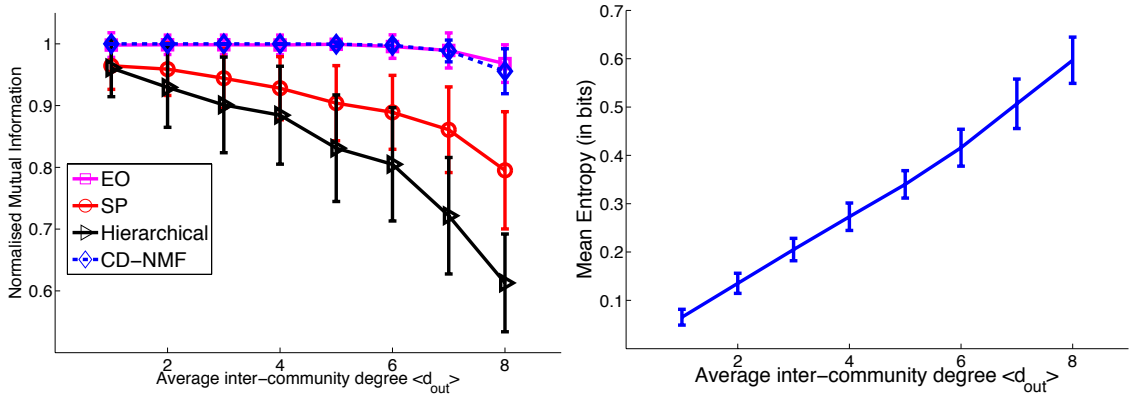


Figure 4.6: One of the CD-NMF basis structures, as extracted from our toy graph. Each atom has a different degree of participation, as it can be seen from the color map on the left. Node 6 is a focal individual, contributing the most energy to the structure along with nodes 7 and 8 while nodes 4, 5, 9 and 10 are peripheral. As in Fig. 4.2, solid line represents strong connection weight, while dashed lines correspond to weaker ties.



(a) Normalised Mutual Information, value range 0–1. (b) Mean entropy of membership distribution.

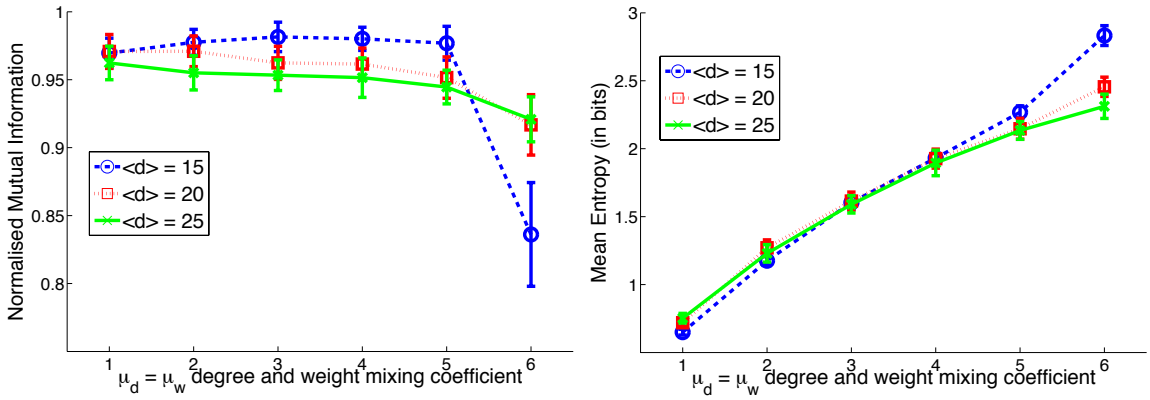
Figure 4.7: Fig. 4.7(a) compares our NMF (dashed ⋄-line at the top) approach against Extremal Optimisation (EO) (pale □-line at the top), Spectral Partitioning (SP) (○-line) and Hierarchical Clustering (Hierarchical) (▷-line) in identifying the communities of Girvan-Newman artificial graphs. Each point is the mean of 100 graph realisations. Fig. 4.7(b) shows the increase in uncertainty in assigning nodes to communities, using NMF, as we increase the fuzziness of modular organisation in GN graphs. Each point is the mean of 100 graph realisations.

In Fig. 4.7(a) we plot our module identification performance based on the Normalised Mutual Information (NMI) criterion [Danon et al., 2005], a real number between 0

and 1 which is maximal when the detected communities exactly meet expectations. In 4.7(b) we monitor our allocation confidence based on the mean entropy (in bits) $H = -\sum_{c=1}^{C} w_{ic} \log_2 w_{ic}$ of each node membership distribution. We can see that as we make the network fuzzier by increasing $\langle d_{out} \rangle$, our method "responds" by increasing the degree of node participation to multiple communities. An attractive aspect of this test is that the increase in entropy (see Fig. 4.7(b)) does not affect the module identification performance (we see from Fig. 4.7(a) that NMI remains close to unity) and is stable for the vast majority of $\langle d_{out} \rangle$ values. For comparison, we also provide in Fig. 4.7(a) the NMI performance of the following hard-partitioning methods: Extremal Optimisation [Duch and Arenas, 2005], Spectral Partitioning [Newman, 2006], and Hierarchical Clustering [Fortunato, 2010]. For hierarchical clustering, Euclidean distance $\sqrt{\sum_{k \neq i,j} (a_{ik} - a_{jk})^2}$ acted as node similarity and complete-linkage clustering acted as group similarity.

We extend the above test to the case of Lancichinetti-Fortunato random graphs (LF) [Lancichinetti and Fortunato, 2009], which reflect more accurately the properties of real-world networks. In this setting, community cohesion is controlled by *mixing parameters* μ_d and μ_w, which express the expected fraction of inter-community degrees and weights per node. Other configuration parameters include the total number of nodes N, the average degree $\langle d \rangle$, the exponent of the degree distribution γ_1, and the exponent of the community-sise distribution γ_2. We tested our method for a (decaying) range of values for μ_d, μ_w (where we set $\mu_d = \mu_w$), in weighted graphs of $N = 1000$ nodes and various values of $\langle d \rangle$, as seen in Fig. 4.8(a). In the same spirit as the GN graph case, in Fig. 4.8(b) we monitor the mean entropy of membership distributions per node (in bits) to quantify the confidence of our node allocations to communities. In Fig. 4.8(a) we can see that our model has an excellent module identification performance and starts to fail only when the mixing coefficients μ have values greater than 0.5. For the same test run, the Louvain method yielded average NMI values of 0.98 ± 0.03 for $\langle d \rangle = 15$, 0.99 ± 0.023 for $\langle d \rangle = 20$ and 0.98 ± 0.023 for $\langle d \rangle = 25$,

for mixing coefficients $1 \leq \mu_d, \mu_w \leq 5$. On the other hand, the increasing fuzziness of the network (based on μ) is captured in the mean entropy of the membership distributions; as the community structure is less cohesive, we are less confident in the allocation of nodes to groups.



(a) Normalised Mutual Information, value range 0–1.

(b) Mean entropy of membership distribution.

Figure 4.8: Results of CD-NMF on realisations of the LF random graphs for $N = 1000$ and different values for the average degree $\langle k \rangle$ and community cohesion μ parameters. Each point represents the mean and standard deviation over 100 graph realisations.

4.3.3 Benchmark data sets

We present the performance of our community detection method on a variety of popular benchmark data sets and compare it against the Extremal Optimisation (EO) [Duch and Arenas, 2005] and Louvain [Blondel et al., 2008] methods. In contrast to the artificial graphs we used above, the absence of an "observed solution" for these problems prevents us from using the Normalised Mutual Information criterion for performance evaluation. Instead we use the popular modularity Q [Newman and Girvan, 2004], discussed in Chapter 2, which measures how "statistically surprising" the intra-community link density is for a proposed network partition. For the purposes of the experiment we remove the overlapping aspect of the CD-NMF solutions, by assigning each node to the community in which it has the maximum membership score. Although this "greedy allocation" scheme omits the wealth of

information provided by our model solutions, it is necessary in order to perform modularity comparisons against hard-partitioning methods. Comparison with Clique Percolation is also absent, as it provides a uniform participation score of nodes to modules, thus no "greedy allocation" can by applied. For each data set, we ran the three methods 100 times, recording the values of modularity Q along with the number of extracted communities $C_\star$. The values are reported in Tables 4.2 and 4.3; because the Louvain method produced consistent solutions across different runs, its standard deviations have been omitted. For CD-NMF initialisation we used $C_0 = N$ with hyper-parameters $\mathtt{y} = 5$ and $\mathtt{x} = 2$, giving a prior close to uniform. It is worth noting that the choice of hyper-parameter values does not affect the resolution of the resulting network partition, but the speed of algorithm convergence. Although the optimal initialisation parameters are problem-dependent, in practice we found that hyper-parameter values with $\mathtt{x} > 1$, for which the Gamma distribution has a unimodal shape, result in faster algorithm convergence.

Table 4.1: Benchmark data sets

Data Set	N	M
Dolphins [Lusseau et al., 2003]	62	159
Books US Politics [Krebs, 2010]	105	441
Les Misérables [Knuth, 1993]	77	254
College Football [Girvan and Newman, 2002]	115	613
Jazz Musicians [Gleiser and Danon, 2003]	198	2742
C. elegans metabolic [Duch and Arenas, 2005]	453	2025
Network Science [Newman and Girvan, 2004]	1589	2742
Facebook Caltech [Traud et al., 2011]	769	16656

From Table 4.2 we can see that our approach performs competitively despite not being designed with the aim of maximising modularity, unlike EO and the Louvain method. Additionally, it has the advantage of providing soft-partitioning solutions and node membership scores to each community. Finally, although our method favours sparse solutions, it does not suffer from the resolution limit [Fortunato and Barthélémy, 2007] of modularity optimisation methods such as EO, where smaller groups are merged [Fortunato and Barthélémy, 2007;

Table 4.2: Modularity results for CD-NMF, EO and Louvain methods

Data Set	CD-NMF	EO	Louvain
Dolphins	0.47 ± 0.03	0.51 ± 0.01	0.52
Books US Politics	$0.52 \pm \epsilon$	0.48 ± 0.01	0.50
Les Misérables	0.53 ± 0.02	0.53 ± 0.01	0.57
College Football	$0.60 \pm \epsilon$	0.58 ± 0.01	0.60
Jazz Musicians	0.43 ± 0.01	0.42 ± 0.01	0.44
C. elegans metabolic	0.36 ± 0.01	0.40 ± 0.09	0.43
Network Science	0.83 ± 0.01	0.86 ± 0.01	0.95
Facebook Caltech	0.38 ± 0.01	0.37 ± 0.01	0.37

Table 4.3: Number of communities from the CD-NMF, EO, and Louvain methods

Data Set	CD-NMF	EO	Louvain
Dolphins	6.67 ± 0.83	4 ± 0	5
Books US Politics	6.23 ± 0.62	4.04 ± 0.4	3
Les Misérables	9.97 ± 0.78	4.96 ± 1.72	6
College Football	8.86 ± 0.79	8 ± 0	10
Jazz Musicians	8.57 ± 8.89	4 ± 0	4
C. elegans metabolic	15.69 ± 1.14	7.96 ± 1.06	10
Network Science	342.53 ± 5.28	58.24 ± 12.36	418
Facebook Caltech	24.28 ± 1.72	6.84 ± 1.82	10

Porter et al., 2009], leading to a smaller number of communities, as seen in Table 4.3.

In Fig. 4.9 we illustrate the first network in Table 4.1, in which vertices are situated according to the Kamada-Kawai technique [Kamada and Kawai, 1988] in Pajek software [Batagelj and Mrvar, 1998]. The hard partitioning of the Louvain method can be contrasted with the soft partitioning of an example run of the CD-NMF, in which vertices near the boundary of two or more communities are represented by pie charts in a manner similar to that used by [Ball et al., 2011]. With the aid of the aforementioned "greedy allocation" scheme, the CD-NMF community assignments agree with the Louvain community assignments for 55 of the 62 nodes. Of the seven mismatches, six correspond to the putative additional community (here coloured dark green, in the dense central portion of the figure) postulated by the Louvain method; CD-NMF replaces this tiny community with soft partitioning among the other communities. The seventh mismatch occurred for a node connected to two red nodes and two pink nodes; the Louvain method allocated it to the pink community whereas CD-NMF allocated it to the red and pink communities in the approximate proportion of 51:49.

4.3.4 Graphs without community structure

We present the behaviour of CD-NMF in cases in which there is no community structure in the network, specifically focusing on the popular Erdős-Rényi (ER) random graphs. In such graphs, each link exists with a probability p which is common for any pair of nodes in the graph. Additionally, the probability of link formation at a given pair of nodes is independent of the presence of other links. This eliminates the tendency to form closed triangles and cliques that characterise real-world networks.

Therefore given various realisations of an Erdős-Rényi graph family $\mathcal{G}(N, p)$ (N number of nodes and p probability of pair connection), we want our method to be able to capture such an absence of community structure, instead of declaring community structure when there is none. In Fig. 4.10 we compare CD-NMF against three modularity-based methods:

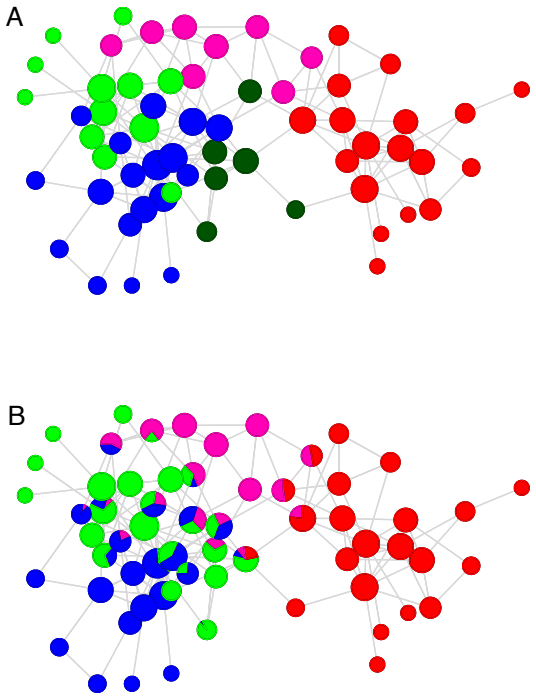


Figure 4.9: The Dolphins network [Lusseau et al., 2003], with (A) hard partitioning as per the Louvain method and (B) soft partitioning as per the CD-NMF method. Node size increases nonlinearly with vertex degree, and soft partitions are shown as pies.

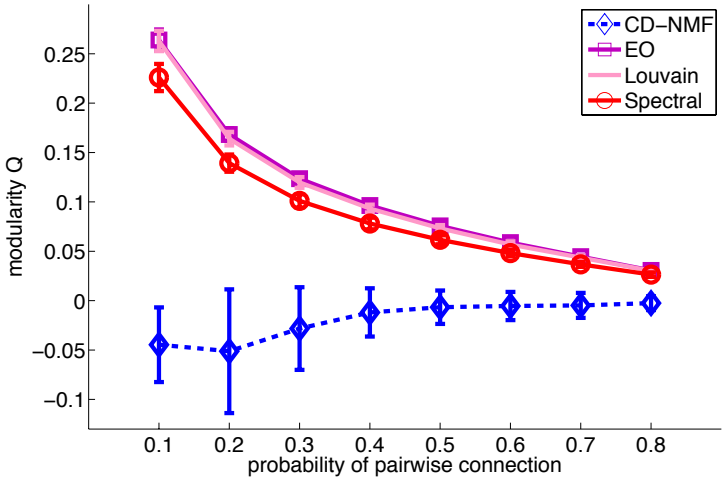


Figure 4.10: Modularity of ER graph $\mathbf{G}(100, p)$ partitions, extracted via four community detection algorithms. Each point represents the mean and standard deviation of modularity over 100 instances of $\mathcal{G}(100, p)$.

Extremal Optimisation (EO), the Louvain method, and Spectral Partitioning, based on the Q value of their extracted network partitions, in realisations of an ER graph class $\mathcal{G}(100, p)$. We control the network density by changing the value of p. For each value of p we generate 100 graphs, run community detection with each algorithm, and record the modularity values. The generated ER graphs we used consist of a single component.

In Fig. 4.10 we can see that EO (purple $\circ$-line), Louvain (light dashed line) and SP ($\square$-line) produce significantly higher modularity values than CD-NMF (bottom $\times$-line), especially for sparse realisations of the Erdős-Rényi random graph, implying the presence of modular organisation. However, those high Q values do not correspond to any community structure, as Erdős-Rényi random graphs do not possess it by design. In contrast to its competing methods, CD-NMF has a more stable behaviour as all modularity values are close to zero, indicating that nodes have no "preference" of being connected with members of the same group or otherwise. Especially for the case of sparse graphs with $p \simeq 0.1$, EO and Louvain achieve higher modularity values; in particular, they are very close to $Q = 0.3$, a threshold above which Newman and Girvan consider community structure to be present [Newman and Girvan, 2004]. This overestimation of modular organisation can be very misleading, especially when studying real-world networks which are often sparse [Faloutsos et al., 2004] due to their heavy-tail degree distribution. Therefore, if certain modularity optimisation methods produce higher Q values than CD-NMF, it might not mean necessarily that they have found a node configuration that corresponds to better community structure.

4.3.5 Real-world applications

In this section, we demonstrate the application of CD-NMF across a range of real-world problems where, in contrast to the traditional benchmark tests presented in Section 4.3.3 and 4.3.4, no observed solution is available. Instead, we show how CD-NMF solutions, which allow overlapping partitions and varying node participation scores, can be utilised in order to

provide valuable insights on the investigation of interdisciplinary research questions.

Application to a large-scale crowdsourcing problem

In modern crowdsourcing applications, there is an urgent need of rigorous statistical and computational methodologies that fuse decisions from large and diverse pools of human and artificial classifiers. Cutting-edge advances in the field [Simpson et al., 2013] have allowed not only optimal combinations of heterogeneous decision agents of varying reliability, but also the incorporation of prior knowledge on problems where the training data are sparse, while possessing excellent computational scalability properties.

In the work of [Simpson et al., 2013], a novel Bayesian classifier combination methodology, VB-iBCC (Variational Bayes - independent Classifier Combination), was introduced and applied to a large citizen science project, Galaxy Zoo Supernovae [Smith et al., 2011]. The aim of the project was to classify candidate supernova images as either supernova or not supernova. The data set contains scores given by individual volunteer citizen scientists (base classifiers) to candidates after answering a series of questions. A set of three linked questions are answered by the users, which are hard-coded in the project repository to scores of $-1, 1$ or 3, corresponding respectively to decisions that the data point is very unlikely to be a supernova, possibly a supernova and very likely a supernova. It has been reported [Simpson et al., 2013] that VB-iBCC optimally aggregates the decisions of multiple agents and outperforms other established approaches, given a data set of known supernovae image examples acquired by full spectroscopic analysis.

Individual decision-makers have their own unique level of technical knowledge, motivation, patience, confidence and overall reliability, which is extracted, via VB-iBCC, in the form of a 2×3 *confusion matrix* $\mathbf{U}^{(u)}$ for each decision-maker n. Each element $u_{i,j}^{(u)}$ denotes the probability that the n-th individual will give answer j given that the correct one is i. For example, $u_{0,3}^{(n)}$ encodes the probability that an individual n will decide that a given image is

"very likely to be a supernova" (hence $j = 3$) when it is not ($i = 0$).

Based on the above, we employed CD-NMF[5] to the problem of determining most likely groupings of base classifiers, given similarities in their decision-making behaviour. In [Simpson et al., 2013] a similarity matrix was calculated over all the citizen scientists participating in the study, based on a Hellinger distance over pairs of confusion matrices. Such a similarity matrix defines a relational structure between individuals, which can be viewed as an affiliation graph. We seek to examine if there are any natural groupings of classifiers, given their performance characteristics.

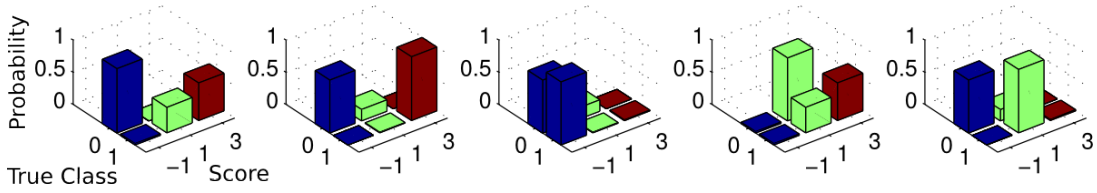


Figure 4.11: Prototypical confusion matrices for each of the five communities inferred using CD-NMF. Each graph corresponds to the individual with the highest community participation score based on $\mathbf{W}_\star$, with bar height indicating probability of producing a particular score for a candidate of the given true class. Source: [Simpson et al., 2013].

Application of CD-NMF to such a similarity matrix robustly gave rise to five distinct groupings of users. In Fig. 4.11 we show the centroid confusion matrices associated with each of these groups of citizen scientists. The centroids are the expected confusion matrices of the individuals with the highest node participation scores for each community. The labels indicate the true class (0 for not supernova or 1 for supernova) and the preference for the three scores offered to each user by the Zooniverse questions ($-1, 1$ and 3). Group 1, for example, indicates users who are clear in their categorisation of not supernova (a score of -1) but who are less certain regarding the possible supernova and likely supernova categories (scores 1 and 3). Group 2 are extremists who use little of the middle score, but who confidently (and correctly) use scores of -1 and 3. By contrast group 3 are users who almost always use score -1 (not supernova) whatever objects they are presented with. Group 4 almost never declares

[5]This joint research work was led by Edwin Simpson in [Simpson et al., 2013].

an object as not supernova ("pessimists") and, finally, group 5 consists of non-committal users who rarely assign a score of 3 to supernova objects, preferring the middle score (possible supernova). It is interesting to note that all five groups have similar numbers of members (several hundred) but each clearly indicates a very different approach to classifying supernova images, which we exploit in order to perform an optimal decision making combination.

Application to vessel tracking

What if we could use community detection to spot pirate ships? Global maritime traffic data has grown to such large volumes over the past decades, due to advances in electronic tracking technologies, so that the process of manually detecting illegal vessels (related to smuggling, terrorism, or unauthorised fishing activity) is now unfeasible for individual analysts. The solution for discovering such an illicit behaviour lies in the automatic detection of anomalous patterns in, otherwise normal appearing, vessel tracks.

In [Smith et al., 2013] we address this issue[6] by modelling vessel tracks using Gaussian Process (GP) regression and then define pairwise similarities via an appropriate Hellinger-based distance metric. Taking the inverse distance between GP tracks allows us to map our data to a relational space, where each pair i, j of vessel tracks is assigned a similarity value s_{ij}. We encode all similarity pairs to a matrix $\mathbf{S}$ so that s_{ij} is the *degree of coupling* between the paths of vessels i and j. By treating $\mathbf{S}$, shown in Fig. 4.12 as an adjacency matrix from a network analysis perspective, in Fig. 4.13 we perform community detection using NMF in order to discover clusters of similar paths that we call *path-based communities*.

Given our CD-NMF solution, with a soft membership score w_{ic} for each node i and cluster c, subsequent tracks can be tested using extreme value theory, by examining how statistically surprising is the deviation of each vessel from a prototypical community member (i.e. the vessel with the highest participation score to a given group). For example a vessel claiming to belong to a given class can be tested against a path-based community from matrix $\mathbf{S}$. Our

[6]This joint research work was led by Mark Smith in [Smith et al., 2013].

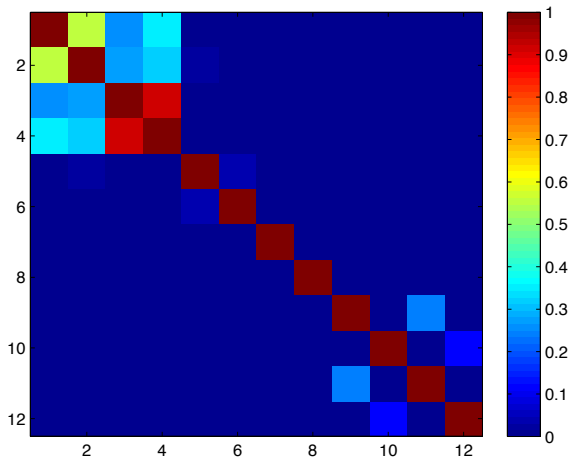


Figure 4.12: The inverse Hellinger distance (adjacency) matrix between inferred functions for the vessel data. Nodes 1 to 12 relate to the different classes of vessel, cargo (1-4), fishing (5-8) and sailing (9-12). Source: [Smith et al., 2013]

empirical experiments on vessel data suggest that the method is capable of detecting anomalies that resemble mooring or drifting, and unexpected departures from regular movements, as shown in the unclassified vessels in Fig. 4.13.

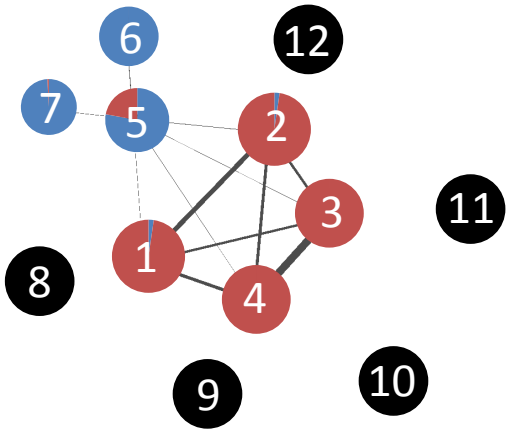


Figure 4.13: Network diagram illustrating the clustering of different vessel tracks. Each edge connection is weighted (illustrated by the varying edge thickness) based on the adjacency of Fig. 4.12. Different node colours relate to the different communities within the data, and the membership score of each node, given by CD-NMF, is defined a portion in the pie. The black nodes are unassigned and do not belong to any given community. Source: [Smith et al., 2013]

The Wytham Woods data set

The body of work developed in this chapter has been motivated by the study of the Wytham Woods great tit population, where we employ a social network analysis approach in understanding the causes of variation in animal social behaviour. In Chapter 8 we use CD-NMF on time-dependent "snapshots" of the bird network, an example shown in Fig. 4.14, across a 2-year study period and examine the form and function of the extracted modules. We examine the cohesiveness of bird social groups throughout the course of each year, we investigate how they compare against null cases and study their importance from the perspective of mating pair formation.

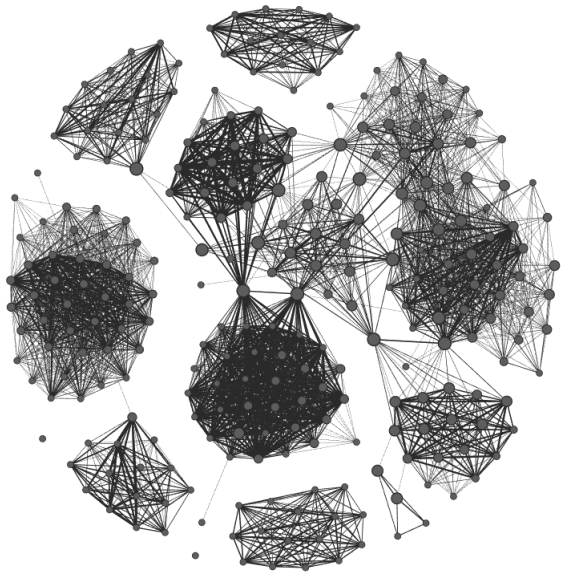


Figure 4.14: The Wytham Woods great tit social network during a particular day of the bird observation period. In this highly modular ($Q \simeq 0.855$) graph of $N = 264$ nodes, $M = 2800$ edges and 5 disconnected components, we applied NMF and identified 12 overlapping foraging flocks.

4.4 Future Extensions

4.4.1 Application to directed graphs

In this chapter we considered cases of undirected networks with symmetric interaction matrices $\mathbf{A}$. Although CD-NMF does not allow negative link weights in the graph, it is still possible to consider the case of asymmetric communication rates that arise in systems such as email or telephone networks.

Following the discussion of Section 4.2.1, we have explained how our NMF framework naturally extends to the directed graph case, where asymmetric interaction rates can exist between node pairs. Consider the example graph of Fig. 4.15. Such a small graph has $N = 8$ nodes and $M = 7$ edges. Ignoring the link directionality and performing community detection will yield the solution shown in the orange dashed line in Fig. 4.15, where each community corresponds to a small "star"-like structure.

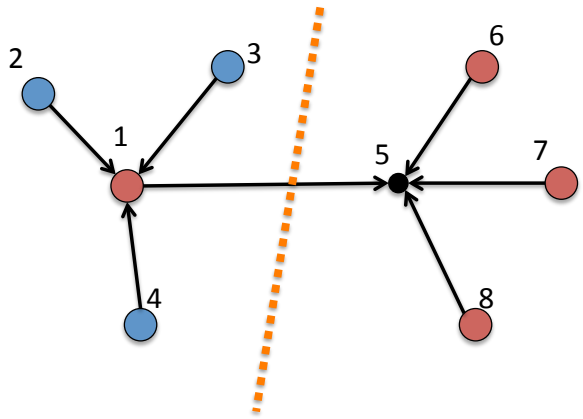


Figure 4.15: An example directed graph of $N = 8$ nodes and $M = 7$ edges. The orange dashed line separates the two communities that were extracted by ignoring the directionality of the links. For the directed case, the three different colours in the nodes correspond to the different communities identified by CD-NMF. Communities in this setting are classes of nodes with a common target, in contrast to the traditional definition of clusters with increased link density.

We run community detection on the toy example of Fig. 4.15 using CD-NMF and illustrate the resulting partition using three different node colourings; community 1 (blue) consists

of the nodes 2, 3, 4, community 2 (red) consists of nodes 1, 6, 7, 8, while node 5 is associated with the singular community 3 (black). In this setting, communities are not comprised of clusters of mutually connected nodes, but correspond to groups where members point to a common target node. In a practical context, consider a directed graph where nodes "follow" each other in a Twitter-like fashion; performing NMF community detection on this network will yield groups of individuals that have a *common information sources*. Based on such an observation, in [Frankel, 2012] we have begun to apply CD-NMF to a large Twitter-like graph of $N = 18174$ investors from the Motley Fool (`http://www.fool.com`) website. In such a network, where individuals share stock-picking information in a social media fashion and their decisions are evaluated through real market data, we seek to find the correspondence between investing performance and an individual's membership in "information communities", e.g. groups of individuals that lie on the receiving end of the same signals in the network. From initial results[7], we have found that over-performing investors do not belong exclusively to a particular information community, but have a rather homogeneous spread over the network [Frankel, 2012].

4.4.2 Temporal community detection

This work addresses the issue of extracting community partitions from a single interaction network defined by $\mathbf{A}$. We acknowledge that in many problems, this matrix describes only a "snapshot" $\mathbf{A}^{(t)}$ of a time-evolving, dynamic complex system. Therefore, we seek to extend our community detection method to allow for a time-evolving solution space. At present we are approaching this via a jump-diffusion model (based around a Markov model), in which rate parameters are allowed to evolve with time and the structure of the community solutions may also have abrupt changepoints [Garnett et al., 2010]. Our aim is to evaluate this approach in time-evolving systems in order to model community drifts and the transitions from one

[7] This joint research work is led by Zach Frankel in [Frankel, 2012].

community structure to another.

In the same dynamic network setting, we plan to apply techniques in changepoint detection [Garnett et al., 2010]. After extracting time-varying sufficient statistics resulting from NMF community detection, we can fit a Gaussian process to that data stream by adjusting the parameter values in a covariance kernel. The likelihood associated with each fit can be assessed; these likelihoods may increase if a constant mean function is subtracted from one contiguous section of the data stream (in particular, for all time beyond a certain point). The ratio of the likelihoods before and after this subtraction can decide whether a changepoint occurred. Ultimately, we are able to calculate the probability that at certain time points, a changepoint occurred in the network dynamics.

At the time of writing, a dynamic extension to the NMF-based community detection was proposed by [Mankad and Michailidis, 2013], which extends the typical constrained optimisation problem for NMF parameter inference, a variation of the one we presented in Section 4.2.3, via the introduction of "slackness conditions" that enforce coupling between time steps.

4.4.3 Further association indices

In Section 4.2.1 we noted that in our CD-NMF model, interaction rates $\hat{a}_{ij}$ between pairs of individuals are influenced by the co-membership of i and j in the same communities. Given a pool of N individuals with community memberships described by $\mathbf{W}, \mathbf{H}^{\mathsf{T}} \in \mathbb{R}_{(+)}^{N \times C}$, each expected interaction rate $\hat{a}_{ij}$ will be given by the linear function of Eq. (4.5), giving rise to the NMF expression of Eq. (4.3) for the adjacency matrix $\hat{\mathbf{A}}$.

Such a function can be extended in various ways, depending on the problem context. Firstly, we can consider non-linear approaches such as hyperbolic tangent-functions that exhibit "saturation" properties [Li et al., 2005], in the sense that as participation scores w_{ic}, h_{jc} increase for the same community c, there is a diminishing-returns effect and $\hat{a}_{ij}$ does not

grow at the same rate.

Additionally, instead of taking into account the absolute co-participation in the same groups, we can also consider the exclusivity of those co-appearances, so that $\hat{a}_{ij}$ is higher for members i, j who appear in the same communities c alone, leading to a structure where smaller communities are (on average) more strongly connected. Such a model is proposed in Chapter 7, where we view association functions from the perspective of one-mode projection.

4.4.4 Improvements to inference

Our current method produces point estimates for the model parameters via a maximum a posteriori (MAP) scheme. A fully Bayesian treatment can be employed via Reversible Jump Markov Chain Monte Carlo as presented in [Zhong and Girolami, 2009], or via the use of variational Bayes as derived in [Cemgil, 2009]. The advantage of a posterior distribution over quantities such as the inner rank dimensionality c is that we can see at which resolutions modular organisation is most prevalent.

We also acknowledge that CD-NMF, along with the majority of community-detection methods, assumes a fully observed adjacency matrix. This is not the case in many real-world applications in which data collection limitations arise; for example when the system under study is sampled or when sensors fail to record every observation. However, CD-NMF can be easily extended to allow for missing data [Cemgil, 2009].

4.5 Discussion

In this chapter we described a novel approach to community detection that adopts a Bayesian nonnegative matrix factorisation model to achieve soft partitioning of a network in a computationally efficient manner. We have demonstrated how community detection can be seen as a generative model in a probabilistic framework in which priors exist over the model param-

eters. This enables model order selection, which in our framework is the number of latent communities (or classes of nodes) in the data. We also showed that the degree of participation of two individuals in various communities is a latent generator of the expected number of interactions between them.

Following the model formulation section, we demonstrated how CD-NMF not only captures the membership of a node in multiple communities, but also quantifies how strongly that individual participates in each of the groups. By using the entropy of the node membership distribution, we can identify "core" nodes in each community or, inversely, "broker" nodes that act as mediators between different groups. At a global level, the mean entropy of the membership distributions can help us quantify the degree of "fuzziness" in the network community structure. Network visualisation tools can also be improved in this manner, as the degree of membership over different communities can be utilised to position an individual in a cloud of nodes.

We also showed that CD-NMF has a competitive performance against popular community detection methods, on various popular network data sets in Section 4.3.3. Although CD-NMF is not a method aiming to maximise modularity, it competes well with methods that directly maximise Q and we have showed that it can even outperform in several module identification problem. More importantly, the soft partitioning solutions along with the node participation scores across communities can give valuable insights into real-world problems, as we have discussed in Section 4.3.5, while an extensive application of CD-NMF to a real-world setting is performed in Chapter 8, where we seek to explore the flocking structure of a wild-bird population.

Chapter 5

Animal Social Networks and the Wytham Woods Data Set

5.1 Introduction

In Chapter 2 we discussed how networks can used for studying various complex systems in nature and technology. In the present chapter, we focus our attention on its application to ecological systems, in particular *animal social networks*. Following a brief overview of the animal social network analysis literature, we introduce our motivating application domain, which is the large-scale study of wild-bird sociality at Wytham Woods, Oxford. We provide the necessary background on the experimental setting and observation gathering scheme, along with the form and statistical properties of the collected data.

In Section 5.2 we discuss the motivation behind the use of network analysis to study animal populations and the advantages of employing such a paradigm, both from a theoretical and computational perspective. In Section 5.3 we present the experimental setting of our motivating application, where a grid of recording locations captures the foraging habits of *Parus major* wild-birds. Following a brief overview of the Parus Major species and its importance for animal behaviour studies, we present our experimental set up, how data are

being collected and various characteristics of the data set.

In summary, this chapter presents the motivation behind the methodological developments we formulate in Chapters 6, 4, 7 and lays the foundation for the zoological analysis we present in Chapter 8. The research questions posed by the Wytham Woods experimental setting are discussed in Section 5.5.

5.2 Animal societies as social networks

Modern technological advances on field equipment have allowed biologists to collect a wealth of observation data on animal behaviour. The principal advantage of applying network analysis tools on animal populations is that we can study their social organisation *at any scale*, without restricting our attention to dyads (pairs) or isolated groups [Krause et al., 2009]. Additionally, the flexibility of the network paradigm allows us to model any time of interaction; sexual, cooperative, competitive, etc, as we discussed in Chapter 2.

One of the key points in behavioural biology is that an individual's social position has important fitness consequences [Krause et al., 2009; McDonald, 2007; Sih et al., 2009]. For example, [McDonald, 2007] showed that node betweeness is crucial for the future mating success of young long-tailed manakins. Additionally, [Krause et al., 2009] argues that there is a feedback loop between population dynamics and individual behavioural strategy, shown diagrammatically in Fig. 5.1. That brings us to the discussion of Section 2.1 on complex systems, where the overall system behaviour is not a mere summation of the individual behaviours, or simply "a flock is not a big bird" [Rosvall, 2006].

The web of interactions in an animal population can help us identify the role each individual plays in a network, as shown by [Lusseau and Newman, 2004] in the case study of the Doubtful Sound bottlenose dolphin society. [McDonald, 2007] also used the graph topology to identify the *alpha* and *beta* status of males in the study of long-tail manakins.

Mapping animal interactions to a graph is not just a way of studying the connectivity

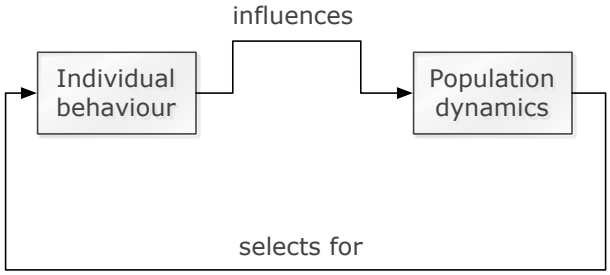


Figure 5.1: A diagram illustrating the feedback loop between population dynamics and individual behavioural strategy, based on the discussion in [Krause et al., 2009].

patterns among individuals (using, for example, the degree or betweeness centrality of each node). Community structure can help us understand how information [Lusseau and Conradt, 2009] or diseases [Krause et al., 2009] may spread in a network while in the temporal setting, the manner in which network connectivity changes over time can help us understand the population response to seasonal and environmental factors [Tantipathananandh et al., 2007].

A critical discussion of social network mapping is presented by [James et al., 2009], focusing on the incomplete nature of the available data sets and the pitfalls of modelling assumptions such as the *Gambit of the Group* (GoG) [Whitehead and Dufault, 1999], where given an observed group of animals we assume that every member of that group is interacting with everyone. Such criticisms are taken into account in the formulation of our probabilistic models presented in the next chapters.

The advantages of the network approach in modelling animal populations are extensively discussed in [Krause et al., 2009] and [Sih et al., 2009], while [Whitehead, 2008] and Darren Croft [Croft, 2008] present a wealth of quantitative methods across various animal societies. Although many results from the contemporary animal social networks literature are bespoke to a particular species and/or experimental setting, their core computational framework is based on social network theory, making such an approach attractive for a wide range of different ecological settings. In the next section, we describe the application motivating this thesis, which is the wild-bird great tit (GT) *Parus major* population at Wytham Woods, Ox-

ford.

5.3 The Wytham Woods experiment

5.3.1 A quantitative approach to studying animal sociality

It is not an understatement to claim that advances in sensor technology have brought a paradigm shift to animal behaviour studies. Miniaturisation of tracking hardware, mainly body-attached Radio-frequency identification (RFID) tags [Gibbons and Andrews, 2004] or Global Positioning System (GPS) transponders [Mann et al., 2011], allows systematic and disturbance-free observation of free-ranging animals at a large scale. A far cry from traditional schemes of manually observing animals with pen & paper and binoculars, this approach leads to a collection of rich sensor-generated data on the daily lives of organisms, allowing researchers to employ powerful data processing and machine learning tools for behavioural analysis [Eagle and Pentland, 2006]. Tapping into such data sources of unprecedented size, quality and resolution, provides the opportunity for obtaining valuable insights on natural systems [Hey, 2009], an approach traditionally reserved for human social and economic behaviour studies [LaValle et al., 2011; Mitchell, 2009]. Towards this goal, in this section we present a data collection scheme that focuses on the digital footprints of a fascinating social agent, the great tit.

5.3.2 The great tit

Let us introduce, in Fig. 5.2, the great tit (GT) *Parus major* wild bird that plays a key role in our data analysis efforts. A small passerine, spread throughout Europe and Asia, the GT is not only one of the most-studied small birds in the world, but also “a near-ideal subject through which to examine questions in ecology and behavioural biology” [Grosler, 1993], as it can forage and nest in man-constructed (and thus monitored) environments without disruption.

The GT is a model system for animal social networks as well, as it expresses a complex and dynamic sociality throughout the year, forming single and mixed species flocks, which bring predator avoidance [Grosler, 1993], information sharing [Aplin et al., 2012] and mate selection benefits [Psorakis et al., 2012].

Figure 5.2: A great tit *Parus major* wild bird. Photo credited to Thor Veen and used with permission.

The GT displays a relatively stable annual cycle; around August bird communities consist of closely related individuals (families and offspring) that start to diffuse into dynamic flocks as the offspring leave their families. Eventually as we approach the mating season around April, male-female pairs emerge from those fuzzy groups forming the basis of the new family communities, as shown in Fig. 5.3. The underlying processes and social phenomena that drive mating partner selection from such winter foraging groups are largely unknown.

Towards the goal of studying the phenomena discussed above, we seek to monitor the GT mobility patterns across their natural habitat, which in the case of our experimental setting is Wytham Woods, Oxfordshire. Wytham Woods (51 46 N, 1 19 W seen in Fig. 5.4) is a mixed deciduous woodland of c. 385ha, with a canopy composed primarily of oak, ash, sycamore and beech. Although the study of GT populations in Wytham dates back to 1947, standardised protocols have been employed since the early 1960s [Grosler, 1993; Perrins, 1965]. In recent

years, advances in sensor technology have allowed the large-scale deployment of wild-bird tracking methods via RFID hardware. The manner with which we collect our observations is presented in the following section.

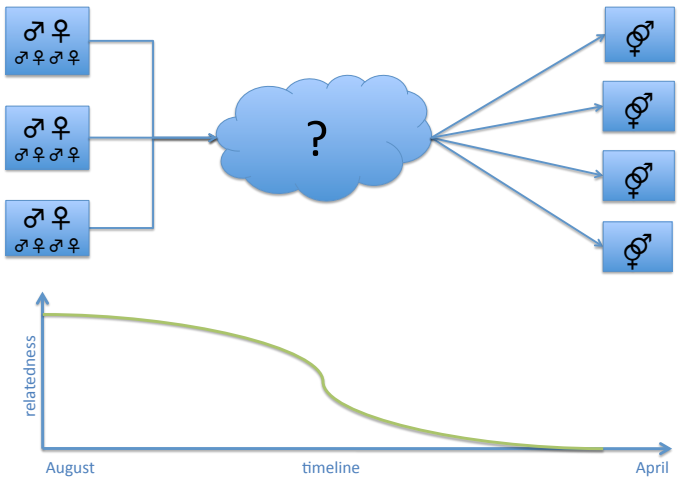


Figure 5.3: The form and function of bird communities undergoes many changes all year around. We illustrate how family groups break down into dynamic communities, from which mating pairs emerge during the breeding season.

5.3.3 Data collection setting

The fact that GTs are residents on their habitat all-year round, along with their willingness to receive food provided by humans [Grosler, 1993], allows us to set up permanent stations that allow observation of their behaviour across the whole range of their lifespans.

Starting from 2007, each nestling great tit born on the study site, and each adult great tit breeding there, were ringed with a device which contained a 125 Hz RFID tag (CoreRFID Ltd), as shown in Fig. 5.5. Such a harmless sensor allows us to monitor the bird's location via an appropriate grid of antennae-enabled feeding stations strategically located across Wytham. Each autumn and winter (beginning in August and ending in early March) during 2007–9, field technicians deployed bird feeding stations, baited with sunflower seeds, at 67

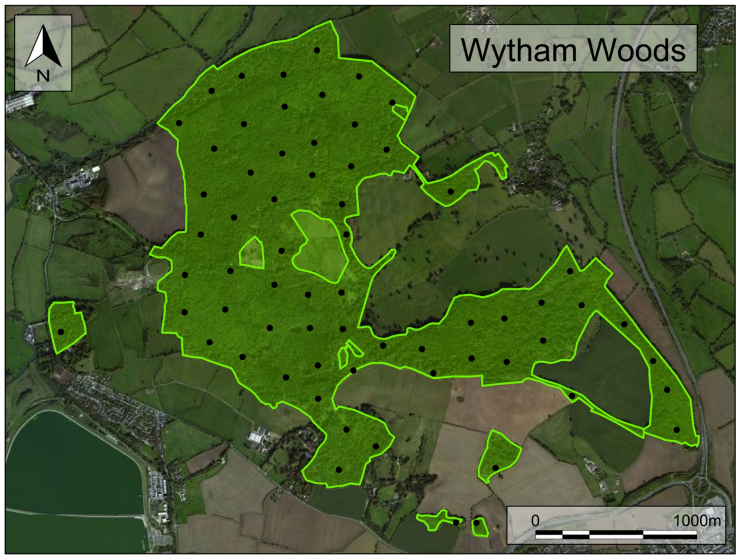


Figure 5.4: Satellite view of Wytham Woods, with borders and feeding locations geotagged. Photo credited to Lucy Aplin and used with permission.

locations, which were spaced regularly on a 250m grid throughout the 385 ha of the study site, shown in Fig. 5.4. The stations were equipped with two antennae (Francis Instruments Ltd, Cambridge), so that every time a GT visits the feeder to collect sunflower seeds, the antenna hardware produces a timestamped record of its visit, to the nearest 15 seconds. The whole data collection process is shown diagrammatically in Fig. 5.6.

Figure 5.5: Birds are "ringed" with a harmless RFID device that generates a sensor observation when the bird comes into proximity of the logger-enabled feeders placed across Wytham Woods, Oxfordshire. Photo credited to Edward Grey Institute and used with permission.

The 16 loggers available at any time were rotated around the 67 locations following a

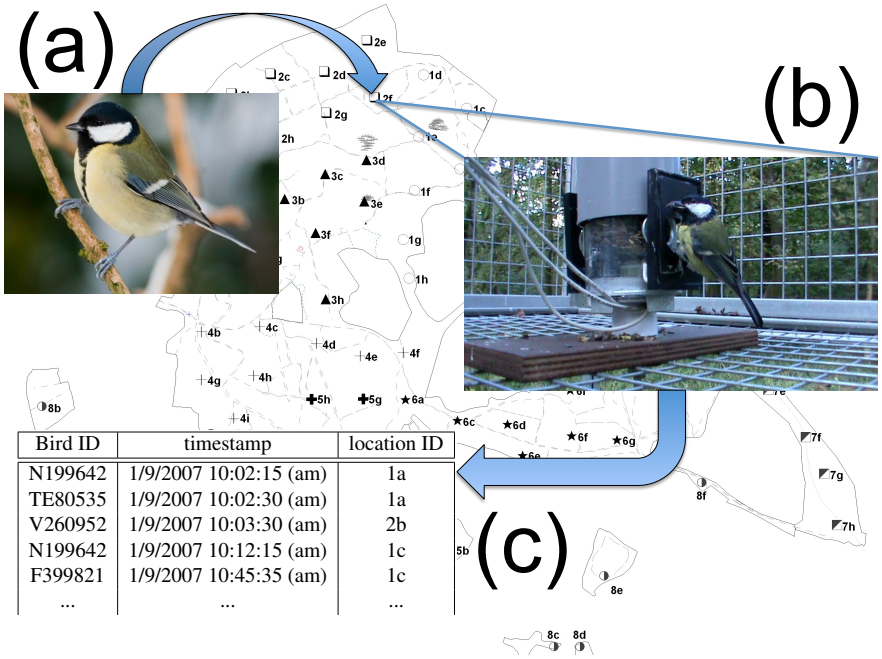


Figure 5.6: An illustration of our data collection scheme. In (a), a pit-tagged great tit visits one of the 67 feeding stations across Wytham Woods, shown in (b). The sensor equipment at the feeder allows us to capture its presence, as shown in (c), in the form of a tuple storing information about its unique bird ID, the timestamp of the visit and the location ID. Photos used with permission from the Edward Grey Institute. The schematic was produced by the author.

randomised scheme, so that each of 8 approximately equally-sized sections of the site always had two active loggers in it. Rotation happened on a 4 day schedule, and feeders were refilled with sunflowers each time they were moved. The data analysed here are taken from the first two winters of this project, 2007–8 and 2008–9, in which there were 548,709 records of 770 individuals and 484,088 records of 753 individuals respectively; in total over the two winters there were 1,032,797 records of 1,217 different individual great tits. In recent years, hardware advances have allowed the collection of a richer data sets, which we describe in Chapter 9.

5.4 Data set details

5.4.1 The format of logger data

In this section we provide details on the format and statistical properties of the data generated by the Wytham Woods experiment. Initially, raw observations from the antennae memory are retrieved in a raw `.csv` format that contains a long stream of `[ID,t]` pairs. Each `ID` is a unique bird identifier code while `t` is the visitation timestamp. Following a series of consistency checks and maintenance tasks (for example, removal of corrupted tuples, usually 1 in 10,000), observations from each location `l` are aggregated to a single datastream, so that each tuple is in the form `[ID,t,l]`. Further post-processing tasks are then performed, such as sorting the datastream in ascending order based on `t`.

The data generated in this fashion consists of a long stream of timestamped tuples, each representing a *unique bird visitation*. In Table 5.1 we illustrate the data set, which can be seen as a *transactions table* in a relational database context. It is important to note that due to the 15-second resolution of the 2007–9 loggers, we can not know the exact length of time a bird stayed at a feeder. That is because great tits may arrive at the feeder, pick a sunflower seed, fly a bit further away and come back within 15 or 30 seconds. Therefore, even successive observations of the same bird (as in the last two rows of Table 5.1) are not indicative of a continuous presence at the feeder.

Table 5.1: Sample format of our data

Bird ID	timestamp	location ID
N199642	1/9/2007 10:02:15 (am)	1a
TE80535	1/9/2007 10:02:30 (am)	1a
V260952	1/9/2007 10:03:30 (am)	2b
N199642	1/9/2007 10:12:15 (am)	1c
F399821	1/9/2007 10:45:30 (am)	1c
F399821	1/9/2007 10:45:45 (am)	1c
...	...	...

In Fig. 5.7 we plot a sample of our data using *clock time*, instead of *event time* as in Table

5.1, for the first 4 days of the 2007–8 data set and focusing on two locations. The large "gaps" of zero observations result from the night period, which acts as a natural separator between days in the data set, as no bird foraging activity takes place and the hardware equipment is switched off. For that reason, in the majority of methodological developments and data analyses presented in this thesis, we process each batch of daily logging data *separately*.

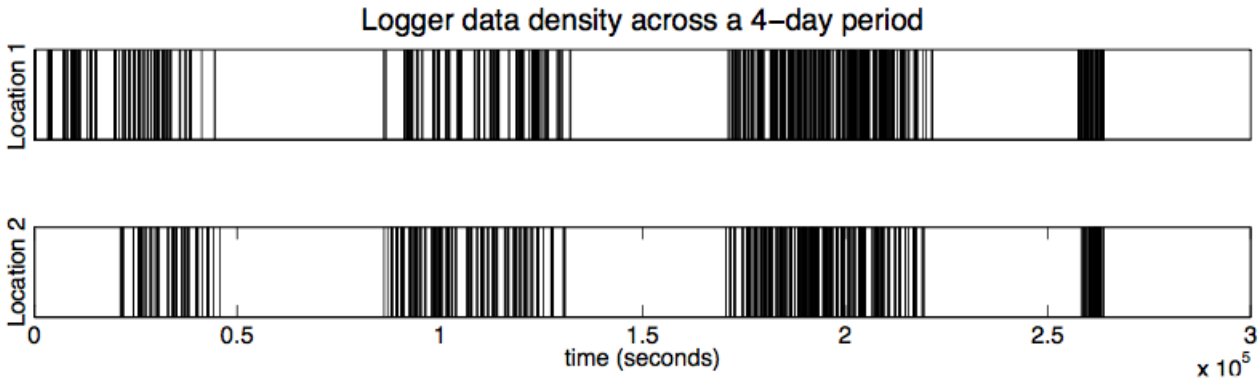


Figure 5.7: An illustration of our bird logging data for the first 4 days of the 2007–8 data set, focusing on two locations. The black vertical lines represent bird records and white space denotes observation-dead periods.

5.4.2 Data set representation and additional zoological information

By collecting the unique entries from the first and third column of Table 5.1 we effectively have two additional lists, containing the active great tits and locations involved in the experiment. We can incorporate further information about each bird (age, gender, birthplace, etc) and location (coordinates, vegetation, temperatures, etc) into those lists, thus building the database presented in Fig. 5.8.

Our Wytham Woods database also includes pedigree information. That is detailed records of bird pairs `ID1, ID2` that became *mating partners* in a particular year (see Table "Pairs" in Fig. 5.8), along with the bird IDs of their children (as in Table "Pedigree" in Fig. 5.8). Such information, collected by field experts, allows us to describe the family structure of wild-

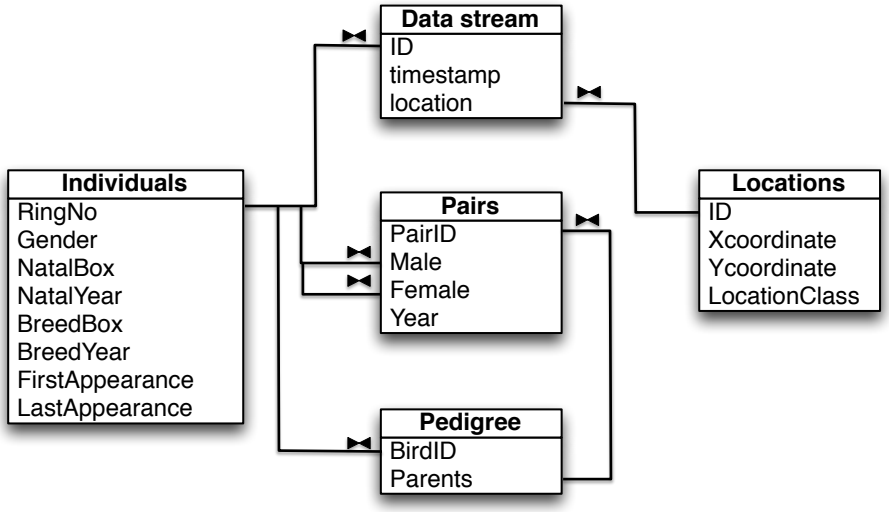


Figure 5.8: The Entity-Relationship (ER) schema of wild-bird database, capturing both the feeder observation stream, along with external information on individual characteristics and pedigree information. The symbol ($\bowtie$) represents one-to-many relationships.

birds during each year and examine how such direct genetic similarity affects the individuals' position in the social network.

From an implementation perspective, the raw information on bird visitations and pedigree records been transcribed from a text-based `.csv` format to SQLite and then ported to a custom-made MATLAB Object-Oriented database for analysis purposes.

5.4.3 Data set statistics

In this section we explore some general statistical characteristics of the Wytham Woods data set, consisting of two big parts; $\mathcal{D}^{(7,8)}$ that covers the activity of $N_{7,8} = 770$ birds from August 2007 to March 2008 and $\mathcal{D}^{(8,9)}$ that spans from August 2008 to March 2009 and contains $N_{8,9} = 753$ birds. Due to migration and fatalities, there are only $N_{7,9} = 360$ common individuals across the two seasons.

We begin by exploring the percentage of the total bird population that is captured during each day of the data collection season. In Fig. 5.9 we plot the fraction of total $N_{7,8}$ and

$N_{8,9}$ birds per season that appear in daily and monthly batches of data. At day-level, the population coverage averages around 11.3% as only 16 feeders are active at any given day across Wytham. At month-level, where the feeder rotation scheme has covered all locations, we observe much higher participation scores that average around 50.2% across both seasons. It is worth noting that we can rarely have total population coverage, not just due to our data collection scheme, but also because of biological factors relating to dispersion, migration and fatalities.

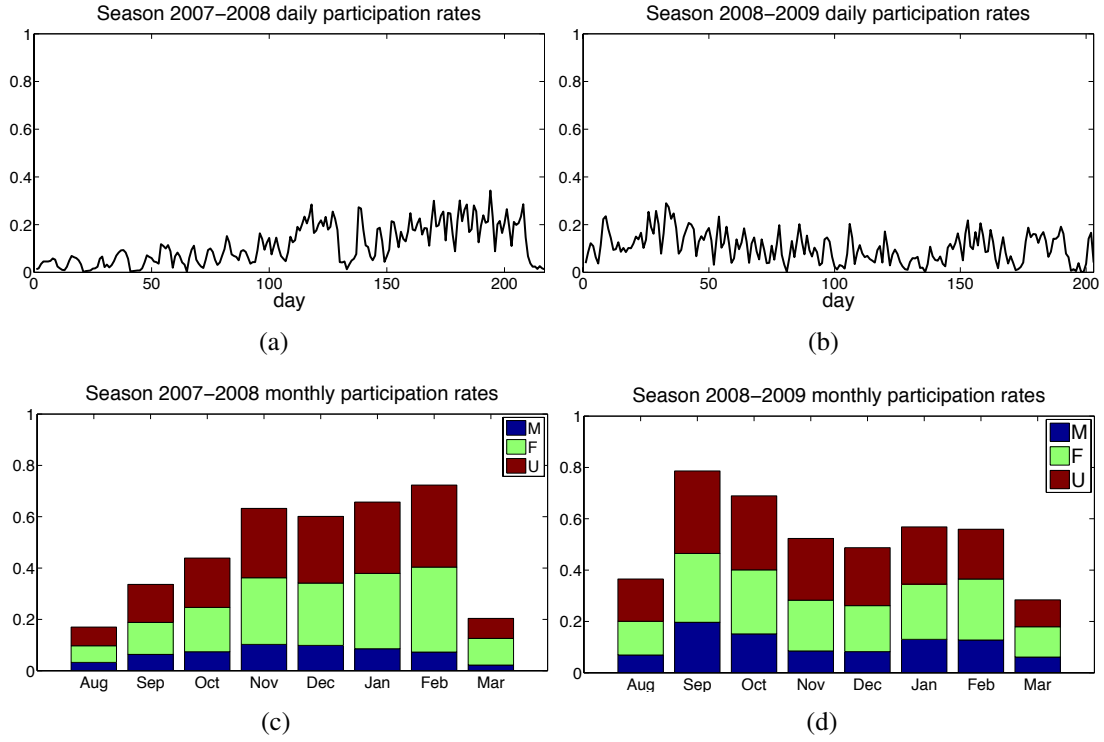


Figure 5.9: We plot the participation rate of great tits in the feeder data, defined as the fraction of population coverage per "time slice". We can see that for month-level resolution we achieve higher coverage, due to more comprehensive antennae rotation scheme. Colouring reflects the gender breakdown of each monthly population group; "M" for males, "F" for females and "U" for unknown.

We also examine the degree of participation of each great tit in the data, by counting its total number of records in the observation stream. In Fig. 5.10(a) and Fig. 5.10(b) we produce a histogram of such counts and show that there is a significant presence of individuals

with small number of records (about 10% of individuals of each season have fewer than 100 records), while most birds are recorded around 400 times in the data stream. Although there are individuals with more than 1000 sensor captures associated with them, we have found no evidence of a "fat tail" in either one of the distributions.

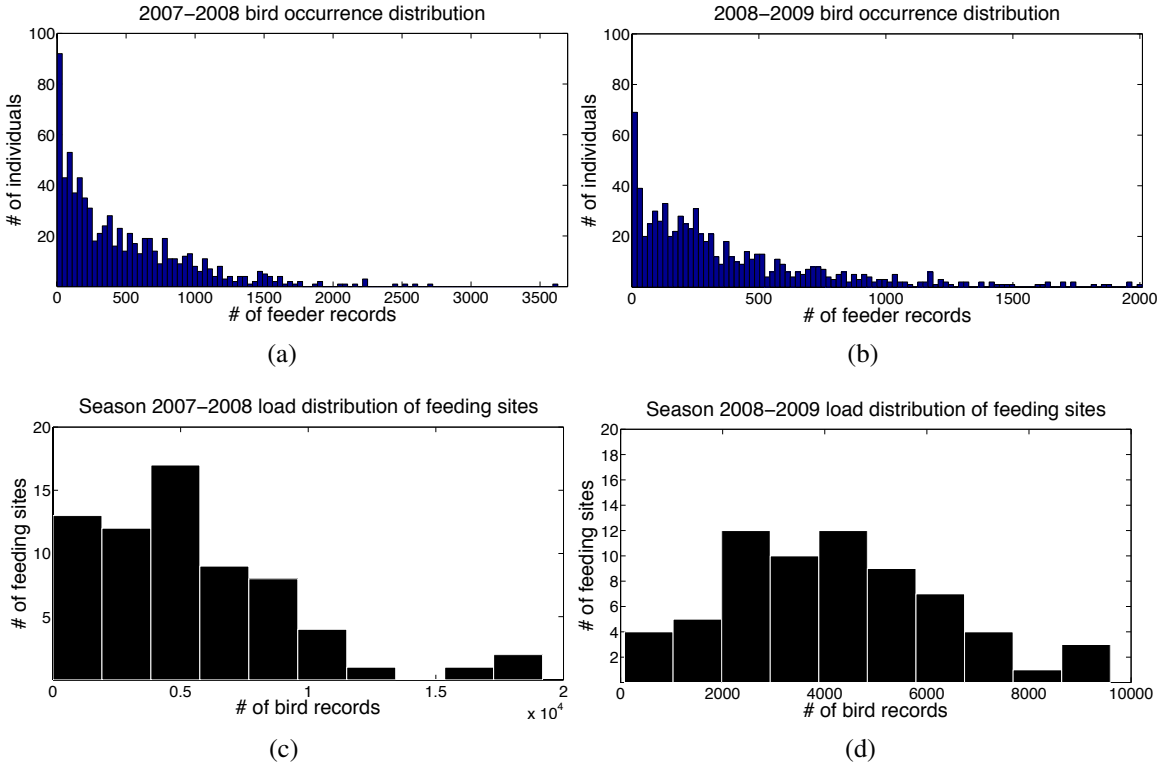


Figure 5.10: We show the histograms of bird occurrences in Fig. 5.10(a) and 5.10(b), which denote how many individuals are associated with a particular number of data set records. In a similar note, in Fig. 5.10(c) and 5.10(d) we plot the histograms associated with how bird records are spread across feeding sites.

We perform a similar computations in Fig. 5.10(c) and Fig. 5.10(d) where we plot the histograms of how bird records are spread across feeding sites. Although there are clear inhomogeneities for both seasons, they are explained by seasonal and ecological characteristics of the local habitats associated with each feeder.

Finally, in Table 5.2 we provide some statistics on both data sets, summarising the quantities we described above, along with some other ones that will prove useful in later analyses.

One of them is the notion of "old" versus "new" mating pairs; we consider a mating pair as "new" if it was formed *during* the particular data collection season, while "old" pairs have already been formed in previous years and persist until the current season. Such information is provided to us by the pedigree data set, described in the previous section.

Table 5.2: Data Set statistics

Season	2007-2008	2008-2009
No. of records Z	548,709	484,088
No. of days	217	203
No. of months	8	8
No. of birds N	770	753
No. of males	300	276
No. of females	321	253
No. of mating pairs	69	58
No. of "new" pairs	49	48
No. of "old" pairs	20	10
Mean participation (per day)	11.33%	10.43%
Mean participation (per month)	47.05%	53.29%
Mean inter-observation time (sec)	103.68	158.49
Variance of inter-observation time (sec^2)	2.9×10^5	5.41×10^5

An important quantity we also consider in Table 5.2 is the *inter-observation time*, which denotes the time difference between consecutive bird visitations at the same location. For the calculation of its mean and variance we have omitted the antennae down-time "gaps" during the night periods (as discussed in previous section), as they artificially affect such statistics of the inter-observation times. From the difference in the level of magnitude between the mean and variance we can see that there is clear indication of over-dispersion in inter-observation times, which forms the basis of the methodological developments of the next chapter.

5.5 Discussion

In this chapter we have discussed how advances in sensor miniaturisation have allowed large-scale deployment of tracking equipment, allowing us to amass large data sets of bird mobility

records. Our application area of focus is the Wytham Woods great tit data set, which we described from various perspectives in the previous sections.

Following on the preliminary analysis we performed in the previous section, we seek to proceed further and describe the great tit population not just via a collection of general descriptive statistics, but from a *social network analysis* perspective. Though such an approach has strong advantages, as discussed in Section 5.2, bird association patterns are not apparent from the raw sensor data. Thus our motivating application poses a series of methodological questions:

- How can we infer the social network among the birds, given the transactional data presented in Table 5.1? In other words, how can we *infer graph structure from non-relational time series data*? In order to study GT sociality and employ the wealth of social network analysis tools, we need to define an appropriate graph topology, which is not obvious from raw sensor data. We address this issue in Chapter 6.

- How do we capture the uncertainty on the inferred graph topologies and communities described above? Can we model a social tie in a fully probabilistic framework, so that our analysis is not so sensitive to noise and missing observations? We address the issue of placing distributions over graph topologies in Chapter 7.

- As flocking structure is a key aspect of GT behaviour across the year, from family groups to mating pairs, how can we discover *communities* in a given GT social network? How can we capture the multi-membership of a given bird across various flocks, foraging across Wytham. Such methodology is presented in Chapter 4.

By addressing the methodological challenges described above, we seek to reveal and explore GT sociality across two annual circles and investigate the relationship of network quantities with fitness quantities relating to reproductive success and survival. We also seek to examine the social aspect of mating pair formation, how and when mating pairs arise from

the fuzzy winter communities and what is their position in the global network. Such issues are addressed in Chapter 8, where we apply all of the methods developed in the present work to the data sets described in this chapter.

Chapter 6

Inferring Graph Structure from Data Streams

6.1 Introduction

The key motivation for employing network analysis tools is that the web of interconnections between individuals can provide us invaluable insights into the underlying mechanisms that govern the system under study [Newman, 2010]. For example, within an ecological context, the position and role of animals in the network can have important fitness consequences [Wey et al., 2008] both for the individual and the population as a whole [Krause et al., 2009]. Additionally, the network paradigm gives us the flexibility to look at the system at various resolutions and model any type of interaction: sexual, cooperative, competitive, etc [Krause et al., 2009].

Despite the advantages of the network paradigm and the wealth of computational tools for network analysis [Barrat et al., 2004; Buchanan and Caldarelli, 2010; Fortunato, 2010; Newman, 2003b], the problem of capturing any given system as a graph is not always trivial. Not all systems possess an obvious "web-like" structure (such as the Internet), where the interconnections between participating entities are apparent from direct observation (com-

puters that are connected through physical cables). Additionally, collected data (from field studies, sensor observations, GPS/transponders etc) may not capture the associations between the observed agents, thus no relational structure can be directly defined. In systems such as animal populations the underlying network of social affiliations needs to be inferred through proxies such as the behaviour (mobility patterns, foraging habits etc) of individual animals. For example, animal social networks are built by observing individuals in the field [Bejder et al., 1998; Croft et al., 2004; Lusseau et al., 2003] and, where possible, place the appropriate link between individuals by recording the type of interaction [Voelkl and Kasper, 2009]. Otherwise, we make the assumption that consistent physical proximity acts as a proxy for social affiliation [Bejder et al., 1998] and members of the same group usually interact with each other [Whitehead and Dufault, 1999]. Based on co-occurrence in different sites, a link is drawn between individuals using various "association indices" that have been proposed [Ginsberg and Young, 1992]. Finally, other studies such as [Walker et al., 2010] make use of sensor equipment to record when a specific action is taking place by an animal and build a correlation network from the behavioural similarity between individuals.

In this chapter, we focus on the problem of discovering a latent social network structure of a population that can only be observed through the mobility patterns of its individual members, an example of which we presented in Chapter 5. We seek to examine under which conditions we can consider a pair of individuals as socially connected, given the similarities in the manner upon which they occur at various locations. In Section 6.2 we provide our problem statement and a brief description of our data format (more details are presented in Chapter 5), along with the necessary nomenclature. In Section 6.3 we present a common approach of discovering network structure from spatio-temporal data, which is based on a discretisation of the observation stream given an appropriate resolution parameter. Based on our criticism of traditional link discovery approaches, we proceed in Section 6.4 by proposing a methodology that exploits key statistical properties of the data stream, in order to reveal

a modular structure that has a direct network interpretation. We introduce the concept of "gathering events", which constitute areas of high observation density in the stream and correspond to flocks of socially affiliated individuals. In order to identify such gathering events, we propose a simple yet efficient algorithm of linear complexity, which we extend in Section 6.5 to the Bayesian setting, via a Gaussian mixture model. In Section 6.6 we discuss how such a gathering event structure allows us to describe the actual bird social network, along with proposing an appropriate significance test for the inferred graph links. Arguments in favour of the proposed methodologies are presented in Section 6.7, where we compare them against conventional approaches via a battery of benchmark tests.

The gathering events methodology, termed GEM for short, which we propose in this chapter plays a crucial role in Chapter 8, as it allows us to extract the wild bird social networks from the sensor data described in Chapter 5. Although such a methodology has been developed in an ecological context, in Chapter 9 we argue that it can be generalised to any setting where agents perform timestamped appearances at various locations (also known as "check-ins" in the social media parlance, for an example see [Kietzmann et al., 2011]).

6.2 Description of data and nomenclature

Recall the discussion of the Wytham Woods data set in Chapter 5. Observations are collected based on the following scheme: every time a tagged bird comes to sufficiently close proximity to a feeder, the recording hardware generates a single data tuple that captures the ID of the bird along with the time and location where the foraging event took place. By aggregating records from all feeding locations, the data generated from this scheme consists of a long stream of timestamped observations, as in the example of Table 6.1.

Let our spatio-temporal data $\mathcal{D}$, a sample of which we show in Table 6.1, be represented in the form $\mathcal{D} = \{\texttt{ID}_z, t_z, \ell_z\}_{z=1}^{Z}$, where Z is the total number of records or *tuples* in our database (e.g. the number of rows of Table 6.1). If we take a single tuple $\{\texttt{ID}_z, t_z, \ell_z\}$,

Table 6.1: Sample format of our data

Bird ID	timestamp	location ID
N199642	1/9/2007 10:02:15 (am)	1a
TE80535	1/9/2007 10:02:30 (am)	1a
V260952	1/9/2007 10:02:30 (am)	2b
V260952	1/9/2007 10:02:45 (am)	2b
N199642	1/9/2007 10:12:15 (am)	1c
...	...	...

we read it as: "Observation # z: bird $\texttt{ID}_z$ appeared at time t_z at the feeding location ℓ_z". From an implementation perspective, the bird and location codes in the first and third column of Table 6.1 are associated with a unique integer (e.g. through a hash-function), so that $\texttt{ID}_z \in \{1, ..., N\}$ and $\ell_z \in \{1, ..., L\}, \forall z \in \{1, ..., Z\}$. Additionally, all timestamps in the second column are converted to seconds (counting from a specific date, e.g. the beginning of data collection) for convenient manipulation. Note that $\{t_z\}_{z=1}^{Z}$ denotes event time, therefore for every timestamp t_z in $\mathcal{D}$ there exists a bird appearance $\texttt{ID}_z$ and each $t_z - t_{z-1}$ is not necessarily constant for all $z, z-1$ pairs. Additionally, given a specific bird i out of total N birds, there can be many records z for which $\texttt{ID}_z = i$, as a single individual may appear many times in the data. Finally, we consider all tuples in $\mathcal{D}$ to be sorted in ascending order based on their timestamp t_z.

In Fig. 6.1 we present a way for visualising a given data stream $\mathcal{D}$, which will be used extensively in this chapter for illustration purposes. We use such plots to represent visitations that take place at the same location. The horizontal axis represents time (in seconds) and data records $\{\texttt{ID}_z, t_z, \ell\}$ referring to a particular location ℓ are positioned based on their timestamp t_z. Each data record $\{\texttt{ID}_z, t_z, \ell\}$ is represented by a stem and, when appropriate, we use different shapes in order to distinguish between different individuals $\texttt{ID}_z$.

Regardless of its representation format, our data stream $\mathcal{D}$ is only a transactions table in a relational database context, which restricts our analysis to a handful of relatively simple counting operations such as finding the total appearances of a given bird, total birds that vis-

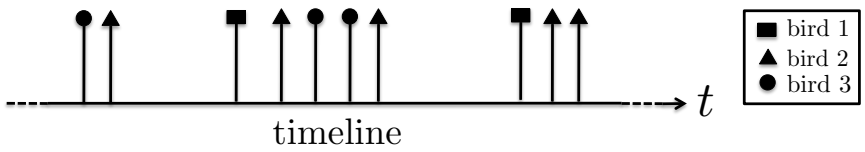


Figure 6.1: We plot a snapshot of an example data stream $\{\texttt{ID}_z, t_z, \ell\}_{z=1}^{Z}$, at a particular location ℓ. Within this snapshot there are 10 records and 3 individual birds (distinguished by the different shape $\square$,$\triangle$, $\circ$ in the stem and indexed by $\texttt{ID}_z \in \{1, ..., 3\}$) that appear at various points in time.

ited a specific feeder, etc. Our goal is to find an appropriate mapping from $\mathcal{D}$ to an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where a_{ij} denotes the extent of social affiliation between birds i and j, based on the similarities in their movement patterns.

To keep the notation uncluttered, from now on we will focus on the case of a single location ℓ (as in the case of Fig. 6.1) and demonstrate, when appropriate, that results can be easily generalised to the multi-site case.

6.3 Network inference via time-windowing

6.3.1 Fixed time window

Most studies in animal social networks involve recording the presence of individuals at specific locations or groups. From the simplest data gathering techniques where animals are simply observed by field experts, to the more sophisticated such as the Wytham Woods experiment in Chapter 5, our core assumption is that physical proximity between two individuals is likely to signify social interaction, while consistent co-occurrences positively correlate with strong social affiliation [Bejder et al., 1998]. Based on this hypothesis, the time window approach [Gero et al., 2009; Krings et al., 2012; Lauw et al., 2005; Oh and Badyaev, 2010; Whitehead, 2008] involves placing a link between two individuals i, j, if they are observed in the data stream within a fixed temporal distance of Δt. The more times they were seen together, the stronger the link weight a_{ij} between them.

To illustrate this, consider the example shown in Fig. 6.2 that shows a snapshot of the data $\mathcal{D}$ under consideration. We discretise the data stream into a series of intervals of length Δt and proceed by scanning the timeline, identifying the observations $\{\texttt{ID}_{z_1}, t_{z_1}, \ell_{z_1}\}, \{\texttt{ID}_{z_2}, t_{z_2}, \ell_{z_2}\}, ..., \{\texttt{ID}_{z_n}, t_{z_n}, \ell_{z_n}\}$ that fall within each bin and place a link between the corresponding pairs of individuals $(\texttt{ID}_{z_1}, \texttt{ID}_{z_2}), ..., (\texttt{ID}_{z_{n-1}}, \texttt{ID}_{z_n})$. The output of the process is an undirected weighted matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ where a_{ij} is the number of time intervals Δt within which individuals i and j co-occurred.

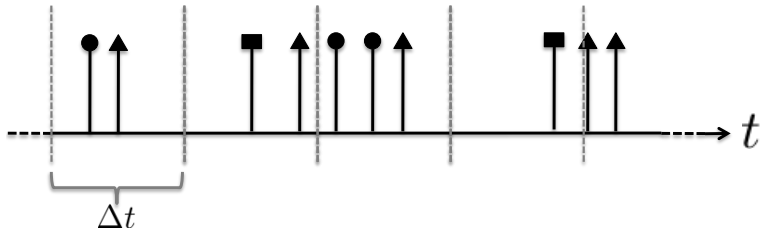


Figure 6.2: An example data stream $\{\texttt{ID}_z, t_z, \ell\}_{z=1}^{Z}$, discretised into a series of fixed intervals of length Δt. Individuals that fall within such bins are assumed to be associated.

If we define $T = t_Z - t_1$ to be the time span of the data collection period in $\mathcal{D}$, the total number of intervals is $n_T = T/\Delta t$. The algorithm performs a linear search $\mathcal{O}(n_T)$ at each interval and within each Δt performs another search, examining which tuples $\{\texttt{ID}_z, t_z, \ell_z\}$ exist within and place links between the corresponding individuals $\texttt{ID}_z$. The number $Z_{\Delta t}$ of observations within Δt can be from 0 to Z, while the number $N_{\Delta t}$ of unique individuals range from 0 to N. The computational cost of the process is $\mathcal{O}(ZNn_T)$, where Z and n_T are "competitive" terms, in the sense that small time windows (n_T large) typically lead to small number of observations and number of individuals within (thus faster searches within each Δt) while large time windows require longer searches within intervals but a smaller number n_T of time windows to consider.

Although the selection of time window size Δt has critical implications on the topology of the inferred graphs (an issue that will be discussed in more detail in Section 6.3.3),

link structure is also affected by the position of interval boundaries (dashed vertical lines in Fig. 6.2) relative to the actual observations. For example, two individuals may appear in very close temporal proximity in the data stream, but fall at different time intervals, as seen in the last two bins of Fig. 6.2. This may lead us to miss important connections in the data stream regardless of the time window size. In order to overcome such shortfalls, in the next section we propose an extension of the fixed time-window method.

6.3.2 Flexible time window

Let us consider the example data stream of Fig. 6.1, where we pick an individual i and we place an "influence zone" of size Δt around each one of its observations (same length as the example of Section 6.2) as seen in Fig. 6.3. Every other individual that falls within such zones is assumed to be connected to i. For example in Fig. 6.3(a) the focal individual $i = 1$ ($\square$) is captured together two times with $j = 2$ ($\triangle$), thus $a_{12} = 2$. In cases where we have successive observations of the same individual that are positioned less than Δt away, we simply merge the intervals, as done in Fig. 6.3(b) and 6.3(c), leading to a *flexible time window* scheme.

Given the example of Fig. 6.3 the 3×3 adjacency matrix would be:

$$\mathbf{A} = \begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & 3 \\ 0 & 3 & 0 \end{pmatrix},$$

where we have ignored the diagonal elements (no self-edges). Similar to the fixed time window case, we end up with an undirected weighted graph described by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where each a_{ij} is the number of times i and j co-occurred within a fixed temporal distance of Δt.

The whole process has similar computational cost as in the fixed time window case; for every individual $n \in \{1, ..., N\}$ we perform a linear scan in $\mathcal{D}$, identify each tuple

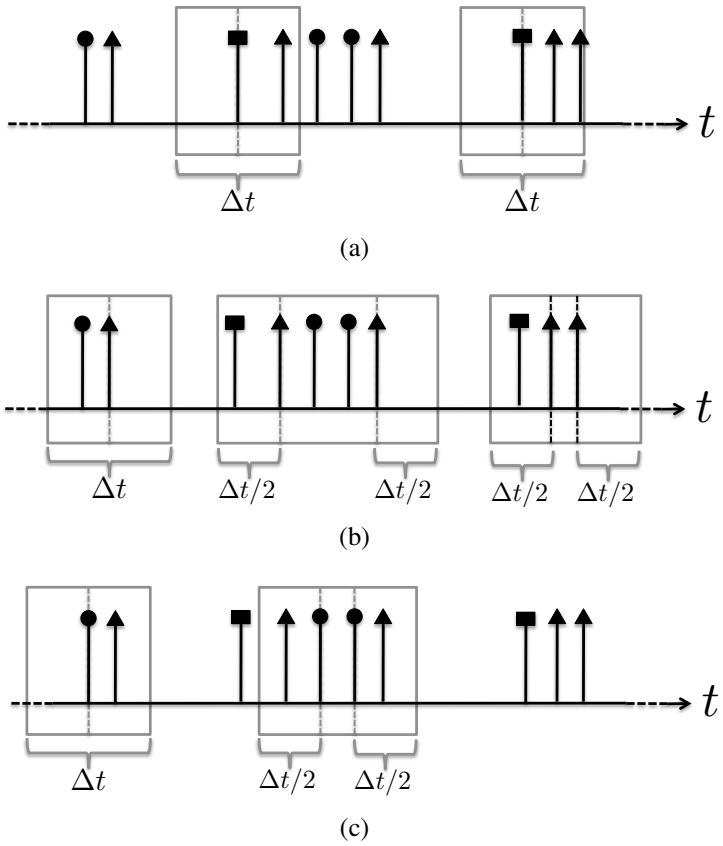


Figure 6.3: We demonstrate how connections between individuals are identified, via the flexible time-window approach. For the first individual $i = 1$ ($\square$) we place around each one of its observations an "interaction zone" of length Δt. Any other bird that is observed within that window is assumed to be associated with i. The weight of link a_{ij} is simply the number of such windows where i co-occurred with j. Overlapping influence zones of the same individual are being merged as in Fig. 6.3(b) and Fig. 6.3(c).

$\{\texttt{ID}_z, t_z, \ell_z\}$ for which $\texttt{ID}_z = i$, place the interaction radius Δt and monitor which other individuals j fall within such an interval. The process of merging time windows has a small computational overhead, implemented efficiently via an appropriate use of pivot variables, thus bringing no additional term in the overall complexity.

In the multiple location setting, we perform the above scheme for all individuals that appear at a particular location ℓ, ending up with an adjacency matrix $\mathbf{A}^{(\ell)}$. Because each link $a_{ij}^{(\ell)}$ denotes the number of co-occurrences between animals i and j, the whole network $\mathbf{A}$

can be reconstructed by simply summing over all locations[1] $\mathbf{A} = \sum_{\ell=1}^{L} \mathbf{A}^{(\ell)}$.

Networks extracted using the above scheme have an intuitive interpretation, as every link is weighted by the total number of times two individuals were observed together. The notion of "togetherness" lies at the core of such models, as it directly relates to the way we define socially meaningful spatio-temporal proximity. Under the time windowing approach, "togetherness" is defined by a fixed resolution parameter Δt, which induces some really strong assumptions on the structure of the data. Those issues are discussed in the following section.

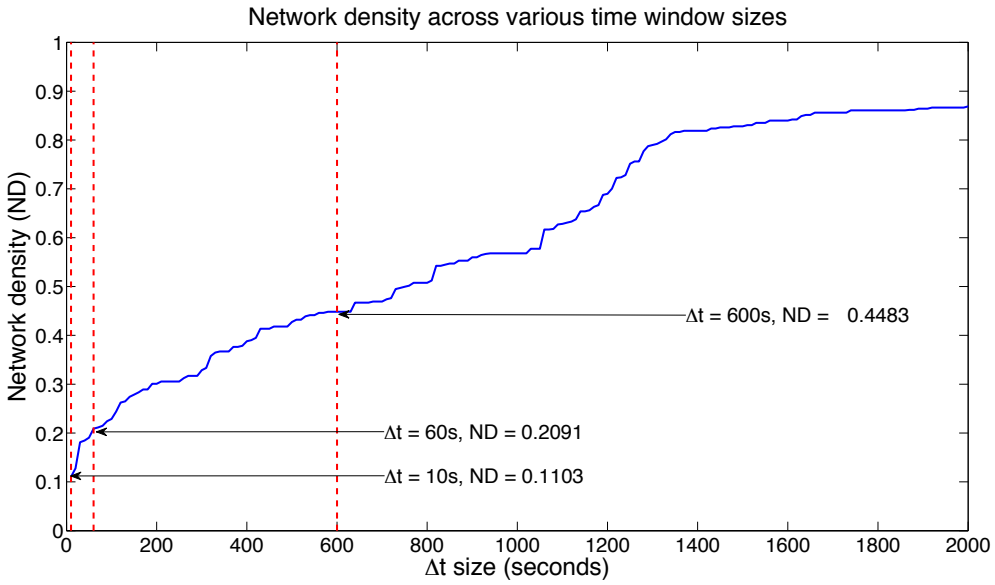
6.3.3 Issues associated with time-windowing

Via a time window approach, interactions a_{ij} are defined within a given temporal distance Δt. The selection of this scale parameter is crucial for the extracted topology of the network; insufficiently small time windows may omit important co-occurrences while unreasonably large ones lead to an over-estimation of the population's social connectivity.

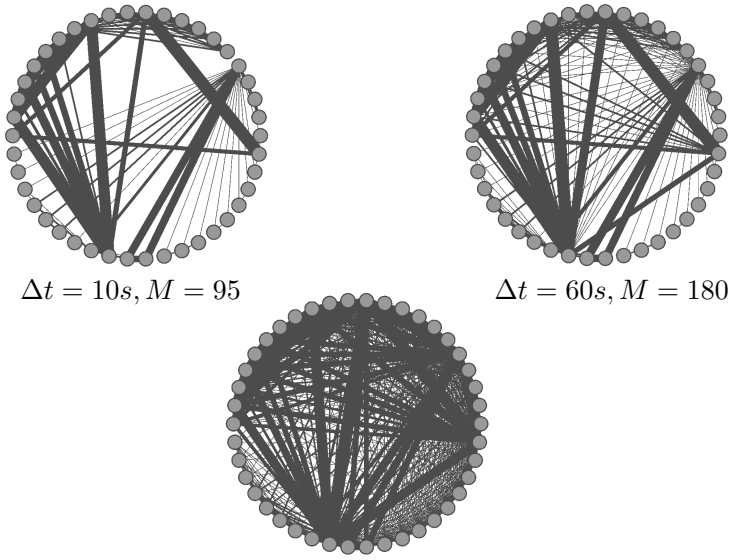
Using our wild-bird data as an example, we take a single day's worth of observations and place links between the $N = 66$ individual birds (nodes) based on the number of times they were recorded within a temporal distance of Δt. We seek to examine the changes that take place in the network as we vary the time window size by monitoring the *network density* (ND), which is the fraction of M links in the network over all possible pair combinations $\frac{1}{2}(N^2 - N)$ of N nodes. We can see in Fig. 6.4(a) that ND increases along with the size of Δt, because more links are placed between nodes. An example of how network topology changes for various selections of time window size is shown in Fig. 6.4(b), while [Krings et al., 2012] have performed similar experimentation considering more network metrics such as average degree, average weight, clustering coefficients, etc.

Between all these different network topologies that result from varying Δt, there is no

[1] That is for cases we are not interested in the contribution of each location to the total co-occurrences. Nevertheless, in our implementations the number of co-occurrences per location is stored as an additional property in an appropriate edge list.



(a)



(b)

Figure 6.4: In Fig. 6.4(a) we plot the network density for various time window sizes, spanning from 10 seconds to half an hour. We can see that especially for early increases of Δt there is a large inclusion of links in the network. We also mark three cases of different time window sizes (dashed vertical line) and show in Fig. 6.4(b) how the graph topology changes based on the Δt value.

direct way of knowing which one is the most appropriate. Therefore, in cases where we have no expert knowledge of the temporal scale of our data, we have to examine multiple time windows and select the one that satisfies some performance metric. Although there are many graph quantities to consider, such as network density, modularity Q, average clustering coefficient, these are more descriptive variables of a particular topological structure, rather than a fitness score on how well a network corresponds to a data stream. Instead, we define our performance metric based on some form of deviation from randomness.

Let us consider a randomised version of the data stream, where for each location ℓ we have performed a shuffling of bird labels $\texttt{ID}_z$ while maintaining the order of timestamps. Such a scheme maintains key characteristics of the data set such as number of observations per individual, location popularity, temporal distribution of records, but breaks all dependences in the observation sequence induced by social structure. From such a randomised instance of the data stream and given a proposed Δt, we can extract a null matrix $\mathbf{A}^{(0)}$ using the time window process described in Section 6.3.2. The dissimilarity of the two networks $\mathcal{L}_{\Delta t}(\mathbf{A}||\mathbf{A}^{(0)})$ can be defined as:

$$\mathcal{L}_{\Delta t}(\mathbf{A}||\mathbf{A}^{(0)}) = -\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (a_{ij} - a_{ij}^{(0)})^2, \tag{6.1}$$

which is the mean square error between the links of the observed and the randomised network[2]. Our task is, given a proposed Δt, to generate randomised versions of the data stream and estimate the average dissimilarity value $\bar{\mathcal{L}}_{\Delta t}(\mathbf{A}||\mathbf{A}^{(0)})$ of the observed versus the null network. For $\Delta t = 0$ we have $\mathcal{L}_0 = 0$ as the time window is so strict that both networks are empty. For the maximum value of $\Delta t = T$ (the experiment time span), both networks are equal ($\mathcal{L}_T = 0$), as every individual is connected to all others that appeared in the same loca-

[2]Note that we have skipped the lower-diagonal part of $\mathbf{A}$ and $\mathbf{A}^{(0)}$, as the networks we are considering at present are undirected.

tion and the actual ordering of $\{\texttt{ID}_z\}_{z=1}^{Z}$ does not matter. We monitor how $\mathcal{L}_{\Delta t}$ changes for different values of time window within $(0, T)$ and select the appropriate Δt that gives rise to the *most dissimilar* network $\mathbf{A}$ compared to the null model. An example is shown in Fig. 6.5 where, given a data stream of wild bird observations, we plotted $\mathcal{L}_{\Delta t}(\mathbf{A}||\mathbf{A}^{(0)})$ across various values of Δt. The suggested time window size using this scheme is $\Delta t \simeq 15$ minutes.

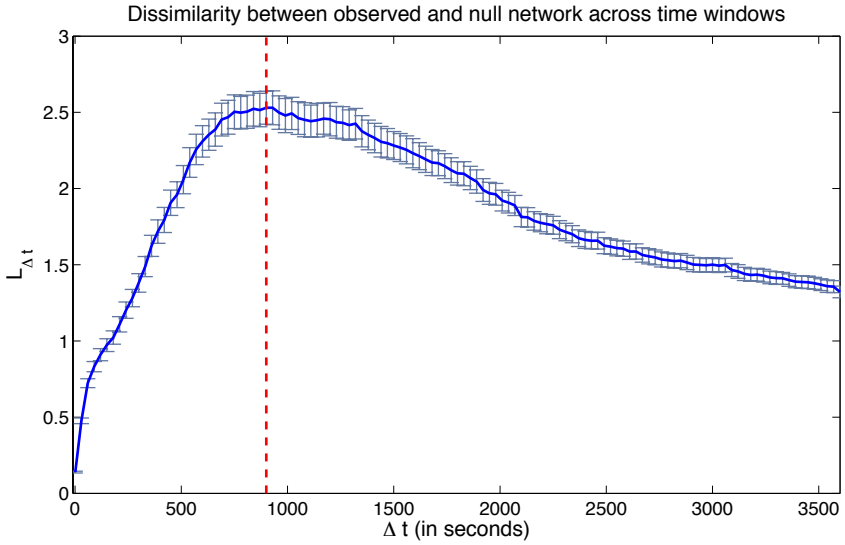


Figure 6.5: We plot the dissimilarity metric $\mathcal{L}_{\Delta t}(\mathbf{A}||\mathbf{A}^{(0)})$ between the observed $\mathbf{A}$ and $r = 1000$ null networks $\mathbf{A}^{(0)}$ that resulted from appropriate randomisations of a data stream with 4989 observations of 166 individuals in 14 sites. The maximum value of $\langle \mathcal{L}_{\Delta t} \rangle \simeq 2.53$ is achieved for a time window of $\Delta t = 900$ seconds.

Although the above scheme allows us to overcome the problem of selecting the appropriate time-resolution parameter in our data, the process of performing multiple runs and network extractions can be computationally demanding, especially in cases of large data streams. Additionally, even if the optimal time window size were given to us a priori (thus not having to evaluate the fitness score of each Δt), we have still made the strong assumption that Δt is fixed throughout the data stream. This corresponds to the belief that the "interaction radius" between individuals is constant across our observation period and is not affected by temporal changes in the overall system.

Additionally, networks that are built via the above scheme assume that all co-occurrences

indicate a social tie. Such a strong assumption, commonly termed in the animal social network literature as "Gambit of the Group" (GoG) [Whitehead and Dufault, 1999], may lead us to overestimate the socially connectivity of the population, ignoring the fact that co-appearances at a particular location can be either coincidental or have no social basis.

Following the above issues and criticisms, in the next section we explore how we can exploit the unique characteristics of spatio-temporal data sets, such as the one presented in Section 6.2, in order to automatically infer an optimal measure of grouping in the data.

6.4 Social network discovery via identification of "gathering events"

In this section we study the statistical properties of spatio-temporal data streams such as the ones presented in Section 6.2, by focusing on the density profile of feeder visitation times. In Section 6.4.1 we introduce the notion of "gathering events", a key pattern in our data that we exploit in order to develop, in Section 6.4.2, a linear-time link discovery algorithm for mapping spatio-temporal data streams to social networks.

6.4.1 Discovering a modular structure in spatio-temporal data streams

Consider the plot in Fig. 6.6, which illustrates how bird arrivals at a particular feeding location are spread throughout a small sample of our observation timeline. Each vertical line represents an sensor capture of a bird $\texttt{ID}_z$ at time t_z. We can see that the records are not uniformly spread across time, but they are "packed" in small observation-dense regions. Indeed, if we study the distribution of time differences $\delta(t_z) = t_z - t_{z-1}$ between every pair of consecutive observations across the data stream, we find broad tails (with exponent $\gamma \simeq 2.5$ for $\delta(t_z) > 10^3$, under a power-law model fit [Clauset et al., 2009]) and clear evidence of over-dispersion, as reported in Fig. 6.7 and Table 6.3. This non-exponential decay of inter-

record timestamps, along with the fact that most $\delta(t_z)$ take small values, implies that the observation profile consists of temporally-focused bursts of recording activity, which can be seen as flocks of foraging individuals. Such a behaviour is observed across two observation seasons: 2007 to 2008 and 2008 to 2009.

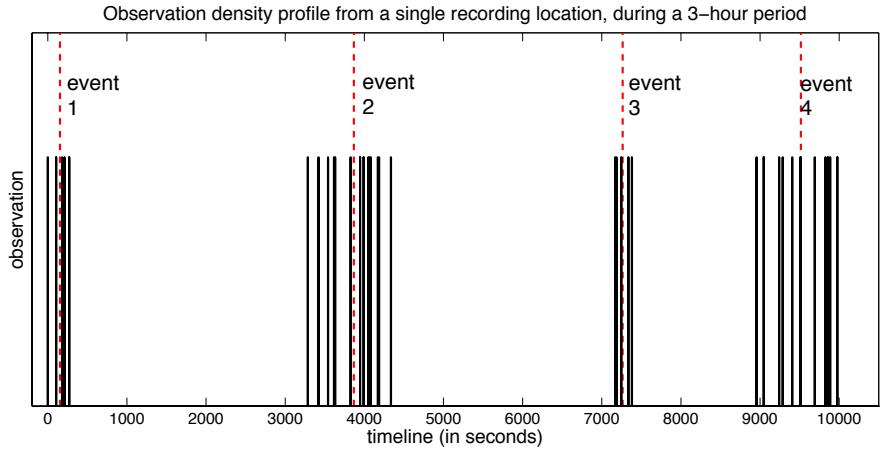


Figure 6.6: We plot bird arrivals as recorded at a specific location over the course of 3-hour period. We can see that the visitation profile is temporally focused, consisting of bursts of bird activity. Our goal is to identify such regions of increased observation density and examine which individuals participate in those gathering events.

Data Set	$\bar{\mu}$ (secs)	σ^2 (secs^2)	$\bar{m}$ (secs)
2007–8	103.68	2.9×10^5	15
2008–9	158.49	5.41×10^5	30

Table 6.2: Sample mean, variance and median inter-arrival time $\delta(t_z)$ for the two different data sets under consideration.The vast difference in the level of magnitude between the sample mean and variance denotes over-dispersion (non-Poissonian arrival times). In all cases, performing log-odds comparison between a exponential decay and power-law yields a very large number in favour of the power law (larger than machine precision), which is not reported here.

Note that for the statistics of $\delta(t_z)$ we consider pairs of consecutive observations $\{\texttt{ID}_z, t_z, \ell_z\}, \{\texttt{ID}_{z+1}, t_{z+1}, \ell_{z+1}\}$ that conform to the following requirements:

1. $\texttt{ID}_z \neq \texttt{ID}_{z-1}$, they refer to different individuals, thus ignoring multiple successive antenna readings of the same RFID tag.

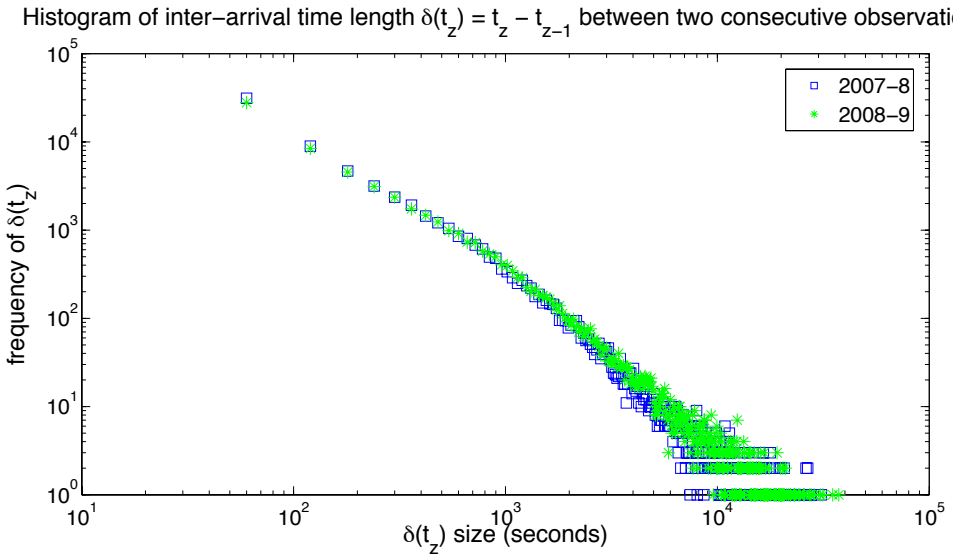


Figure 6.7: We calculate the time difference $\delta(t_z) = t_z - t_{z-1}$ between every valid pair of consecutive observations at each location in two data streams (seasons 2007–8 and 2008–9) and plot the histogram of those values on a logarithmic scale. The two data sets exhibit a very similar fat tail distribution, that approximates a power law with $\gamma \simeq 2.5$ for $\delta(t_z) > 10^3$. Further statistics are presented in Table 6.3.

2. $\ell_z = \ell_{z-1}$, take place at the same feeding location, as proximity is defined both in space and time.

3. $t_z - t_{z-1} \leq$ 12 hours, so that they belong to the same day. Antenna hardware is switched off during the night, so we ignore the artificially generated time difference in observation couples $\{\texttt{ID}_z, t_z, \ell_z\}, \{\texttt{ID}_{z-1}, t_{z-1}, \ell_{z-1}\}$ where $z-1$ took place e.g. Monday evening and z on Tuesday morning.

Pairs $\{\texttt{ID}_z, t_z, \ell_z\}, \{\texttt{ID}_{z-1}, t_{z-1}, \ell_{z-1}\}$ that satisfy the above will be termed from now on as *valid*.

The above statistical characteristics of our data streams, relating to inter-event time distribution and over-dispersion, are consistent with the notion of *burstiness*: that is, short periods of intense activity followed by long times of no or reduced activity [Goh and Barabási, 2008; Neuts, 1993]. There is a wide range of real-world systems that exhibit such property, from

human behavioural patterns in online communication or stock trading [Vázquez et al., 2006] to natural phenomena relating to earthquakes [Bak et al., 2002] or gene expression [Golding et al., 2005]. Additionally, the activity profile of bursts is a signature of the underlying system and directly relates to its function and behaviour [Goh and Barabási, 2008].

Based on the above and given the statistical findings presented earlier in this section, we formulate the following hypotheses:

- The bursts in the feeder observation profile correspond to small flocks of foraging individuals.
- Birds not only visit the feeder as part of such small flocks, but also have a preference to the members of the flock with whom they choose to forage.

We view such regions of increased observation density as K *gathering events* of socially affiliated birds. By clustering our data stream $\mathcal{D}$, we effectively group individuals $\texttt{ID}_z$ that appear in close temporal proximity (based on their arrival timestamp t_z) into the same gathering event k. Based on such a clustering, we seek to build a *bipartite network of individuals to gathering events* and define social connections between pairs of individuals based on the degree of co-occurrence in such foraging groups.

6.4.2 Clustering data streams: a simple and efficient algorithm

In clustering, a given data set $\{\mathbf{x}_n\}_{n=1}^{N}$ of N observations described by a D-dimensional feature vector $\mathbf{x} \in \mathbb{R}^{D\times 1}$, is partitioned to K modules. Grouping is performed based on some pre-defined notion of "similarity", or *inverse distance*, so that members of the same cluster are placed closer to each other than the rest of the data set. Note the direct correspondence between clustering and community detection, discussed in Chapter 2, with the key difference that data points here are described by their position on a D-dimensional space instead of their connectivity patterns through a graph structure $\mathcal{G}$.

In our case, the definition of similarity is based on the notion of temporal proximity, which is the time difference $|t_{z_1} - t_{z_2}|$ between any pair of observations $\{\texttt{ID}_{z_1}, t_{z_1}, \ell\}, \{\texttt{ID}_{z_2}, t_{z_2}, \ell\}$. We seek to extract a grouping where each tuple $\{\texttt{ID}_z, t_z, \ell\}$ is assigned to a particular gathering event [3], as seen in Fig. 6.6. Performing this clustering scheme is the most fundamental step towards our graph discovery goal, as uncovering the community structure of the observation stream leads to a natural relational structure in our data.

The sample statistics of $\delta(t_z)$, presented in Section 6.4.1, imply that most observations are positioned at close temporal proximity, while there is a significant presence of "gaps" (long periods of inactivity) in the data stream, as illustrated in Fig. 6.6. Let us employ a different view of the data, where we take each valid pair of consecutive tuples $\{\texttt{ID}_{z-1}, t_{z-1}, \ell_{z-1}\}, \{\texttt{ID}_z, t_z, \ell_z\}$ and plot its $\delta(t_z) = t_z - t_{z-1}$ in Fig. 6.8 as a bar with height equal to the corresponding time difference in seconds.

We can see in Fig. 6.8 that the inter-observation time profile consists of sequences of small $\delta(t_z)$ corresponding to distances of *intra-cluster points*, along with relatively larger $\delta(t_z)$ that denote *inter-cluster gaps*. As illustrated in Fig. 6.8, such a large $\delta(t_z)$ act as separators in the data stream and we use them to perform our clustering scheme. This can be summarised as a linear scan of $\mathcal{D} = \{\texttt{ID}_z, t_z, \ell_z\}_{z=1}^{Z}$, where for each valid pair of successive observations $\{\texttt{ID}_{z-1}, t_{z-1}, \ell_{z-1}\}, \{\texttt{ID}_z, t_z, \ell_z\}$ we calculate $\delta(t_z)$. If $\delta(t_z) < T_g$ (pre-defined threshold parameter), then assign $\{\texttt{ID}_{z-1}, t_{z-1}, \ell_{z-1}\}, \{\texttt{ID}_z, t_z, \ell_z\}$ to the same cluster. If not, create a new cluster with the first member being $\{\texttt{ID}_z, t_z, \ell_z\}$ and continue until $z = Z$ (end of data stream).

Following the above clustering scheme, we build a bipartite network described by an $N \times K$ incidence matrix where N individual birds are linked to K foraging groups. Each element b_{ik} denotes how many observations $\{\texttt{ID}_z, t_z, \ell_z\}$ of a particular individual $\texttt{ID}_z = i$ have appeared at each cluster k. The bird-to-bird social network can then be extracted via an appropriate *one-mode projection scheme*, which we discuss in Section 6.6.

[3] Note that we are currently focusing on tuples $\{\texttt{ID}_z, t_z, \ell\}$ that take place in the same location ℓ.

Algorithm 2 Gathering event identification via parametric clustering

Require: Data stream $\mathcal{D} = \{\texttt{ID}_z, t_z, \ell\}_{z=1}^{Z}$, with $\texttt{ID}_z \in \{1, ..., N\}$.
Require: Gap threshold T_g.
Ensure: $K = 1$ the initial number of clusters.
Ensure: `CM` an empty array of length Z.
Set $z = 1, i = 1, \texttt{CM[1]} = 1$
while $z < Z$ **do**
  $z := z + 1$
  if $\texttt{ID}_i \neq \texttt{ID}_z$ **then**
    $\delta(t_z) := t_z - t_i$
    if $\delta(t_z) < T_g$ **then**
      $\texttt{CM[z]} := K$
    else
      $K := K + 1$
      $\texttt{CM[z]} := K$
    end if
    $i := z$
  else
    $\texttt{CM[z]} := K$
  end if
end while
Allocate memory for $N \times K$ matrix $\mathbf{B}$.
for $z = 1, ..., Z$ **do**
  $n := \texttt{ID}_z$.
  $k := \texttt{CM[z]}$.
  $b_{nk} := b_{nk} + 1$.
end for
return Assignment array `CM`, where `CM[z]` the cluster id of observation z.
return $N \times K$ incidence matrix $\mathbf{B}$.

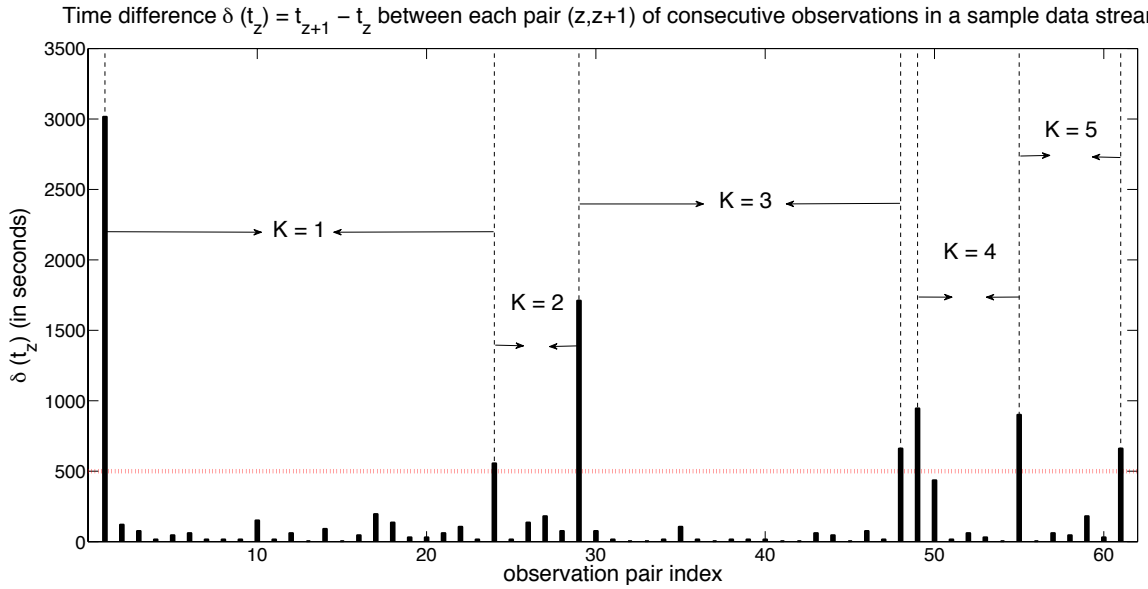


Figure 6.8: We plot the time difference for each pair of valid consecutive tuples as a bar of height $\delta(t_z) = t_z - t_{z-1}$. Our hypothesis is that gathering events take place between "gaps" or sufficiently large inter-observation times $\delta(t_z)$. For that reason, we perform clustering given a gap threshold T_g (horizontal dashed line) and group together observations that fall within such gaps. In our example, we used a threshold $T_g = 500$ seconds, which led to a partition with $K = 5$ clusters.

Such a methodology, illustrated in Algorithm 2, has the key advantage of linear scaling $\mathcal{O}(Z)$ to the number of observations Z, both in terms of computational and memory cost. Additionally, it does not require a priori knowledge of the effective cluster number K in the data, though it forces us to make a choice on the value of the gap threshold T_g.

Although such as approach alleviates us from the computational load of time window approaches, it still forces us either make a choice on the granularity with which we view the data, or perform multiple runs to infer an appropriate T_g given some performance function. In the next section, we propose a fully probabilistic treatment to our clustering problem, where through the use of Bayesian inference we automatically identify gathering events without having to commit to a particular a priori resolution. The scheme presented in the current section, will prove useful for initialisation and parameter tuning, due to its excellent computational scalability.

6.5 Bayesian gathering event extraction

6.5.1 A mixture model for data streams

Having defined a notion of similarity (or inverse distance), we assume a structure with K "centres of mass" around which data points are placed, as discussed in [Lloyd, 1982]. We illustrate those centroids in Fig. 6.6 as dashed vertical lines in the middle of each burst of foraging activity. This assumption provides an excellent intuition for the problem, where each centroid is viewed as a prototypical data point for a particular cluster, while the real data points are "corrupted" observations of that archetype, given a particular noise model.

Based on the above and given our data $\mathcal{D}$, where observations $\texttt{ID}_z$ are positioned on the one-dimensional space based on their timestamps, we model each t_z as a draw from a mixture of Gaussian distributions:

$$p(t_z|\mathbf{y}_z, \boldsymbol{\mu}, \boldsymbol{\beta}) = \prod_{k=1}^{K} \mathcal{N}(t_z; \mu_k, \beta_k^{-1})^{y_{zk}}, \tag{6.2}$$

where each t_z is associated with a latent K-dimensional binary vector $\mathbf{y}_z$ that encodes its membership to a particular cluster; there is only one $k_\star$ for which $y_{zk_\star} = 1$ and $y_{zk} = 0$ for $k \neq k_\star$, so that $\sum_{k=1}^{K} y_{zk} = 1$. Via $y_{zk_\star} = 1$, the model effectively "activates" the mean and precision $\mu_{k\star}, \beta_{k\star}^{-1}$, so that t_z is drawn from the $k_\star$-th Gaussian component.

Equation 6.2 implies that there are K "centres of mass" in the data stream, such as the ones in Fig. 6.6, around which observations are concentrated. Each k of those centroids is placed in the data stream with a timestamp μ_k. The precision term β_k controls the density of each gathering event, in terms of the temporal distance of observations around its centroid. As K is initially unknown, we start with a proposed value (for example Z, the maximum number of possible clusters) and define an appropriate model formulation that automatically "shrinks" to the effective number of mixtures required by the data.

Our choice of the Gaussian has both statistical and biological justification. From an in-

ference viewpoint, assuming a Gaussian structure per cluster is the approach that induces the least amount of structure (the most "non-committal") in the model, as it is the maximum entropy distribution [Bishop, 2007] for continuous spaces. Although timestamps t_z are measured in seconds, adopting a Poisson noise model (maximum entropy distribution for nonnegative quantities) makes our model less expressive, as there is no variance/precision term that would allow us to describe the density of each cluster (gathering event). Additionally, we seek to make the model "future-proof", as advances in antenna hardware might allow us to capture bird arrivals in fractions of the second.

From a biological standpoint, the shape of the Gaussian expresses our belief that during each event k the feeder is initially occupied by the most exploratory birds (alphas or "innovators"), which then trigger a gradual occupation by the rest of the flock. The overall feeder visitation reaches a "crescendo" around a given point in time μ_k and then it is being gradually deserted once all individuals have been fed. Such a pattern of tit feeding behaviour is also reported in the work of [Aplin et al., 2012], where it has been shown that the discovery of a food patch by a single individual triggers an information cascade within its flock. In such a process the feeder visitations gradually increase, as each bird that discovers the food source transmits this information to its network neighbours. The time point around which the feeder reaches the maximum number of occupants is viewed as the centroid μ_k of a gathering event k, from a data clustering perspective. The bird visitation timestamps are then assumed to be placed around such a mode μ_k under a Gaussian noise model, due to our ignorance of the mechanisms that drive the exact timing and sequence of individual appearances at the feeder.

From a practical perspective, a Gaussian mixture model with the appropriate latent variable structure allows us to perform clustering on the feeder data without having to commit to an a priori number K of gathering events. Such a framework, which we present in the next section, overcomes the difficulties associated with pre-specifying the resolution under which we view the data (defined as time window size in Section 6.3.1 and threshold parame-

ter in Section 6.4.2) by selecting the appropriate number K of flocks based on the statistical properties of the data stream.

Based on the latent structure of our data stream, described in Eq. (6.2), our goal is to infer:

1. The assignment $\mathbf{y}_z$ of each observation z to one of the K clusters or "events".

2. The effective number of clusters $K_\star$ in the data stream, given an initial value K for which $K_\star \leq K$.

3. The position μ_k of each one of their respective centroids, along with the "density" parameter β_k.

In the next section we propose an appropriate probabilistic model, which allows us to efficiently cluster the data stream and assign each observation to a gathering event.

6.5.2 Graphical model and prior structure

Consider the graphical model of Fig. 6.9, where the observed variable t_z denotes the timestamp of the z-th observation.

As we have already seen in Eq. (6.2), each t_z is dependent on a membership vector $\mathbf{y}_z$, along with the mean μ_k and precision β_k of the corresponding Gaussian component. From the graphical model of Fig. 6.9 and Eq. (6.2), we can write the likelihood of all of the timestamps, contained in a single vector $\mathbf{t} = [t_1, t_2, ..., t_Z]^\mathsf{T}$, as:

$$p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta}) = \prod_{z=1}^{Z} \prod_{k=1}^{K} \mathcal{N}(t_z; \mu_k, \beta_k^{-1})^{y_{zk}}, \tag{6.3}$$

where $\mathbf{Y}$ is a $Z \times K$ binary matrix and its z-th row is the membership vector $\mathbf{y}_z^\mathsf{T}$. Following the standard formulation in Gaussian mixture models [Bishop, 2007], for each $\mathbf{y}_z$ we place a prior:

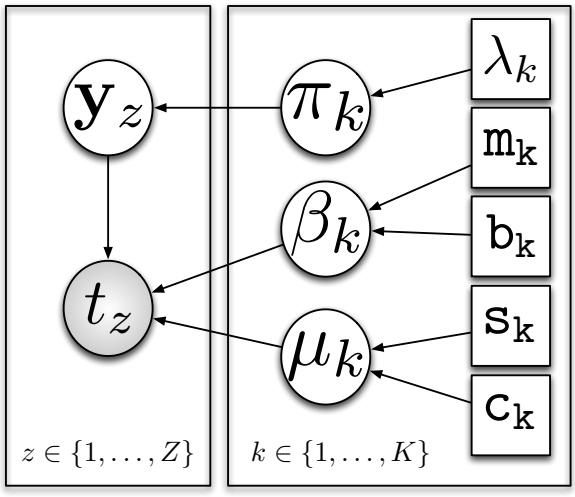


Figure 6.9: The graphical model expressing the generation of an observation t_z via a mixture of K Gaussians. The membership of t_z to a particular mixture is controlled by the latent $1 \times K$ binary vector $\mathbf{y}_z$.

$$p(\mathbf{y}_z|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{y_{zk}}, \tag{6.4}$$

which corresponds to a single draw from a multinomial distribution $\text{Multi}(\mathbf{y}_z; \boldsymbol{\pi}, 1)$ parameterised by a set of coefficients $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ for which $\pi_k \in [0, 1]$ and $\sum_{k=1}^K \pi_k = 1$, while the prior over all $\mathbf{y}_z$ is given by:

$$p(\mathbf{Y}|\boldsymbol{\pi}) = \prod_{z=1}^{Z} \prod_{k=1}^{K} \pi_k^{y_{zk}}. \tag{6.5}$$

Under this model, the expected membership score of point z to cluster k is given by the multinomial mean:

$$\mathbb{E}[y_{zk}|\boldsymbol{\pi}] = \pi_k. \tag{6.6}$$

Before seeing any data, all points z have the same expected membership score π_k for a given cluster k. This parameter π_k can be seen as an expression of the prevalence of each mix-

ture k in the data stream; it expresses an a priori bias of data points to belong to a particular cluster k.

As $\boldsymbol{\pi}$ is also an unknown stochastic parameter in our model, we need to describe it via an appropriate probabilistic form. Because $\pi_k \in [0,1]$ and $\sum_{k=1}^{K} \pi_k = 1$, we can view $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K}$ as a discrete probability distribution over k "outcomes". Based on that, we follow the standard modelling choice [Jaynes, 2003] of placing a Dirichlet prior $\text{Dir}(\boldsymbol{\pi}; \boldsymbol{\lambda})$:

$$p(\boldsymbol{\pi}) = \frac{1}{B(\boldsymbol{\lambda})} \prod_{k=1}^{K} \pi_k^{\lambda_k - 1}, \tag{6.7}$$

where $B(\boldsymbol{\lambda})$, the standard Beta function over $\boldsymbol{\lambda}$, acts as a normalisation term. The hyper-parameters $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^{K}$ can be seen as counts of "successes" associated with each "outcome" k, under the discrete probability distribution denoted by $\boldsymbol{\pi}$. In settings where no prior knowledge is available on the prevalence π_k of each cluster k, they can be set to a common starting value $\lambda_k = \lambda_0, \forall k \in \{1, ..., K\}$, thus providing a uniform prior over $\boldsymbol{\pi}$. In practice, we will make use of Algorithm 2, presented in Section 6.4.2, as a computationally inexpensive pre-processing step that will give us an initial value for each λ_k. More details on hyper-parameter initialisation are presented in Section 6.5.5.

Based on the graphical model of Fig. 6.9 and the data likelihood function in Eq. (6.3), the observations t_z are also controlled by the mean μ_k and precision β_k^{-1} of the k-th Gaussian component, which are also unknown stochastic variables. We consider each mean μ_k as a draw from another Gaussian distribution:

$$p(\mu_k) = \mathcal{N}(\mu_k; \mathtt{m}_k, \mathtt{b}_k^{-1}), \tag{6.8}$$

where $\mathtt{m}_k$ and $\mathtt{b}_k$ are the corresponding hyper-parameters for the mean and precision of μ_k. Assuming independence between each cluster k, the prior for all $\boldsymbol{\mu} = \{\mu_k\}_{k=1}^{K}$ is given by:

$$p(\boldsymbol{\mu}) = \prod_{k=1}^{K} \mathcal{N}(\mu_k; \mathtt{m}_k, \mathtt{b}_k^{-1}). \tag{6.9}$$

We have chosen a Gaussian prior over the mean μ_k of each cluster as it is conjugate to the Gaussian of each mixture $\mathcal{N}(t_z; \mu_k, \beta_k^{-1})$, thus allowing analytical tractability in our model. Although this choice is largely based on mathematical and computational convenience, it does not violate nonnegativity constraints for μ_k, assuming a reasonable initialisation of hyper-parameters $\mathtt{m}, \mathtt{b}$ as discussed in Section 6.5.5.

In the spirit of ensuring analytical tractability in our model and also ensuring the nonnegativity, we adopt the approach of [Roberts et al., 1998] by placing a Gamma prior over each precision term β_k:

$$p(\beta_k) = \text{Ga}(\beta_k; \mathtt{s}_k, \mathtt{c}_k), \tag{6.10}$$

where $\mathtt{s}_k, \mathtt{c}_k$ are the shape and scale parameters of the Gamma. Given the independence between the different k mixtures, the prior over $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^{K}$ is:

$$p(\boldsymbol{\beta}) = \prod_{k=1}^{K} \text{Ga}(\beta_k; \mathtt{s}_k, \mathtt{c}_k), \tag{6.11}$$

where the shape and scale hyper-parameters an initialised either arbitrarily or based on pre-processing scheme (Section 6.5.5).

From the probabilistic structure of Fig. 6.9 we have fully specified the dependencies between all variables and parameters of our mixture model. Additionally, in Eq. (6.3) we have specified our data likelihood function, where in equations (6.5), (6.7), (6.9), (6.11) we provided the functional form of the priors of our latent stochastic variables $\mathbf{Y}$, $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$. Based on this information, in the next section we present an inference scheme that discovers the appropriate observation-to-cluster assignment via a series of computationally inexpensive update equations.

6.5.3 Probabilistic inference via Variational Bayes

Based on the requirements of our clustering problem, the key inference task is to discover an appropriate allocation of observations t_z to clusters k, given the probabilistic dependency structure of Fig. 6.9, the likelihood function of Eq. (6.3), the priors of Eq. (6.5), (6.7), (6.9) and the observed data $\mathbf{t}$. In practical terms, that means finding the posterior distribution $p(\mathbf{Y}|\mathbf{t}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta})$ over the cluster assignments, where $\mathbb{E}[y_{zk}] = r_{zk}$ denotes the expected membership score of observation z to cluster k.

From the graphical model structure in Fig 6.9, we can see that the following factorisation holds:

$$p(\mathbf{t}, \mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\pi}) = p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta})p(\mathbf{Y}|\boldsymbol{\pi})p(\boldsymbol{\beta})p(\boldsymbol{\mu}). \tag{6.12}$$

Let us now apply the sum and product rules of probability on Eq. (6.12), in order to express the posterior $p(\mathbf{Y}|\mathbf{t}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta})$:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{t}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\pi})p(\mathbf{t}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\pi}) &= p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta})p(\mathbf{Y}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta}) \\ \Leftrightarrow p(\mathbf{Y}|\mathbf{t}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\pi}) &= \frac{p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta})p(\mathbf{Y}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta})}{p(\mathbf{t}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\pi})} \\ \Leftrightarrow p(\mathbf{Y}|\mathbf{t}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\pi}) &= \frac{p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta})p(\mathbf{Y}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta})}{p(\mathbf{t}|\boldsymbol{\mu}, \boldsymbol{\beta})p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta})}, \end{aligned} \tag{6.13}$$

where the enumerator on the right-hand side of Eq. (6.13), is given by the likelihood from Eq. (6.3) times the priors from equations (6.5), (6.7), (6.9) and (6.11). Due to the combinatorial summation over $\mathbf{Y}$ in the denominator:

$$
\begin{aligned}
p(\mathbf{t}|\boldsymbol{\mu},\boldsymbol{\beta})p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta}) &= \left(\sum_{\mathbf{Y}} p(\mathbf{t},\mathbf{Y}|\boldsymbol{\mu},\boldsymbol{\beta})\right) p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta}) \\
&= \left(\sum_{\mathbf{Y}} p(\mathbf{t}|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\beta})p(\mathbf{Y})\right) p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta}) \\
&= \left\{\sum_{\mathbf{Y}} \left[p(\mathbf{t}|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\beta}) \sum_{\boldsymbol{\pi}} p(\mathbf{Y}|\boldsymbol{\pi})p(\boldsymbol{\pi})\right]\right\} \times \\
&\quad p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\beta}), \qquad (6.14)
\end{aligned}
$$

the posterior $p(\mathbf{Y}|\mathbf{t},\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\pi})$ in Eq. (6.13) cannot be expressed in closed form. In order to avoid computationally expensive sampling techniques [Diebolt and Robert, 1994], we follow a different approach by employing a Variational Bayes (VB) [Bishop, 2007; Fox and Roberts, 2012] framework for approximating the posterior $p(\mathbf{Y}|\mathbf{t},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta})$.

In the typical VB setting we consider the marginal likelihood $p(\mathbf{t})$ of our data, where all latent variables $\{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}\}$ and parameters have been integrated out. To keep the notation uncluttered, we set $\mathcal{X} = \{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}\}$ and propose the following decomposition:

$$
\begin{aligned}
\ln p(\mathbf{t}) &= \mathcal{L}(q) + \mathrm{KL}(q(\mathcal{X})||p(\mathcal{X}|\mathbf{t})) \\
&= \int_{\mathcal{X}} q(\mathcal{X}) \ln \frac{p(\mathbf{t},\mathcal{X})}{q(\mathcal{X})} - \int_{\mathcal{X}} q(\mathcal{X}) \ln \frac{p(\mathcal{X}|\mathbf{t})}{q(\mathcal{X})} d\mathcal{X}, \qquad (6.15)
\end{aligned}
$$

where $q(\mathcal{X}) = q(\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta})$ is a new distribution over all latent stochastic variables in model. In Section A.1 we provide the derivations that show that this decomposition holds, along with more details on the theoretical aspects of Variational Bayes.

As the marginal likelihood of the data $p(\mathbf{t})$ is a fixed quantity given our model structure[4], by maximising the "free energy" term $\mathcal{L}(q)$ with respect to q, the divergence

[4]Recall from Chapter 3 that the marginal likelihood expresses the probability of the observed data when all latent variables and parameters have been integrated out.

$\mathrm{KL}(q(\mathcal{X})||p(\mathcal{X}|\mathbf{t}))$ vanishes, so that $q(\mathcal{X}) \rightarrow p(\mathcal{X}|\mathbf{t})$.

Because $q(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta})$ is an arbitrary joint distribution over the latent variables[5], we are allowed to express it in the form of the following factorisation:

$$q(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}) = q(\mathbf{Y})q(\boldsymbol{\pi})q(\boldsymbol{\mu})q(\boldsymbol{\beta}). \tag{6.16}$$

Minimisation of the divergence $\mathrm{KL}(q(\mathcal{X})||p(\mathcal{X}|\mathbf{t}))$ can now be expressed as a sequential maximisation of the free energy $\mathcal{L}(q)$ with respect to each one of the factors from Eq. (6.16) in turn. In Section A.1 (Appendix) and based on [Fox and Roberts, 2012], we show that the optimal expression for each factor for maximising $\mathcal{L}(q)$ is given by:

$$\ln q(\mathbf{Y}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}}[\ln p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{t})] + \text{const}, \tag{6.17}$$

$$\ln q(\boldsymbol{\pi}) = \mathbb{E}_{\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\beta}}[\ln p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{t})] + \text{const}, \tag{6.18}$$

$$\ln q(\boldsymbol{\mu}) = \mathbb{E}_{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\beta}}[\ln p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{t})] + \text{const}, \tag{6.19}$$

$$\ln q(\boldsymbol{\beta}) = \mathbb{E}_{\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\mu}}[\ln p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{t})] + \text{const}. \tag{6.20}$$

It is worth noting that we have provided no specification on the functional form of the factors $q(.)$, as it will arise naturally from the right hand side expressions above. More specifically, in Section A.1 we show that each one of the approximate posteriors of $\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}$ in Eqs. (6.17), (6.18), (6.19), (6.20), when exponentiated, has the same functional form (governed by the same type of distribution) as its corresponding prior $p(\mathbf{Y}|\boldsymbol{\pi}), p(\boldsymbol{\pi}), p(\boldsymbol{\mu}), p(\boldsymbol{\beta})$ in Eq. (6.5), (6.7), (6.9) and (6.11).

In particular, by expanding Eq. (6.17) and exponentiating both sides (see Section A.2 for the full derivation), the posterior over the membership variables $\mathbf{Y}$ (our inference objective) takes the form:

[5]The joint distribution $q(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta})$ is not part of the graphical model of Fig. 6.9, thus it is not required to conform with the factorisation $p(\mathbf{t}, \mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\pi}) = p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta})p(\mathbf{Y}|\boldsymbol{\pi})p(\boldsymbol{\beta})p(\boldsymbol{\mu})$ the model implies.

$$q(\mathbf{Y}) = \prod_{z=1}^{Z} \prod_{k=1}^{K} r_{zk}^{y_{zk}}, \tag{6.21}$$

where r_{zk} is often termed the *responsibility* [Bishop, 2007] of mixture k into explaining observation z and takes the form:

$$r_{zk} = \frac{\mathbb{E}_{\boldsymbol{\pi}}[\ln \pi_k] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\beta}}[\ln \mathcal{N}(t_z; \mu_k, \beta_k^{-1})]}{\sum_{\kappa} \{\mathbb{E}_{\boldsymbol{\pi}}[\ln \pi_\kappa] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\beta}}[\ln \mathcal{N}(t_z; \mu_\kappa, \beta_\kappa^{-1})]\}}. \tag{6.22}$$

Note that our posterior over $\mathbf{Y}$ in Eq. (6.21) has the same functional form as the prior in Eq. (6.5); it is an updated multinomial distribution $\mathbf{y}_z \sim \text{Multi}(\mathbf{y}_z; \mathbf{r}_z, 1)$ with $\mathbf{r}_z = \{r_{zk}\}_{k=1}^{K}$. They key difference between our prior and posterior is that the expected membership score of a given point z to cluster k is now data-point specific:

$$\mathbb{E}[y_{zk}|\mathbf{r}_z] = r_{zk}, \tag{6.23}$$

in contrast to $\mathbb{E}[y_{zk}] = \pi_k$ from the prior in Eq. (6.6).

This result shows that the expected membership score of z to k is not only dependent on an a priori bias towards a mixture k, in contrast to the prior case in Eq. (6.6), but also on the ability of k to produce observation t_z, as we can see in the second summand of the numerator in Eq. (6.22).

From Eq. (6.22) it is apparent that in order to have a fully defined expression for the membership scores r_{zk}, we need the expected values $\mathbb{E}[.]$ of the other stochastic variables π_k, μ_k, β_k in our model, which we need to extract from the corresponding posteriors in Eq. (6.18),(6.19),(6.20). As shown in Section A.2 (Appendix), by expanding each one of those expressions and exponentiating, we have:

$$q(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}; \lambda_k + Z_k), \tag{6.24}$$

$$q(\boldsymbol{\mu}) = \prod_{k=1}^{K} \mathcal{N}\left(\mu_k; \mathtt{m}_k \tilde{\beta}_k^{-1} \mathtt{b}_k + Z_k \bar{t}_k, \mathtt{b}_k + \tilde{\beta}_k Z_k\right), \tag{6.25}$$

$$q(\boldsymbol{\beta}) = \prod_{k=1}^{K} \text{Ga}\left(\beta_k; \mathtt{s}_k + Z_k, \mathtt{c}_k - Z_k \bar{t}_k \tilde{\mu}_k\right), \tag{6.26}$$

where we have defined the following auxiliary variables to simplify the notation:

$$Z_k = \sum_z r_{zk}, \tag{6.27}$$

$$\bar{t}_k = Z_k^{-1} \sum_{z=1}^{Z} r_{zk} t_z, \tag{6.28}$$

$$\tilde{\mu}_k = \mathbb{E}[\mu_k], \tag{6.29}$$

$$\tilde{\beta}_k = \mathbb{E}[\beta_k]. \tag{6.30}$$

As we can see in Eq. (6.21), (6.24), (6.25), (6.26), the posterior of each stochastic variable in our model is expressed in relation to the moments of the other variables. As these posterior equations maintain the original functional form from the priors, our inference process involves cycling through the following equations, which update the sufficient statistics[6] of each distribution:

[6]Sufficient statistics are the parameters required to fully describe a given distribution. For example, the sufficient statistics of a Gaussian distribution are its mean and variance, while for a Poisson distribution only its rate.

$$\lambda_k \leftarrow \lambda_k + Z_k, \tag{6.31}$$

$$\mathtt{m}_k \leftarrow \mathtt{m}_k \tilde{\beta}_k^{-1} \mathtt{b}_k + Z_k \bar{t}_k, \tag{6.32}$$

$$\mathtt{b}_k \leftarrow \mathtt{b}_k + \tilde{\beta}_k Z_k, \tag{6.33}$$

$$\mathtt{s}_k \leftarrow \mathtt{s}_k + Z_k, \tag{6.34}$$

$$\mathtt{c}_k \leftarrow \mathtt{c}_k - Z_k \bar{t}_k \tilde{\mu}_k, \tag{6.35}$$

where Eq. (6.31) results from Eq. (6.24), while Eq. (6.32) and (6.33) result from Eq. (6.25) and Eq. (6.34) and (6.35) from Eq. (6.26). The above equations, along with the calculation of the responsibilities from Eq. (6.22), constitute a single iteration in our inference process. In the full algorithm, which we illustrate in Algorithm 3, we perform those updates in an iterated scheme until there is no further improvement (under a precision threshold) of the free energy term $\mathcal{L}(q)$ in Eq. (6.15).

It is important to note that after such a clustering algorithm has been run on a given data stream $\mathcal{D}$, the mixture model parameters $\mathcal{X} = \{\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}\}$ can be used to reproduce the original visitation data. From the perspective of the inter-observation times $\delta(t_z)$ and their summary statistics (mean, variance and median values), we have found that data streams generated from the Gaussian mixture model approximate well the mean and median of $\delta(t_z)$ but underestimate its variance. So from one hand, each Gaussian component is able to approximate the centroid of its corresponding cluster, thus reproducing the high observation density within a given gathering event[7]. On the other hand, each Gaussian component has an infinite support over the real space and a non-zero probability of generating observations that fall within the inter-event "gaps". The presence of such observations leads to smaller values of $\delta(t_z) = t_z - t_{z-1}$, in cases where data points t_z and t_{z-1} belong to different clusters, thus

[7]Recall that in a Gaussian distribution, we expect 95.4% of the observations to lie within $[\mu - 2\sigma, \mu + 2\sigma]$, where μ the mean and σ the standard deviation.

yielding a smaller variance of $\delta(t_z)$ in the reproduced data stream. Such a result is shown in Table 6.3.

Data Set	$\bar{\mu}$	$\bar{\mu}$ (GMM)	σ^2	σ^2 (GMM)	$\bar{m}$	$\bar{m}$ (GMM)
2007–8	103.68	151.74	2.9×10^5	1.5×10^4	15	15
2008–9	158.49	188.29	5.41×10^5	9.3×10^4	30	30

Table 6.3: Sample mean, variance and median inter-arrival time $\delta(t_z)$ for the two different data sets under consideration. For each summary statistic, the first column represents the empirical values from the data while the second one (GMM) denotes the value from the reproduced data stream under a Gaussian mixture model fit. Inter-observation times are measured in seconds, while the results from the artificially generated data streams correspond to the average over 100 realisations.

6.5.4 Result of the clustering scheme

The main output of interest from the scheme presented in Section 6.5.3 is the responsibilities r_{zk}, which we encode in an $Z \times K$ observation-to-cluster *responsibility matrix* $\mathbf{R}$. Given an observation z, each row $[r_{z1}, r_{z2}, ..., r_{zK}]$ of $\mathbf{R}$ denotes the *expected membership score* of observation t_z across the K events. Columns of $\mathbf{R}$ representing clusters that hold no primary responsibility for any of the Z data points are pruned, so that for every given column k', there is at least one observation t_z for which $r_{zk'} > r_{zk},\ \forall\ k \neq k'$. Such a scheme allows us to to remove redundant clusters without having to impose an arbitrary numerical threshold on the responsibility values r_{zk}.

As a single bird i can be recorded at many time points t_z in the data stream, there can be many tuples $\{\texttt{ID}_z, t_z, \ell_z\}$ for which $\texttt{ID}_z = i$. Therefore, we can map the $Z \times K$ observation-to-cluster matrix $\mathbf{R}$ to an $N \times K$ bird-to-cluster matrix $\mathbf{B}$. We achieve this firstly by taking each row $\mathbf{r}_z = [r_{z1}, r_{z2}, ..., r_{zK}]$ of $\mathbf{R}$ and setting the largest element to 1 and all the others to 0, so that $\mathbf{r}_z \to \mathbf{r}'_z$ and $\mathbf{R} \to \mathbf{R}'$. Such a binarisation step on one hand provides a satisfactory approximation of $\mathbf{R}$, as it is based on our empirical observation that for every t_z there is one very dominant gathering event k, for which $r_{zk} \to 1$ and $r_{zk'} \to 0,\ \forall\ k \neq k'$. On the other

hand, it greatly facilitates the subsequent calculations at the rest of this chapter, as gathering event membership is quantified by integer-valued occurrences.

We proceed by taking each individual bird $i \in \{1, ..., N\}$ and identifying the subset $\mathcal{Z}_i$ of rows $\mathbf{r}'_z$ of $\mathbf{R}'$ that correspond to observations $\{\texttt{ID}_z, t_z, \ell_z\}$ for which $\texttt{ID}_z = i$. We set each row $\mathbf{b}_i$ of $\mathbf{B}$ as the sum $\mathbf{b}_i = \sum_{z \in \mathcal{Z}_i} \mathbf{r}'_z$. The resulting $N \times K$ incidence matrix $\mathbf{B}$ can be seen as a representation of a bipartite or two-mode network, where each element b_{ik} denotes the number of times bird i was observed at a specific foraging group [8] k.

It is worth noting that although we performed a "winner-takes-all" scheme for each row of the responsibility matrix $\mathbf{R}$, due to the one-to-many relationship between birds and observations, the incidence matrix $\mathbf{B}$ describes "soft membership", as each bird is allowed to belong to more than one foraging group, with varying participation weights. Such a bipartite network of birds to gathering events is the key structure extracted through our clustering scheme, which we use in Section 6.6 in order to extract the bird-to-bird social network.

In the next section we conclude this description of our probabilistic clustering model, by discussing various implementation and complexity issues.

6.5.5 Notes on implementation and initialisation

The inference scheme presented in Section 6.5.3 involved a series of consecutive updates on the model hyper-parameters, in order to approximate the true posterior distribution over its latent variables. In Algorithm 3 we have transcribed the formal model equations from Section 6.5.3 into a pseudocode format. The computational effort of the algorithm is dominated by the double loop over Z and K_0 in order to update the responsibilities r_{zk}, yielding a $\mathcal{O}(ZK)$ complexity.

The first initialisation requirement for Algorithm 3 is the provision of an estimation re-

[8] An alternative approach to representing the grouping of individuals to flocks can be implemented via the use of hypergraphs [Newman, 2010]. Under this framework, there exist N nodes and K hyperedges, with each hyperedge k linking all individuals that participate in the k-th gathering event.

Algorithm 3 Gathering event identification using Variational Bayes

Require: Data stream $\mathcal{D} = \{\mathtt{ID}_z, t_z, \ell\}_{z=1}^{Z}$, with $\mathtt{ID}_z \in \{1, ..., N\}$.
Require: K_0 the initial number of clusters.
Require: Initial values for hyper-parameters $\lambda_k, \mathtt{m}_k, \mathtt{b}_k, \mathtt{s}_k, \mathtt{c}_k,\ \forall k \in \{1, ..., K_0\}$.
Allocate memory for the $Z \times K_0$ responsibility matrix $\mathbf{R}$.
Update responsibilities r_{zk} using Eq. (6.22), based on initial values of $\lambda_k, \mathtt{m}_k, \mathtt{b}_k, \mathtt{s}_k, \mathtt{c}_k$.
while $\Delta\mathcal{L}(q) > \epsilon$ **do**
  for $k = 1, ..., K_0$ **do**
    Update $\boldsymbol{\pi}$ hyper-parameter λ_k using Eq. (6.31).
    Update $\boldsymbol{\mu}$ hyper-parameters $\mathtt{m}_k, \mathtt{b}_k$ using Eq. (6.32) and Eq. (6.33).
    Update $\boldsymbol{\beta}$ hyper-parameters $\mathtt{s}_k, \mathtt{c}_k$ using Eq. (6.34) and Eq. (6.35).
    Evaluate expectations $\tilde{\mu}_k, \tilde{\beta}_k, \tilde{\pi}_k, \mathbb{E}[\ln \beta_k], \mathbb{E}[\ln \pi_k]$ using the updated posteriors.
    for $z = 1, ..., Z$ **do**
      Update responsibilities r_{zk} using Eq. (6.22), based on the current hyper-parameter values $\lambda_k, \mathtt{m}_k, \mathtt{b}_k, \mathtt{s}_k, \mathtt{c}_k$.
    end for
  end for
  Evaluate free-energy gain $\Delta\mathcal{L}(q)$.
end while
Allocate memory for the $N \times K_0$ incidence matrix $\mathbf{B}$.
for $z = 1, ..., Z$ **do**
  $k_{\max} = \underset{k}{\operatorname{argmax}}\{r_{zk}\}_{k=1}^{K_0}$.
  Set $r_{zk_{\max}} := 1$ and $r_{zk} := 0,\ \forall k \neq k_{\max}$.
  Set $i := \mathtt{ID}_z$.
  Set $B_{i:} := B_{i:} + R_{z:}$ (add the z-th row of $\mathbf{R}$ to the i-th row of $\mathbf{B}$).
end for
Remove 0-valued columns of $\mathbf{B}$, reducing dimensionality from $Z \times K_0$ to $Z \times K$.
return $N \times K$ incidence matrix $\mathbf{B}$.

garding the number of clusters in the data stream. This choice is very important from an implementation perspective, as it directly affects the computational cost of the inference scheme via $\mathcal{O}(ZK)$. In settings where we have no prior knowledge on the number of gathering events K in the observation stream, a naive implementation would be to set $K_0 = Z$, which is the maximum possible number of clusters in the data. Such an approach would yield a quadratic computational cost $\mathcal{O}(Z^2)$, making the algorithm unscalable to large problems.

For that reason, we will make use of the efficient algorithm presented in Section 6.4.2 and Algorithm 2, where we can get an initial estimation of the cluster number K_0 in linear time. By setting the gap parameter T_g to a low value such as the mean of inter-observation times $\delta(t_z)$, we can get a "liberal" estimation of K_0, in the sense that it will always be $K_0 \geq K$ (allowing the model to shrink to the effective cluster number K) but not large enough $K_0 \simeq Z$ so that it makes the model unscalable.

Using the same rationale, we can use the clustering result provided by Algorithm 2 in order to initialise the model hyper-parameters $\lambda_k, \mathtt{m}_k, \mathtt{b}_k$ as:

- λ_k: the number of observations per cluster k, that is $\lambda_k = \sum_{z=1}^{Z} \texttt{CM[z]}\,\delta(\texttt{CM[z]}, k)$ where `CM[z]` the output of Algorithm 2 and $\delta(.,.)$ is the Kronecker delta,
- $\mathtt{m}_k$: the mean of all timestamps t_z, for which $\texttt{CM[z]} = k$,
- $\mathtt{b}_k$: the variance of all timestamps t_z, for which $\texttt{CM[z]} = k$,

while the hyper-parameters of the Gamma $\mathtt{s}_k, \mathtt{c}_k$ are set to uninformative priors $\mathtt{s}_k = 10^3, \mathtt{c}_k = 10^{-3}$, as in [Bishop, 2007].

Initialising Algorithm 3 using the scheme described above allows us to encode our initial belief over the gathering event structure of the data stream, based on the preliminary analysis we performed using Algorithm 2. Although the same data stream $\mathcal{D}$ is used to initialise and update the priors, such a scheme is deemed appropriate for the following reason: it provides a "better than random" starting point for Algorithm 3, thus allowing us to avoid

the quadratic computational cost $\mathcal{O}(Z^2)$ of setting $K_0 = Z$, along with random values for the hyper-parameters $\lambda_k, \mathtt{m}_k, \mathtt{b}_k$. This two-step clustering approach is analogous to a two-step community detection scheme, where a linear-time algorithm such as Label Propagation [Raghavan et al., 2007] is used to provide an initial partition of a network, in a computationally efficient manner. This solution is subsequently refined by the more computationally demanding Kernighan Lin algorithm [Kernighan and Lin, 1970], which performs such a fine-tuning with a $\mathcal{O}(N^3)$ computational cost [Newman, 2010]. In both cases, the overall performance is based on a reasonable parameterisation of the algorithm in the first step. For example, in Algorithm 2 we would not pick an inappropriate value for the T_g threshold, such as the time length of the experiment, as it will result to a $K_0 = 1$ value for the maximum number of clusters in the data.

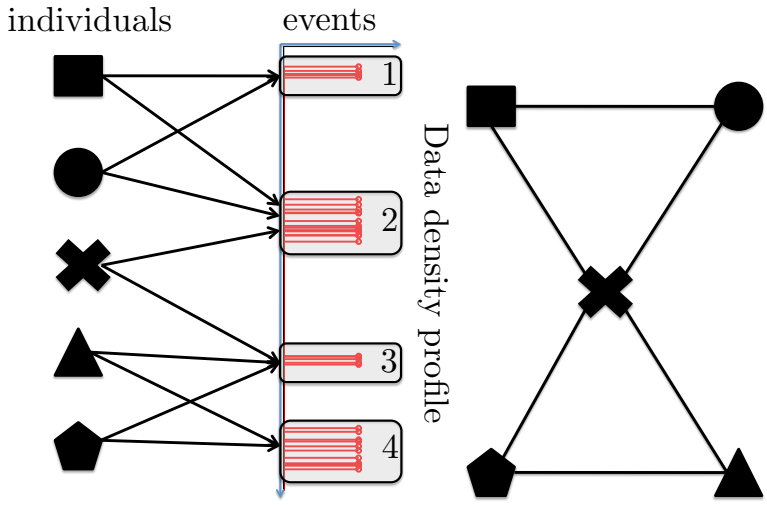
The Gaussian mixture model from Algorithm 3 combined with Algorithm 2 as a pre-processing step constitutes a single computational methodology, which from now on we call the "Gathering-Event model", or GEM for short.

6.6 Building the social network

By making use of the clustering models developed in Sections 6.4.2 and 6.5, we are now able to automatically extract gathering events from a given data stream $\mathcal{D}$ and describe the membership of individuals to such flocks via a bipartite network. Such a bird-to-event graph is an important finding by itself as it provides us an insight into the temporal grouping structure of the population under consideration. For the purposes of this section, we focus our attention on the more theoretical problem of extracting graph structure from spatio-temporal data.

Having extracted the bird-to-event bipartite graph from $\mathcal{D}$, we seek to move one step further and extract a bird-to-bird social network, based on the mutual participation of individuals to gathering events. This is accomplished via an appropriate one-mode projection $\mathbf{B} \in \mathbb{R}^{N \times K} \rightarrow \mathbf{A} \in \mathbb{R}^{N \times N}$, shown in Fig. 6.10, so that a link a_{ij} between a pair i, j in the

resulting network will express how frequently the two birds forage together.



(a) The bird-to-event bipartite network $\mathbf{B} \in \mathbb{R}^{N \times K}$ (b) The bird-to-bird social network $\mathbf{A} \in \mathbb{R}^{N \times N}$

Figure 6.10: Our method identifies gathering events from the bursts in our observation stream as seen in Fig. 6.10(a). Then individuals are assigned to such events creating a bipartite network. In Fig. 6.10(b) we recover the bird-to-bird social network, via an appropriate one-mode projection, based on the co-participation of individuals to these events.

There is a wealth of literature regarding one-mode projection methods, addressing both graph-theoretic [Zweig and Kaufmann, 2011] and application-side [Lambiotte and Ausloos, 2005; Newman, 2001a; Zhou et al., 2007] problems. In our framework, we have considered the following approaches:

1. Association indices. One of the most attractive properties of the gathering events model is that it is compatible with the association indices [Whitehead et al., 2005], which are extensively used in Animal Behaviour studies; individuals are being observed in groups across "sites" and a affiliation score is calculated based on various parameters that account for frequency and exclusivity of co-occurrences.

2. Bayesian One-Mode Projection (BOMP) methodology, which provides a probability distribution over the presence of each link. The method is based on a) an appropriate

noise model for the co-occurrence counts, b) a measure that quantifies the exclusivity of co-occurrences and c) a Bayesian update scheme that incorporates knowledge from past data. This model will be formally presented in Chapter 7.

3. Co-occurrence graph, where links a_{ij} in the projected network are weighted by the number gathering events at which both i and j participate.

For the purposes of this chapter, we will focus on the 3rd approach, while a discussion on association indices and Bayesian One-Mode Projection is presented in Chapter 7.

Following the above discussion, we project the $N \times K$ bird-event matrix $\mathbf{B}$ to the $N \times N$ bird-bird adjacency matrix $\mathbf{A}$ using the following scheme: first, we binarise $\mathbf{B}$ so that $b_{ik} \in \{0, 1\}$ simply denotes if individual i occurred in event k or not. Then, we perform $\mathbf{A} = \mathbf{B}\mathbf{B}^{\mathsf{T}}$ so that a_{ij} is the total number of gathering events where i, j co-appeared.

Although such a projection omits important information given to us by the weighted links b_{ik} of the incidence matrix, it gives rise to an intuitive interpretation of the one-mode network weights as integer-valued co-occurrences. This allows us to perform comparisons of this method against traditional time-window approaches, as we shall see in Section 6.7.

6.6.1 GEM as a clique-rolling process

In this section we briefly comment on how GEM discovers the underlying network structure, implied by the data stream $\mathcal{D}$, by performing multiple partial observations of the whole graph, where each observation corresponds to an N_k-clique.

Let us assume that the ground-truth network is known to us, and one of its communities is shown in Fig. 6.11. We also have the corresponding data stream, shown in Fig. 6.12, where our algorithm has identified various gathering events. Based on the one-mode projection scheme discussed in Section 6.6, all individuals that participate in the same event are connected. Therefore, each event k with N_k members corresponds to an N_k- clique in the resulting network. Due to the fact that many gathering events in the data stream can have

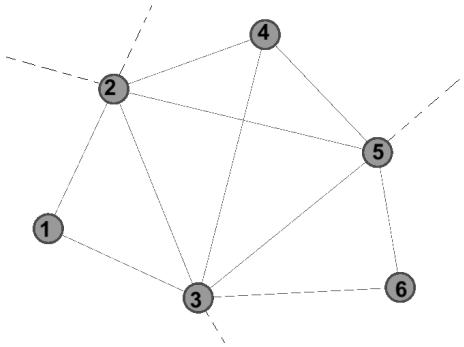


Figure 6.11: An example community of 6 nodes. Dashed lines represent links that allow connections with the rest of the network.

common members (as individual birds do not have a fixed set of companions that join them during every single feeder visitation) our model naturally extracts a series of fully-connected subgraphs that share nodes. The whole network is then reconstructed as an aggregation of such partially overlapping or adjacent cliques.

Community structure in such a process arises naturally, as there can be collections of fully-connected subgraphs that share members. In fact, [Palla et al., 2005] have shown that network communities can be seen as aggregations of adjacent cliques. The way our algorithm in Fig. 6.12 performs multiple "partial observations" of the community in Fig. 6.11 (one per gathering event), corresponds to what the authors in [Palla et al., 2005] describe as a "clique rolling process", which they use to identify communities.

From an application perspective, in a data stream such as the one shown in Fig. 6.12 there can be individuals that appear in gathering events coincidentally, without having any social connection with any of the other members. Those co-occurrences need to be identified and removed based on an appropriate null model. It is worth noting that the issue of deciding if co-occurrences are due to social affiliation or coincidence is not a drawback of our approach, or any graph discovery scheme presented in this chapter, but an inherent challenge of our data collection scheme. The advantage of our gathering events approach is that it allows us to build an appropriate null model, along with a computationally efficient significance test for

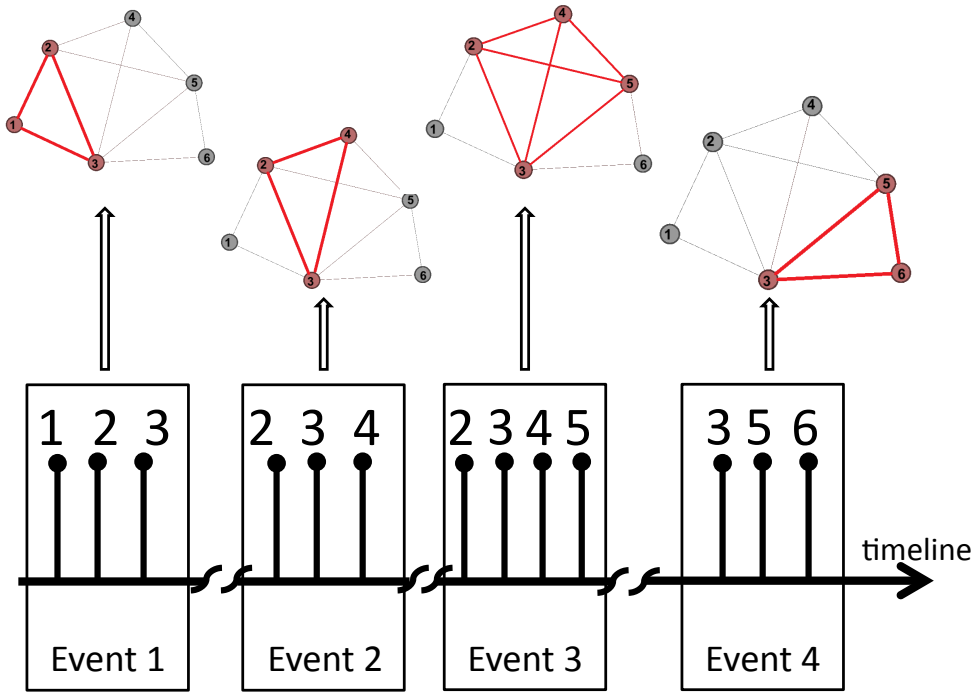


Figure 6.12: An example data stream of 4 gathering events. Each gathering event with k participants corresponds to a k-clique (fully connected subgraph with k nodes), which can be seen as a "partial view" of the overall community.

the observed link weight, which we present in the next section.

6.6.2 Co-occurrences: social tie versus coincidence

The next issue we seek to address is the statistical significance of the extracted link weights. Building the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ in the manner described in Section 6.6 makes the very strong assumption that if two individuals participate in the same gathering event, they have some form of social affiliation. This assumption, known in the animal social network literature as the *Gambit of the Group* (GoG) [Whitehead and Dufault, 1999], may lead us to adjacency matrices encumbered with "junk" links, produced by co-occurrences that happened by chance. Such coincidences are also frequent in settings where there are natural peak-hours

in the data collection period and also when the sensor hardware act as attraction points, as for example the bird feeders in our study. Hence, we seek to define an appropriate null model that describes how "statistically surprising" a given link weight would be, if there was no underlying social preference in the foraging habits of the bird population. In previous sections we have discussed that observations occur in bursts (as shown in Fig. 6.6) that imply small foraging groups of birds that arrive together at the feeders. This is captured by the bird-to-event matrix $\mathbf{B} \in \mathbb{R}^{N \times K}$, where each element b_{ik} in the row vector $\mathbf{b}_i$ expresses the number of times bird i appeared at the gathering event k.

Consider each row vector $\mathbf{b}_i$ as a draw from a multinomial distribution $\mathcal{M}(n_i, \mathbf{p}_i)$, with parameters $n_i = \sum_{k=1}^{K} b_{ik}$ and $p_{ik} = b_{ik}/n_i$. The values of the parameter vector $\{p_{ik}\}_{k=1}^{K}$ can be viewed as a *preference profile* of a bird i to each foraging event k. If our hypothesis that social affiliation between birds affects event membership holds, then closely interacting birds i, j will have similar preference profiles $\mathbf{p}_i$ and $\mathbf{p}_j$.

Let us now propose an element shuffling σ of $\mathbf{p}_i$ so that $\mathbf{p}_i \to \sigma(\mathbf{p}_i)$ and draw a new event occurrence vector $\mathbf{b}_i^{(0)}$ from the multinomial distribution $\mathcal{M}(n_i, \sigma(\mathbf{p}_i))$. Performing this permutation and sampling scheme independently for all birds $i \in \{1, ..., N\}$ leads to a new bird-to-event bipartite network described by $\mathbf{B}^{(0)} \in \mathbb{R}^{N \times K}$. This new matrix $\mathbf{B}^{(0)}$ preserves many key characteristics of the original data, among them the event membership structure, because bird appearances remain concentrated in K regions of increased observation density. Quantities such as the number of individuals N, and the total records n_i, of bird i in the data are also retained.

The key difference introduced in $\mathbf{B}^{(0)}$ is that, although a bird's uneven participation preference $\mathbf{p}_i$ across foraging groups is preserved (as the permutation $\sigma(\mathbf{p}_i)$ has the same entropy as $\mathbf{p}_i$), the shuffling σ "breaks" all correlations between $\mathbf{b}_i$ and $\mathbf{b}_j$ induced by latent social affiliation between individuals i and j. In other words, under our null model birds still forage in small groups, but with *no social preference to which other members of the group with*

whom they will forage. We repeat this process R times and for each generated bird-to-event matrix $\mathbf{B}^{(0)}$ we extract the bird-to-bird matrix $\mathbf{A}^{(0)}$ using the "conventional" one-mode projection presented in Section 6.6. By generating multiple instances of $\mathbf{A}^{(0)}$ in this manner, we are effectively drawing samples from the ensemble or family of graphs $\mathcal{G}^{(0)}$ that contains all possible network configurations generated by the null model. Our goal is to examine if our observed network $\mathbf{A}$ is an unlikely case of $\mathcal{G}^{(0)}$.

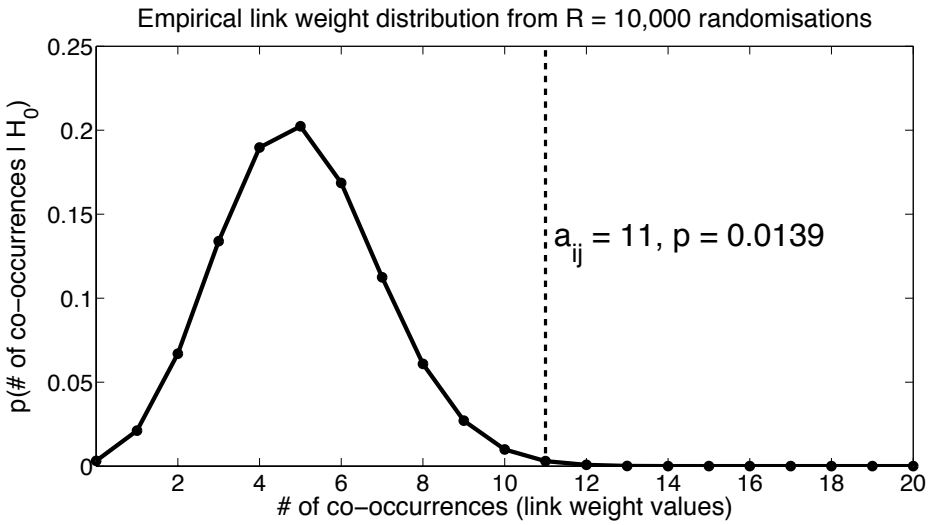


Figure 6.13: We generate $R = 10,000$ draws from the null ensemble $\mathcal{G}^{(0)}$ and given a pair i, j of individuals, we draw a histogram of their co-occurrence values under the null hypothesis. We then examine if the observed co-occurrences a_{ij} (i.e. the link weight in the extracted graph) are a special case of the null model. In the case presented above, observing a link weight value as extreme as a_{ij} (vertical dashed line) under the null hypothesis yields a probability mass below our significance threshold, thus the link a_{ij} is retained.

The randomisation process generates R values of the weight of each link between i and j. From the histogram we get the empirical distribution $p(a_{ij}|H_0)$, presented in Fig. 6.13, which denotes the probability of having a link of weight a_{ij} given that the null hypothesis H_0 holds. We examine the likelihood of each observed co-occurrence a_{ij} by performing a hypothesis test, given an appropriate significance level α, by examining the likelihood $p = p(x \geq a_{ij}|H_0)$ of co-occurrences as large as a_{ij}. Note that the key point of a null model is that co-occurrences happen between individuals, but not as a result of an under-

lying social structure. In other words, the links in $\mathbf{A}^{(0)}$ are independent under H_0, hence $p(\mathbf{A}|H_0) = \prod_{ij} p(a_{ij}|H_0)$. Thus our significance test lies in examining how well this independence assumption can explain the observed co-occurrences encoded in each link of $\mathbf{A}$.

6.6.3 Integrating information from multiple locations

We briefly expand our graph inference scheme to the multi-location setting. For each record $\{t_z, \texttt{ID}_z, \ell_z\}$ in our data stream, we now have an additional term $\ell_z \in \{1, ..., L\}$ that denotes the index of the location where observation z took place.

We start by segmenting our data $\mathcal{D} = \{t_z, \texttt{ID}_z, \ell_z\}_{z=1}^{Z}$ into L streams, so that each $\mathcal{D}^{(\ell)}$ contains records referring only to location ℓ. For each $\mathcal{D}^{(\ell)}$ we perform the gathering event extraction process as presented in Section 6.4.1 leading to L incidence matrices $\mathbf{B}^{(\ell)} \in \mathbb{R}_{(+)}^{N_\ell \times K_\ell}$, where $N_\ell \leq N$ is the subset of birds recorded at location ℓ. As we consider all events independent, we concatenate all $\mathbf{B}^{(\ell)}$ into a global $N \times K_g$ incidence matrix $\mathbf{B}$:

$$\mathbf{B} = [\mathbf{B}^{(1)}|\mathbf{B}^{(2)}|...|\mathbf{B}^{(K)}]$$

with $K_g = \sum_{\ell=1}^{L} K_\ell$. One-mode projection and significance tests are then performed in exactly the same manner as presented in Section 6.6 and 6.6.2.

6.7 Results

In this section we compare the various graph discovery algorithms presented in this chapter against a particular class of benchmark data streams with observed network structure. We start in Section 6.7.1 by proposing an algorithm that, given a graph topology, generates data streams similar to the one presented in Section 6.2. We then proceed in Section 6.7.2 by using such artificial problems in order to assess how well various topological characteristics of the ground-truth graphs are being recovered by the link discovery algorithms we have presented.

6.7.1 Benchmark data stream generator

Consider the discussion of Section 6.6.1, where our link discovery methodology can be seen as a sequence of partial observations of the underlying graph and each observation corresponds to a clique, so that connected individuals are placed in close spatio-temporal proximity. Based on this discussion we build our data stream $\mathcal{D}$ generator in the following way; from a given graph described by $\mathbf{A} \in \mathbb{R}_{(+)}^{N \times N}$, we extract its collection of cliques $\{\mathcal{C}_u\}_{u=1}^{U}$, where $\mathcal{C}_u = \{i, j, ...\}$ a set containing the node indices of all members $\{i, j, ...\}$ of the u-th clique. We then proceed by picking uniformly at random a clique $\mathcal{C}_u$ and placing the $|\mathcal{C}_u|$ nodes "close" to each other, given a temporal distance drawn from a Poisson distribution with rate $\delta_{in}(t_z)$. Once all $|\mathcal{C}_u|$ nodes of the u-th clique have been placed, we add an inter-cluster gap of expected length $\delta_{out}(t_z)$ (again under a Poisson random model) and proceed iteratively by picking another clique. The algorithm is finalised once the desired number Z of observations have been added to $\mathcal{D}$.

The above scheme is flexible enough both in mimicking the modular structure of real-world data and in generating uniformly-spaced observation streams. For the first case, by setting $\delta_{in}(t_z) \ll \delta_{out}(t_z)$, well-connected groups of individuals are positioned in close temporal proximity, while the corresponding inter-group distances are large, as in the case shown in Fig. 6.6. On the other hand, we can set $\delta_{in}(t_z) \simeq \delta_{out}(t_z)$ so that the algorithm produces a data stream where observations are placed apart using the same mean temporal distance.

We also consider the case discussed in Section 6.6.2, where gathering events can be "contaminated" by individuals that participate without having any sort of social affiliation to the other members. In order to emulate such coincidental co-occurrences in our artificial data streams, we introduce a noise term `NOISE` that expresses the probability of adding a random node in the gathering event, instead of a member of clique $\mathcal{C}_u$.

All of the steps of the above process are presented in Algorithm 4, which we will use in the next section in order to perform comparisons between gathering event and time window

Algorithm 4 Benchmark data stream generator

Require: Ground-truth graph, described by $N \times N$ adjacency matrix $\mathbf{A}$.
Require: Data stream length Z.
Require: Mean intra-cluster distance $\delta_{in}(t_z)$, mean inter-cluster distance $\delta_{out}(t_z)$.
Require: Noise level $\texttt{NOISE} \in [0, 1]$.
Extract all U cliques $\{\mathcal{C}_u\}_{u=1}^{U}$ from $\mathbf{A}$, where $\mathcal{C}_u$ a clique of size $|\mathcal{C}_u|$.
Initialise empty data stream $\mathcal{D} = \{\}$.
Set $t := 1$ and $z := 0$.
while $z < Z$ **do**
  Pick uniformly at random a clique $\mathcal{C}_u$ from $\{\mathcal{C}_u\}_{u=1}^{U}$.
  for $n = 1 : |\mathcal{C}_u|$ **do**
    $z := z + 1$.
    Draw a random value $\texttt{RAND} \in [0, 1]$.
    if $\texttt{RAND} \geq \texttt{NOISE}$ **then**
      $\texttt{ID}_z :=$ n-th member of $\mathcal{C}_u$.
    else
      $\texttt{ID}_z :=$ a node i uniformly at random, for which $i \notin \mathcal{C}_u$.
    end if
    Add tuple $\{\texttt{ID}_z, t, \ell\}$ to data stream $\mathcal{D}$.
    if $n < |\mathcal{C}_u|$ **then**
      $t := t + \text{Pois}(\delta_{in}(t_z))$.
    else
      $t := t + \text{Pois}(\delta_{out}(t_z))$.
    end if
  end for
end while
return Benchmark data stream $\mathcal{D} = \{\texttt{ID}_z, t_z, \ell\}_{z=1}^{Z}$.

based graph extraction methodologies. Note that we are generating visitations to a single location ℓ, for the purposes of simplicity. Our testing process can be trivially expanded to the multi-site case, with one run of Algorithm 4 per location ℓ and aggregation of results to a single $\mathcal{D}$.

6.7.2 Method comparison

In this section we will use Algorithm 4 in order to generate artificial data streams given a ground-truth graph, and assess how well different methodologies extract the underlying network structure of interest.

As the computational load of Algorithm 4 is dominated by the NP-complete clique extraction scheme [Karp, 2010], in order to perform efficient data stream generation we have to use sufficiently small ground-truth graphs. For our testing purposes, we have used the Girvan-Newman (GN) random graphs [Girvan and Newman, 2002], which have $N = 128$ nodes and a fully observed community structure of $C = 4$ modules with $N_c = 32$ members each. Although such graphs do not capture the properties of real-world networks (in terms of size, degree distribution, etc), this is not an issue in our application; our goal is to assess how well each graph discovery scheme recovers a given ground-truth adjacency matrix.

For our tests, we consider three performance measures, which assess the quality of extracted graphs at three scales: i) link-level, ii) local neighbourhood and iii) mesoscopic scale.

The first is the mean square error (MSE) $\mathcal{L}^2$ of the *binarised* adjacency matrix $\breve{\mathbf{A}}^{(m)}$ extracted from method m, versus the ground-truth one $\mathbf{A}$. This is defined as:

$$\mathcal{L}^2(\mathbf{A}||\breve{\mathbf{A}}^{(m)}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (a_{ij} - \breve{a}_{ij}^{(m)})^2, \tag{6.36}$$

where $\breve{\mathbf{A}}^{(m)}$ denotes the binarised adjacency matrix with elements $\breve{a}_{ij}^{(m)}$ extracted from data stream $\mathcal{D}$ by a given method m (e.g. GEM or time-window method). The double summations run only on the top diagonal parts of each matrix, as we are working with undirected graphs.

It is important to note that as the ground-truth graphs are unweighted, the reason we binarised $\mathbf{A}^{(m)}$ is to perform comparison only on topological basis.

The second quality measure is the MSE between the observed and extracted clustering coefficients. We consider a vector $\mathbf{c} \in \mathbb{R}^{N\times 1}$ where each element c_i denotes the clustering coefficient of node-i, calculated using the approach presented in [Barrat et al., 2004]:

$$c_i = \frac{1}{d_i(d_i-1)} \sum_{j,k} \check{a}_{ij}\check{a}_{ik}\check{a}_{jk}. \tag{6.37}$$

Based on the above, we can now define the following error measure:

$$\mathcal{L}^{CC} = N^{-1}||\mathbf{c} - \mathbf{c}^{(m)}||_2^2, \tag{6.38}$$

where $||\mathbf{c} - \mathbf{c}^{(m)}||_2$ the Euclidean norm of the vector difference $\mathbf{c} - \mathbf{c}^{(m)}$. Such an error measure allows us to assess how well algorithm m extracts the local link density around each node neighbourhood. As in the previous case, we have considered binarised versions of the extracted adjacency matrices $\mathbf{A}^{(m)}$, in order to calculate the clustering coefficients.

The third measure under consideration is the quality of community structure of the extracted networks, versus the ground-truth one. This is achieved by performing community detection on $\mathbf{A}^{(m)}$ using the nonnegative matrix factorisation algorithm presented in Chapter 4 and compare the solution against the ground-truth one, using Normalised Mutual Information (NMI) [Danon et al., 2005]. This measure allows us to assess how well each method m discovers the community organisation of a ground-truth network.

Let us begin our assessment by considering a noiseless case (by setting $\texttt{NOISE} = 0$ in Algorithm 4), where all intra-cluster members form a clique in the ground-truth network. We compare the performance of our gathering events (GEM) based approach against the traditional time window (TW), in the three measures defined above, across different levels of data stream "burstiness"; starting with an average intra-cluster distance $\delta_{in}(t_z) = 30$, we define $\delta_{out}(t_z) = \texttt{r}\delta_{in}(t_z)$ where we allow $\texttt{r}$ to vary between 1 and 10. Large $\texttt{r}$ values denote

that the data stream consists of a series of observation-dense regions separated by large gaps, while values close to 1 denote that we have evenly spread records. For the TW approach, due to lack of any prior knowledge, we pick the "optimal" time-window based on the scheme we proposed in Section 6.3.3. For every different value of $\delta_{out}(t_z)/\delta_{in}(t_z)$, we have generated 100 instances of $\mathcal{D}$, run GEM and TW on each instance, and took the mean and standard deviation of each performance measure across those instances. For the generation of NG graphs we used a d_{in} parameter value of 15, yielding a "crisp" group structure, so that we avoid errors that arise from the community detection algorithm itself.

In Fig. 6.14(a) we compare TW versus GEM on the MSE of the recovered adjacency matrix, starting from a very modular data stream ($\delta_{out}(t_z)/\delta_{in}(t_z) = 10$) and moving to a completely uniform one ($\delta_{out}(t_z)/\delta_{in}(t_z) = 1$). In Fig. 6.14(a) we can see that our GEM approach yields a consistently lower MSE on the recovered adjacency matrix (based on Eq. (6.36)) than the TW approach, while it starts to fail when the data stream becomes uniform. Similar results are given, in Fig. 6.14(b), for the MSE of the recovered clustering coefficient. We find that TW approaches underestimate the local density of the ground-truth network as they consider, regardless of the set time window value, consistently shorter "interaction zones" between the data points in $\mathcal{D}$, thus omitting important links.

We also consider the quality of community structure in the networks extracted by GEM and TW across different levels of data stream modularity. In the same experimental setting as before, we consider different values of $\delta_{out}(t_z)/\delta_{in}(t_z)$ and for each one we extract the adjacency matrix using GEM and TW. We then run our community detection algorithm against those matrices and compare the inferred groups against the ground-truth ones of NG. As we can see in Fig. 6.14(c), the GEM approach yields a consistently higher NMI event in settings with low data stream burstiness. The similarity in terms of performance between GEM and TW for the initial $\delta_{out}(t_z)/\delta_{in}(t_z)$ values can ben explained from the fact that this community-based measure considers a more macroscopic view of the network (in contrast to the MSE

on links and clustering coefficients), without being sensitive to dissimilarities in local node neighbourhoods.

We continue our analysis by considering the more realistic case of noisy data streams. We consider a range of NOISE values, from 0 to 0.6, that express different levels of co-occurrence between unaffiliated individuals in the data stream. For the purposes of this test, we have included the extended version of our GEM approach, presented in Section 6.6.2, where a significance test of is performed in order to prune noisy links in the inferred graph. For the test, we used $R = 1000$ samples of the null network in order to calculate the empirical distribution of link probability between nodes i, j. We then maintain, or prune, each observed link a_{ij} based on how well it can be explained by the null hypothesis, under a significance threshold of 0.05. This approach is termed GEM(N), where N denotes that the algorithm is "null-hypothesis aware". For our testing purposes we have considered an expected intra-cluster distance of $\delta_{in}(t_z) = 30$, along with a gap distance of $\delta_{out}(t_z) = 90$, so that the data stream has a gathering event structure.

In Fig. 6.15(a) we plot the MSE of the extracted adjacency matrix versus the ground-truth one from the GN random graph. As a first observation, we can see that both GEM and TW perform worse than the noiseless case, as we increase NOISE above 0. An interesting point is that, in contrast to the noiseless case, TW performs better than GEM for NOISE > 0.2 as it considers shorter interaction zones, making it less prone to add links between unaffiliated individuals. Most importantly, the null-model enabled GEM(N) yields the best performance as it manages to skip co-occurrences that result from noise in the data stream. Good performance is also being achieved on the quality of recovered community structure, as shown in Fig. 6.15(c), although we have found no evidence of strongly statistically significant improvement of GEM(N) over GEM.

An interesting behaviour is presented in Fig. 6.15(b), where although both GEM and GEM(N) outperform TW, the "noise-naive" GEM yields a better performance for large noise

values in terms of recovering the correct clustering coefficient. This is a compromise we are making, when having a removal mechanism for co-occurrences that are being deemed statistically insignificant; there are cases of false negatives and "real" links are being pruned, thus reducing the clustering coefficient in the inferred graph.

In Fig. 6.16, we expand the above tests by considering various noise and data stream modularity values in the same benchmark. In Fig. 6.16(a) we show the performance of GEM across a range of benchmark test parameters $\delta_{out}(t_z)/\delta_{in}(t_z)$ and `NOISE` level, while in Fig. 6.16(b) we perform a comparison against the TW method. Each point on the colormaps of Fig. 6.16(b) denotes the ratio of GEM(N)/TW performance scores across a range of measures. We can see that GEM(N) consistently yields a lower MSE of the extracted adjacency matrix and clustering coefficients, along with a higher NMI for the community structure, for all but the most extreme experiment configurations.

Although the presented methods exhibit attractive performance in recovering a ground-truth graph from spatio-temporal data, it is worth noting that in real-world applications, such as the one presented in Chapter 5, there is not a "real" underlying graph to extract (or even approximate). That is because we have used network structure as an explanatory variable for the observed sequence of data points and inference over such a variable may reveal important information on the relational structure of individual entities. Thus there is no way to know the ground truth and compare it against the extracted network structures in real-world settings, unless some domain-specific knowledge is employed. Such issues are discussed in more detail in Chapter 8, where we apply the developed methodologies to the Wytham Woods data set.

6.8 Discussion

Network analysis is a powerful tool for studying real-world complex systems. As there is an extensive collection of methods and algorithms for network analysis, in this work we have

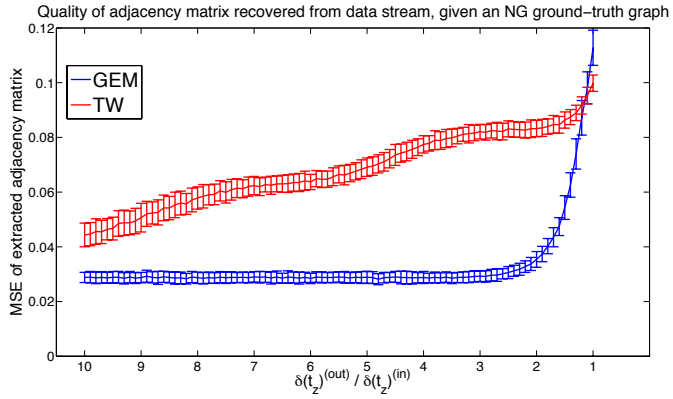
focused on the problem of constructing a network in the first place. In many applications, the collected data capture the behaviour of the system in some manner, like the spatial trajectories of participating agents, but not the underlying relations between them. We address this issue by assuming that mobility patterns of individuals can be correlated based on some form of underlying social connection. By identifying observation-dense regions in the data stream, which can be seen as gathering events of affiliated individuals, we propose a methodology of drawing links between agents based on their co-participation into those events.

Traditional approaches [Gero et al., 2009; Lauw et al., 2005; Oh and Badyaev, 2010; Whitehead, 2008] in constructing social networks from spatio-temporal data involve discretising the observation stream based on some fixed time window Δt and drawing links between individuals when they lie within such an "interaction radius". Alternative approaches, based upon Poisson point-process models, still require conditioning on the selection of an appropriate time interval (or rolling window) size to estimate the feeding rates per individual and the co-feeding rates for every bird pair. Our method overcomes the practical difficulties of selecting the appropriate Δt size and also allows a direct interpretation of the activity bursts in our data stream, as they are viewed as small foraging flocks. We regard this as not only intuitive for ecologists, but also as providing an additional layer of information regarding how birds group during feeding times. Additionally, it allows the incorporation of an appropriate null model, which we use in order to investigate if the co-occurrences of individuals into gathering events are a result of a latent social tie or coincidence. Such a null model retains the "bursty" nature of the data stream but breaks all correlations between the individuals' appearance patterns through an appropriate randomisation.

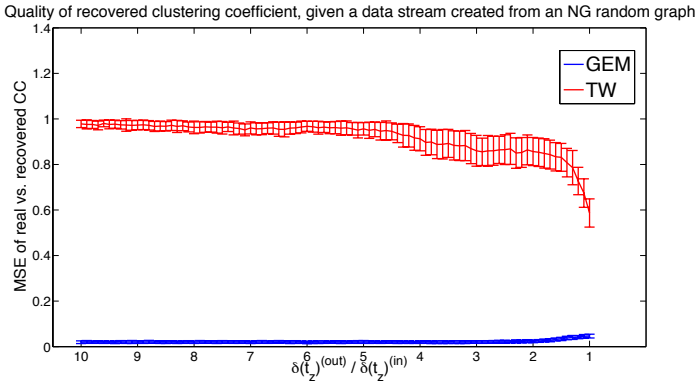
Although the methodologies we presented in this section have been developed for the purposes of analysing the Wytham Woods data set, they can be applied to any social system where we seek to infer the affiliations between agents based on their recorded appearances at various locations.

An important issue that has not been adequately addressed in this chapter is the one-mode projection methodology, briefly mentioned in Section 6.6. The reason we have skipped further analysis on the way we determine link weights based on the incidence matrix $\mathbf{B}$ is that such a scheme is very application-dependent; the way we define a link is based on what type of social network analysis question we seek to address. For the purposes of our experimentation in Section 6.7 we have used a simple one-mode projection scheme where links express the co-occurrence across either gathering events or time windows. More sophisticated schemes are considered in Chapter 7, where we present a novel one-mode projection methodology that produces probability distributions over the presence and weight of each link in the inferred graph.

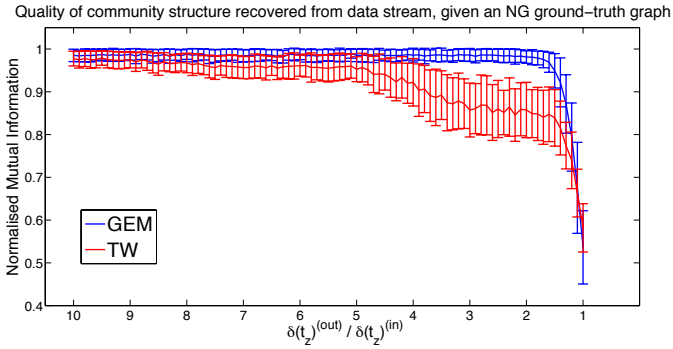
We need to underline that in real-world data streams there is no ground-truth graph to approximate. What we are merely seeking, is the most plausible relational structure that can explain a given observation sequence $\mathcal{D}$. In fact, we argue that there can be infinitely many graphs which can, based on a process such as the one described in Algorithm 4, generate a given data stream $\mathcal{D}$. Based on such graphs, we harness the tools of network analysis in order to gain further insights to the social characteristics of the population, originally hidden to us by the non-relational structure of the data stream. Some of those graphs may provide important insights on the relational structure of interacting individuals, some others do not. What we are aiming for is to discover a maximally informative network structure, from which the observed sequence of spatio-temporal records is more likely to rise.



(a)

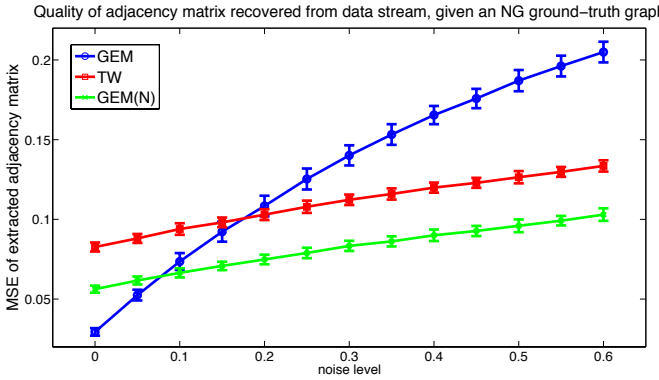


(b)

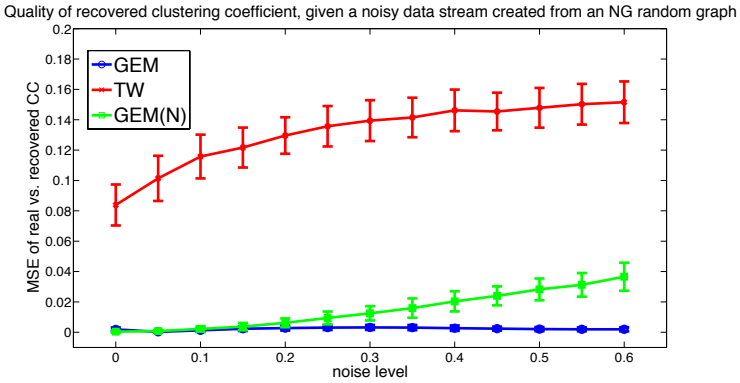


(c)

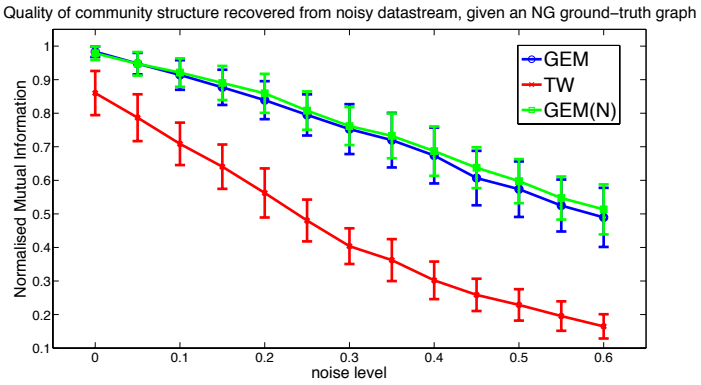
Figure 6.14: We illustrate the performance of Gathering Events (GEM) and Time Window (TW) approaches based on three different performance measures. Using NG random graphs as a ground-truth solution, we have run Algorithm 4 across a range of parameter values $\delta_{out}(t_z)/\delta_{in}(t_z)$, in order to generate data streams $\mathcal{D}$ of varying level of "modularity". Then for each data stream, we assess the accuracy of each method in recovering the original NG graph at link (Fig. 6.14(a)), local (Fig. 6.14(b)) and community level (Fig, 6.14(c)). Each data point in the above curves represents the mean performance value (with standard deviations) over 100 different data streams produced by a given NG adjacency matrix $\mathbf{A}$ and $\delta_{out}(t_z)/\delta_{in}(t_z)$ parameter value.



(a)

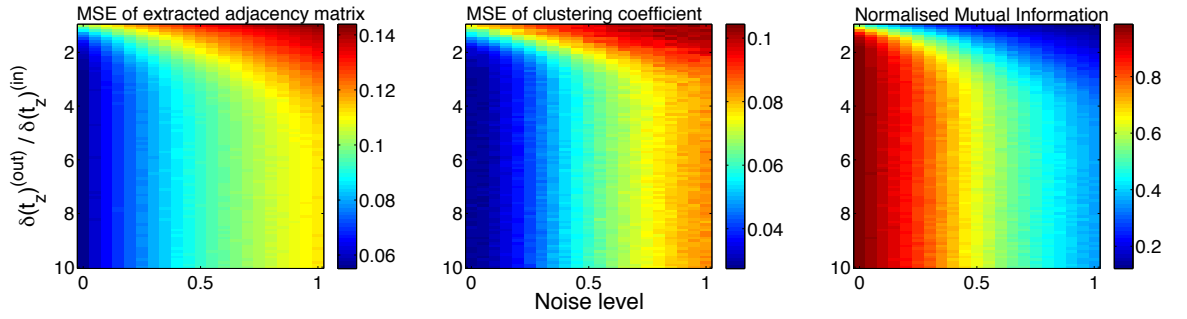


(b)

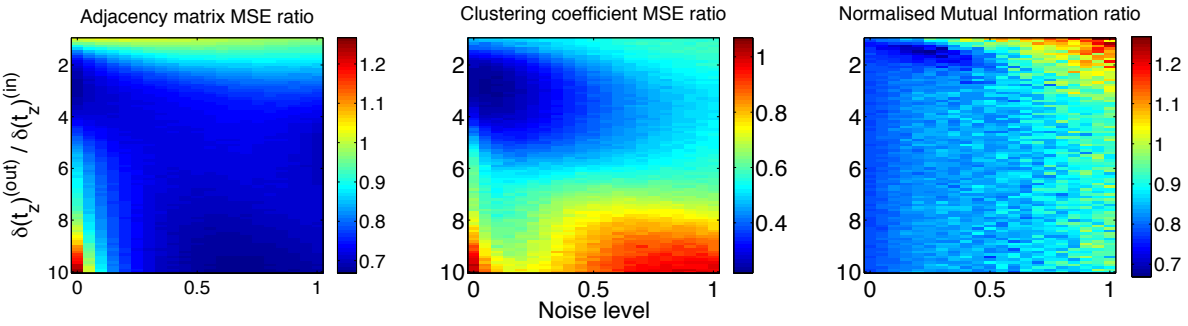


(c)

Figure 6.15: We illustrate the performance of Gathering Events (GEM), Gathering Events with null-model (GEM(N)) and Time Window (TW) approaches based on three different performance measures. Using NG random graphs as a ground-truth solution, we have run Algorithm 4 across a range of values of NOISE, so that we generate data streams $\mathcal{D}$ of varying rate of co-occurrence between unassociated individuals. Then for each data stream, we assess the accuracy of each method in recovering the original NG graph at link (Fig. 6.15(a)), local (Fig. 6.15(b)) and community level (Fig, 6.15(c)). Each data point in the above curves represents the mean performance value (with standard deviations) over 100 different data streams produced by a given NG adjacency matrix $\mathbf{A}$ and NOISE value.



(a)



(b)

Figure 6.16: In Fig. 6.16(a) we illustrate the performance of GEM(N) across a range of benchmark test parameters $\delta_{out}(t_z)/\delta_{in}(t_z)$ and NOISE levels. Each point in the heat map represents the average over 100 NG random graph instances. The three heat maps correspond to our three different performance measures under consideration. In Fig. 6.16(b) we consider the same performance measures but instead plot their ratio GEM(N)/TW in order to compare the null-model enabled Gathering Events approach with the Time Window one.

Chapter 7

Bayesian One-mode Projection

7.1 Introduction

In Chapter 6 we addressed the problem of extracting relational structure from spatio-temporal data, by deriving a model that groups wild bird feeder visitations into gathering events of foraging activity. Such a membership of N individuals to K flocks can be viewed as a bipartite graph from a Graph Theory perspective. In the same theme, quantifying the social tie strength between N birds, based on the similarities in their participation profile across K gathering events, can be seen as performing a one-mode projection of the implied bipartite graph.

In this chapter, we present a Bayesian methodology for one-mode projecting a bipartite network that is being observed across a series of discrete time steps. The resulting one-mode network captures the uncertainty over the presence/absence of each link and provides a probability distribution over its possible weight values. Additionally, the incorporation of prior knowledge over previous states makes the resulting network less sensitive to noise and missing observations that usually take place during the data collection process described in Chapter 5.

Finally, we show how our approach can be seen as a Bayesian extension of the widely

adopted *Simple Ratio Index* (SRI), from a traditional animal social network perspective [Whitehead, 2008; Whitehead et al., 2005] and can be directly applied to a wide range of similar settings.

7.2 Bipartite networks

A bipartite or two-mode network is a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{U}, \mathcal{E}\}$ with two sets of nodes, $\mathcal{V}$ and $\mathcal{U}$, where connections $\mathcal{E}$ exist only between nodes that belong to different sets. The overall connectivity is described by the $N \times K$ incidence matrix $\mathbf{B}$, where $N = |\mathcal{V}|$ and $K = |\mathcal{U}|$ and $b_{ik} = 1$ if there exists a link[1] between a given pair of nodes i, k for which $i \in \mathcal{V}, k \in \mathcal{U}$ and 0 otherwise. We use bipartite networks to describe a diverse range of complex systems; scientific collaboration networks [Newman, 2001b], animal visitation patterns to various sites [Psorakis et al., 2012; Whitehead et al., 2005], gene-to-disease associations [Goh et al., 2007] social media [Konstas et al., 2009], product co-purchasing networks [Leskovec et al., 2007], and many more [Zhou et al., 2007; Zweig and Kaufmann, 2011].

One-mode projection is an operation where a bipartite network $\mathcal{G} = \{\mathcal{V}, \mathcal{U}, \mathcal{E}\}$ described by the $N \times K$ incidence matrix $\mathbf{B}$ is mapped to a graph with only one class of nodes, $\mathcal{G}_U = \{\mathcal{V}, \mathcal{E}_U\}$ via $\mathbf{B} : N \times K \rightarrow \mathbf{A} : N \times N$. The new connections are now placed between nodes of the set $\mathcal{V}$, which we shall call from now on the "agent" set, based on the way they are linked to nodes of the vanished "location" set. In our Wytham Woods application, the agent set describes the wild birds while the location set the gathering events, as extracted by the methodology presented in Chapter 6.

The most trivial way to build the adjacency matrix $\mathbf{A}$ of $\mathcal{G}_U$ would be to set $a_{ij} = 1$ if nodes $i, j \in \mathcal{V}$ have at least one common target k in $\mathcal{G}$ and zero otherwise [Barabási et al., 2002; Newman, 2001a]. A reasonable refinement [Newman, 2010] involves setting the

[1] Each element b_{ik} can take any value in $\mathbb{R}$, indicating weight or participation strength. However, in this chapter we consider only the Boolean case.

weight of each link as the total number common targets, or co-occurrences $a_{ij} = \sum_{k=1}^{K} b_{ik} b_{jk}$ that i and j have across nodes in $\mathcal{U}$. We can also set the diagonal elements a_{ii} to 0, in settings where self-edges are not appropriate for our problem. Further extensions have been considered, such as moderating the weight by taking into account the exclusivity of co-occurrences [Newman, 2001a], the number of agents connected to the site [Porter et al., 2007] or introducing a saturation function [Li et al., 2005], which moderate the projected link weight a_{ij}.

For example, consider a simple graph of $N = 5$ individual birds linked to $K = 3$ foraging flocks (gathering events), where $b_{ik} = 1$ encodes that bird i has appeared at flock k. A typical one-mode projection [Newman, 2010] would be the 5-by-5 bird network, where $a_{ij} = \sum_k b_{ik} b_{jk},\ \forall i \neq j$ expresses the number of flocks where i and j co-appeared. The same process can be applied to recover the 3-by-3 flock network, with $a'_{k\ell} = \sum_i b_{ik} b_{i\ell}$ the number of common members of flocks k and ℓ. Although such a process can be seen as a lossy compression of the original bipartite network [Zhou et al., 2007], it allows us to exploit the wealth of computational methods that have been developed for unipartite graphs [Nacher and Akutsu, 2011], thus revealing interesting associations such as communities of socially affiliated birds, which often break and merge with other ones in a fission-fusion manner [Conradt and Roper, 2000].

It is worth noting that the "naive" one-mode projection $\mathbf{A} = \mathbf{B}\mathbf{B}^{\mathsf{T}}$ forces all nodes from $\mathcal{V}$ that point to a particular location node $k \in \mathcal{U}$ to form a fully connected subgraph. Thus each location node k corresponds to a d_k-clique in $\mathcal{G}_U$, where $d_k = \sum_{i=1}^{N} a_{ik}$ the degree of k. Therefore, although the original bipartite graph is locally dense and globally sparse [Guillaume and Latapy, 2006], due to the heavy-tail degree distribution on the location set [Nacher and Akutsu, 2011] there is a non-trivial number of nodes in $\mathcal{U}$ with such a high degree d_k, which make the projected network almost fully connected [Lambiotte and Ausloos, 2005]. Methods that can be employed to regulate such a densification in the resulting graph $\mathcal{G}_U$, range from information filtering [Radicchi et al., 2011; Tumminello et al., 2005] to defining

appropriate null models that examine the statistical significance of the observed weights (see Section 6.6.2 of Chapter 6) or network motifs [Zweig and Kaufmann, 2011]. In the present work, we formulate a model that describes the plausibility of a social tie in a fully probabilistic manner; by placing a distribution over the link presence and another one over the connection strength.

In this chapter, we seek to one-mode project a temporal bipartite network $\mathcal{G}^{(t)} = \{\mathcal{U}, \mathcal{V}, \mathcal{E}^{(t)}\}, t \in \{1, ..., T\}$ that is described by a sequence of incidence matrices $\{\mathbf{B}^{(t)}\}_{t=1}^{T}$. The one-mode projection at any given time point t captures the associations between nodes $i, j \in \mathcal{V}$ by taking into account past and present link information from all steps 1 to t. We require that all projected connections between nodes i, j are weighted appropriately so that we take into account both the strength and the statistical significance of the association. Finally, we seek to model the uncertainty over the resulting topology, by placing probability distributions over the presence of each link. The model is formally presented in Section 7.3 and we illustrate its use in Section 7.4. In Section 7.5 we conclude with a short discussion on theoretical extensions, while application of the model on the wild bird data of Chapter 5 is presented in Chapter 8.

7.3 Bayesian one-mode projection

7.3.1 Problem statement

Consider a setting where we observe a temporal bipartite network as a sequence of "snapshots" $\{\mathbf{B}^{(t)}\}_{t=1}^{T}$, where each incidence matrix $\mathbf{B}^{(t)} : N \times K$ describes the linkage of N *agents* to K *locations*. For the sake of simplicity in notation, from now on we will assume that N and K are fixed for each t, although such a constraint can be relaxed. In our application setting described in Chapter 5, such a data set encodes the participation of N individual wild birds in K foraging flocks, extracted using the GEM method from Chapter 6 on the t-th

day of data collection. In applications outside of the Wytham Woods experiment, we can view $\{\mathbf{B}^{(t)}\}_{t=1}^{T}$ as a bipartite sequence describing the buying habits of N customers, who are performing purchases among a set of K products each month t, or the daily mobility patterns of N social media users who appear (or equivalently, perform "check-ins") at K locations.

Our key assumption is that there is an underlying *association* or *similarity* network $\mathbf{\Pi} \in \mathbb{R}^{N\times N}$ between agents, which directly affects the structure of $\{\mathbf{B}^{(t)}\}_{t=1}^{T}$, in the sense that strongly-associated agents consistently point to the same locations and vice-versa. Our goal is to learn the structure of $\mathbf{\Pi} \in \mathbb{R}^{N\times N}$ at every step t, by defining a Bayesian model that captures our belief about the presence (or absence) of each link along with a probability distribution over the strength of connections.

7.3.2 Probabilistic model for graph links

Given the observation sequence $\{\mathbf{B}^{(t)}\}_{t=1}^{T}$ described in Section 7.3.1, let us isolate one particular timestamp t, so that $\mathbf{B} = \mathbf{B}^{(t)}$. Each element b_{ik} is 1 if agent i links to a location k and zero otherwise while the sum $d_i = \sum_{k=1}^{K} b_{ik}$ is the total locations or out degree of i. Let us now define an additional variable x_{ij} that we will call *opportunities*, which is the number of locations linked to either node i or node j; that is obtained by performing an element-by-element logical disjunction on the rows of $\mathbf{B}$ and summing the elements of the resulting vector:

$$x_{ij} = \sum_{k=1}^{K} \text{OR}(b_{ik}, b_{jk}). \tag{7.1}$$

A list of variables used in this chapter is presented in Table 7.1.

Given the observed $N \times K$ incidence matrix $\mathbf{B}$, we begin by performing the standard weighted one-mode projection, getting the co-occurrence matrix $\mathbf{A} = \mathbf{B}\mathbf{B}^{\mathsf{T}}$. Each $a_{ij} = \sum_{k=1}^{K} b_{ik}b_{jk}$ represents integer-valued counts (co-occurrences) that we can view as the number of "successes" in x_{ij} "trials". This quantity is typically modelled as a draw from a

Table 7.1: Notation

Variable	Interpretation
N	# of "agent nodes".
K	# of "location nodes".
$\mathbf{B}$	$N \times K$ incidence matrix of bipartite graph.
$\mathbf{A}$	$N \times N$ projection matrix.
a_{ij}	# of co-occurrences of agents i and j.
d_i	degree of agent i based on $\mathbf{B}$.
x_{ij}	# of locations linked to by either i or j (opportunities)
π_{ij}	attraction coefficient of i, j, with $\pi_{ij} \in [0, 1]$
$\kappa_{ij}, \lambda_{ij}$	Beta distribution parameters.

binomial distribution [Bishop, 2007; Jaynes, 2003]:

$$a_{ij} \sim \text{Binom}(\pi_{ij}; x_{ij}), \tag{7.2}$$

with two parameters; the number of opportunities x_{ij} and a bias term $\pi_{ij} \in [0, 1]$ that corresponds to our modelling assumption that there is a latent *attraction coefficient* between all pairs i, j, which controls the extent to which opportunities x_{ij} are manifested as co-occurrences a_{ij} across locations. We view π_{ij} as a measure of similarity or association between i and j and it is the key variable in our model, encoded in matrix form $\mathbf{\Pi} \in \mathbb{R}^{N \times N}$ for all pairs. For the special case of $x_{ij} = 1$, the variable π_{ij} becomes a Bernoulli parameter that can be viewed as the probability of a link between i and j.

Based on Eq. (7.2), the probability of observing a particular number of co-occurrences, or link weight, a_{ij} is given by:

$$p(a_{ij}|\pi_{ij}, x_{ij}) = \binom{x_{ij}}{a_{ij}} \pi_{ij}^{a_{ij}} (1 - \pi_{ij})^{x_{ij} - a_{ij}}, \tag{7.3}$$

which is the likelihood function of the observed weights a_{ij}. As our inference task is to describe the attraction coefficient π_{ij} given the known a_{ij}, x_{ij}, a first approach would consist of maximising Eq. (7.3) with respect to π_{ij}. The trivial maximum likelihood (ML) solution

to Eq. (7.3), which yields the Simple Ratio Index (SRI) $\hat{\pi}_{ij} = a_{ij}/x_{ij}$ from the animal social network literature [Whitehead, 2008; Whitehead et al., 2005], is deemed inappropriate for the following reasons:

- it makes our model sensitive to degenerate values of a_{ij} and x_{ij}, which result from imperfect observations of the incidence matrix $\mathbf{B}$.
- it provides a point estimate of π_{ij}, thus not capturing the uncertainty on the attraction coefficient due to noise and missing observations.
- it does not provide a systematic framework for learning π_{ij}, by exploiting both past and future observations of the bipartite networks.

To overcome the above difficulties, we employ a Bayesian approach by working with the probability distribution over π_{ij}; we start with a prior $p(\pi_{ij})$ and revise at each time step t as we observe new values for a_{ij} and x_{ij}.

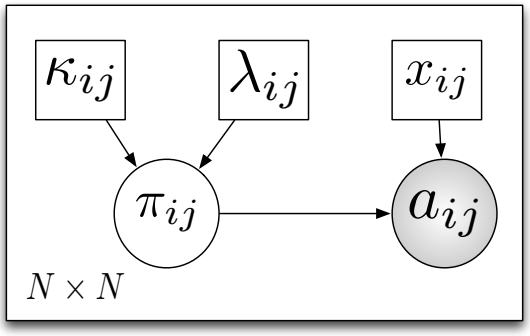


Figure 7.1: Our graphical model, expressing how the observed (shaded circle) co-occurrences a_{ij} between individuals i and j depend on the number of opportunities x_{ij} (locations where either i or j link to in the original bipartite graph) and an unobserved (unshaded circle) attraction coefficient π_{ij}. The square plates represent deterministic parameters of the model.

Recall that in Eq. (7.2) and (7.3) we have stated that the co-occurrences i and j depend on the opportunities and the attraction coefficient. This can be expressed via a graphical model

in Fig. 7.1, where the probabilistic dependencies are indicated via arrows from nodes x_{ij} and π_{ij} pointing to a_{ij}. This allows us to express the probability of π_{ij} as:

$$p(\pi_{ij}|a_{ij}, x_{ij}) = \frac{p(a_{ij}|\pi_{ij}, x_{ij})p(\pi_{ij})}{\int_0^1 p(a_{ij}, \pi_{ij}|x_{ij})d\pi_{ij}}, \tag{7.4}$$

where $p(\pi_{ij})$ is the prior and expresses our belief on how the attraction coefficient for i, j varies before observing a_{ij} and x_{ij}, while the posterior $p(\pi_{ij}|a_{ij}, x_{ij})$ is the revised belief on π_{ij} in the light of these observations. We model $p(\pi_{ij})$ as a Beta distribution:

$$\pi_{ij} \sim \text{Beta}(\kappa_{ij}, \lambda_{ij}), \tag{7.5}$$

parameterised by κ_{ij} and λ_{ij}, so that:

$$p(\pi_{ij}) = \frac{\pi_{ij}^{\kappa_{ij}-1}(1-\pi_{ij})^{\lambda_{ij}-1}}{\int_0^1 u^{\kappa_{ij}-1}(1-u)^{\lambda_{ij}-1}du}. \tag{7.6}$$

The Beta distribution is appropriate for modelling π_{ij}, as it is supported on the bounded interval $[0, 1]$ and provides us (as we will demostrate later) the flexibility to update our belief over the, initially unknown, attraction coefficient π_{ij} via a computationally inexpensive update of the κ_{ij} and λ_{ij} parameters in Eq. (7.6).

Having a functional form for our prior in Eq. (7.6), we combine it with the likelihood from Eq. (7.3) based on Eq. (7.4) to obtain the posterior:

$$\begin{aligned}
p(\pi_{ij}|a_{ij}, x_{ij}) &= \frac{p(a_{ij}|\pi_{ij}, x_{ij})p(\pi_{ij})}{\int_0^1 p(a_{ij}, \pi_{ij}|x_{ij})d\pi_{ij}} \\
&= \frac{\binom{x_{ij}}{a_{ij}}\pi_{ij}^{a_{ij}}(1-\pi_{ij})^{x_{ij}-a_{ij}}}{\int_0^1 \binom{x_{ij}}{a_{ij}}\pi_{ij}^{a_{ij}}(1-\pi_{ij})^{x_{ij}-a_{ij}}d\pi_{ij}} \times \frac{\pi_{ij}^{\kappa_{ij}-1}(1-\pi_{ij})^{\lambda_{ij}-1}}{\int_0^1 u^{\kappa_{ij}-1}(1-u)^{\lambda_{ij}-1}du} \\
&= \frac{\pi_{ij}^{a_{ij}+\kappa_{ij}-1}(1-\pi_{ij})^{x_{ij}-a_{ij}+\lambda_{ij}-1}}{\binom{x_{ij}}{a_{ij}}^{-1} \times \int_0^1 \binom{x_{ij}}{a_{ij}}\pi_{ij}^{a_{ij}}(1-\pi_{ij})^{x_{ij}-a_{ij}}d\pi_{ij} \times \int_0^1 u^{\kappa_{ij}-1}(1-u)^{\lambda_{ij}-1}du} \\
&= \text{Beta}(\kappa_{ij} + a_{ij}, \lambda_{ij} + x_{ij} - a_{ij}),
\end{aligned} \tag{7.7}$$

which is a revised Beta distribution over π_{ij}, with updated parameters:

$$\kappa'_{ij} = \kappa_{ij} + a_{ij}, \tag{7.8}$$

$$\lambda'_{ij} = \lambda_{ij} + x_{ij} - a_{ij}. \tag{7.9}$$

The posterior distribution $p(\pi_{ij}|a_{ij}, x_{ij}) = \text{Beta}(\kappa'_{ij}, \lambda'_{ij})$ provides all of the information we need to describe the attraction coefficient π_{ij}, capturing the uncertainty over each possible value in $[0, 1]$, while all dependencies between links are encoded in the a_{ij} and x_{ij} terms.

Having a fully probabilistic formulation for the attraction coefficient from Eq. (7.7), we can proceed one step further and "integrate out" π_{ij} from the likelihood function in Eq. (7.3) in order to obtain the predictive distribution over the connection weight a_{ij}:

$$\begin{aligned} p(a_{ij}|x_{ij}, \kappa'_{ij}, \lambda'_{ij}) &= \int_0^1 p(a_{ij}, \pi_{ij}|x_{ij}, \kappa'_{ij}, \lambda'_{ij}) d\pi_{ij} \\ &= \binom{x_{ij}}{a_{ij}} B^{-1}(\kappa'_{ij}, \lambda'_{ij}) \int_0^1 \pi_{ij}^{a_{ij}+\kappa'_{ij}-1} (1-\pi_{ij})^{x_{ij}-a_{ij}+\lambda'_{ij}-1} d\pi_{ij} \\ &= \binom{x_{ij}}{a_{ij}} \frac{B(a_{ij}+\kappa'_{ij}, x_{ij}-a_{ij}+\lambda'_{ij})}{B(\kappa'_{ij}, \lambda'_{ij})}, \end{aligned} \tag{7.10}$$

which is a Beta-binomial probability density function and $B(.)$ is the standard beta function. Such a distribution captures the variability of co-occurrences a_{ij} given our noise model.

We have now described the theoretical foundation of our model along with the one-mode projection scheme for a single learning step t. The full process involves cycling through the update equations:

$$\kappa_{ij}^{(t)} = \kappa_{ij}^{(t-1)} + a_{ij}^{(t-1)} \tag{7.11}$$

$$\lambda_{ij}^{(t)} = \lambda_{ij}^{(t-1)} + x_{ij}^{(t-1)} - a_{ij}^{(t-1)} \tag{7.12}$$

and revising our distributions over the attraction coefficients and link weights. Details of the full learning scheme are presented in the following section.

7.3.3 Algorithm overview and implementation details

Consider the state of the system at time $t = 0$, before receiving the first network observation $\mathbf{B}^{(1)}$. At this stage, we have no observations regarding the bipartite graph and any prior beliefs on the agent pair associations i, j are encoded in the Beta parameters $\kappa_{ij}^{(0)}, \lambda_{ij}^{(0)}$. These can be initialised, for example, to vanilla values $\kappa_{ij}^{(0)} = \lambda_{ij}^{(0)} = 0$ expressing a uniform prior or $\kappa_{ij}^{(0)} = \lambda_{ij}^{(0)} = 10$ that centre the attraction coefficients $\pi_{ij}^{(0)}$ around 0.5.

Upon receiving the first $\mathbf{B}^{(1)}$ we calculate the opportunities x_{ij} and then the co-occurrences $a_{ij}^{(1)} = \sum_{k=1}^{K} b_{ik}^{(1)} b_{jk}^{(1)}$ for all i, j. We then update $\kappa_{ij}^{(1)}, \lambda_{ij}^{(1)}$ based on Eq. (7.11) and (7.12).

We build the projection matrix $\mathbf{\Pi} \in \mathbb{R}^{N \times N}$ via the use of smooth fixed-point estimates, which can be directly derived from the posterior distributions. For this particular study we have used the expected value $\mathbb{E}[\pi_{ij}] = \frac{\kappa_{ij}}{\kappa_{ij}+\lambda_{ij}}$ for each element of $\Pi \in \mathbb{R}^{N \times N}$, while for the expected co-occurrences $\tilde{a}_{ij}$ in $\tilde{\mathbf{A}} \in \mathbb{R}^{N \times N}$, we use $\mathbb{E}[a_{ij}] = \frac{x_{ij}\kappa_{ij}}{\kappa_{ij}+\lambda_{ij}}$. The process, which we call Bayesian One-Mode Projection (BOMP), is presented in Algorithm 5.

The computational cost of Algorithm 5 can be moderated via an appropriate distributed implementation. Matrix operations such as the multiplication $\mathbf{B}\mathbf{B}^{\mathsf{T}}$ can be parallelised (examples for Map-Reduce are shown in [Rajaraman and Ullman, 2011]) while $\kappa_{ij}, \lambda_{ij}$ updates for each pair i, j can be performed at different processing units. The benign computational

Algorithm 5 Bayesian One-Mode Projection (BOMP)

Require: bipartite sequence $\{\mathbf{B}^{(t)}\}_{t=1}^{T}$
1: Initialise $\kappa_{ij}^{(0)}, \lambda_{ij}^{(0)}, \ \forall\, i, j \in \{1, ..., N\}$
2: **for** $t = t_0$ to T **do**
3: Set $\mathbf{B} = \mathbf{B}^{(t)}$
4: Get opportunities $x_{ij}^{(t)}$ from Eq. (7.1)
5: Get co-occurrences via $\mathbf{A}^{(t)} = \mathbf{B}\mathbf{B}^{\mathsf{T}}$
6: **for** $i, j \in \{1, ..., N\}$ **do**
7: update $\kappa_{ij}^{(t)}$ from Eq. (7.8)
8: update $\lambda_{ij}^{(t)}$ from Eq. (7.9)
9: $\mathbb{E}^{(t)}[\pi_{ij}] = \frac{\kappa_{ij}^{(t)}}{\kappa_{ij}^{(t)} + \lambda_{ij}^{(t)}}$
10: $\mathbb{E}^{(t)}[a_{ij}] = \frac{x_{ij}^{(t)} \kappa_{ij}^{(t)}}{\kappa_{ij}^{(t)} + \lambda_{ij}^{(t)}}$
11: **end for**
12: **end for**
13: **return** $\mathbf{\Pi}^{(t)} = \{\mathbb{E}^{(t)}[\pi_{ij}]\}_{i,j \in N}$ and $\tilde{\mathbf{A}}^{(t)} = \{\mathbb{E}^{(t)}[a_{ij}]\}_{i,j \in N}, \forall\, t \in \{1, ..., T\}$

scalability of the method relies on the structure of the probabilistic model itself; the conjugacy of our Beta prior in Eq. (7.6) with our Binomial likelihood function in Eq. (7.3) makes the posterior in Eq. (7.7) an updated Beta, thus no sampling (such as Markov Chain Monte Carlo) schemes need to be employed. In the next section we illustrate the application of the proposed method in a working example, using an artificial data set.

7.4 Experimentation on benchmark data sets

In this section we illustrate BOMP via an artificial experimentation scheme; we start with a fully observed adjacency matrix $\mathbf{A}^{(\text{obs})}$, from which we generate a noise-contaminated gathering event sequence $\mathbf{B}^{(t)}$, using Algorithm 4 of Chapter 6. By applying BOMP on $\mathbf{B}^{(t)}$, we illustrate how the proposed method uncovers the ground-truth association network.

Such an artificial data set mimics the real Wytham Woods data, where at every day t we receive a batch of sensor observations $\mathcal{D}_t$ and run GEM in order to extract the gathering events matrix $\mathbf{B}^{(t)}$. The flocking structure for the whole season is thus described by the sequence

of gathering events matrices $\mathbf{B}^{(t)}$, from which we seek to infer the underlying social ties between the participating birds.

7.4.1 Artificial data generation scheme

Consider the simple "bowtie" graph of Fig. 7.2 that consists of $N = 5$ nodes and two 3-cliques, overlapping on a single node $i = 3$. We use algorithm 4 in order to generate $T = 100$ noise-contaminated data streams (one per "observation day" t) based on its adjacency matrix $\mathbf{A}^{(\text{obs})}$, each one having $Z = 1000$ records. We use a noise parameter $\texttt{NOISE} = 0.25$, expressing that, on average, one of every four individuals at each gathering event is not connected with the others in $\mathbf{A}^{(\text{obs})}$.

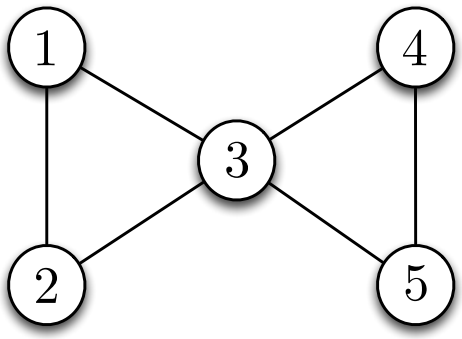


Figure 7.2: A small graph with $N = 5$ nodes and two triangles (3-cliques) that overlap on the "broker" node 3.

In our example, we have set individuals 1-2 and 3-4 to be associated (possibly because they are a mating pair), while individual 3 is highly gregarious, thus appearing in all foraging flocks generated by Algorithm 4. For each day t, our goal is to learn the associations between each pair i, j, by exploiting current and past observations.

Note that the individual-to-event matrices $\mathbf{B}^{(t)}$ are automatically given by Algorithm 4, therefore for this example we do not need to perform a gathering event extraction scheme, via GEM, on the generated data streams.

7.4.2 Applying the method

We start at $t = 0$ by assuming no prior knowledge of any link structure. At this stage we have not seen any observations, thus our model parameters must reflect our ignorance regarding pairwise associations. For each pair i, j we set $\kappa_{ij}^{(0)} = \lambda_{ij}^{(0)} = 10$ that gives $\mathbb{E}[\pi_{ij}^{(0)}] = 0.5\ \forall i, j$, implying that we are unable to tell if there should be an association link between i, j before observing any data.

At $t = 1$ we receive our first batch of data, namely the incidence matrix $\mathbf{B}^{(1)}$. The opportunity values $x_{ij}^{(1)}$ are retrieved from $\mathbf{B}^{(1)}$ via Eq. (7.1), while the co-occurrences $a_{ij}^{(1)}$ are retrieved via a standard weighted one-mode projection of $\mathbf{B}^{(1)}$ to $\mathbf{A}^{(1)}$. These values of $a_{ij}^{(1)}, x_{ij}^{(1)}$, along with $\kappa_{ij}^{(0)}, \lambda_{ij}^{(0)}$ from initialisation, are all that we need to describe the distribution of the attraction coefficients via the update equations (7.11) and (7.12). Such an update of the sufficient statistics of π_{ij} allows us to fully describe the posterior distribution of the attraction coefficient π_{ij} between i and j. Fixed-value estimates of π_{ij} can be obtained using, for example, the expected value under the posterior distribution $\mathbb{E}^{(t)}[\pi_{ij}] = \frac{\kappa_{ij}^{(t)}}{\kappa_{ij}^{(t)} + \lambda_{ij}^{(t)}}$. The posterior distribution over $a_{ij}^{(1)}$ (i.e. the number of gathering events in which both i and j appear) is obtained via Eq. (7.4)

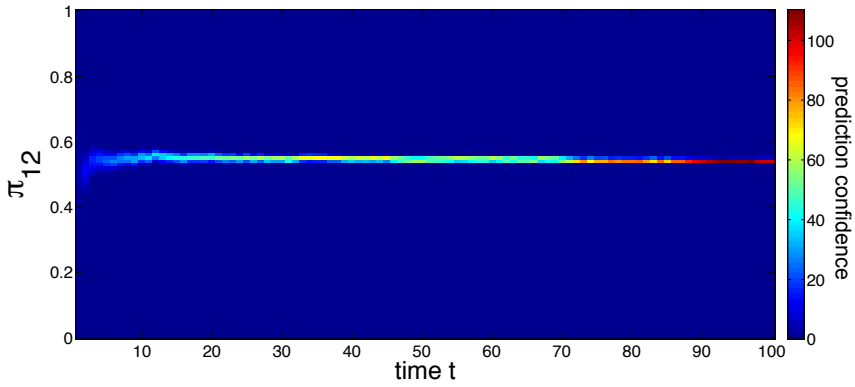
In Fig. 7.3(a) we plot how the posterior distribution $p(\pi_{12}|a_{12}, x_{12}, \kappa_{12}, \lambda_{12})$ of the attraction coefficient $\pi_{12}^{(t)}$ progresses during each iteration. For $t = 0$ the distribution is our flat prior centered around 0.5, as we have no evidence to support the presence of an association between nodes 1 and 2. As we start observing non-zero link weights a_{12}, shown as a red line in Fig. 7.3(c), this prior belief is updated in order to explain the incoming data, effectively shifting the distribution so that more probability mass is concentrated around larger values of π_{12}. It is important to note that the increase of $\mathbb{E}[\pi_{12}^{(t)}]$, shown as a blue line in Fig. 7.3(b), is less steep for later t, as the impact of new co-occurrences i, j is not so strong as in the beginning of data collection. Such an important saturation or "diminishing returns" property arises naturally in Bayesian learning models, without the need to explicitly induce it via additional

machinery such as hyperbolic tangent functions [Li et al., 2005]. Additionally, our fixed-point estimate $\mathbb{E}[\pi_{12}]$ of the association score between individuals 1 and 2 is smoother that the traditional Simple-Ratio (with standard deviation $\sigma(\tilde{a}_{12}) \simeq 6.07$ versus $\sigma(a_{12}) \simeq 9.43$), as shown in Fig. 7.3(b), as SR does not take into account past observations and relies only the current a_{ij} and x_{ij} at each t.

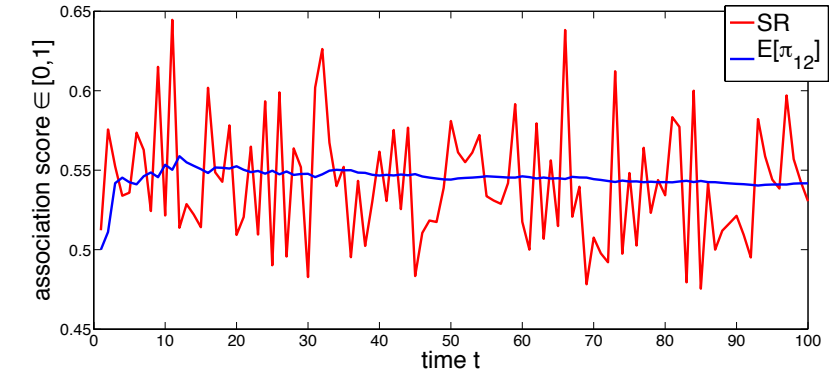
In Fig. 7.4(a) we plot the posterior distribution $p(\pi_{24}|a_{24}, x_{24}, \kappa_{24}, \lambda_{24})$ that expresses our belief regarding the presence of a link between nodes 2 and 4. As before, he distribution is initially centered around 0.5, due to the lack of evidence, while upon receiving data it rapidly shifts towards small values which can also be seen by plotting $\mathbb{E}[\pi_{24}]$ in Fig. 7.4(b). We also note that although the distribution is flatter during the initial iterations, we are constantly making more and more confident predictions (decreasing posterior entropy) as we keep observing $a_{24}^{(t)}$.

Now let us examine how our method models associations between $i = 3$ and other nodes in the graph. Node 3 is an exceptional case in our example, as it reflects the example of a gregarious bird, connected to every other individual in the toy graph and thus appearing across all generated gathering events.

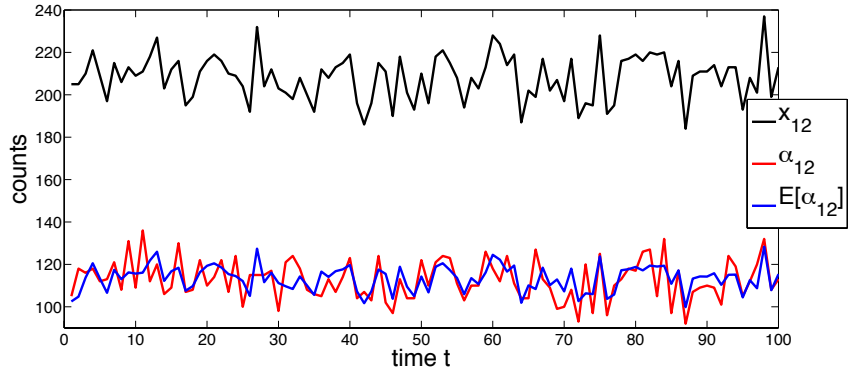
In Fig. 7.5(a) we plot the posterior density curves of $\pi_{23}^{(t)}$ across t, along with the expected values $\mathbb{E}[\pi_{23}^{(t)}]$ in Fig. 7.5(b) and $\mathbb{E}[a_{23}^{(t)}]$ in Fig. 7.5(c) as we did in the previous cases above. We can see that although we consistently observe non-zero link weights $a_{23}^{(t)}$ (red line in Fig. 7.5(c)), the association score or attraction coefficient $\pi_{23}^{(t)}$ decreases to lower values (seen in Fig. 7.5(b)). This is in complete disagreement with the case of association between individuals 1 and 2, where consistently non-zero observations $a_{12}^{(t)}$ (red line in Fig. 7.3(c)) led to an increase in the attraction coefficient $\pi_{12}^{(t)}$ (seen in Fig. 7.3(b)). The reason why our posterior belief over the presence of the link between 2 and 3 is reduced, although we observe non-zero co-occurrences between individuals 2 and 3, is because of the role the opportunities variable x_{ij} plays in the model. Recall Eq. (7.2):



(a)

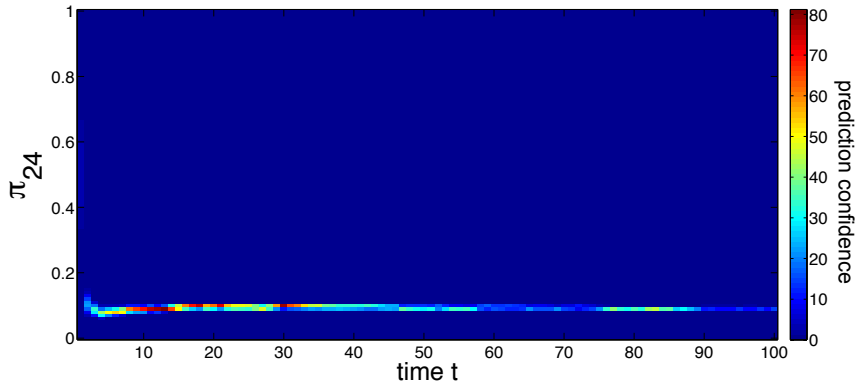


(b)

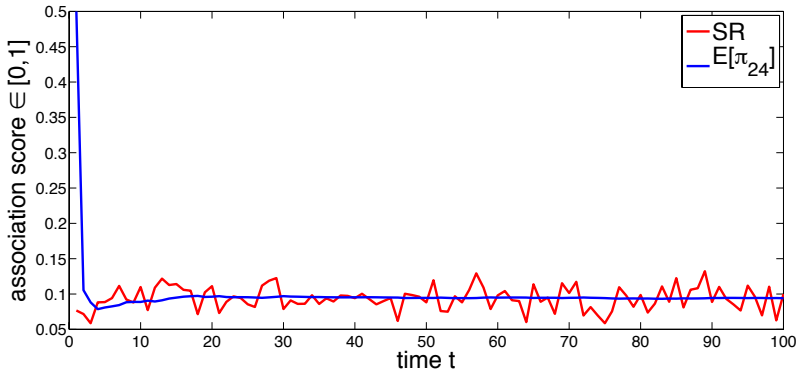


(c)

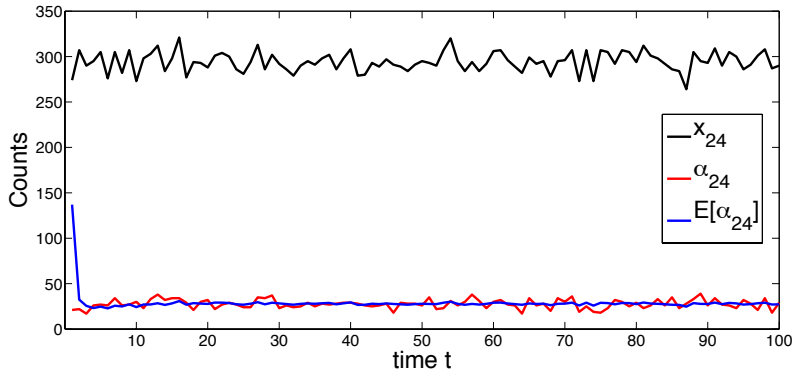
Figure 7.3: In Fig. 7.3(a) we illustrate how the distribution over the attraction coefficient $\pi_{12}^{(t)}$ changes across 100 step, in the form of a heat map. The color intensity (shown in the color bar) expresses the confidence of the π_{12} estimate. In Fig. 7.3(b) we show how BOMP produces smooth estimates of the association score $\mathbb{E}[\pi_{12}]$ compared to the Simple Ratio index. The expected number of co-occurrences $\tilde{a}_{12} = \mathbb{E}[a_{12}]$ is also smoother than the observed co-occurrences a_{12}, with standard deviation $\sigma(\tilde{a}_{12}) \simeq 6.07$ versus $\sigma(a_{12}) \simeq 9.43$.



(a)

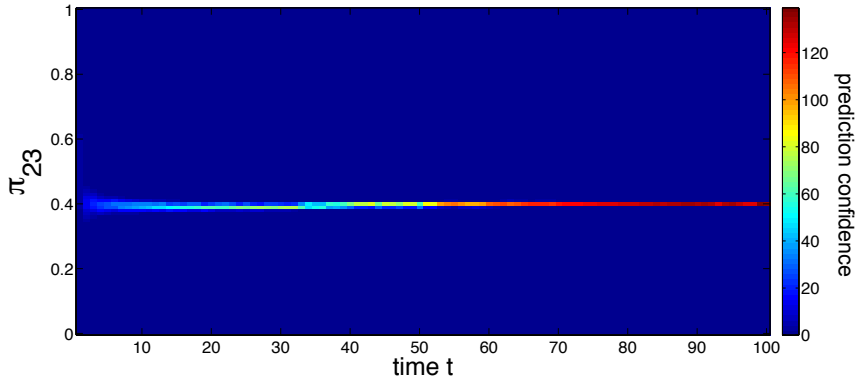


(b)

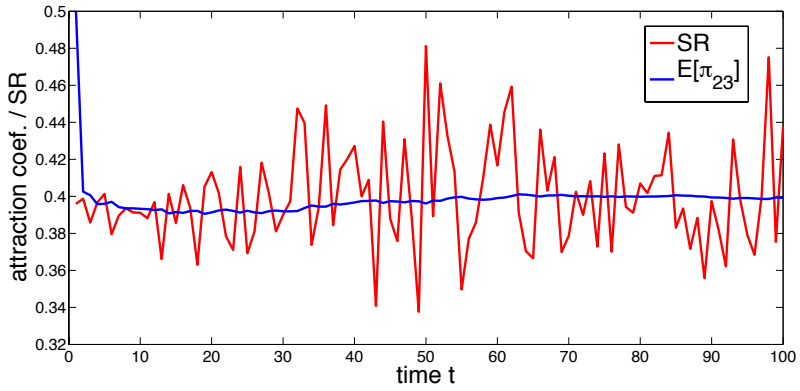


(c)

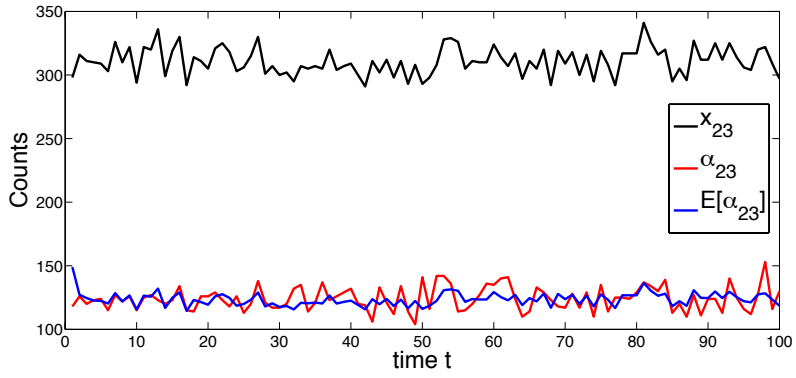
Figure 7.4: We plot various association scores for the case of the unconnected individuals 2 and 4. We can see in Fig. 7.4(a) that our posterior over π_{24} rapidly converges to small values, with smoother behaviour than the SR index, as seen in Fig. 7.4(c). The expected co-occurrences estimate $\tilde{a}_{24} = \mathbb{E}[a_{24}]$, shown in Fig. 7.4(c), is also smoother than the observed a_{24} with standard deviation $\sigma(\tilde{a}_{24}) \simeq 4.99$ versus $\sigma(a_{24}) \simeq 11.02$.



(a)



(b)



(c)

Figure 7.5: We plot association metrics for the node pair 2 and 3. Although the pair has similar number of co-occurrences across t as 1 and 2 (see red lines in Fig. 7.3(c) and Fig. 7.5(c)), the estimated attraction $\mathbb{E}[\pi_{23}]$ coefficient, shown as the blue line in Fig. 7.5(a), is smaller than $\mathbb{E}[\pi_{12}]$ due the large number of opportunities x_{23} (the black line in Fig. 7.5(c)) that imply lack of exclusivity.

$$a_{ij} \sim \text{Binom}(\pi_{ij}; x_{ij}),$$

which models the observed link weight a_{ij} as the result of a latent attraction term π_{ij} that controls how many of the opportunities x_{ij} are manifested as co-occurrences. These opportunities are the total number of events where either i or j appeared. On one hand, birds 1 and 2 belong to the same clique and no other, therefore their event membership, as generated from Algorithm 4, almost completely overlaps (apart from the inconsistencies induced by our noise parameter). On the other hand, although nodes 2 and 3 have similar number of co-occurrences with 1 and 2 (red lines in Fig. 7.3(c) and Fig. 7.5(c)), for the number of opportunities we have $x_{23}^{(t)} > x_{12}^{(t)},\ \forall t \in \{1, ..., 100\}$ (see black lines in Fig. 7.3(c) and Fig. 7.5(c)). That leads our model, based on the binomial expression in Eq. (7.2), to infer a lower value of the attraction coefficient π_{23} compared to π_{12}, effectively penalising such a lack of exclusivity in the co-appearances of 2 with 3.

This is a very attractive property of the model, which not only regulates the link weights between "gregarious" individuals (who tend to link to everywhere) and "selective" ones (with a small set of gathering events at which they appear), but also allows the model to downplay the effect of purely coincidental co-appearances on the attraction coefficient, which would otherwise introduce "junk" associations in the projection network.

7.4.3 Future work on changepoint detection

We have shown that our method learns the association patterns of nodes, by making more and more confident predictions on the model quantities of interest while being resilient to perturbations induced by noise. The question is, what happens in cases where the underlying system dramatically changes at some given time point t_c, making all prior knowledge from $t = 0, ..., t_c - 1$ obsolete? In order to illustrate this, let us revisit the example of Section 7.4, where at a given point t_c our ground-truth network, originally shown in Fig. 7.2, becomes

rewired, as in Fig. 7.6.

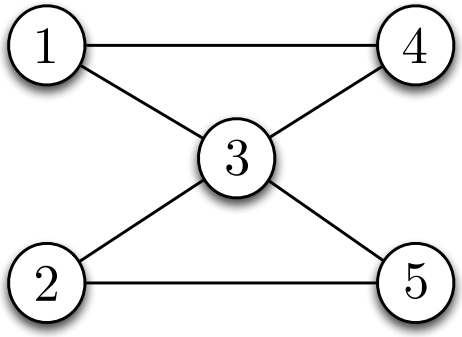


Figure 7.6: A rewired version of the graph presented in Fig. 7.2. It maintains the key topological properties of $N = 5$ nodes and two triangles (3-cliques) overlapping on the "broker" node 3.

In the gathering-event sequence generated by Algorithm 4 for $t > t_c$, individuals 1 and 4 now co-appear in the first two gathering events, 2 and 5 in the last two while 3 remains common across all $K = 4$ events.

Assume now that we run our methodology on a data set of $T = 200$ instances of $\mathbf{B}^{(t)}$, where the first 100 are generated based on Fig. 7.2 (as in Section 7.4.1) and at $t_c = 101$ to T we use the new ground-truth graph from Fig. 7.6. In Fig. 7.7 we plot the expectation of the attraction coefficient $\mathbb{E}[\pi_{ij}]$ between pairs 1 and 2 along with 1 and 4.

We can see that for the first 100 steps, before the changepoint, the model has identical behaviour to that presented in Section 7.4; as $a_{12}^{(t)}$ tends to be non-zero for $t < t_c$, the expected value of our posterior $\mathbb{E}[\pi_{12}^{(t)}]$ increases as we observe more data (blue line in Fig. 7.7). Similarly, as $a_{24}^{(t)} = 0$ for $t < t_c$ there is a steep drop of $\mathbb{E}[\pi_{24}^{(t)}]$ (dashed red line in Fig. 7.7). For $t > t_c$, we can see that the model responds by slowly shifting the posterior mean. In fact, even though the number of observations after the changepoint is the same as the one before t_c, the model fails to reach an appropriate value in both cases; there exists a weight of prior knowledge that forces the model to expect new data that conform with the system state before t_c.

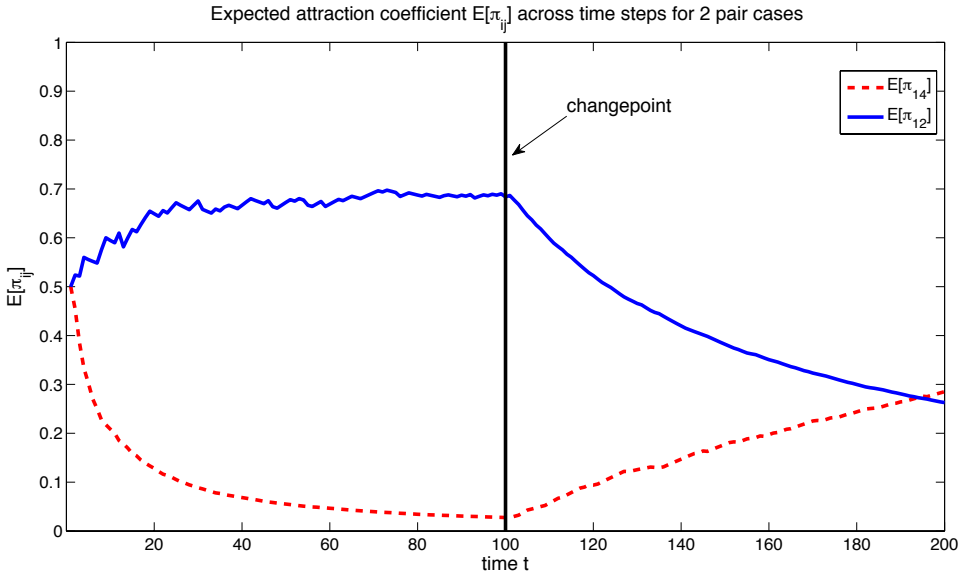


Figure 7.7: We demonstrate the responsiveness of the model at the presence of a particular changepoint. For the first 100 time steps, nodes 1 and 2 tend to point to the same locations in the bipartite network and model behaves exactly as in Fig. 7.3(b). After a particular step t_c, 1 and 2 stop having common locations, so the the attraction coefficient starts to drop. The reason for such a slow drop after t_c is the fact that past observations (before t_c) are strongly weighted in the model.

The limitation described above is not a drawback of Bayesian learning in general; in fact, the model behaves exactly as it should in settings where the underlying mechanism that generates the observations is *stationary*. By implying such a stationarity in the system, we are effectively making our model *changepoint-naive* and our inferences very conservative as the system is changing. In order to handle such a case, we have to control the way prior and current information is fused, by introducing a mixing coefficient γ_{ij} that maps each update equation (7.8,7.9) to a convex sum:

$$\kappa'_{ij} = \gamma_{ij}\kappa_{ij} + (1-\gamma_{ij})a_{ij} \tag{7.13}$$

$$\lambda'_{ij} = \gamma_{ij}\lambda_{ij} + (1-\gamma_{ij})(x_{ij} - a_{ij}). \tag{7.14}$$

At this stage, the parameter γ is determined manually and depends on the application. To automate the selection of γ the simplest approach is to use a bank of filters each with different values for γ. The choice of γ can then be determined using the probability of the next observation $a_{ij}^{(t+1)}$ under the posterior predictive distribution [Cesa-Bianchi and Lugosi, 2006] in Eq. 7.10.

Whilst simple to implement, the approach could quickly become computationally expensive compared to alternative approaches based on generalised linear models, which update the log-odds of the attraction coefficients using dynamic linear model recursions [Penny and Roberts, 1999]. Both approaches avoid ad hoc heuristics for the selection of γ and allow us to detect changepoints and dynamically control the degree to which we mix past and present information in order to perform optimal predictions.

For the purposes of our analysis in Chapter 8, we break down our observation period $\{1, ..., T\}$ into segments of zoologically meaningful duration (a month of data collection) and at the beginning of each segment we reset the BOMP parameters $\kappa_{ij}, \lambda_{ij}$ to the initial value of 10, for all i, j.

7.5 Discussion

In this chapter we have presented a probabilistic approach for one-mode projection temporal bipartite networks, in order to infer latent associations between the node class of interest. Such inferred associations π_{ij} are parameters of a binomial noise model, which can be viewed as link probabilities for all pairs of nodes, effectively mapping the temporal bipartite network to an ensemble of possible graphs. This is a very attractive aspect of our method, as, along with the distributions over π_{ij}, it fully captures the uncertainty over connectivity patterns. Additionally, the model benefits from constant influx of new information by updating our current beliefs over the network connectivity based on more recent observations.

Our approach consists of processing the data stream in a T-number of time slices and

updating the model parameters based on new observations received at t and prior knowledge from previous steps $0, ..., t-1$. Such a fusion of information from both current observations and past experience lies at the heart of every Bayesian learning model and allows us to perform rigorous inference in real-world settings where noise and missing observations are prevalent. Indeed, we have already shown in benchmark tests that although the Simple Ratio index $\hat{\pi}_{ij} = a_{ij}/x_{ij}$ (red line) fluctuates, our probabilistic treatment allows us to extract a smooth trend of how the association score π_{ij} progresses through time.

Capturing uncertainty and performing smooth estimates over temporal link weights is only one aspect of the method. The probability of connection or association between any pair of nodes can also be used for link completion tasks or personal recommendation tasks, while macroscopic topological properties of the inferred networks can now be described in the form of distributions, for example, for clustering coefficients, geodesic distances and diameters. That allows us to study the stability of the overall structure, in terms of the variability of properties such as community structure, homophily, small-world effect.

From a zoological perspective, our methodology allows us to extend the traditional association indices from the animal social network literature, to a probabilistic setting, where:

- Our beliefs over the presence/absence of an association and its intensity are formally described via the use of probability distributions.
- Any estimation of social tie between two individuals takes into account past data and the extend to which past observations are considered can be fully controlled.
- Our estimates are more robust to noise and missing observations, problems prevalent in sensor-generated data sets.

In Chapter 8, we apply the developed methodology to the Wytham Woods data set introduced in Chapter 5, in order to explore wild bird social structure. Extracting gathering events from logging data via the GEM model presented in Chapter 6 and seeking to define

associations between common members of a foraging flock is a problem well suited for the concept of one-mode projection and BOMP will be essential for our zoological analyses.

Chapter 8

Social Behaviour of the Great Tit

8.1 Introduction

In this chapter we focus on analysing the great tit (GT) wild bird population from a social network perspective. Our goal is to show how the methodological advances introduced in previous chapters can be applied to study the biological correlates of sociality, by examining the relationship between social network quantities and zoological properties at individual and population level. We make use of GEM (Gathering Events Method) from Chapter 6 in order to extract flocking structure (termed "gathering events") from the raw tracking data presented in Chapter 5, where we identify regions of statistically significant observation density as small foraging groups. We define the social network between birds based on their co-occurrence in such foraging flocks, using the one-mode projection methodology BOMP, discussed in Chapter 7. The community structure of such networks is then extracted via the nonnegative matrix factorisation CD-NMF approach proposed in Chapter 4 and used to explore the wild bird social circles.

In Section 8.2, we present the process upon which we apply the proposed methods on the wild bird tracking data, in order to extract a sequence of daily and monthly network "snapshots" that express the evolution of wild bird sociality across the season. We proceed in

Section 8.3 by examining various properties of the extracted graphs in terms of link structure, local densification and community structure. In Section 8.4 we investigate the relationship between network quantities and bird characteristics such as mobility, sex and dispersion. Based on our findings we proceed in Section 8.5 by investigating the connection weights between mating pairs, across time. We show how co-occurrences between breeding partners are concentrated towards the breeding period and use the extracted community organisation of the inferred graphs to seek evidence of pre-existing social bonds earlier in the season.

8.2 Social graph extraction from wild-bird sensor records

Before performing our analyses, we have to extract the social graphs by applying GEM on the data set of wild-bird foraging records presented in Chapter 5. Our observations consist of two main streams; $\mathcal{D}^{(7,8)}$ that covers the activity of $N_{7,8} = 770$ birds from August 2007 to March 2008 and $\mathcal{D}^{(8,9)}$ that spans from August 2008 to March 2009 and contains $N_{8,9} = 753$ birds.

Instead of applying our method to the whole multi-season data stream directly, we break it down into 24-hour segments. An example of the observation data is illustrated in Fig. 5.7 of Chapter 5, where we have shown that the isolated observation-rich regions correspond to a particular day of data collection and they are separated by the night period (no-observation zones in between days), where no bird foraging activity takes place.

Our task is to produce a series of network "time slices" that allow us to study the day-by-day changes in the population's sociality. Due to population-coverage issues relating to the feeder rotation scheme, discussed in Chapter 5, we also produce network slices at a lower month-by-month resolution, by taking aggregates of the daily networks through an appropriate "stacking" of their gathering event matrices.

Having a batch $\mathcal{D}^{(t)}$ of data that correspond to a given day t, we proceed by breaking down each $\mathcal{D}^{(t)}$ into sub-streams that correspond to L different feeding locations, as shown

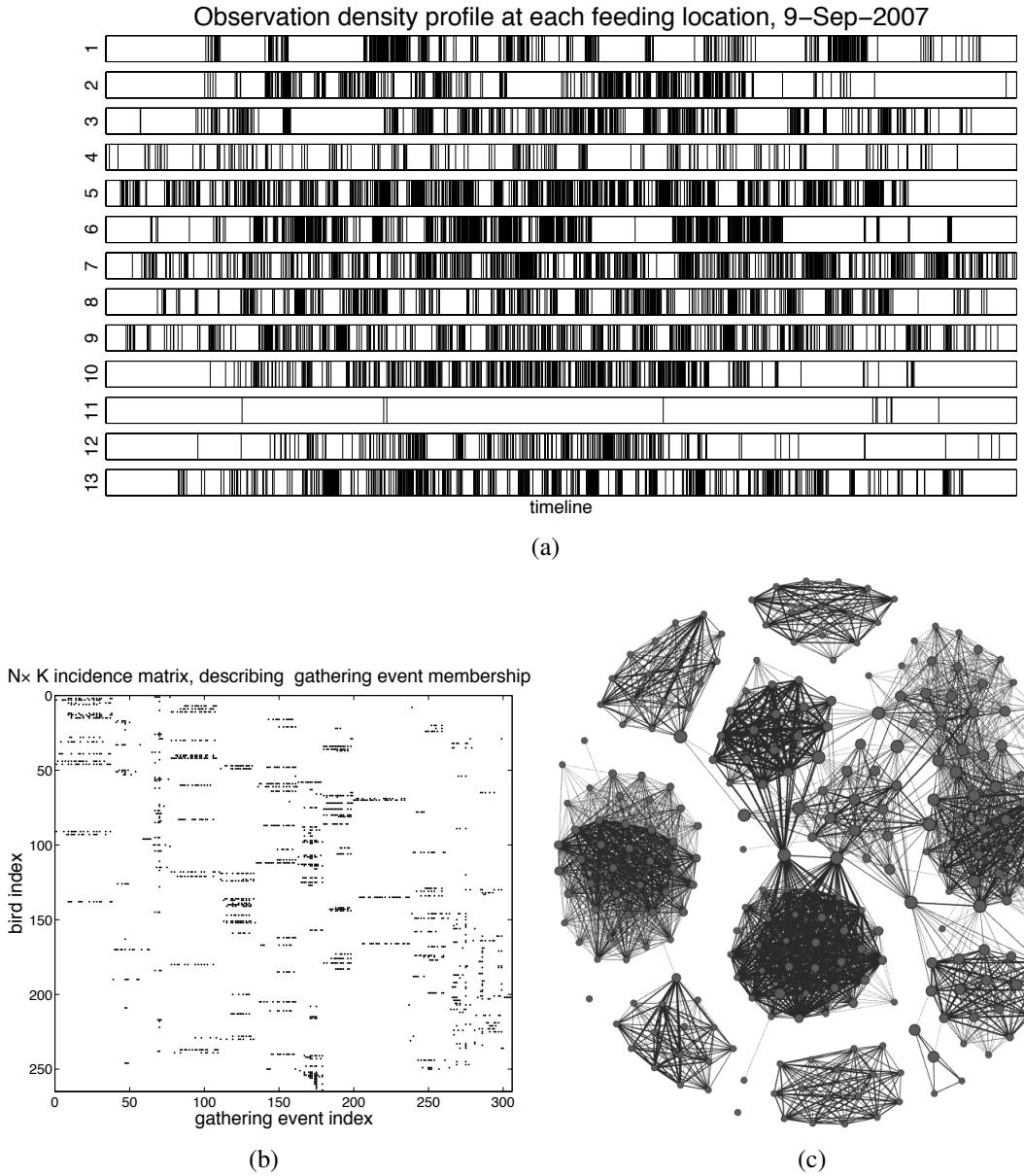


Figure 8.1: We take every daily batch of observations and break it down into separate streams that refer to bird records at each particular location, shown in Fig. 8.1(a). For each location-specific stream, we apply GEM in order to extract the bird-to-event bipartite graph, illustrated in Fig. 8.1(b). We then use BOMP to project the bird-to-event bipartite graph of Fig. 8.1(b) into an one-mode network, shown in Fig. 8.1(c), based on co-occurrences in gathering events.

in Fig. 8.1(a). We apply GEM at each location ℓ separately, as gathering events need to be defined both in terms of temporal and spatial proximity. On each feeder-specific stream $\mathcal{D}_\ell^{(t)}$ of day t, our method identifies $K_\ell^{(t)}$ bursts of foraging activity and builds a bipartite network,

described by $\mathbf{B}_\ell^{(t)}$, between N birds and $K_\ell^{(t)}$ gathering events. The $N \times K_\ell^{(t)}$ incidence matrices $\mathbf{B}_\ell^{(t)}$ across all sites L are then aggregated to a single $N \times K^{(t)}$ matrix $\mathbf{B}^{(t)} = [\mathbf{B}_1^{(t)}|\mathbf{B}_2^{(t)}|...|\mathbf{B}_L^{(t)}]$ with $K^{(t)} = \sum_{\ell=1}^{L} K_\ell^{(t)}$, based on the concatenation scheme discussed in Section 6.6.3 of Chapter 6. Each element $b_{ik}^{(t)}$ of $\mathbf{B}^{(t)}$ is 1 if bird i appeared at gathering event k and 0 otherwise. For 2007–8, the median number of gathering events per day is 534, with a mean value of 712.08 and standard deviation of 632.34. For the 2008–9 data set, the corresponding numbers are 129 for the median value, 148.82 for the mean and 114.26 for the standard deviation. Figures relating to the duration and number of birds per gathering event are provided in Table 8.1.

Table 8.1: Gathering event statistics

Statistic	No. of birds			Duration (secs)		
	Median	Mean	Std. dev.	Median	Mean	Std. dev.
2007–8	70	85.33	63.43	330	615.02	761.72
2008–9	76	76.9	47.78	450	657.14	711.93

For the purposes of our analyses, we also make use of the null model described in Section 6.6.2 of Chapter 6 in order to extract null gathering event matrix sequences $\mathbf{B}_0^{(t)}$, which we use for various comparative analyses.

Following the discussion of 6.6 in Chapter 6, along with the model introduced in Chapter 7, we extract the $N \times N$ great tit social network at time slice t, starting with the simple co-occurrence matrix $\mathbf{A}^{(t)}$ where for each link $a_{ij}^{(t)}$ we have $a_{ij}^{(t)} = \sum_{k=1}^{K^{(t)}} b_{ik}^{(t)} b_{ij}^{(t)}$. We then use $\mathbf{A}^{(t)}$ to calculate the smooth co-occurrence matrix $\tilde{\mathbf{A}}^{(t)}$, derived from the expected values $\mathbb{E}[a_{ij}^{(t)}]$ of the link weight posterior distribution under the BOMP model. This projection incorporates a "memory" term, by taking into account past data in order to make our inferred co-occurrence estimates $\tilde{a}_{ij}^{(t)}$ more robust to missing observations. Additionally, BOMP up- and down-weights bird co-occurrences from $\mathbf{A}^{(t)}$, based on their exclusivity in gathering event participation as discussed in Section 7.4.2 of Chapter 7, allowing the new adjacency matrix $\tilde{\mathbf{A}}^{(t)}$ to "magnify" important associations.

For the purposes of our analyses, the above projection schemes are performed both on the observed gathering event matrices $\mathbf{B}^{(t)}$ and the null model ones $\mathbf{B}_0^{(t)}$. In Fig. 8.1(c) we show an example of a network that describes bird social structure on a given day of 2007, built using the above scheme. We repeat the process for all T 24-hour segments of our data stream and build a stack of adjacency matrices $\{\mathbf{A}^{(t)}\}_{t=1}^{T}$ that represent daily snapshots of the wild-bird social network.

8.3 Analyses of network connectivity

We begin our analysis by presenting some global topological properties of our temporal networks. In Fig. 8.2 we present the network density (ND) of each co-occurrence matrix $\mathbf{A}^{(t)}$ throughout the two seasons. Network density is defined as the ratio of existing links M divided by the maximum number of possible connections $\frac{1}{2}N^{(t)}(N^{(t)} - N^{(t)})$, where $N^{(t)}$ denotes the number of birds that appear in the feeder data at time t. We also monitor the participation rate defined as $\mathtt{PT}^{(t)}_{\text{total}} = N^{(t)}/N_{\text{total}}$, defining the fraction of population coverage of our data at time t.

We are considering two temporal resolutions, "daily" in Fig. 8.2(a) and 8.2(b) and "monthly" in Fig. 8.2(c) and 8.2(d), both of which yield sparse network sequences with average network density $\bar{\mathtt{ND}}^{(7,8)}_{\text{day}} = 0.15$ and $\bar{\mathtt{ND}}^{(7,8)}_{\text{month}} = 0.08$ for 2007–8, along with $\bar{\mathtt{ND}}^{(8,9)}_{\text{day}} = 0.10$ and $\bar{\mathtt{ND}}^{(7,8)}_{\text{month}} = 0.05$ for 2008–9.

The linear correlation coefficient between participation score and population coverage is $\rho^{(7,8)}_{\text{day}}(\mathtt{PT}, \mathtt{ND}) = -0.45$ and $\rho^{(7,8)}_{\text{month}}(\mathtt{PT}, \mathtt{ND}) = -0.13$ for season 2007–8, along with $\rho^{(8,9)}_{\text{day}}(\mathtt{PT}, \mathtt{ND}) = -0.39$ and $\rho^{(8,9)}_{\text{month}}(\mathtt{PT}, \mathtt{ND}) = -0.18$ for season 2008–9. The reported correlations are deemed statistically significant with p-values $< 10^{-6}$, across all tests. Thus observing a higher percentage on individuals in the population (higher number of nodes N) is not associated with an increase in social ties (links M), implying a sparse topology with

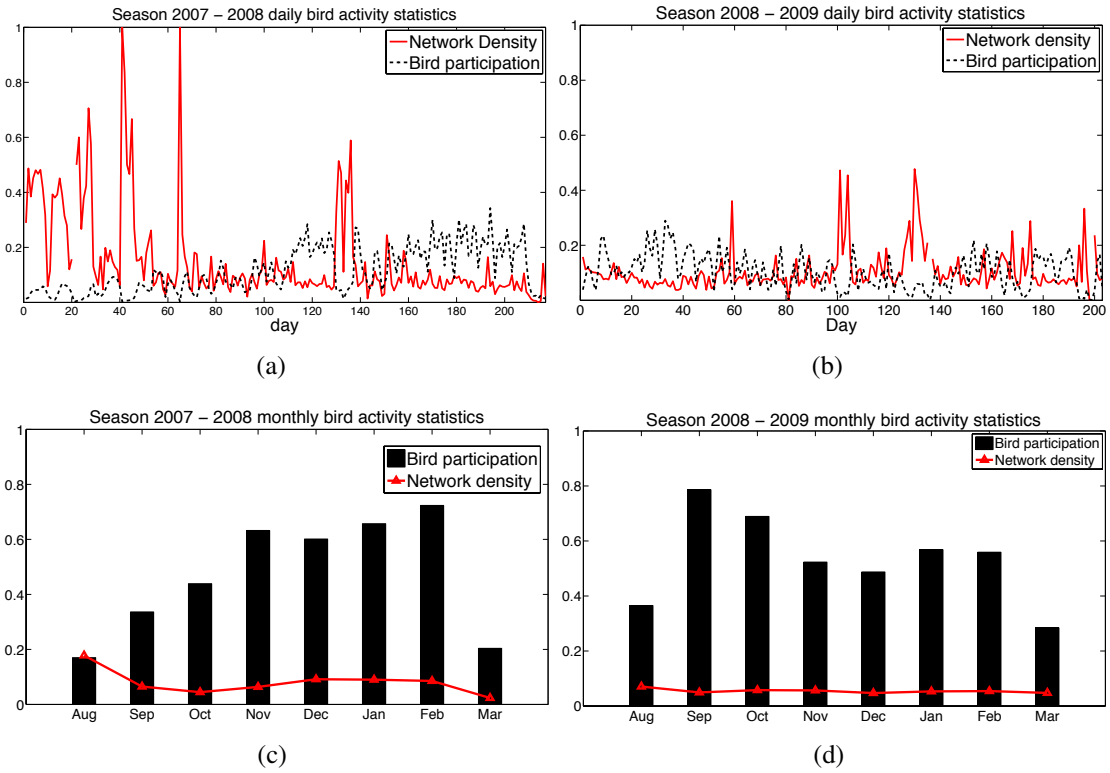


Figure 8.2: We plot the network density (ND) of each co-occurrence matrix across the season. We show that both on a daily and monthly resolution, the extracted networks are sparse, yielding a low ND that is not correlated with the bird participation score.

evidence of strong preferential component[1], which we investigate below.

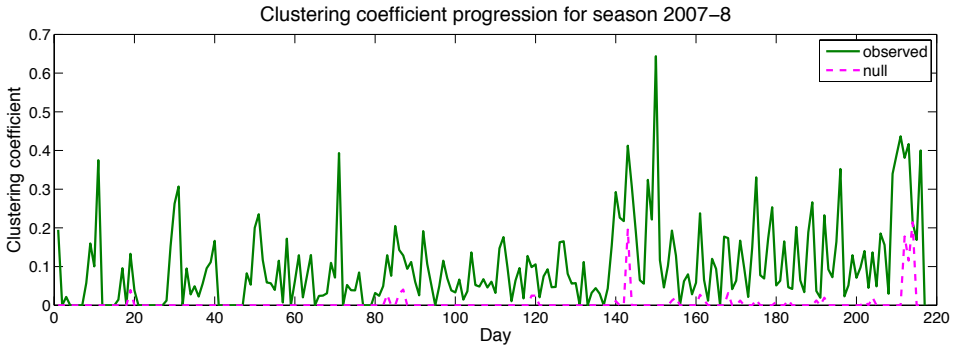
We now seek to examine if preferential structure exists, or if it is a mere artefact of the spatial component of our data. By looking at local network "neighbourhoods", we calculate the *average clustering coefficient* CC by following the approach presented in [Barrat et al., 2004]:

$$\mathrm{CC} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{d_i(d_i-1)}\sum_{j,k}\breve{a}_{ij}\breve{a}_{ik}\breve{a}_{jk}\right), \tag{8.1}$$

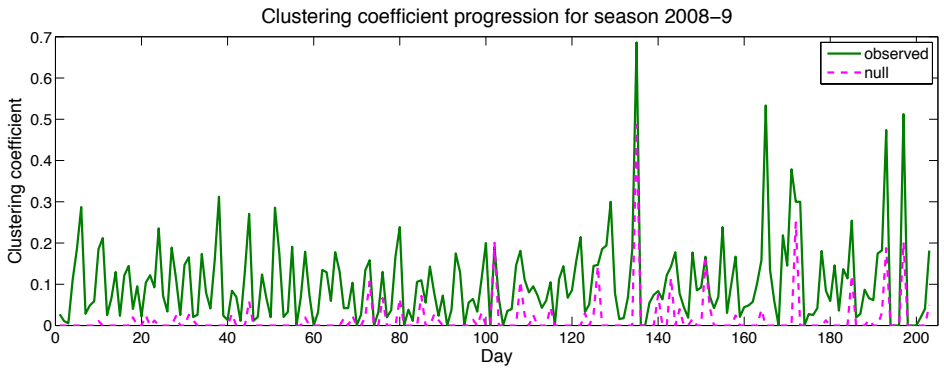
where d_i is the degree of node i and $\breve{a}_{ij}$ is the binarised element a_{ij} of $\mathbf{A}$. The average clus-

[1]We use the term "preferential" not in terms of the preferential attachment growth model by [Barabási and Albert, 1999], but in the context of individuals being stringent on how to allocate their social connections, by preferring a particular subset of the population.

tering coefficient can be seen as a good indicator of preferential structure, as it implies local densification of the network where tightly knit groups cluster together; if two individuals i and j are connected to the same neighbour k, $\mathtt{CC}$ defines the empirical likelihood that i and j will be connected too. In Fig. 8.3 we show the progression of $\mathtt{CC}^{(t)}$ for our two data sets, for the 2007–8 (Fig. 8.3(a)) and 2008–9 period (Fig. 8.3(b)) respectively, comparing it with the corresponding $\mathtt{CC}_0^{(t)}$ for the null network, which, based on the discussion in Section 6.6.2 of Chapter 6, retains the spatial clustering component of the data. Results show a consistently higher presence of triangle formation in the inferred networks, compared to the null networks, implying a strong preference between individuals to form closely connected cliques, a common property across real-world networks, including ecological ones [Wey et al., 2008].



(a)



(b)

Figure 8.3: For each daily snapshot of the network, we monitor the average clustering coefficient $\mathtt{CC}$, which measures the degree of triangle presence in the network. Results for both data sets show a strong 3-clique presence, which is consistently higher than the null model across both seasons.

We also study densification in our temporal networks at mesoscopic scale. As discussed in Chapter 2, communities in networks represent regions of increased observation density, usually with important functional properties for the graph as a whole. We apply CD-NMF on each daily adjacency matrix $\tilde{\mathbf{A}}^{(t)}$, extracted via GEM and BOMP, and calculate the cohesiveness of each solution via the Newman-Girvan modularity Q. In Fig. 8.4 we plot Q at each step t (blue line) and compare it against the modularity score of $R = 10,000$ instances of the corresponding null graphs (red line). The observed and null graph partitions at each step are compared via the use of a Normalised Mutual Information (NMI) score by [Danon et al., 2005] (green line). We can see in both Fig. 8.4(a) and Fig. 8.4(b) that the wild bird networks possess a highly modular organisation, persisting throughout the season, apart from days t of low population coverage.

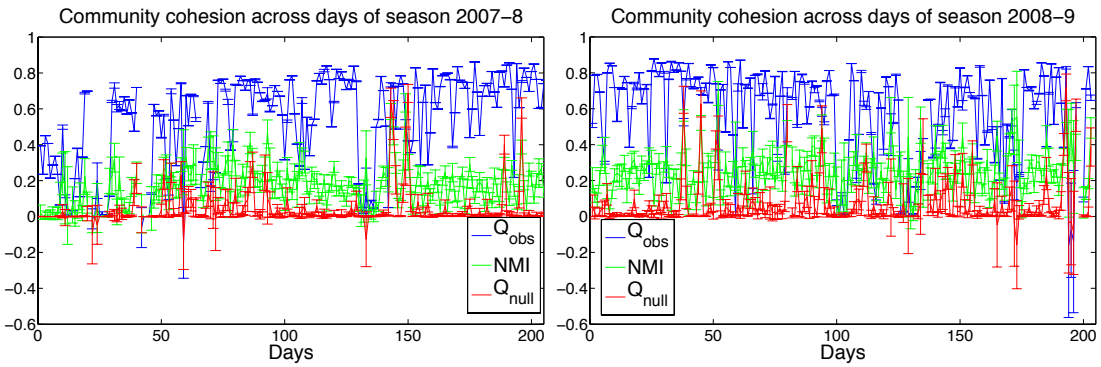


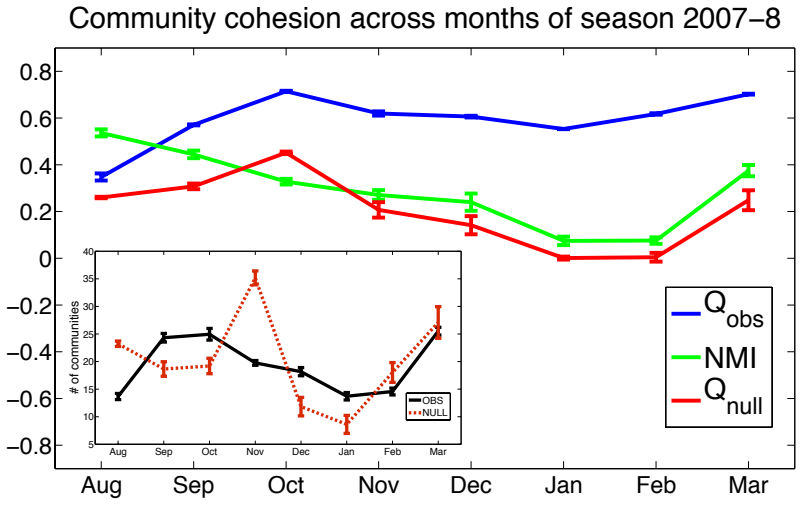
Figure 8.4: We plot the modularity scores for the observed (blue line) and null networks (red line) across all days of our two observation seasons. For each day t, we also compare the observed and null partition via the use of Normalised Mutual Information score (green line). "Gaps" represent days where community structure can not be defined due to low population coverage (less than 3 birds or all extracted communities have one member).

For the above reason, we also investigate community organisation at a monthly resolution, in order to avoid modularity biases derived from isolated bird co-occurrences at days of low sensor volume. We begin by "stacking" the gathering events matrices $\mathbf{B}^{(m)} = [\mathbf{B}^{(t_1)}|\mathbf{B}^{(t_2)}|...]$ for all days within month m and perform BOMP in order to extract $\tilde{\mathbf{A}}^{(m)}$. For extracting the

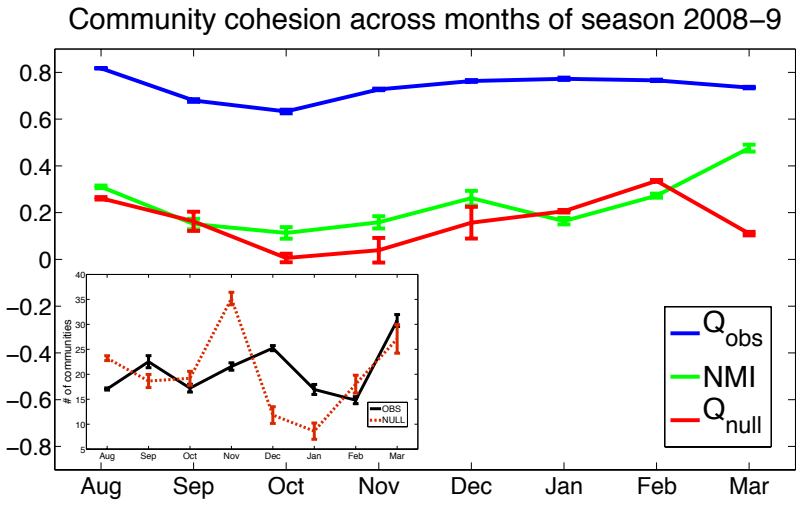
null network, we start by randomising each $\mathbf{B}^{(t)}$ separately in order to preserve the temporal and spatial habits of birds, but at the same time "break" correlations in gathering event membership structure that are induced by social affiliation. We proceed stacking all null gathering events matrices $\mathbf{B}_0^{(m)} = [\mathbf{B}_0^{(t_1)}|\mathbf{B}_0^{(t_2)}|...]$ and project, via BOMP, to $\tilde{\mathbf{A}}_0^{(m)}$. For each $\tilde{\mathbf{A}}^{(m)}$ and each instance of $\tilde{\mathbf{A}}_0^{(m)}$ we calculate the modularities and the NMI score in the same experimental setup as previously. We can see in both Fig. 8.5(a) and Fig. 8.5(b) that the results are consistent with the daily case examined previously, where the observed graphs possess a very modular community structure compared to the null case. Such a difference between the observed and null graphs implies evidence that birds consistently"focus" their co-occurrences to particular members of the network, both in terms of linkage and weight. Such a hypothesis is investigated below.

We now narrow our investigation to link-level, by examining how the total co-occurrences $s_i^{(t)} = \sum_{j=1}^{N} \tilde{a}_{ij}^{(t)}$ of an individual i on day t, also known as *strength* in network analysis jargon, are distributed across adjacent nodes. Our goal is to investigate if co-occurrences are spread evenly between foraging birds or if there is some form of preference to certain individuals. Thus given an individual i and day t, we take the i-th row (or column) of the adjacency matrix $\tilde{\mathbf{A}}^{(t)}$ and divide each element by the strength $s_i^{(t)}$ of the node. This leads us to an empirical *preference distribution* $\mathbf{q}_i^{(t)} \in \mathbb{R}^N$ over nodes, where each element $q_{ij}^{(t)}$ expresses the fraction of total connection strength $s_i^{(t)}$ of i absorbed by j.

Based on the above, we quantify the "uneven-ness" of $\mathbf{q}_i(t)$ using the Shannon entropy $H^{(t)}(i) = -\sum_{j=1}^{N} q_{ij}^{(t)} \log q_{ij}^{(t)}$ that becomes zero when only one neighbour of i absorbs all node strength and increases as $\mathbf{q}_i^{(t)}$ approaches the uniform distribution. We summarise the link weight entropy for all nodes in the network $\tilde{\mathbf{A}}^{(t)}$ as the mean $H^{(t)} = N^{-1} \sum_{i=1}^{N} H^{(t)}(i)$. This measure of "preference" in the network is then compared against the null case, where $H_0^{(t)}(i)$ and $H_0^{(t)}$ are built using the null adjacency matrix $\mathbf{A}_0^{(t)}$ presented in Section 6.6.2 of Chapter 6. In Fig. 8.6, we present how the mean entropy ratio $H^{(t)}/H_0^{(t)}$ progresses during



(a)



(b)

Figure 8.5: We plot the modularity scores for partitions of the observed (blue line) and null networks (red line), for all monthly aggregate networks of our two observation seasons. For each month m, we also compare the observed and null partition via the use of Normalised Mutual Information score (green line). Embedded figures show the number of extracted groups for the observed (black line) and null graph (red line).

each season. We can see that the observed $H^{(t)}$ is consistently smaller than $H_0^{(t)}$ for both seasons, implying the presence of a bias between individuals to forage more often with particular members of their immediate social circle.

Up to this point, we have performed our analyses by analysing the link and weight struc-

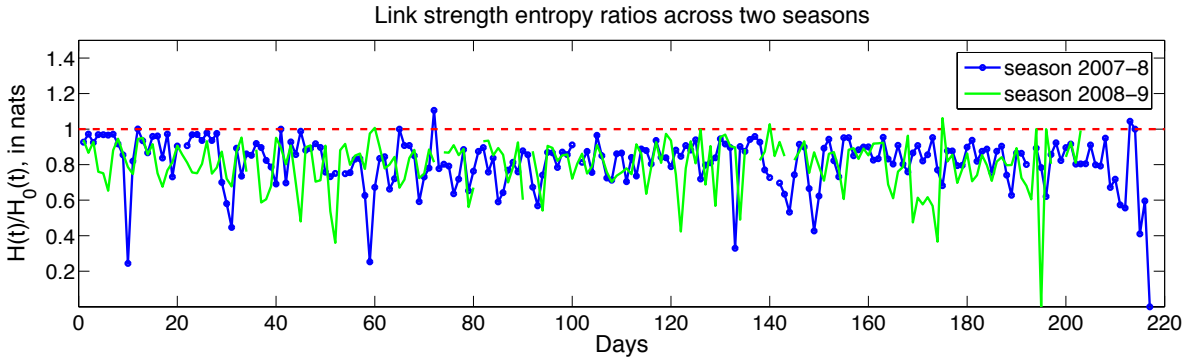


Figure 8.6: For each daily network, we monitor the mean link weight entropy $H^{(t)}$ of nodes, where higher values express the preference of birds to have more co-occurrences with a specific subset of their network neighbours. We compare our results with the mean entropy $H_0(t)$ of the respective networks created under our null model, via the ratio $H^{(t)}/H_0^{(t)}$. Values below 1 imply a consistent trend of birds to have a preference towards particular network neighbours (foraging partners).

ture of the inferred graphs. In the next section, we examine how network quantities relate to quantities associated with bird behaviour and demographics.

8.4 Network quantities versus bird features

In the previous section we investigated various topological characteristics of the inferred networks, which relate to localised densification and preferential allocation of link weight. In this section we seek to examine the relationship of network connectivity with bird behavioural and demographical properties. In the same theme as in the previous section, we use the monthly networks $\tilde{\mathbf{A}}^{(m)}$ for both seasons and examine various assortativity scores [Newman, 2003a], which we compare against instances our null model.

We start by considering the monthly bird occurrences; that is, total number of times $o_i^{(m)}$ bird i was recorded across Wytham. In Fig. 8.7(a) and Fig. 8.7(b) we illustrate to what extent birds with similar occurrence number links together, with results showing higher assortativity score than expected from the null hypothesis. Such increased assortativity, although expected

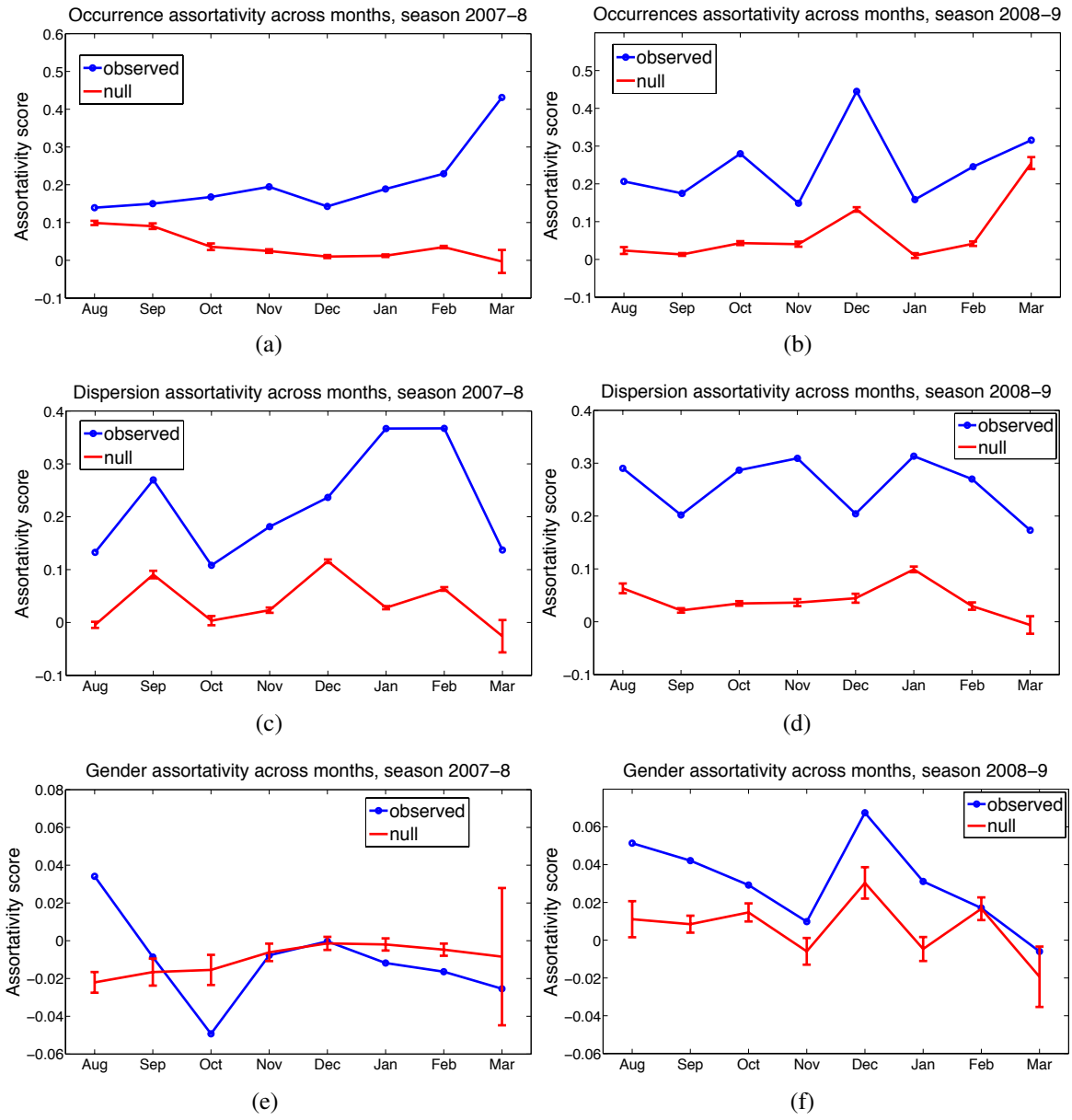


Figure 8.7: We plot various feature-based assortativity quantities for the inferred graphs and compare them against the null model. For each experiment we use networks generated via monthly aggregates, both for the observed and null topologies.

on the basis that individuals are linked because they are recorded at the feeders in the first place, may be important when considering issues relating to dominance and exploration-exploitation. Regarding the former, great tits arrive at the feeder and compete in a small space at collecting sunflower seeds. Thus on one hand, it is reasonable to assume that individuals

that make the effort to crowd around the feeder and achieving more food intake, are dominant to the others. On the other hand, individuals that are captured very regularly from the feeder sensors are birds which effectively exploit a "free food" source in their environment, instead of foraging naturally. Those individuals may possess personality traits that make them more prone to exploit a direct food source instead of exploring their environment. Both of such hypotheses are to be tested with additional tracking and genetic data, as we discuss in Chapter 9.

We also seek to examine the relationship between link structure and the area covered by individual birds. Towards this goal, we introduce the notion of (monthly) *dispersion entropy* $\mathtt{DH}_i^{(m)}$ given an individual i, defined as follows: for each bird i we produce a histogram of its visitation counts, during month m, at each one of the $L = 67$ locations across Wytham. By considering the normalised version of such a histogram, we define $\mathtt{DH}_i^{(m)}$ as the Shannon entropy of such a empirical distribution over locations given i, which measures how uniformly i spreads its occurrences across the $L = 67$ sites. In Fig. 8.7(c) and Fig. 8.7(d) we plot its assortativity score and show that in both seasons, there is evidence that individuals with similar site fidelity patterns tend to be connected. This relates to our previous discussion of homophily among exploratory individuals and provides additional weight of evidence to such an assumption.

Finally, we seek to examine the sex-based assortativity of our social graphs. In Fig. 8.7(e) and Fig. 8.7(f) we plot the corresponding assortativity score across months for both seasons. In contrast to our previous analyses, we find minimal evidence of homophily, with scores close to zero and variation that is inconsistent across the two seasons. This may result from looking at such correlations at a global network scale, where the loose connections between random pairs "cancel-out" the effect of strong male-female mating bonds that we know a priori they exist in such a society. For that reason, in the next section we expand on such a sex analysis by investigating the relationship between graph connectivity and mating pair

formation, via the extracted community organisation of the bird temporal networks.

8.5 Mating pair formation

We seek to examine how the mixed-sex structure, discussed in the previous section, relates to mating-partner selection. We begin our analysis on a season-wide scale, by defining a *summary* adjacency matrix $\mathbf{A}^{(y)} \in \mathbb{R}^{N\times N}$ as the sum of all "time slices" $\mathbf{A}^{(t)}$ of $\{\mathbf{A}^{(t)}\}_{t=1}^{T}$ across t, $\mathbf{A}^{(y)} = \sum_{t=1}^{T} \mathbf{A}_t$, so that each element $a_{ij}^{(y)}$ denotes the total co-occurrences between bird i and j during the course of the whole season. For each individual i, we extract an annualised version of the preference distribution $\mathbf{q}_i^{(y)}$ presented previously and examine characteristics of neighbours which are connected via the top 25% link weights, absorbing the most strength s_i of node i. By exploiting the pedigree data set, we find that for the 70.34% of individuals in season 2007–8 and 75.49% in season 2008–9 for whom we have mating information, connections with their mating partners belong to the top 25% link values per node. Such results, presented for various other thresholds in Table 8.2, imply that mating partners tend to be strongly connected in the extracted networks. It is important to note that in this section, although we focus our investigation on mating pairs, all calculations have been performed using all network links, regardless of their connection strength.

Table 8.2: Percentage of mating dyads across top % of connection weights

Threshold	50%	25%	15%	10%
2007–8	88.11	70.34	68.02	56.27
2008–9	91.63	75.49	74.91	67.81

We continue our analysis at a more temporal setting, by examining how the total bird co-occurrences are spread throughout the year. For both seasons, given the total annual co-occurrences matrix $\mathbf{A}^{(y)}$ and the sequence $\{\mathbf{A}^{(t)}\}_{t=1}^{T}$ of daily network slices, we apply the following scheme:

1. We calculate the co-occurrences $a_{ip}^{(m)} = \sum_{t\in T_m} a_{ip}^{(t)}$ that took place during month m,

where p is the mating partner of individual i and T_m is the set of days in month m.

2. We normalise $a_{ip}^{(m)}$ by taking the fraction $\mathtt{f}^{(m)}(i,p) = a_{ip}^{(t)}/a_{ip}^{(y)}$ of monthly versus annual co-occurrences, for each bird i and its mating partner p.

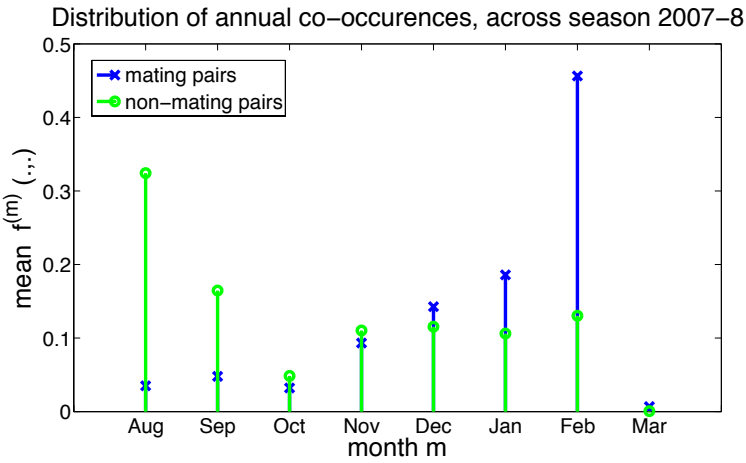
3. We take the average $\bar{\mathtt{f}}^{(m)}(i,p)$ across all mating pairs at month m.

4. We compare the above quantity against $\bar{\mathtt{f}}_0^{(m)}(i,j)$, which is calculated in the same manner as $\bar{\mathtt{f}}^{(m)}(i,p)$ but considering only "random" (non-mating) pairs i,j, for which $j \neq p$. This quantity expresses how the annual co-occurrences $\mathbf{A}_{ij}^{(y)}$ between a non-mating pair are spread, on average, across each month m.

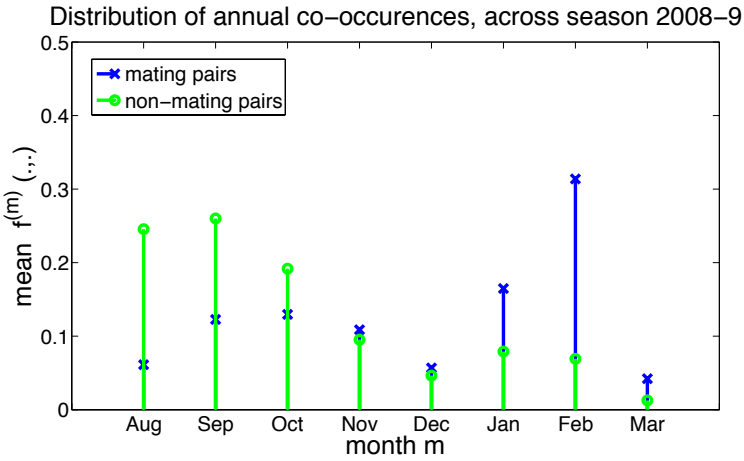
In Fig. 8.8 we plot the $\bar{\mathtt{f}}^{(m)}(i,p)$ (blue $\times$-stem) and $\bar{\mathtt{f}}_0^{(m)}(i,j)$ (green $\circ$-stem) values across the 8 months of season 2007–8 and 2008–9 respectively. We can see that co-occurrences between non-mating individuals are concentrated at the beginning of the season, while the ones between mating individuals increase as we approach the breeding season. Note that the small values at the end are the result of incomplete data during March, as only a very small fraction of birds with mating information appeared.

Based on the above analysis, we have shown that the inferred networks capture the fact that mating pairs forage together more often as we approach the breeding season. We now seek to investigate if pair formation, which fully materialises towards March, results from some kind of prior social affiliation. For that reason, we examine if "future partners" previously interacted with each other within the same social circles.

We extract such social circles via the use of CD-NMF, by applying the community detection algorithm at each daily network described by $\tilde{\mathbf{A}}^{(t)}$ and we examine the membership of mating pairs in such a mesoscopic structures. We find that the majority of mated pairs in network communities are connected through a direct link in 77.26% of cases for the 2007–8 data and 71.57% of cases for the 2008–9 data. Reachability through a path of two links is reported for the 14.74% of cases in 2007–8 and 17.06% of cases in 2008–9. The average path



(a)



(b)

Figure 8.8: We show how the annual co-occurrences between individuals allocated throughout two data collection seasons. We examine two different kinds of co-occurrences; mating pairs (blue $\times$-stem) and random (non-mating) pairs (green $\circ$-stem). We can see that in both cases, co-occurrences between mating pairs are more concentrated towards the end of each data collection period (the breeding season).

length between two members, for the cases where both of them are observed in the data, is 1.33 (2007–8) and 1.46 (2008-9) with median value of 1 in both data sets. Finally, there are still cases (8% in 2007-8 and 11.37% in 2008-9) of pairs where their geodesic distance spans from 3 to 6 edges but still belong to the same community.

In a temporal setting, we monitor bird membership in these groups using a binary matrix $\mathbf{C}_t$, where each element $c_{ijt} = 1$ encodes that birds i, j appeared in the same community

at day t. This leads us to a new collection of co-membership matrices $\{\mathbf{C}_t\}_{t=1}^{T}$ that encode temporal changes in the way birds participate with each other in communities. From a summation across t we get a matrix $\mathbf{C}^{(y)} \in \mathbb{R}^{N\times N}$ where each element $c_{ij}^{(y)}$ expresses the total number of days in the season where the pair i, j participated in the same community. In Fig. 8.9 we plot a histogram of all co-membership values based on two matrices $\mathbf{C}^{(y)}$ that refer to bird co-membership values in field seasons 2007-8 and 2008-9 respectively. We can see that for both seasons, the vast majority of dyads have never participated in the same group and the distribution is heavily skewed. This implies a strong preferential mechanism in the population, where random individuals rarely belong to the same social circle.

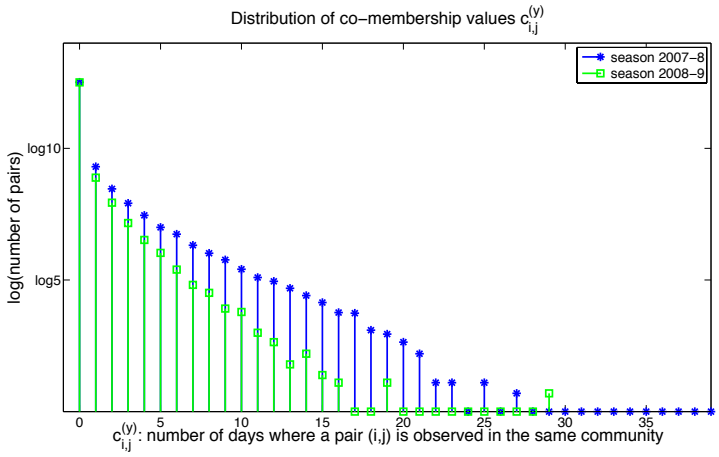


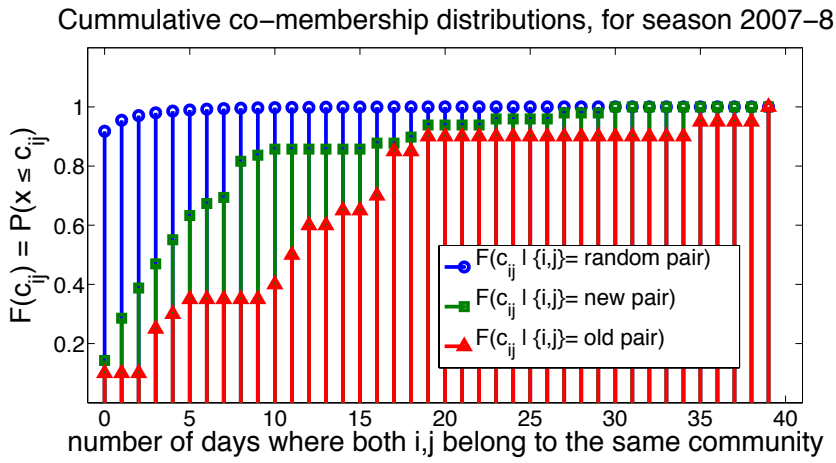
Figure 8.9: We plot the co-membership values of $\mathbf{C}^{(y)}$ on a logarithmic scale. On the horizontal axis we have the total number of days a random (non-mating) pair is observed in the same community. We can see that $\mathbf{C}^{(y)}$ is sparse and the vast majority of co-membership values are zero. This shows that if we pick a random dyad in the population, it will most likely be never seen in the same social circle.

We now examine if the above distribution holds for particular sub-category of pairs in the network, which we know a priori that are connected with actual social ties. This prior information is provided by the pedigree data set we mentioned previously, gives a list of node dyads i, j that are breeding individuals. In this list we also distinguish between mated pairs that were formed during our observation season, called *new pairs*, and others that already existed before, called *old pairs*. In Fig. 8.10 we plot the cumulative distributions $F(c_{ij})$,

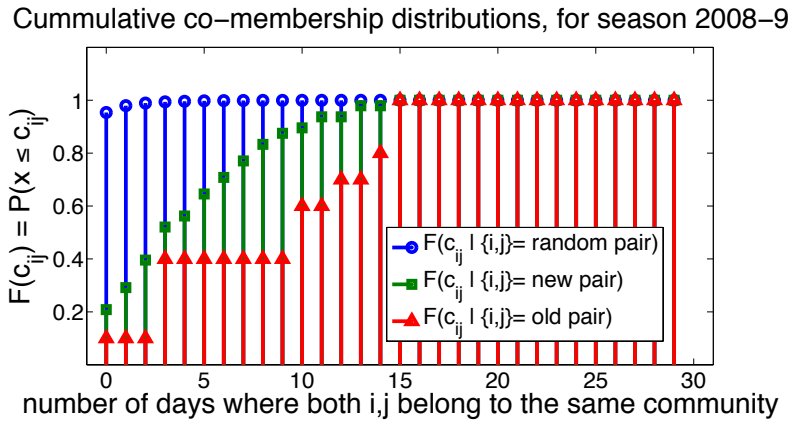
where c_{ij} are values co-membership matrix $\mathbf{C}^{(y)}$ and i, j can be a) any node pair (blue $\circ$-stem), b) a new pair (green $\square$-stem) and c) old pair (red $\triangle$-stem). In Fig. 8.10(a) we plot the distributions that refer to the 2007-08 season, with $N_{7,8} = 217$ individuals, from which we have 49 new pairs and 20 old pairs. For season 2008-9, shown in Fig. 8.10(b), we have $N_{8,9} = 203$ individuals that include 48 new pairs and 10 old pairs.

We can see that for both seasons presented in Fig. 8.10 the distributions that refer to mated pairs differ significantly from the one for random ones, with p-values $p < 10^{-15}$ under a Kolmogorov-Smirnov test [Lilliefors, 1967] with 5% precision level for both seasons. In contrast to the random case where values c_{ij} are mostly zero, co-membership for mated pairs achieves larger values thus implying stronger and consistent graph proximity. The differences between old and new pairs are also revealed between their respective cumulative distributions (green $\square$-stem and red $\triangle$-stem), where old pairs achieve higher co-membership values due to the fact that they existed before new pairs were formed, thus they had more opportunities during the season to participate in the same foraging flocks.

We have already seen that co-membership distributions differ between various pair types. We now examine when such a differentiation takes place during the observation season. We start by breaking down the observation period into 8 months. For each month, we used the respective daily networks in order to find the three co-membership distributions of interest. We then compared $p(c_{ij}|\{i,j\} = \text{random pair})$ versus $p(c_{ij}|\{i,j\} = \text{old pair})$ and $p(c_{ij}|\{i,j\} = \text{random pair})$ versus $p(c_{ij}|\{i,j\} = \text{new pair})$, by calculating the p-value under a Kolmogorov-Smirnov test with a proposed significance level 0.05. In Fig. 8.11 we can see that on one hand, at the beginning of the season, new pairs have similar co-membership patterns to random ones, as they have not been formed at such an early point. But as we move through the year, this similarity drops and from the "cloud" of random associations, breeding relationships emerge. On the other hand, old pairs that have already been formed from previous seasons have a consistent non-random co-membership pattern, even from very



(a)

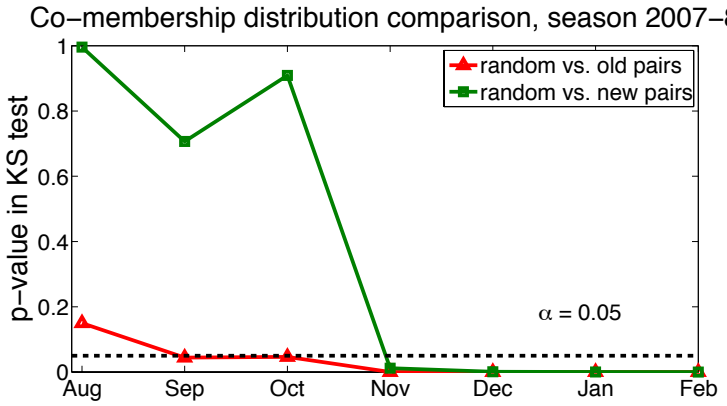


(b)

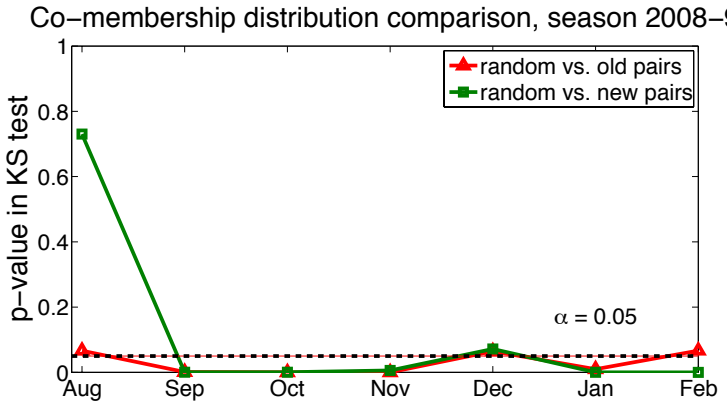
Figure 8.10: We plot the cumulative co-membership distributions for three different dyad types; random pairs, mating pairs formed in previous seasons and pairs that formed in the current season. Although for the majority of random bird pairs in the network co-membership values are concentrated around zero, breeding individuals tend to participate much more frequently into the same flocks.

early points in the season.

Therefore, we have showed that although breeding pairs start foraging together more often than random after December, new-pair partners belong to the same mixed-sex social circles earlier; from November during the 2007-8 season and September during the 2008-9 season. In contrast to new pairs, old pairs exhibit a quite stable co-membership profile, foraging together across the whole season.



(a)



(b)

Figure 8.11: We compare the co-membership distributions $p(c_{ij}|\{i,j\} = \text{random pair})$ versus $p(c_{ij}|\{i,j\} = \text{old pair})$ (red $\triangle$-line) and $p(c_{ij}|\{i,j\} = \text{random pair})$ versus $p(c_{ij}|\{i,j\} = \text{new pair})$ (green $\square$-line) in a month-by-month basis, using a Kolmogorov-Smirnov test. Values above the proposed $\alpha = 0.05$ significance threshold imply that the two distributions under comparison are similar. We can see that on one hand, from very early in the year old pairs differentiate themselves from random, by starting to participate frequently in the same communities. On the other hand, members of new pairs in the beginning of the year treat each other as random, while preferential mechanism that makes them flock together, starts to build-up during early winter.

8.6 Discussion and future work

In this chapter, we applied the methodological developments introduced in this thesis, in order to explore the Wytham Woods great tit population. We ran GEM on the two data sets from 2007-8 and 2008-9 and extracted the network topologies at day and month granularity.

We showed that the bird social networks display a range of properties, such as high clustering coefficient and strong community organisation, that are prevalent in real-world networks. By examining the connectivity structure in relation to biological quantities, we have reported evidence of homophily in terms of feeder use and site fidelity, while no such an assortativity is evident regarding the individuals' sex.

Following up on this finding of "sex-neutral" topology, we have examined the social ties of breeding pairs, known to us a priori from the pedigree data set. Such pairs forage together more often towards the breeding season and are positioned at close graph proximity throughout the year. Regarding new pairs, evidence of co-membership in the same social circles, before the formation of an actual mating bond, is reported for both observation seasons in early winter (September to November). Old pairs exhibit a more stable pattern, belonging to the same groups from the beginning of the season.

At this stage, our findings on mating bonds are general and focus solely on the inferred co-occurrence structure, without taking into account the heterogeneity among male-female relationships. For example, one can imagine a scenario in which any given male could have a very strong association with one female but weaker than expected association with neighbouring females, possibly because these are prevented by their own mates from meeting other nearby females. Such a heterogeneity can be captured via the use of multiplex ties, where co-occurrences a_{ij} are coupled with an indicator variable $v_{ij} \in \{-1, 0, 1\}$, denoting 1 for "attraction", -1 for "avoidance" and 0 for "neutral".

An initial approach for extracting v_{ij} would be to exploit the significance test described Section 6.6.2 of Chapter 6. We firstly generate R null graph instances and produce the empirical link weight distribution given a node pair i, j, as seen in Fig. 8.12. Given an observed co-occurrence value a_{ij} between i and j, we can classify the relationship type as of "attraction" ($v_{ij} = 1$) if co-occurrences a_{ij} are higher than we would expect given a significance threshold α_1, or as of "avoidance" ($v_{ij} = -1$) if i, j co-occur less that we would expect given

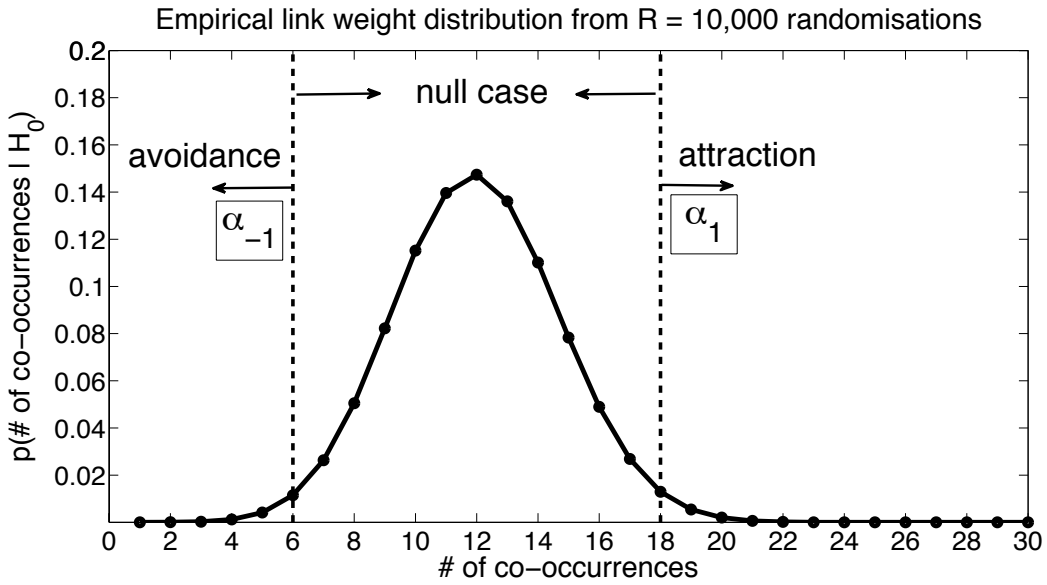


Figure 8.12: The empirical link weight distribution given a node pair i, j, produced by $R = 10,000$ null graph instances. The observed co-occurrences a_{ij} may correspond to relationship types such as "avoidance" (left side), "attraction" (right side), or be considered statistically insignificant (middle part) under our null hypothesis.

a threshold α_{-1}. For weights a_{ij} that can be explained by the null model, we set $v_{ij} = 0$. Repeating the above process for all dyads i, j we end up with a new adjacency $\mathbf{V}$, coupled with the co-occurrences matrix $\mathbf{A}$, which allows us to fine-tune our investigation both from a link analysis and community detection [Mucha et al., 2010] perspective, by taking into account this additional association type.

The communities identified here are based on temporal occurrence at feeding stations, and although the data analysed here are extensive, they are not entirely complete, since the observations were made only for a proportion of time, and only for the feeding-related activity. Therefore, the extracted networks describe social connectivity only for a subset of the great tit population at Wytham Woods, consisting exclusively of the RFID-tagged birds that visit the feeding stations. This directly relates to the concept of *node sampling* [Lee et al., 2006] and *boundary specification* [Laumann et al., 1989], according to which the observed network represents a subgraph of the ground-truth network, formed by a given subset of nodes for

which certain conditions are satisfied. In our Wytham Woods application, such nodes correspond to individuals that have been "ringed" with an RFID tag and have visited the feeder during the data collection season. Although the omission of nodes can create discrepancies between the observed and ground-truth networks, in terms of their degree and betweeness centrality distributions [Lee et al., 2006], mean node degree [Kossinets, 2006] and mean path length [Kossinets, 2006], quantities relating to degree assortativity and clustering coefficient remain robust [Kossinets, 2006; Lee et al., 2006]. In addition, the authors in [Costenbader and Valente, 2003] report that the ranking of nodes across various centrality measures can be retained in a network sample, despite the absolute differences in the centrality values themselves.

Our imperfect view of the great tit social structure is not only a result of the partial observation of the bird population. In chapters 5 and 6, we have already discussed how the inferred connections can be affected by missing logger observations or coincidental co-occurrences. Although the issue of pruning redundant links has been addressed in Section 6.6.2 and extensions to Nonnegative Matrix Factorisation [Cemgil, 2009; Zhong and Girolami, 2009] allow us to handle missing data in a principlined manner via CD-NMF, we may want to employ a more unified approach to the problem of reconstructing an adjacency matrix from sampled data. Such an approach is presented in [Guimerà and Sales-Pardo, 2009], where the authors formulate a probabilistic model that performs an exploration on the space of potential stochastic block models (SBMs), in order to find the one under which certain network estimates have higher likelihood scores. The exploration of such a large combinatorial space, performed via a Metropolis-Hastings sampling scheme [Metropolis et al., 1953], can be improved in our Wytham Woods application setting if we restrict our attention to SBMs that reflect the extracted gathering event co-membership. In particular, each "block" of a candidate SBM corresponds to a gathering event extracted via GEM and node co-membership scores can be configured by making use of the responsibility values discussed in Section

6.5.3. Although such a scheme would improve the performance of the method presented in [Guimerà and Sales-Pardo, 2009] (which yields excellent network reconstruction accurrancy across various datasets), the main drawback is that it can be applied only to the unweighted network case, thus restricting our analyses to the topology of bird connectivity.

Of particular relevance to our application is the work of [Franks et al., 2009], where the authors recommend the investigation of sampling effects via the use of artificially generated networks, which reproduce the ecological social system of interest in some meaningful way. This relates to our benchmark data stream generator (Algorithm 4 in Section 6.7.1), which is capable of reproducing "bursty" bird visitations in logger data, given a ground-truth graph. The use of such an algorithm in combination with the sampling strategies described by [Costenbader and Valente, 2003; Kossinets, 2006; Lee et al., 2006] can allow us to evaluate the robustness of network estimates not only within the context of coincidental co-occurrences (as in Section 6.7.2), but also in settings where individuals, links, or even locations are under-observed.

Following to the present discussion, in the next and final chapter of this thesis we comment on how the aforementioned methods and analyses are currently being extended not just for the purposes of analysing bird data sets of a larger scale and volume, but also in terms of incorporating information from a wider range of external data sources. We discuss ongoing and future work on advanced data collection systems, mixed-species bird networks and genetic information data sets, all of which pose exciting and promising research questions.

Chapter 9

Conclusions

9.1 Summary

The study of complex ecological systems, such as animal societies, requires a rigorous and systematic framework that can account for the multi-scaled structure and intricacy of social affiliation patterns. Network analysis has been widely adopted as an appropriate modelling framework for such settings, as it provides a wealth of methodologies for exploring the web of interconnections between individuals, in order to reveal invaluable insights on the fitness consequences of social organisation.

Advances in sensor technology have fundamentally changed the way zoologists collect data on ecological systems and the study of animal sociality is becoming more and more data-driven. RFID tag miniaturisation allows systematic and disturbance-free observation of free-ranging animals at a large scale, giving rise to data sets of unprecedented volume and resolution. This approach requires powerful statistical and computational models in order to harness the wealth of information contained in the "digital footprints" animals produce during the course of their life.

Based on the above, in this thesis we have developed a collection of models that extract the social structure of a wild bird *Parus major* population at Wytham Woods, from a large

data set of tracking records. In the spatio-temporal data the social ties between individuals are not directly apparent, as the observations consist of timestamped bird occurrences at various feeding sites across Wytham. By analysing the statistical properties of such data streams we introduced the concept of gathering events, which correspond to regions of temporally-focused bird feeding activity that can be seen as foraging flocks. In Chapter 6 we proposed and implemented a model termed Gathering Events Method (GEM) that identifies such events on the data stream and builds a bipartite network of individuals to flocks. Network links between individuals are then placed based on their co-occurrence in such events and evaluated based on their statistical significance. We have shown that GEM outperforms traditional "time-windowing" methods, which force a particular parametric structure in the data, on a variety of benchmark problems with observed network structure.

Following the extraction of bird-to-flock bipartite graphs from the spatio-temporal data, in Chapter 7 we have proposed a probabilistic methodology for one-mode projection, which places probability distributions over the presence and weight of each social tie. This Bayesian One-mode Projection (BOMP) model, can be seen as a probabilistic extension of the traditional Simple Ratio Index. It captures uncertainty over associations in a principled manner and makes the inferred graphs more robust to missing data by exploiting prior knowledge from past observations.

As bird populations usually consist of dynamic groups that split and merge throughout the course of the day (a phenomenon that is called fission-fusion [Conradt and Roper, 2000; Grosler, 1993]), we seek to move beyond the small foraging flocks of gathering events and identify larger social aggregations by exploiting the inferred network structure. Such structure consists of a sequence of undirected graphs (one per day), where each link is weighted by the number of co-occurrences (or expected co-occurrences under BOMP) of two individuals across gathering events. By considering such integer-valued counts as a emissions from a Poisson count process, in Chapter 4 we have proposed a nonnegative matrix factorisation

approach for community detection (CD-NMF) that we use to in order to identify overlapping network modules. The approach possesses excellent module identification properties, as demonstrated by a series of benchmark evaluations and it has been applied successfully to other community detection problems outside animal social networks.

The methodological advances we introduced in Chapters 6, 7 and 4, were applied to the Wytham Woods wild bird data set in Chapter 8. We analysed the inferred bird networks and showed that their topologies display a range of properties, such as high clustering coefficient and strong community organisation, typical in real-world networks [Newman, 2003b; Porter et al., 2009]. Particular focus has been given to the process of mating pair formation, a key biological process of interest, where we showed that co-occurrences among mating pairs start to form during the late winter and dominate the network as we approach the breeding season. Pairs being formed from previous years tend to forage together throughout the season, while newly-formed pairs are formed mid-winter via common participation across the same communities.

The work presented in this thesis by no means constitutes (or aims to be) a theoretical contribution to the field of animal behaviour. Instead, we have explored a series of interesting technical questions, which arise from data sets such as the ones produced from the Wytham Woods experiment, and proposed novel ways of looking at those problems. In that sense, this thesis fulfils the general data analysis aims we outlined at the beginning and provides an adequate baseline for exciting future extensions. In the next section, we discuss ongoing and future research directions, which build upon the presented material.

9.2 Future work

Each of the preceding chapters introducing novel material has a section discussing ideas for technical extensions to the proposed methodologies. Therefore in this section we discuss the potential of this work in a broader setting, from the perspective of its applicability to a wider

range of research questions. Such extensions are separated between the ones that concern the Wytham Woods study and others that refer to applications of our models to a range of different complex systems.

9.2.1 Wytham Woods

Data collection

As mentioned at the beginning of this thesis, the work presented here is part of a large and ongoing zoological experiment at Wytham Woods, Oxford. Since the deployment of the first antenna-enabled feeding stations in 2007, the scheme has been improving by the adoption of more sophisticated equipment and more comprehensive data collection schemes. Since 2011, all feeding stations were equipped with two antennae (Dorset ID, The Netherlands) which logged visits of RFID-tagged great tits to collect sunflower seeds. As before, birds are ringed with a plastic colour rings with RFID tags encased within the mold (IB Technology, UK), containing a 125 Hz RFID tag. For a single winter (2011–12), 10,649,407 records of 4,223 individuals have been collected. The main difference, from our data perspective, is the time resolution of their internal clock. Observations retrieved from the conventional loggers can be 15 seconds apart at minimum, while new loggers possess more advanced hardware that allows them to capture the timestamp of bird visitations at 1-second resolution. This obviously affects the richness of the observations retrieved, allowing us to perform analyses by looking at the data at finer levels of granularity. Interestingly, from preliminary calculations we have found that this new data set exhibits a very similar bursty structure compared to 2007–8 and 2008–9, based on the fat tail of the inter-event distribution we show in Fig. 9.1.

Genetic information

One of the most promising research directions of the Wytham Woods study is the investigation of the genetic foundations of sociality. The relationship between genetics and networks

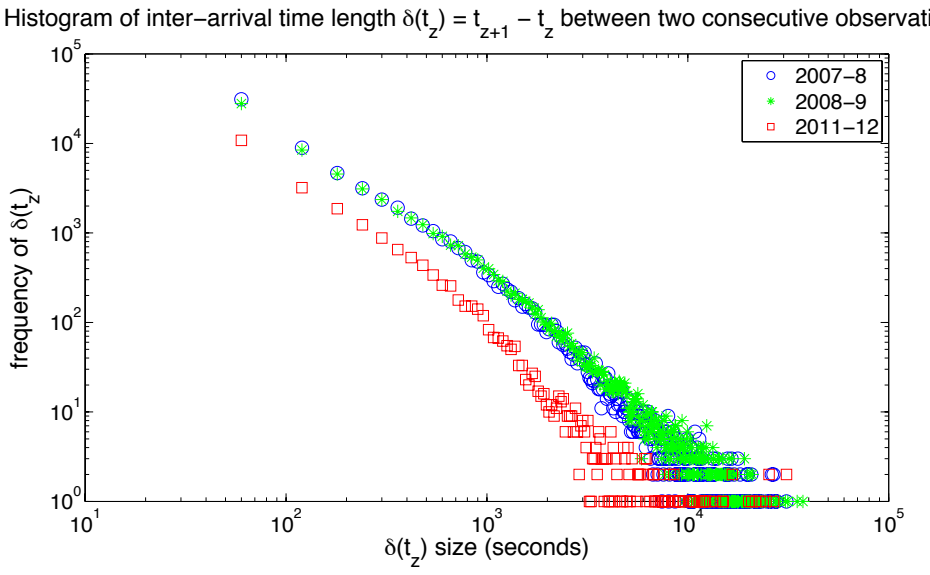


Figure 9.1: We plot the histogram of $\delta(t_z)$ values for three data streams: seasons 2007–8, 2008–9 and 2011–12. Similar to the results of the analysis we performed in Section 6.4.1, using the approach of [Clauset et al., 2009], all distributions exhibit a fat tail and can be approximated by a power law with $\gamma \simeq 2.5$ for $\delta(t_z) > 10^3$.

has been investigated both in human [Fowler et al., 2011] and primate [Brent et al., 2013] networks, providing deep insights on how fitness and behaviour depend not only on an individual's genes, but also on genes of his/her social circle.

Understanding genetic foundations of sociality can shed new light on the underlying processes giving rise to complex network structures. Towards this goal, we plan to combine the inferred network structures with genetic information, obtained via taking blood samples from the wild birds. We currently have available a large data set where 664 out of 770 individuals from the 2007–8 season and 584 out of 753 individuals from the 2008–9 season have been genotyped across 4878 different single nucleotide polymorphisms (SNPs). In our preliminary analysis we have extracted all bird communities throughout the season using CD-NMF and calculated group cohesion via the NMF membership scores. We then estimated the SNP frequencies and compared them to the overall population SNP frequency. In Fig. 9.2, each data point corresponds to a community, placed on the figure based on its community cohesion and

divergence from the population SNP frequency. We can see preliminary evidence of genetic clustering, where the more strongly connected the community is, the more differentiated its members are from the baseline genetic frequency.

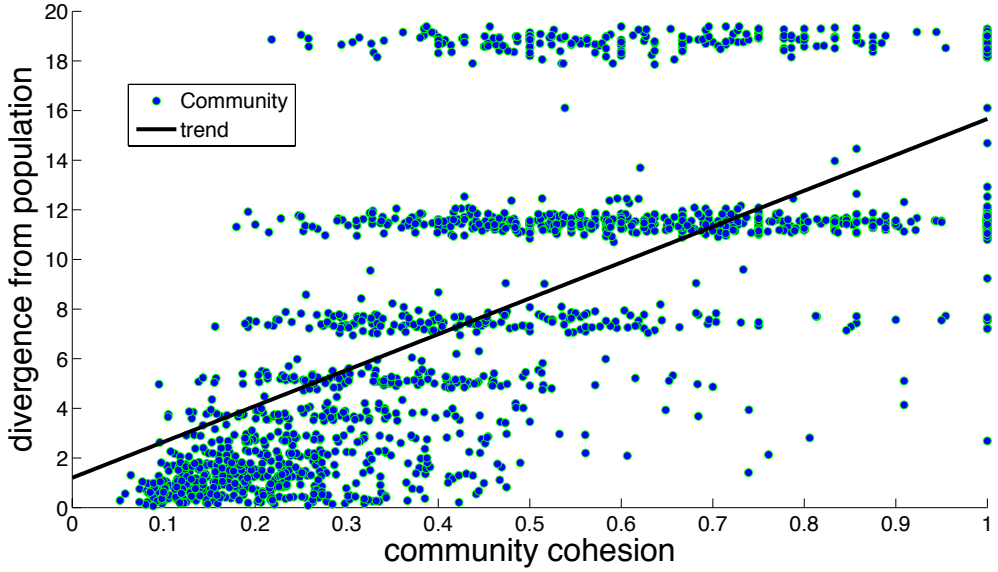


Figure 9.2: We plot each bird community intra-cohesion score (average node membership score given by CD-NMF) versus the KL-divergence of its SNP frequency to the overall SNP frequency of the whole population.

Mixed-species networks

In the present work we have focused on modelling the sociality of *Parus major* great tits as an isolated population at Wytham Woods. In reality, no group of conspecifics lives in a "vacuum", as ecological systems are formed by the rich and complex interactions of a wide range of animal species. For that reason, the methodologies developed in this thesis are already being used for the study of mixed-species bird networks [Farine et al., 2012] in order to explore interspecific social behaviour relating to dominance hierarchies, environmental effects and foraging success. Future research, armed with data sets of greater detail and wider time horizon, will also examine the effects relating to information diffusion (in particular social learning) and disease spreading.

Validating the extracted networks

Throughout the present thesis, we have underlined the fact that there is no way to compare the inferred wild bird networks with some kind of "ground truth" solution. Instead, the validity of the extracted graphs is to be evaluated in terms of their ability to produce ecological insights of interest, an example of which we have presented in Chapter 8. As mentioned in previous sections, ongoing advancements in data collection technologies and the incorporation of external data sources will allow the investigation of more sophisticated ecological questions. This will enable us, through such investigations, to assess how well the inferred networks reflect the actual social structure of the wild bird population. In particular, current work at Wytham Woods focuses on investigating if social network structure can predict the sequence of individuals that discover a particular food source (social learning) and how certain diseases (such as avian pox) spread among the population. For the purposes of this thesis and given the limitations of the 2007–9 data sets, we have driven our methodological developments towards a Bayesian formalism in order to handle noise and missing observations more rigorously. We aspire that the methodologies introduced in this thesis, applied to data sets of larger volume and resolution, will provide a solid baseline for rigorous ecological investigations, which will in turn assess the explanatory power of the inferred social structures.

9.2.2 Other applications

The development of all theoretical and algorithmic tools in this thesis has been motivated by the Wytham Woods wild bird study. Nevertheless, the proposed models are general enough to be applied to a variety of other areas, which we describe in this section.

Stock market data

We are considering the problem of inferring a similarity network among investors, based on their stock-picking decisions on the popular online community "Motley Fool". In this

website users play a stock selection game, where they attempt to predict which stocks will outperform or under-perform the S&P 500 Index. Users can select any stock and make a "call" as to whether or not it will out or under-perform. The data consists of a long stream of timestamped "calls", where individuals up-vote or down-vote particular stocks. Such data stream is complemented by a "Twitter-like" graph, where individuals "follow" each other.

The stock-call information can be seen as a temporal data stream, similar to the ones we have examined in this thesis, with N individuals and 1 location. Current analysis has shown a strong modular organisation of the stock-call data stream, where investor activities occur in "bursts" that signify some kind of market event. The key difference with the examples we have considered in this thesis, from a purely data-driven point of view, is that individuals now may have a negative participation to a "gathering event", as some investors tend to down-vote stocks. Therefore, we seek to build appropriate one-mode projection methodologies that quantify the degree of any disagreement between investors in order to define an appropriate pairwise similarity value. After defining a similarity network between investors, we seek to compare it versus the "Twitter-like" graph and compare the CD-NMF groups extracted in order to investigate the relationship between information flow and stock returns.

Moreover, we plan to develop the method further to capture non-network related variables which vary over time, but impact the likelihood of association (in the case of financial markets, one can imagine general market sentiment, or stock correlation). The hope with this work would be to understand times at which joint appearance at "gathering events" represents stronger associations - if two investors make the same decision at times of limited liquidity, we should be able to capture this information systematically.

Database security

We consider a large online DataBase Management System (DBMS), where various SQL queries are being applied to it over time. Some of the queries consist of a series of SQL com-

mands that perform a legitimate user case, while others may perform a harmful operation (SQL injection). We seek to exploit a large proprietary data set of query logs[1], by viewing the "bursty" structure of query bouts interrogating the database as gathering events of complementary commands. Our goal is then to infer the underlying association network of SQL queries and perform CD-NMF community detection, cross-examining the extracted groups with known malicious command sequences, in order to predict future intrusions.

9.3 Closing remarks

The methodologies developed in this thesis provide a methodological baseline for future analysis of a variety of complex systems, including the Wytham Woods data set. We place special emphasis on the latter, as insights from studying a model system such as the great tit population not only contribute towards better understanding of issues relating to animal behaviour, evolutionary biology and conservation but also have impact on a greater scale; animal social networks can bring powerful insights to complex systems in general, as they allow us to observe the sociality of individuals during the course of their whole lifetime and are not limited by privacy-related issues that are prominent in human systems. Therefore, fundamental research questions relating to the genetic foundations of social organisation, the processes that drive information diffusion and social learning, disease spreading and containment, tie formation and dissolution, can be addressed using powerful machine learning tools applied to data sets of great scale and resolution.

[1]The SQL query data have been provided to us by Steve Moyle, The Global Identity Foundation.

9.4 Network analysis of network analysis software

All models and computational tools presented in this work have been implemented in the form of MATLAB scripts[2], which have been used for analysing the wild bird data and performing various experimentation schemes. In total, this research has produced 575 unique MATLAB scripts that perform a wide range of tasks, from Bayesian inference to network analysis and from text parsing to database interrogation.

Such a diverse computational ecosystem can be seen as a directed graph, where nodes represent unique MATLAB scripts and edges represent function calls; we consider a link starting from script `A` and pointing to script `B`, if `A[B[x]]` for some input `x`. In order to build such a network, a "crawler" has been implemented that explores the directory tree where all codes reside and extracts the names of all `.m` files (MATLAB script extension). All N file names are added to a hashtable, where the key-value pair contains the script name, along with a unique integer that denotes the node index in the resulting adjacency matrix. The algorithm proceeds by using `grep` to parse every file we have retrieved, looking for function calls to other files. For each script i that calls another script j within its code, a link is placed starting from i and ending in j.

In Fig. 9.3 we show such a graph, where the size of each node expresses its in-degree and its brightness corresponds to its Pagerank centrality. The graph possesses 15 weakly connected and 567 strongly connected components, with a diameter of 7 and average path length of 1.9 hops. There is a strong presence of community structure, with 27 groups that yield $Q \simeq 0.768$. Various centrality measures have been evaluated in order to evaluate the popularity of each module in the network, ranging from in-degree to betweeness and Pagerank centralities. We have found that the most central links in the graph correspond to "utility" scripts such as a custom-made hashtable structure, the calculation of Shannon entropy, calculation of modularity, along with some ad hoc plotting scripts. It is worth noting that the

[2]GEM and CD-NMF are made available online, on `http://www.robots.ox.ac.uk/~parg/doku.php?id=software_page`

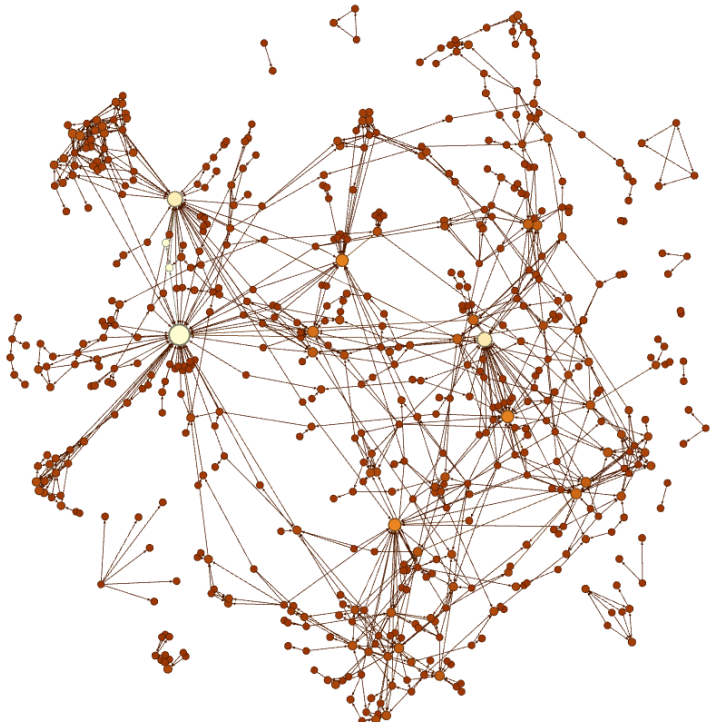


Figure 9.3: The network of the network analysis software used in this thesis. The $N = 575$ nodes represent MATLAB scripts and the $M = 1085$ links represent calls from one scripts to another. The size of each node expresses its in-degree and its brightness corresponds to its Pagerank centrality.

bottom-ranked script, under all importance metrics, is the one that performed such analysis, thus raising serious scepticism regarding the utility of the whole experiment.

It is worth noting that the popularity of a node, calculated based on various network metrics, does not necessarily correlate with its importance in a network of this kind. For example,"master scripts" that only perform calls of other scripts and coordinate the execution of a large-scale experiment, are not ranked important by such network topology although they are crucial to the goal of such code base.

Appendix A

Variational Inference for Gathering Event Extraction

A.1 Background on Variational Bayes

Given a probabilistic model with $\mathbf{X}$ observed and $\mathbf{Z} = \{Z_1, Z_2, ..., Z_K\}$ latent variables, we seek to approximate the posterior $p(\mathbf{Z}|\mathbf{X})$ in settings where the distribution is intractable. Variational Bayes (VB) proposes a new distribution $q(\mathbf{Z})$ over the latent variables $\mathbf{Z}$, along with the following decomposition of the marginal likelihood $p(\mathbf{X})$:

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p), \tag{A.1}$$

where:

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}, \tag{A.2}$$

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}, \tag{A.3}$$

$\mathcal{L}(q)$ is a *free energy term* and $\mathrm{KL}(q||p)$ is the KL-divergence between the proposed distribution $q(\mathbf{Z})$ and the posterior $p(\mathbf{Z}|\mathbf{X})$ we seek to approximate. We prove the decomposition from Eq. (A.1) holds, by expanding Eq. (A.2) and Eq. (A.3) respectively:

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\ &= \int q(\mathbf{Z}) \ln p(\mathbf{X},\mathbf{Z}) \mathrm{d}\mathbf{Z} - \int q(\mathbf{Z}) \ln(\mathbf{Z}) \mathrm{d}\mathbf{Z} \\ &= \int \ln p(\mathbf{Z}|\mathbf{X}) \mathrm{d}\mathbf{Z} + \int q(\mathbf{Z}) \ln p(\mathbf{X}) \mathrm{d}\mathbf{Z} - \int q(\mathbf{Z}) \ln q(\mathbf{Z}) \mathrm{d}\mathbf{Z}, \end{aligned} \tag{A.4}$$

while for the KL-divergence we have:

$$\begin{aligned}\mathrm{KL}(q||p) &= -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\ &= -\int q(\mathbf{Z}) \ln p(\mathbf{Z}|\mathbf{X}) \mathrm{d}\mathbf{Z} + \int q(\mathbf{Z}) \ln q(\mathbf{Z}) \mathrm{d}\mathbf{Z}. \end{aligned} \tag{A.5}$$

By summing the expressions from Eq. (A.4) and Eq. (A.5) we have:

$$\begin{aligned}\mathcal{L}(q) + \mathrm{KL}(q||p) &= \int q(\mathbf{Z}) \ln p(\mathbf{X}) \mathrm{d}\mathbf{Z} \\ &= \ln p(\mathbf{X}) \int q(\mathbf{Z}) \mathrm{d}\mathbf{Z} \\ &= \ln p(\mathbf{X}), \end{aligned} \tag{A.6}$$

thus recovering the expression of Eq. (A.1).

As $\ln p(\mathbf{X})$ is *fixed* given a particular choice of model, our goal is to find the appropriate function $q(\mathbf{Z})$ that increases the lower bound $\mathcal{L}(q)$, thus *minimising the KL-divergence* between the approximate $q(\mathbf{Z})$ and true posterior $p(\mathbf{Z}|\mathbf{X})$. Towards this goal, we follow [Bishop,

2007; Fox and Roberts, 2012] by proposing a factorised distribution of M components:

$$q(Z_i) = \prod_{i=1}^{M} q_i(Z_i), \tag{A.7}$$

where each Z_i of $\mathbf{Z}$ corresponds to a collection of latent variables that we consider to be *statistically independent* to the others in our model. Among all possible $q(\mathbf{Z})$, we seek the distribution that maximises the lower bound (free energy) $\mathcal{L}(q)$ with respect to all $q_i(Z_i)$. By expressing the dependence of $\mathcal{L}(q)$ on the j-th factor $q_j(Z_j)$, or q_j for short, we have:

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} = \int \left(q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z}) - q(\mathbf{Z}) \ln q(\mathbf{Z}) \right) \mathrm{d}\mathbf{Z} \\
&= \int \left(\prod_{i=1}^{M} q_i \ln p(\mathbf{X}, \mathbf{Z}) - \prod_{i=1}^{M} q_i \int \ln q_i \mathrm{d}Z_i \right) \mathrm{d}\mathbf{Z} \\
&= \int q_j \prod_{i \neq j}^{M} q_i \ln p(\mathbf{X}, \mathbf{Z}) \mathrm{d}\mathbf{Z} - \int q_j \prod_{i \neq j}^{M} q_i (\ln q_j + \int \ln q_i \mathrm{d}Z_i) \mathrm{d}\mathbf{Z} \\
&= \int q_j \left[\int \prod_{i \neq j}^{M} q_i \ln p(\mathbf{X}, \mathbf{Z}) \mathrm{d}Z_i \right] \mathrm{d}Z_j - \int q_j \ln q_j \mathrm{d}Z_j + \text{const} \\
&= \int q_j \ln \tilde{p}(\mathbf{X}, Z_j) \mathrm{d}Z_j - \int q_j \ln q_j \mathrm{d}Z_j + \text{const},
\end{aligned} \tag{A.8}$$

where

$$\begin{aligned}
\ln \tilde{p}(\mathbf{X}, Z_j) &= \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, Z_i)] + \text{const} \\
&= \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \mathrm{d}Z_i,
\end{aligned} \tag{A.9}$$

the expectation over all $\mathbf{Z}$ except Z_j. We can view the dependency of $\mathcal{L}(q)$ on the j-th factor, expressed in Eq. (A.8), as the *negative KL-divergence*:

$$\begin{aligned}\mathcal{L}(q) &= \int q_j \ln \tilde{p}(\mathbf{X}, Z_j)\mathrm{d}Z_j - \int q_j \ln q_j \mathrm{d}Z_j + \text{const}, \\ &= \int q_j \ln \frac{\tilde{p}(\mathbf{X}, Z_j)}{q_j}\mathrm{d}Z_j + \text{const}, \\ &= \text{KL}(q_j||\tilde{p}(\mathbf{X}, Z_j)), \end{aligned} \tag{A.10}$$

between q_j and $\tilde{p}(\mathbf{X}, Z_j)$, so that maximising $\mathcal{L}(q)$ is the equivalent of minimising $\text{KL}(q_j||\tilde{p}(\mathbf{X}, Z_j))$. Therefore, among all possible distributions $q(.)$ for approximating the true posterior $P(\mathbf{Z}|\mathbf{X})$ we choose the one that minimises Eq. (A.10) so that $q_j^\star = \tilde{p}(\mathbf{X}, Z_j)$, or:

$$\ln q_j^\star = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, Z_i)] + \text{const}. \tag{A.11}$$

In settings where the distributions over our stochastic variables belong to the exponential family, the solution to our inference problem of approximating the posterior $p(\mathbf{Z}|\mathbf{X})$ consists of cycling through the factors $j \in \{1, ..., M\}$ and update their sufficient statistics. A direct application of such update equations to the inference scheme of the GEM model proposed in Chapter 6 is shown in the next section.

A.2 Posterior derivations

In this section we present the derivations of the posterior distributions, over the latent variables and parameters, of the clustering model presented in Section 6.5.3 of Chapter 6. We begin by providing the natural logarithm of the three distributions of interest used throughout this section, the Gaussian, the Gamma and the Dirichlet, to which we shall refer throughout our analysis:

$$\ln \mathcal{N}(x; \mu, \beta^{-1}) = \ln \beta - \frac{1}{2} \ln 2\pi - \frac{\beta}{2} x^2 + \beta \mu x - \frac{\beta}{2} \mu^2, \tag{A.12}$$

$$\ln \text{Ga}(x; \mathtt{s}, \mathtt{c}) = \mathtt{s} \ln \mathtt{c} - \ln \Gamma(\mathtt{s}) + (\mathtt{s} - 1) \ln x - \beta x, \tag{A.13}$$

$$\ln \text{Dir}(\mathbf{x}; \boldsymbol{\lambda}) = \ln C(\boldsymbol{\lambda}) + \sum_{k=1}^{K} (\lambda_k - 1) \ln x_k, \tag{A.14}$$

where $\Gamma()$ the gamma function and $C()$ the normalisation term of the Dirichlet.

Recall the following factorisation that we have proposed, concerning the approximate joint distribution $q(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}) \simeq p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta})$ over the latent variables:

$$q(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}) = q(\mathbf{Y}) q(\boldsymbol{\pi}) q(\boldsymbol{\mu}) q(\boldsymbol{\beta}). \tag{A.15}$$

Minimisation of the divergence $\text{KL}(q(\mathcal{X}) || p(\mathcal{X}|\mathbf{t}))$ can now be expressed as a sequential maximisation of the free energy $\mathcal{L}(q)$ with respect to each one of the factors from Eq. (A.15) in turn. Approximating the true posterior over $\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}$ is the equivalent of finding the *optimal factors* in Eq. (A.15), so that their product maximises $\mathcal{L}(q)$. Following the VB scheme presented in [Bishop, 2007; Fox and Roberts, 2012] and based on the previous section, this is achieved by expressing each factor in Eq. (6.16) in relation to the other variables in the model:

$$\ln q^\star(\mathbf{Y}) = \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}}[\ln p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{t})] + \text{const}, \tag{A.16}$$

$$\ln q^\star(\boldsymbol{\pi}) = \mathbb{E}_{\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta}}[\ln p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{t})] + \text{const}, \tag{A.17}$$

$$\ln q^\star(\boldsymbol{\mu}) = \mathbb{E}_{\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\beta}}[\ln p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{t})] + \text{const}, \tag{A.18}$$

$$\ln q^\star(\boldsymbol{\beta}) = \mathbb{E}_{\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}}[\ln p(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{t})] + \text{const}, \tag{A.19}$$

where each one of the above expressions is analysed in the following sections.

A.2.1 Posterior of $\mathbf{Y}$

We seek to expand Eq. (A.16) in order to get an functional form for the approximate posterior over $\mathbf{Y}$. Based on the factorisation from Eq. (A.15) we have:

$$\begin{aligned}\ln q^{\star}(\mathbf{Y}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}}[\ln p(\mathbf{t},\mathbf{Y},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta})] + \text{const} \\ &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}}[\ln p(\mathbf{t}|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\beta}) + \ln p(\mathbf{Y}|\boldsymbol{\pi}) + \ln p(\boldsymbol{\pi}) + \ln p(\boldsymbol{\mu}) \ln p(\boldsymbol{\beta})] + \text{const},\end{aligned}$$

as the expected value runs over $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}$, the term $\ln p(\boldsymbol{\pi}) + \ln p(\boldsymbol{\mu}) + \ln p(\boldsymbol{\beta})$ is absorbed by the constant. By expanding the rest of the terms we have:

$$\begin{aligned}\ln q^{\star}(\mathbf{Y}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}}\left[\sum_z \sum_k \ln\left\{\mathcal{N}(t_z;\mu_k,\beta_k^{-1})^{y_{zk}}\right\}\right] + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}}\left[\sum_z \sum_k y_{zk} \ln \pi_{zk}\right] + \text{const} \\ &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}}\left[\sum_z \sum_k y_{zk}(\ln \beta_k - \frac{1}{2}\ln 2\pi - \frac{\beta_k}{2}t_z^2 + \beta_k t_z \mu_k - \frac{\beta_k}{2}\mu_k^2)\right] + \\ &\quad + \sum_z \sum_k y_{zk}\mathbb{E}_{\pi}[\ln \pi_k] + \text{const} \\ &= \sum_z \sum_k y_{zk}\left[\mathbb{E}_{\beta}[\ln \beta_k] - \frac{t_z^2}{2}\mathbb{E}_{\beta}[\beta_k] + t_z \mathbb{E}_{\mu\beta}[\mu_k\beta_k] - \frac{1}{2}\mathbb{E}_{\beta\mu}[\beta_k\mu_k^2] + \mathbb{E}_{\pi}[\ln \pi_k]\right] + \\ &\quad + \text{const}.\end{aligned}$$

By following [Bishop, 2007], we consider the term inside the brackets of $\sum_z \sum_k y_{zk}[.]$ as:

$$\ln \rho_{zk} = \mathbb{E}_{\beta}[\ln \beta_k] - \frac{t_z^2}{2}\mathbb{E}_{\beta}[\beta_k] + t_z \mathbb{E}_{\mu\beta}[\mu_k\beta_k] - \frac{1}{2}\mathbb{E}_{\beta\mu}[\beta_k\mu_k^2] + \mathbb{E}_{\pi}[\ln \pi_l], \tag{A.20}$$

thus having the following expression for our approximate posterior:

$$\begin{aligned}
\ln q^{\star}(\mathbf{Y}) &= \sum_{z}\sum_{k} y_{zk} \ln \rho_{zk} + \text{const} \\
\Rightarrow q^{\star}(\mathbf{Y}) &= \prod_{z}\prod_{k} \rho_{zk} \times \text{const} \\
\Rightarrow q^{\star}(\mathbf{Y}) &= \prod_{z}\prod_{k} r_{zk}^{y_{zk}}, \qquad \text{(A.21)}
\end{aligned}$$

by setting $r_{zk} = \frac{\rho_{zk}}{\sum_z \rho_{zk}}$. Note that the posterior in Eq. (A.21) has the *same functional form* as the prior $p(\mathbf{Y}|\boldsymbol{\pi}) = \prod_z \prod_k \pi_k^{y_{zk}}$, with $\mathbb{E}[y_{zk}] = r_{zk}$ as from Eq. (A.21). The r_{zk} terms, or *responsibilities*, define the membership score of each observation z to gathering event k and constitute the solution to our clustering problem.

Based on Eq. (A.20), in order to calculate the responsibilities, we need $\mathbb{E}[\ln \beta_k], \mathbb{E}[\beta_k], \mathbb{E}[\mu_k], \mathbb{E}[\mu_k^2], \mathbb{E}[\ln \pi_k]$. Such terms are calculated from the posteriors of our model parameters, which we derive in the next sections.

A.2.2 Posterior of $\boldsymbol{\pi}$

In the same spirit, we expand Eq. (A.17):

$$\begin{aligned}
\ln q^{\star}(\boldsymbol{\pi}) &= \mathbb{E}_{\mathbf{y},\mu,\beta}[\ln p(\mathbf{t}|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\beta}) + \ln p(\mathbf{Y}|\boldsymbol{\pi}) + \ln p(\boldsymbol{\pi}) + \ln p(\boldsymbol{\mu}) \ln p(\boldsymbol{\beta})] + \text{const} \\
&= \mathbb{E}_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\beta}}[\ln p(\mathbf{Y}|\pi) + \ln(\boldsymbol{\pi})] + \text{const} \\
&= \mathbb{E}_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\beta}}\left[\sum_{z}\sum_{k} y_{zk} \ln \pi_k\right] + \ln \text{Dir}(\boldsymbol{\pi};\lambda_0) + \text{const} \\
&= \mathbb{E}_{\mathbf{y},\boldsymbol{\mu},\boldsymbol{\beta}}\left[\sum_{z}\sum_{k} y_{zk} \ln \pi_k\right] + \ln C(\lambda_0) + (\lambda_0 - 1)\sum_{k} \ln \pi_k + \text{const} \\
&= \sum_{z}\sum_{k} \mathbb{E}_{\mathbf{y},\mu,\beta}[y_{zk}] \ln \pi_k + (\lambda_0 - 1)\sum_{k} \ln \pi_k + \text{const.} \qquad \text{(A.22)}
\end{aligned}$$

By exponentiating both sides of Eq. (A.22), having $\mathbb{E}_{\mathbf{y},\mu,\beta}[y_{zk}] = r_{zk}$ from the previous

section and considering $Z_k = \sum_z r_{zk}$, we have:

$$\begin{aligned} q^\star(\boldsymbol{\pi}) &= \exp\left\{\sum_z \sum_k r_{zk} \ln \pi_k\right\} \times \exp\left\{\sum_k \ln \pi_k\right\}^{\lambda_0 - 1} \times \text{const} \\ &= \left[\prod_k \pi_k\right]^{Z_k} \times \prod_k \pi_k^{\lambda_0 - 1} \times \text{const} \\ &= \text{const} \times \prod_k \pi_k^{\lambda_0 + Z_k - 1}, \end{aligned} \tag{A.23}$$

which has the same Dirichlet form as the prior $p(\boldsymbol{\pi}) = \frac{1}{B(\boldsymbol{\lambda})} \prod_{k=1}^K \pi_k^{\lambda_k - 1}$, with an updated parameter $\lambda_k = \lambda_0 + Z_k - 1$.

A.2.3 Posterior of $\boldsymbol{\mu}$

We use a similar process to expand Eq. (A.18):

$$\begin{aligned} \ln q^\star(\boldsymbol{\mu}) &= \mathbb{E}_{\mathbf{y},\boldsymbol{\pi},\boldsymbol{\beta}}[\ln p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta}) + \ln p(\mathbf{Y}|\boldsymbol{\pi}) + \ln p(\boldsymbol{\pi}) + \ln p(\boldsymbol{\mu}) \ln p(\boldsymbol{\beta})] + \text{const} \\ &= \mathbb{E}_{\mathbf{y},\boldsymbol{\pi},\boldsymbol{\beta}}[\ln p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta}) + \ln p(\boldsymbol{\mu})] + \text{const} \\ &= \mathbb{E}_{\mathbf{y},\boldsymbol{\pi},\boldsymbol{\beta}}\left[\sum_z \sum_k y_{zk} \ln \mathcal{N}(t_z; \mu_k, \beta_k^{-1}) + \sum_k \ln \mathcal{N}(\mu_k; \mathtt{m}_0, \mathtt{b}_0^{-1})\right] + \text{const} \\ &= \mathbb{E}_{\mathbf{y},\boldsymbol{\pi},\boldsymbol{\beta}}\left[\sum_z \sum_k y_{zk} \left(\ln \beta_k - \frac{1}{2} \ln 2\pi - \frac{t_z^2}{2}\beta_k + t_z \beta_k \mu_k - \frac{1}{2}\beta_k \mu_k^2\right)\right] \\ &\quad + \mathbb{E}_{\mathbf{y},\boldsymbol{\pi},\boldsymbol{\beta}}\left[\sum_k \left(\ln \mathtt{b}_0 - \frac{1}{2} \ln 2\pi - \frac{\mathtt{b}_0}{2}\mu_k^2 + \mu_k \mathtt{m}_0 \mathtt{b}_0 - \frac{1}{2}\mathtt{m}_0^2 \mathtt{b}_0\right)\right] + \text{const}. \end{aligned}$$

By setting $\mathbb{E}[\beta_k] = \tilde{\beta}_k, \mathbb{E}[\ln \beta_k] = \ln \tilde{\beta}_k, \bar{t} = \frac{1}{Z}\sum_z t_z, \bar{t^2} = \frac{1}{2}\sum_z t_z^2$ to unclutter the

notation, we have:

$$\begin{aligned}
\ln q^{\star}(\boldsymbol{\mu}) &= \sum_{z}\sum_{k}\left[r_{zk}\ln\tilde{\beta}_k - \frac{1}{2}r_{zk}\ln 2\pi - \frac{1}{2}r_{zk}t_z^2\tilde{\beta}_k + r_{zk}t_z\tilde{\beta}_k\mu_k - \frac{1}{2}r_{zk}\tilde{\beta}_k\mu_k^2\right] + \\
&\quad + \sum_{k}\left[\ln \mathtt{b}_0 - \frac{1}{2}\ln 2\pi - \frac{\mathtt{b}_0}{2}\mu_k^2 + \mu_k\mathtt{m}_0\mathtt{b}_0 - \frac{1}{2}\mathtt{m}_0^2\mathtt{b}_0\right] + \text{const} \\
&= \sum_{k}[Z_k\ln\tilde{\beta}_k - \frac{1}{2}Z_k\ln 2\pi - \frac{\tilde{\beta}_k}{2}Z_k\bar{t^2}_k + \tilde{\beta}_k\mu_k Z_k\bar{t}_k - \frac{Z_k}{2}\tilde{\beta}_k\mu_k^2 + \\
&\quad + \ln \mathtt{b}_0 - \frac{1}{2}\ln 2\pi - \frac{\mathtt{b}_0}{2}\mu_k^2 + \mathtt{b}_0\mathtt{m}_0\mu_k - \frac{\mathtt{b}_0}{2}\mu_0^2] + \text{const} \\
&= \sum_{k}[-(\frac{1}{2}Z_k\tilde{\beta}_k + \frac{\mathtt{b}_0}{2})\mu_k^2 + (\tilde{\beta}_k Z_k\bar{t} + \mathtt{b}_0\mathtt{m}_0)\mu_k + \ln\tilde{\beta}_k Z_k \\
&\quad - \frac{1}{2}\ln 2\pi Z_k - \frac{\tilde{\beta}_k}{2}Z_k\bar{t^2}_k + \ln \mathtt{b}_0 - \frac{1}{2}\ln 2\pi - \frac{\mathtt{b}_0}{2}\mathtt{m}_0^2] + \text{const}.. \qquad \text{(A.24)}
\end{aligned}$$

By re-organising the terms of the right-hand side of Eq. (A.24) based on Eq. (A.12) we have an expression for the approximate posterior of $\boldsymbol{\mu}$:

$$\begin{aligned}
\ln q^{\star}(\boldsymbol{\mu}) &= \ln\left\{\prod_k \mathcal{N}\left(\mu_k; \beta_k^{-1}(Z_k\bar{t}_k\tilde{\beta}_k + \mathtt{b}_0\mathtt{m}_0), Z_k\tilde{\beta}_k + \mathtt{b}_0\right)\right\} \\
\Rightarrow q^{\star}(\boldsymbol{\mu}) &= \prod_k \mathcal{N}\left(\mu_k; \beta_k^{-1}(Z_k\bar{t}_k\tilde{\beta}_k + \mathtt{b}_0\mathtt{m}_0), Z_k\tilde{\beta}_k + \mathtt{b}_0\right), \qquad \text{(A.25)}
\end{aligned}$$

which has the same Gaussian form as the prior $p(\boldsymbol{\mu}) = \prod_k \mathcal{N}(\mu_k; \mathtt{m}_0, \mathtt{b}_0)$, with updated parameters:

$$\mathtt{m}_k = \beta_k^{-1}(Z_k\bar{t}_k\tilde{\beta}_k + \mathtt{b}_0\mathtt{m}_0), \qquad \text{(A.26)}$$

$$\mathtt{b}_k = Z_k\tilde{\beta}_k + \mathtt{b}_0. \qquad \text{(A.27)}$$

A.2.4 Posterior of $\boldsymbol{\beta}$

We finally expand Eq. (A.19):

$$\begin{aligned}
\ln q^{\star}(\boldsymbol{\beta}) &= \mathbb{E}_{\mathbf{y},\boldsymbol{\pi},\boldsymbol{\mu}}[\ln p(\mathbf{t}|\mathbf{Y},\boldsymbol{\mu},\boldsymbol{\beta}) + \ln p(\mathbf{Y}|\boldsymbol{\pi}) + \ln p(\boldsymbol{\pi}) + \ln p(\boldsymbol{\mu}) \ln p(\boldsymbol{\beta})] + \text{const} \\
&= \mathbb{E}_{\mathbf{y},\boldsymbol{\pi},\boldsymbol{\mu}}\left[\sum_z \sum_k y_{zk} \ln\{\mathcal{N}(\mathbf{t};\mu_k,\beta_k)\} + \sum_k \ln\{\text{Ga}(\beta_k;\mathtt{s}_0,\mathtt{c}_0)\}\right] + \text{const} \\
&= \mathbb{E}_{\mathbf{y},\boldsymbol{\pi},\boldsymbol{\mu}}[\sum_z \sum_k y_{zk}\left(\ln\beta_k - \frac{1}{2}\ln 2\pi - \frac{1}{2}\beta_k t_z^2 + \beta_k t_z \mu_k - \frac{1}{2}\beta_k \mu_k^2\right) \\
&\quad + \sum_k (\mathtt{s}_0 \ln \mathtt{c}_0 - \ln\Gamma(\mathtt{s}_0) + (\mathtt{s}_0 - 1)\ln\beta_k - \mathtt{c}_0\beta_k)] + \text{const},
\end{aligned}$$

where by setting $\mathbb{E}[\mu_k] = \tilde{\mu}_k$ and re-organising the right-hand side of Eq. (A.28) based on Eq. (A.13), we express the approximate log-posterior over $\boldsymbol{\beta}$ as:

$$\begin{aligned}
\ln q^{\star}(\boldsymbol{\beta}) &= \sum_k (Z_k \ln\beta_k - \frac{1}{2} Z_k \ln 2\pi - \frac{1}{2}\beta_k Z_k \bar{t^2}_k + \beta_k \tilde{\mu}_k Z_k \bar{t}_k - \frac{1}{2}\beta_k \tilde{\mu}_k^2 \\
&\quad + \mathtt{s}_0 \ln \mathtt{c}_0 - \ln\Gamma(\mathtt{s}_0) + (\mathtt{s}_0 - 1)\ln\beta_k - \mathtt{c}_0\beta_k) + \text{const}, \\
&= \sum_k [-\beta_k(-\tilde{\mu}_k Z_k \bar{t}_k + \mathtt{c}_0) + \ln\beta_k(Z_k + \mathtt{s}_0 - 1) + \text{const}] \\
&= \ln\left\{\prod_k \text{Ga}(\beta_k; \mathtt{s}_0 + Z_k, \mathtt{c}_0 - Z_k \bar{t}_k \tilde{\mu}_k)\right\},
\end{aligned}$$

where by exponentiating both sides we have:

$$q^{\star}(\boldsymbol{\beta}) = \prod_k \text{Ga}(\beta_k; \mathtt{s}_0 + Z_k, \mathtt{c}_0 - Z_k \bar{t}_k \tilde{\mu}_k), \tag{A.28}$$

the same functional form as the prior $p(\boldsymbol{\beta}) = \prod_k \text{Ga}(\beta_k; \mathtt{s}_0, \mathtt{c}_0)$, which is a Gamma distribution with updated parameters:

$$\mathsf{s}_k = \mathsf{s}_0 + Z_k, \tag{A.29}$$

$$\mathsf{c}_k = \mathsf{c}_0 - Z_k \bar{t}_k \tilde{\mu}_k. \tag{A.30}$$

A.3 Convergence diagnostics

Algorithm 3 in Chapter 6 consists of cycling through the update rules based on the derivations of the previous section. In order to define an appropriate termination criterion, we follow the derivations of [Bishop, 2007] and use the free energy term of Eq. (A.2), given by:

$$\begin{aligned}
\mathcal{L}(q) &= \sum_{\mathbf{Y}} \int_{\boldsymbol{\pi}\in\mathbb{R}^K_{(+)}} \int_{\boldsymbol{\mu}\in\mathbb{R}^K} \int_{\boldsymbol{\beta}\in\mathbb{R}^K_{(+)}} q(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}) \ln \left\{ \frac{p(\mathbf{t}, \mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}}{q(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta})} \right\} \mathrm{d}\boldsymbol{\pi}\ \mathrm{d}\boldsymbol{\mu}\ \mathrm{d}\boldsymbol{\beta} \\
&= \mathbb{E}[\ln p(\mathbf{t}, \mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta})] - \mathbb{E}[\ln q(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta})] \\
&= \mathbb{E}[\ln p(\mathbf{t}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta})] + \mathbb{E}[\ln p(\mathbf{Y}|\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\mu})] + \mathbb{E}[\ln p(\boldsymbol{\beta})] \\
&\quad -\mathbb{E}[\ln q^\star(\mathbf{Y})] - \mathbb{E}[\ln q^\star(\boldsymbol{\pi})] - \mathbb{E}[\ln q^\star(\boldsymbol{\mu})] - \mathbb{E}[\ln q^\star(\boldsymbol{\beta})],
\end{aligned} \tag{A.31}$$

where we can see that each term on the r.h.s. of the above expressions, denotes the Shannon entropy of distributions in Eq. (6.3), Eq. (6.5), Eq. (6.7), Eq. (6.9), Eq. (6.11) and the approximate distributions in Eq. (A.21), Eq. (A.23), Eq. (A.25) and Eq. (A.28).

Each one of the above equations is known, allowing us to calculate the free energy $\mathcal{L}(q)$ at each iteration and terminate if the increase from one step to the next is less that 1%.

Bibliography

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

Albert, R. and Barabási, A. L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

Albert, R.; Jeong, H., and Barabási, A. L. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

Aplin, L. M.; Farine, D. R.; Morand-Ferron, J., and Sheldon, B. C. Social networks predict patch discovery in a wild population of songbirds. *Proceedings of the Royal Society B: Biological Sciences*, 279(1745):4199–4205, 2012.

Bak, P.; Christensen, K.; Danon, L., and Scanlon, T. Unified scaling law for earthquakes. *Physical Review Letters*, 88(17):178501, 2002.

Ball, B.; Karrer, B., and Newman, M. E. J. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.

Barabási, A. L. and Albert, R. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

Barabási, A. L. and Bonabeau, E. Scale free networks. *Scientific American*, 288:60 – 69, 2003.

Barabási, A. L; Jeong, H.; Néda, Z.; Ravasz, E.; Schubert, A., and Vicsek, T. Evolution of the social network of scientific collaborations. *Physica A*, 311(34):590 – 614, 2002.

Barrat, A.; Barthélémy, M., and Vespignani, A. Modeling the evolution of weighted networks. *Physical Review E*, 70(6):066149, 2004.

Batagelj, V. and Mrvar, A. Pajek – program for large network analysis. *Connections*, 21(2): 47–57, 1998.

Bejder, L.; Fletcher, D., and Brager, S. A method for testing association patterns of social animals. *Animal Behaviour*, 56(3):719 – 725, 1998.

Bernardo, J. M. and Smith, A. F. M. *Bayesian Theory*. John Wiley, 1st edition, 1994.

Berry, M. W.; Browne, M.; Langville, A. N.; Pauca, V. P., and Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, 52:155–173, 2007.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 1st edition, 2007.

Blondel, V. D.; Guillaume, J. L.; Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008(10):P10008, 2008.

Brent, L. J. N.; Heilbronner, S. R; Horvath, J. E.; Gonzalez-Martinez, J.; Ruiz-Lambides, A.; Robinson, A. G.; Skene, J. H. P., and Platt, M. L. Genetic origins of social networks in rhesus macaques. *Scientific Reports*, 3, 2013.

Buchanan, M. and Caldarelli, G. A networked world. *Physics World*, 23(2):22–24, 2010.

Cemgil, A. T. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009:4:1–4:17, 2009.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 1st edition, 2006.

Clauset, A.; Newman, M. E. J., and Moore, C. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

Clauset, A.; Shalizi, C. R., and Newman, M. E. J. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Cohen, R. and Havlin, S. Scale-free networks are ultrasmall. *Physical Review Letters*, 90(5): 058701, 2003.

Conradt, L. and Roper, T. J. Activity synchrony and social cohesion: a fission-fusion model. *Proceedings of the Royal Society Series B: Biological Sciences*, 267(1458):2213–2218, 2000.

Costenbader, E. and Valente, T. W. The stability of centrality measures when networks are sampled. *Social networks*, 25(4):283–307, 2003.

Croft, D. P. *Exploring Animal Social Networks*. Princeton University Press, 1st edition, 2008.

Croft, D. P.; Krause, J., and James, R. Social networks in the guppy (*Poecilia reticulata*). *Proceedings of the Royal Society Series B: Biological Sciences*, 271(Suppl 6):S516, 2004.

Danon, L.; Diaz-Guilera, A.; Duch, J., and Arenas, A. Comparing community structure identification. *Journal of Statistical Mechanics*, 2005(09):P09008, 2005.

Diebolt, J. and Robert, C. P. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society Series B: Methodological*, 56(2):pp. 363–375, 1994.

Donetti, L. and Muñoz, M. A. Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics*, 2004(10):P10012, 2004.

Duch, J. and Arenas, A. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104, 2005.

Eagle, N. and Pentland, A. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

Evans, T. S. and Lambiotte, R. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1):016105, 2009.

Faloutsos, C.; McCurley, K. S., and Tomkins, A. Connection subgraphs in social networks. In *SIAM International Conference on Data Mining, Workshop on Link Analysis, Counterterrorism and Security*, 2004.

Faloutsos, M.; Faloutsos, P., and Faloutsos, C. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, SIGCOMM '99, pages 251–262, New York, NY, USA, 1999. ACM.

Farine, D. R.; Garroway, C. J., and Sheldon, B. C. Social network analysis of mixed-species flocks: exploring the structure and evolution of interspecific social behaviour. *Animal Behaviour*, 84(5):1271 – 1277, 2012.

Farkas, I.; Abel, D.; Palla, G., and Vicsek, T. Weighted network modules. *New Journal of Physics*, 9(6):180, 2007.

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

Fortunato, S. and Barthélémy, M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

Fowler, J. H.; Settle, J. E., and Christakis, N. A. Correlated genotypes in friendship networks. *Proceedings of the National Academy of Sciences*, 108(5):1993–1997, 2011.

Fox, C. W. and Roberts, S. J. A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 38(2):85–95, 2012.

Frankel, Z. Probational Research Student (PRS) transfer thesis. Technical report, Department of Engineering Science, University of Oxford, 2012.

Franks, D. W.; James, R.; Noble, J., and Ruxton, G. D. A foundation for developing a methodology for social network sampling. *Behavioral Ecology and Sociobiology*, 63(7): 1079–1088, 2009.

Garnett, R.; Osborne, M. A.; Reece, S.; Rogers, A., and Roberts, S. J. Sequential Bayesian prediction in the presence of changepoints and faults. *The Computer Journal*, 53(9):1430–46, 2010.

Gero, S.; Engelhaupt, D.; Rendell, L., and Whitehead, H. Who cares? between-group variation in alloparental caregiving in sperm whales. *Behavioral Ecology*, 20(4):838, 2009.

Gibbons, J. W. and Andrews, K. M. Pit tagging: simple technology at its best. *Bioscience*, 54(5):447–454, 2004.

Ginsberg, J. R. and Young, T. P. Measuring association between individuals or groups in behavioural studies. *Animal Behaviour*, 44:377–379, 1992.

Girvan, M. and Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

Gleiser, P. and Danon, L. Community structure in jazz. *Advances in Complex Systems*, 6(4): 565–573, 2003.

Goh, K. I. and Barabási, A. L. Burstiness and memory in complex systems. *Europhysics Letters*, 81(4):48002, 2008.

Goh, K.-I.; Cusick, M. E.; Valle, D.; Childs, B.; Vidal, M., and Barabási, A.L. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

Golding, I.; Paulsson, J.; Zawilski, S. M., and Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

Good, B. H.; de Montjoye, Y.-A., and Clauset, A. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

Granovetter, M. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

Grosler, A. *The Great Tit*. Hamlyn, 1st edition, 1993.

Grunwald, P.D. *The Minimum Description Length Principle*. MIT Press, 1st edition, 2007.

Guillaume, J.-L. and Latapy, M. Bipartite graphs as models of complex networks. *Physica A*, 371(2):795 – 813, 2006. ISSN 0378-4371.

Guimerà, R. and Amaral, L. A. N. Cartography of complex networks: modules and universal roles. *Journal of Statistical Mechanics*, 2005(02):P02001, 2005.

Guimerà, R. and Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

Hey, T. The fourth paradigm: data-intensive scientific discovery. Microsoft Research Redmond, WA, 2009.

Holmstrom, E.; Bock, N., and Brannlund, J. Modularity density of network community divisions. *Physica D*, 238(14):1161 – 1167, 2009.

James, R.; Croft, D. P., and Krause, J. Potential banana skins in animal social network analysis. *Behavioral Ecology and Sociobiology*, 63:989–997, 2009.

Jaynes, E. T. *Probability Theory: The Logic of Science*. Cambridge University Press, 1st edition, 2003.

Johnsen, T. S.; Zuk, M., and Fessler, E. A. Social dominance, male behaviour and mating in mixed-sex flocks of red jungle fowl. *Behaviour*, 138(1):1–18, 2001.

Kamada, T. and Kawai, S. A simple method for computing general position in displaying three-dimensional objects. *Computer Vision, Graphics, and Image Processing*, 41(1):43–56, 1988.

Karp, R. M. Reducibility among combinatorial problems. *50 Years of Integer Programming 1958-2008*, pages 219–241, 2010.

Kernighan, B. W. and Lin, S. An Efficient Heuristic Procedure for Partitioning Graphs. *The Bell system technical journal*, 49(1):291–307, 1970.

Kietzmann, J. H.; Hermkens, K.; McCarthy, I. P., and Silvestre, B. S. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 54(3):241 – 251, 2011.

Kirkpatrick, S.; Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science*, 220(4598):pp. 671–680, 1983.

Knuth, D. E. *The Stanford GraphBase: A Platform for Combinatorial Computing*. ACM Press, 1993.

Konstas, I.; Stathopoulos, V., and Jose, J. M. On social networks and collaborative recommendation. In *Proceedings of the 32nd International Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 195–202, New York, NY, USA, 2009. ACM.

Kossinets, G. Effects of missing data in social networks. *Social Networks*, 28(3):247 – 268, 2006.

Krause, J.; Lusseau, D., and James, R. Animal social networks: an introduction. *Behavioral Ecology and Sociobiology*, 63:967–973, 2009.

Krebs, V. http://www.orgnet.com/, 2010.

Krings, G.; Karsai, M.; Bernhardsson, S.; Blondel, V. D., and Saramäki, J. Effects of time window size and placement on the structure of an aggregated communication network. *European Physical Journal Data Science*, 1(1):1–16, 2012.

Lambiotte, R. Multi-scale modularity in complex networks. In *Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, pages 546–553. IEEE, 2010.

Lambiotte, R. and Ausloos, M. Uncovering collective listening habits and music genres in bipartite networks. *Physical Review E*, 72:066107, 2005.

Lancichinetti, A. and Fortunato, S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80 (1):016118, 2009.

Lancichinetti, A.; Fortunato, S., and Kertsz, J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

Lanczos, C. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office, 1st edition, 1950.

Laumann, E. O.; Marsden, P. V., and Prensky, D. The boundary specification problem in network analysis. *Research methods in social network analysis*, 61:87, 1989.

Lauw, H. W.; Lim, E. P.; Pang, H., and Tan, T. T. Social network discovery by mining spatio-temporal events. *Computational & Mathematical Organization Theory*, 11(2):97–118, 2005.

LaValle, S.; Lesser, E.; Shockley, R.; Hopkins, M. S., and Kruschwitz, N. Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 52(2):21–32, 2011.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorisation. *Nature*, 401:788–791, 1999.

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562. MIT Press, 2000.

Lee, S. H.; Kim, P.-J., and Jeong, H. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102, 2006.

Leskovec, J.; Kleinberg, J., and Faloutsos, C. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 177–187, New York, NY, USA, 2005. ACM.

Leskovec, J.; Adamic, L. A., and Huberman, B. A. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), 2007.

Leung, I. X. Y.; Hui, P.; Liò, P., and Crowcroft, J. Towards real-time community detection in large networks. *Physical Review E*, 79:066107, 2009.

Li, M.; Fan, Y.; Chen, J.; Gao, L.; Di, Z., and Wu, J. Weighted networks of scientific communication: the measurement and topological role of weight. *Physica A*, 350(24):643 – 656, 2005.

Lilliefors, H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, pages 399–402, 1967.

Lloyd, S. Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28 (2):129 – 137, 1982.

Lusseau, D. and Conradt, L. The emergence of unshared consensus decisions in bottlenose dolphins. *Behavioral Ecology and Sociobiology*, 63(7):1067–1077, 2009.

Lusseau, D. and Newman, M. E. J. Identifying the role that animals play in their social networks. *Proceedings of the Royal Society Series B: Biological Sciences*, 271(6):S477–S481, 2004.

Lusseau, D.; Schneider, K.; Boisseau, O. J.; Haase, P.; Slooten, E., and Dawson, S. M. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

MacKay, D. J. C. Probable networks and plausible predictions a review of practical Bayesian models for supervised neural networks. *Network: Computation in Neural Systems*, 6(3): 469–505, 1995.

Mankad, S. and Michailidis, G. Structural and functional discovery in dynamic networks with non-negative matrix factorization. *arXiv:1305.7169*, 2013.

Mann, R.; Freeman, R.; Osborne, M.; Garnett, R.; Armstrong, C.; Meade, J.; Biro, D.; Guilford, T., and Roberts, S. J. Objectively identifying landmark use and predicting flight trajectories of the homing pigeon using Gaussian processes. *Journal of The Royal Society Interface*, 8(55):210–219, 2011.

McDonald, D. B. Predicting fate from early connectivity in a social network. *Proceedings of the National Academy of Sciences*, 104(26):10910–10914, 2007.

Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21: 1087, 1953.

Milgram, S. The small world problem. *Psychology Today*, 1:61–67, 1967.

Mitchell, T. M. Mining our reality. *Science*, 326(5960):1644–1645, 2009.

Mucha, P. J.; Richardson, T.; Macon, K.; Porter, M. A., and Onnela, J-P. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

Nacher, J. C. and Akutsu, T. On the degree distribution of projected networks mapped from bipartite networks. *Physica A*, 390(2324):4636 – 4651, 2011.

Nepusz, T.; Petróczi, A.; Négyessy, L., and Bazsó, F. Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1):016107, 2008.

Neuts, M. F. The burstiness of point processes. *Stochastic Models*, 9(3):445–466, 1993.

Newman, M. E. J. Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64:016131, 2001a.

Newman, M. E. J. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64:016132, 2001b.

Newman, M. E. J. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

Newman, M. E. J. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003a.

Newman, M. E. J. The structure and function of complex networks. *Society for Industrial and Applied Mathematics Review*, 45(2):167–256, 2003b.

Newman, M. E. J. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

Newman, M. E. J. *Networks: an Introduction*. Oxford University Press, 1st edition, 2010.

Newman, M. E. J and Girvan, M. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

Newman, M. E. J.; Strogatz, S. H., and Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

Newman, M. E. J.; Barabási, A. L., and Watts, D. J. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

Nicosia, V.; Mangioni, G.; Carchiolo, V., and Malgeri, M. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics*, 2009(03):P03024, 2009.

Oh, K. P. and Badyaev, A. V. Structure of social networks in a passerine bird: consequences for sexual selection and the evolution of mating strategies. *The American Naturalist*, 176 (3):E80–E89, 2010.

Onnela, J.-P.; Fenn, D. J.; Reid, S.; Porter, M. A.; Mucha, P. J.; Fricker, M. D., and Jones, N. S. Taxonomies of networks from community structure. *Physical Review E*, 86(3): 036104, 2012.

Palla, G.; Derenyi, I.; Farkas, I, and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature Letters*, 435(7043):814–818, 2005.

Penny, W. and Roberts, S. J. Bayesian multivariate autoregressive models with structured priors. *IEEE Proceedings on Vision, Image and Signal Processing*, 149(1):33–41, 2002.

Penny, W. D. and Roberts, S. J. Dynamic logistic regression. In *International Joint Conference on Neural Networks*, volume 3, pages 1562–1567, 1999.

Perrins, C. M. Population fluctuations and clutch-size in the great tit, parus major l. *The Journal of Animal Ecology*, pages 601–647, 1965.

Porter, M. A.; Mucha, P. J.; Newman, M. E. J., and Friend, A. J. Community structure in the united states house of representatives. *Physica A*, 386(1):414–438, 2007.

Porter, M. A.; Onnela, J.-P., and Mucha, P. J. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097,1164–1166, 2009.

Psorakis, I.; Rezek, I.; Roberts, S. J., and Sheldon, B. C. Inferring social network structure in ecological systems from spatio-temporal data streams. *Journal of the Royal Society Interface*, 9(76):3055–3066, 2012.

Radicchi, F.; Ramasco, J. J., and Fortunato, S. Information filtering in complex weighted networks. *Physical Review E*, 83:046101, 2011.

Raghavan, U. N.; Albert, R., and Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.

Rajaraman, A. and Ullman, D. J. *Mining of Massive Datasets*. Cambridge University Press, 1st edition, 2011.

Reichardt, J. and Bornholdt, S. Detecting fuzzy community structures in complex networks with a potts model. *Physical Review Letters*, 93(21):218701, 2004.

Reichardt, J. and Bornholdt, S. When are networks truly modular? *Physica D*, 224(1-2):20 – 26, 2006.

Roberts, S. J.; Husmeier, D.; Rezek, I., and Penny, W. Bayesian approaches to Gaussian mixture modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20 (11):1133–1142, 1998.

Robertson, G. J.; Cooke, F.; Goudie, R. I., and Boyd, W. S. The timing of pair formation in harlequin ducks. *Condor*, pages 551–555, 1998.

Rosvall, M. *Information Horizons in a Complex World*. PhD thesis, Department of Physics, Umea University, 2006.

Rosvall, M. and Bergstrom, C. T. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331, 2007.

Rosvall, M. and Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

Sandell, M. and Smith, H. G. Dominance, prior occupancy, and winter residency in the great tit (Parus major). *Behavioral Ecology and Sociobiology*, 29(2):pp. 147–152, 1991. ISSN 03405443. URL `http://www.jstor.org/stable/4600597`.

Service, R. Complex systems: Exploring the systems of life. *Science*, 284(5411):80–83, 1999.

Sih, A.; Hanser, S., and McHugh, K. Social network theory: new insights and issues for behavioral ecologists. *Behavioral Ecology and Sociobiology*, 63:975–988, 2009.

Simpson, E.; Roberts, S. J.; Psorakis, I., and Smith, A. Dynamic Bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, pages 1–35. Springer, 2013.

Smith, A. M.; Lynn, S.; Sullivan, M.; Lintott, C. J.; Nugent, P. E.; Botyanszki, J.; Kasliwal, M.; Quimby, R.; Bamford, S. P.; Fortson, L. F.; Schawinski, K.; Hook, I.; Blake, S.; Podsiadlowski, P.; Jnsson, J.; Gal-Yam, A.; Arcavi, I.; Howell, D. A.; Bloom, J. S.; Jacobsen, J.;

Kulkarni, S. R.; Law, N. M.; Ofek, E. O., and Walters, R. Galaxy zoo supernovae. *Monthly Notices of the Royal Astronomical Society*, 412(2):1309–1319, 2011.

Smith, M.; Reece, S.; Roberts, S. J.; Psorakis, I., and Rezek, I. Maritime abnormality detection using Gaussian processes. *Knowledge and Information Systems (KAIS)*, 2013. (in press).

Stumpf, Michael P. H. and Porter, M. A. Critical truths about power laws. *Science*, 335 (6069):665–666, 2012.

Tan, V. and Févotte, C. Automatic relevance determination in nonnegative matrix factorization. In *SPARS09 - Signal Processing with Adaptive Sparse Structured Representations*, pages 1–19, 2009.

Tantipathananandh, C.; Berger-Wolf, T., and Kempe, D. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 717–726, New York, NY, USA, 2007. ACM.

Traud, A. L.; Kelsic, E. D.; Mucha, P. J., and Porter, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

Tumminello, M.; Aste, T.; Di Matteo, T., and Mantegna, R. N. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426, 2005.

Vázquez, A.; Oliveira, J. G.; Dezsö, Z.; Goh, K.-I.; Kondor, I., and Barabási, A. L. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.

Voelkl, B. and Kasper, C. Social structure of primate interaction networks facilitates the emergence of cooperation. *Biology Letters*, 5(4):462, 2009.

Wagner, G. P.; Pavlicev, M., and Cheverud, J. M. A networked world. *Nature Reviews Genetics*, 8(12):921–931, 2007.

Walker, D. M.; Carmeli, C.; Pérez-Barbería, F. J.; Small, M., and Pérez-Fernández, E. Inferring networks from multivariate symbolic time series to unravel behavioural interactions among animals. *Animal Behaviour*, 79(2):351–359, 2010.

Wang, R. S.; Zhang, S.; Wang, Y.; Zhang, X. S., and Chen, L. Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures. *Neurocomputing*, 72(1-3):134 – 141, 2008.

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

Wey, T.; Blumstein, D. T.; Shen, W., and Jordn, F. Social network analysis of animal behaviour: a promising tool for the study of sociality. *Animal Behaviour*, 75(2):333 – 344, 2008.

Whitehead, H. *Analyzing Animal Societies: Quantitative Methods for Vertebrate Social Analysis*. Chicago University Press, 1st edition, 2008.

Whitehead, H. and Dufault, S. Techniques for analyzing vertebrate social structure using identified individuals: review and recommendations. *Advances in the Study of Behavior*, 28:33–74, 1999.

Whitehead, H.; Bejder, L., and Ottensmeyer, C.A. Testing association patterns: issues arising and extensions. *Animal Behaviour*, 69(5):e1–e6, 2005.

Zhang, S.; Wang, R.S., and Zhang, X. S. Uncovering fuzzy community structure in complex networks. *Physical Review E*, 76(4):046103, 2007.

Zhong, M. and Girolami, M. Reversible jump MCMC for non-negative matrix factorization. In *12th International Conference on Artificial Intelligence and Statistics*, page 8, 2009.

Zhou, T.; Ren, J.; Medo, M., and Zhang, Y. C. Bipartite network projection and personal recommendation. *Physical Review E*, 76:046115, 2007.

Zweig, K. and Kaufmann, M. A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*, 1:187–218, 2011.