UNCERTAINTY QUANTIFICATION IN COMPLEX NETWORKS WITH APPLICATIONS TO BRAIN CONNECTOMICS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Claire Donnat
June 2020

This dissertation is online at: http://purl.stanford.edu/pz822dg7210

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Susan Holmes, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jerome Friedman

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jure Leskovec

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

From social networks to neurosciences, researchers across all disciplines are now faced with the challenge of adapting statistical methods to vast, high-dimensional arrays of data: how can we extract patterns and make sense of such large arrays of numbers? In this setting, graphs have become ubiquitous by offering a versatile modeling framework in which data points are represented as nodes, while edges capture various aspects of the underlying organization of the data. These edges typically denote some flexible notion of proximity, ranging from affinities between users and products in recommendation networks, to causal links in directed acyclic graphs or neuronal coactivation patterns in brain datasets. Whereas most of the current literature has focused on using graphs for learning nodes' properties at an atomic level (community detection [59, 113], link prediction [159], etc.), in a variety of applications, the object of interest is the graph itself. In brain connectomics for instance — one of the areas of application at the center of this thesis —, the focus is on understanding the functional and anatomical "wiring" of the brain and its association with cognitive processes and psychiatric diseases. This process requires the extension of traditional statistical notions (mean, variance, etc.) to the graph setting, which currently appear as the missing, albeit crucial building blocks for principled inference on complex systems.

This PhD thesis focuses on providing some methodological tools for extending statistical inference and uncertainty quantification to graph-structured data — whether these graphs are observed or latent. In particular, we motivate this problem by its application to the analysis of fMRI data. Chapter 1 describes some of the properties of this extremely rich data source, as well as the many interesting challenges (low signal-to-noise ratio, scalability, multi-resolution behavior, non-stationarity, etc.) posed by its analysis [14] and its representation as a graph. Building upon this overview of the multiple facets of the challenges which arise when working with real-life graph-structured data, we organize this thesis around the three following themes:

1. **Inference and variability quantification on observed graphs.** Starting with the setting where the graphs are observed and aligned, the first question that we try to tackle consists in quantifying their similarity. We divide this analysis task along two axes of variation:

   (a) **Horizontal: comparing multiple graphs** [48]. The first main block of our work

(Chapter 2) concentrates around the definition of an appropriate distance for contrasting and comparing aligned networks.

(b) **Vertical: comparing multiscale representations of graphs** [46]. In some instances, the comparison of coarsened representations of graphs can be more informative than the comparison of the original ones. We thus turn in Chapter 3 to the extraction of robust multiscale representations of graphs, which we address here by an adaptation of convex clustering.

2. **Inference for latent graphs**: The second main part of our work [45] tackles the case where the graphs are unobserved, and need to be simultaneously inferred and contrasted. In particular, Chapter 4 is centered around the extraction of reliable brain connectome networks through the lens of Bayesian Independent Component Analysis — an approach which allows the flexible integration of multiple sources of information while providing Bayesian uncertainty estimates.

3. **Inference with graphs** [47]: Finally, Chapter 5 opens our discussion to the analysis of data and signals on graphs –rather than the graphs themselves. Indeed, in a number of settings, the underlying organization of a complex system as a graph is crucial in understanding its behavior. In epidemiological studies for instance, social networks have been shown to influence the outcome of an epidemic, its propagation speed, or the variability in the transmission rate. In this context, it becomes essential to try to impute and integrate characteristics of the network structure in the analysis. We focus here on accounting for the potential heterogeneity of the contact network on predictive scenarios for epidemics.

In particular, rather than giving a holistic approach, we strive to tackle and motivate each of these aspects through the lens of their application on a real, concrete dataset. While most of our motivation is provided by fMRI analysis, we emphasize that the use of graphs transcends the neuroscience realm and can be extended to many biomedical applications — which we exemplify in Chapter 2 by considering microbial networks, and in Chapter 5, by highlighting the usefulness of graphs as modeling tool in the study of contagion and infectious diseases.

Acknowledgments

"Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference."
– Robert Frost

Grad school has been an interesting experience — a road that I chose to take five years ago, without really realizing the consequences of moving across the world nor really pondering what working in research could be like. Nonetheless, these past five years have been a formidable, intellectually-shaping experience, both — it goes without saying — from the statistics perspective and on a personal level. I am extremely grateful to my mentor, Professor Susan Holmes, for the faith she put in me. Professor Holmes has consistently been there in every step of the process, guiding me in my research, giving me bursts of motivation whenever my will was starting to falter — whether it was to continue upon a project that I had stopped believing in or to give a talk to a conference when I felt my work was still unsatisfactory. She has been an example of true scholarship, and has shown me how research should be approached. I would also like to thank Professor Jure Leskovec, who has provided me with an invaluable experience by introducing me to another, more Machine-Learning focused facet of graph-related problems — I have learned a lot by collaborating with him and the SNAP group — and Professor Jerome Friedman for kindly agreeing to be on my reading committee

I would also like to thank the members of Holmes Lab, past and present, for providing me with a dynamic and intellectually-stimulating research environment. In particular, I would like to give a special thanks to Kris, who has been a source of motivation and inspiration, and to whom I have always looked up to. To Lan and her crazy Thanksgivings. To Laura and Nina, my spiritual big sisters, whose dedication and approach to research I find inspiring. To Anne-Maud, who has been there for me and fed me chocolate bunnies when I most needed it.

As my friend Keli once said, grad school is not only about the research — it's also about the long hours, the weekends spent in the lab, and the emotional roller-coaster. I would thus like to dedicate this thesis to my friends who have helped my ship stay afloat and on course amidst the occasionally raging PhD storm. In particular, to Mona, who has been a real pillar for me, and whose tea and snacks have kept me through the long hours. To "His Bateness" Stephen, who has been a true source

of inspiration, both as a leader of our Statistics Peloton and because he knocks off all these p-values so well. To Sherrie, and all our bike rides. To Nima, despite his unfortunate fixation with white kitchen appliances. To Keli, for braving these steep uphills together, going ghost hunting with me in the Tenderloin at night or spontaneously singing along in the office — the first years won't miss us.

Finally, I would also like to extend a huge thank you to everyone who has supported me throughout this journey from the other side of the Atlantic. To my parents, my younger brother and sister, Pierre-Olivier and Hélène. To all the "jeunes bernaches" from the Raid 2012 squadron at École Polytechnique, whose camaraderie I know I can depend on despite the time and distance. To Sophie and Virgile, whose friendship has supported me through the ups and downs of life and the PhD, and has lifted my spirits so many times. And of course, to Mathilde, with whom I have shared so much over the past ten years — from suffering under DPM in "classes prépa", to pretending to be officers in undergrad and ranting about research in grad school.

To all my mentors and friends, thank you. The words escape me to convey all the gratitude I have for everyone who has helped me over these past five years. I will thus finish these acknowledgments by sharing some words of wisdom that my math professor in prep school once impressed upon me. May these seem as random as they did back then.

"Rien ne sert de courir; il faut partir à point." — DPM, quoting Jean de la Fontaine.

Contents

Abstract **iv**

Acknowledgments **vi**

1 Networks, Statistics and Brain Connectomics **1**

1.1 Brain Connectomics 2

1.2 Illustration: the Neuroanalysis Reproducibility Study 4

1.2.1 The Data 4

1.2.2 Our analysis pipeline 6

1.2.3 Results of the analysis. 7

1.2.4 Main take-aways from the NARPS study 8

1.3 Objectives for this thesis 9

2 Inference on Observed Graphs: Defining Distances **11**

2.1 Motivation: quantifying network similarity 12

2.1.1 Applications: microbiome and fMRI data 12

2.1.2 Problem statement and notation 14

2.2 Quantifying local changes via structural distances 15

2.2.1 The Hamming distance 16

2.2.2 The Jaccard distance 20

2.2.3 Shortcomings of local approaches 23

2.3 Comparing graph structures: a spectral approach 24

2.3.1 ℓ_p distances on the eigenvalues 25

2.3.2 Spanning tree similarities 28

2.3.3 Distances based on the eigenspectrum distributions 30

2.3.4 The Polynomial Approach 33

2.4 Quantifying change at the mesoscale 35

2.4.1 Quantifying interactions: connectivity-based distances 36

2.4.2 Heat spectral wavelets 37

2.4.3 Application to the microbiome and fMRI study . . . 40
2.5 Synthetic Experiments . . . 41
2.6 Case study for spatial dynamics: worldwide recipe networks . . . 44
2.7 Conclusion: which distance should we use? . . . 49

3 Inference on Observed Graphs: Convex Clustering 53
3.1 Motivation: why do we need robust multiscale representations for graphs? . . . 54
3.2 Mathematical Problem Statement . . . 55
3.3 Algorithm: a dual FISTA approach . . . 58
3.4 Performance Analysis . . . 64
3.5 Validation with Real-life Experiments . . . 66
3.6 Conclusion: how can we compare hierarchical structures? . . . 69

4 Inference on Latent Graphs: Bayesian ICA 70
4.1 Motivation: extracting relevant subnetworks in Brain Connectomics data . . . 71
4.2 A Bayesian ICA model . . . 74
4.3 Testing: Synthetic Experiments . . . 77
4.4 Validation: Real-life experiments . . . 83
4.5 Conclusion: advantages, disadvantages and further research directions . . . 88

5 Inference on Graphs: An Epidemics Example 92
5.1 Motivation: contagion, graphs and heterogeneity . . . 93
5.2 Model and Theory: towards a heterogeneous R . . . 98
5.3 Evaluating the impact of adding heterogeneity in predictive scenarios. . . . 114
5.4 Conclusion: . . . 127

6 Conclusion 129

Bibliography 130

Appendices 140

A Complements for the distances between networks 141
Appendices . . . 141
A.1 The 2011 Relman microbiome study : a network perspective . . . 141
A.2 From fMRI data to brain connectomics . . . 143
A.3 The recipes network . . . 144
A.4 Results for the synthetic experiments . . . 147
A.5 Understanding the HIM parameters . . . 150

B Convex Clustering **155**
B.1 Derivation of the ADMM updates 155
B.2 Large Scale Computations: derivation of the linearization algorithm 158

C Bayesian ICA **159**
C.1 fMRI-Studies 159
A Functional fMRI Pre-processing 159
A Structural Processing Pipeline: NDMG 162
C.2 Numerical approximation 163

D Inference on Graphs: An Epidemics Example **170**
D.1 Validation: Synthetic Experiments 170
D.2 Comparison of the Bayesian model with its deterministic counterpart 172
D.3 Appendix: Added variability in the infectious profile 174

List of Tables

1.1 Variability in the reported results across the NARPS analysis study 8

3.1 Performance of kmeans clustering on the raw embeddings, with respectively 4 or 28 classes as ground truth labels, for $\alpha = 0.95$. 66

4.1 Comparison of the different algorithms' results on our synthetic problem 82
4.2 Comparison of the different algorithms on synthetic problem: Evaluation of the different components 83

5.1 **Spatial Random-Effects Model:** Comparison of performances of the different prediction (over 5 days).For the Bayesian method, we compare predictions that make full use of the spatial heterogeneity of the method (1st row), vs one that uses the fixed average group value (2nd row), and a third that uses the mean R_0 over all groups (3rd row) to make the predictions. We compare these results to the coefficients obtained using `earlyR` fitted independently on each cluster (4th row) or over the aggregated data (5th row). 110
5.2 **Spatial Random-Effects:**Comparison of the stopping times associated to the scenarios drawn using different R (group-specific or another group's). The N/A values indicate that the stopping time has not been reached in any of our 800 simulations. 121

A.1 Identification of the ingredients that change the most from one graph to another . . 146

List of Figures

1.2 Example of a mixed-gamble task. 5
1.3 Localization of the brain regions tested in the NARPS study. 6
1.4 Consistency between team results, as reported in [14] 7

2.1 Hamming distance between bacterial graphs (top rows), and brain graphs (bottom row). Heatmap of the Hamming distances between Kendall-correlation-based bacterial graphs **(A)** and MDS projections on the first two principal components. **(B)**.Colors denote treatment phases, and shapes represent different subjects . Plots of the consecutive distances between bacterial graphs **(C)**. Minimum Spanning tree between bacterial graphs induced by the Hamming distance **(G)**. Friedman-Rafsky test for significance for the different datasets **(F)**. Clustermap of the fMRI graphs **(E)**. Minimum spanning tree between brain connectomes induced by the Hamming distance **(G)**. 19
2.2 Application of the Jaccard distance to the microbiome study. Cluster-map of the Jaccard distances between Kendall-correlation-based bacterial graphs **(A)** . Plots of the consecutive distances between bacterial graphs **(B)**. MDS projection of the bacterial **(C)** graphs on the first two principal axes. Colors denote treatment phases, and shapes represent different subjects. Minimum spanning tree between bacterial graphs **(D)** 22
2.3 Two modifications of the same initial graphs (displayed in figure 2.3a, such that the Hamming distance with the original is $d_H(G_1, G_2) = d_H(G_1, G_3) = \frac{4}{21}$, and the Jaccard distances are $d_J(G_1, G_2) = \frac{2}{5}$ and $d_J(G_1, G_3) = \frac{1}{4}$. The average shortest path length are 1.71 for the initial graph G_1, 1.51 for G_2 and 2 for G_3 24

2.4 Application of ℓ_2 spectral distances using two functions of the Laplacian eigenspectra in Eq. 2.3.1: low-pass filters $f(\lambda) = \epsilon^{-0.1\lambda}$ on the Microbiome (**top row**) and $f(\lambda) = \epsilon^{-1.2\lambda}$ on the fMRI dataset(**bottom row**). Clustermap of the corresponding distances between bacterial graphs **(A)**/ brain connectomes **(E)** . MDS projection of the bacterial **(B)**/ fMRI **(C)** graphs on the first two principal axes. Colors denote treatment phases/ years of dependency. Pvalue of the FR-test for the 1-nn metagraph across the different datasets, for the low-pass filter $f(\lambda) = \epsilon^{-1.2\lambda}$ **(D)**. 27

2.5 Application of the Spanning tree dissimilarity. **(Top) Microbiome Data/(Bottom) fMRI Data.** Heatmap of the corresponding dissimilarity between Kendall-correlation-based bacterial graphs **(A)**. MDS projection of the bacterial graphs on the first two principal axes **(B)**. Colors denote treatment phases, and shapes represent different subjects. p-values associated to Friedman-Rafsky test of the consistency of the 3-nn metagraph with the labeling of the nodes and analysis-of-variance test **(C)** Clustermap for the fMRI data **(D)**. MDS projections of the brain connectomes on the first two principal components **(E)**. 29

2.6 Application of the Hamming-Ipsen-Mikhailov distance. MDS projection of the bacterial graphs on the first two principal axes**(A)** . Colors denote treatment phases, and shapes represent different subjects. Minimum Spanning Tree induced on the bacterial graphs by the HIM distance **(B)**. Application of the HIM distance to the fMRI data set: clustermap of the different distances between connectomes **(C)**. 33

2.7 Application of the polynomial dissimilarity to the microbiome bacterial graphs, for $K = 3, \alpha = 0.9$. Heatmap of the corresponding dissimilarity **(A)** . MDS projection of the bacterial graphs **(B)** on the first two principal axes. Colors denote treatment phases, and shapes represent different subjects. Plots of the consecutive distances between bacterial graphs **(C)**. 35

2.8 Representation of the heat kernel: the source node (red) has the highest temperature (interpreted here as a "signal value") – illustrated by the high red bar. The heat wave diffuses over the neighborhood (illustrated by the varying intensities of the nodes' colors and height of the bars representing the dwindling temperature/signal strengths.) 38

2.9 Application of the heat wavelet characteristic distance to the microbiome bacterial graphs, for $\tau = 1.2$. Heatmap of the corresponding dissimilarity **(A)** . P-values of the Freidman-Rafksy test on the 5-nearest-neighbor graphs induced by the heat distance, on each dataset. Plots of the consecutive distances between bacterial graphs **(C)**. MDS projection of the bacterial graphs **(D)** on the first two principal axes. Colors denote treatment phases, and shapes represent different subjects. Minimum Spanning tree induced on the bacterial graphs **(E)**. 40

2.10 Results for the Stochastic Block Model topology. **Top Row:** Comparison of the smooth dynamics (no change point), with 0.05% edges rewired at each time step. **Bottom Row:** Change point detection experiment. 43
2.11 Comparison of the pairwise distances and three-nearest cuisine summary graphs (**Left column:** Hamming distance. **Right column:** Jaccard). The three-nearest cuisine summary graph is constructed by representing each co-occurrence network by a node and linking it to its three nearest neighbors according to a given pairwise similarity matrix. 46
2.12 **Ipsen-Mikhailov distance. (right)** 3-nearest-neighbor Proximity Graph between cuisines **(left)** Pairwise distances between cuisines. 47
2.13 Pairwise distances between cuisines for various spectral distances. **Top row:** Ipsen-Mikhailov distance. 2^{nd} **row:** Polynomial distance (section 2.3.4), with $\alpha = 0.9$ and $K = 5$. **Bottom row:** Eigenspectrum-based distance (section 2.3.1) with $f(x) = e^{-0.9x}$. 48
2.14 Proximity Graph between cuisines (heat-wavelet based distance) 49
2.15 Summary of the distances detailed in this chapter. 51

3.1 Application of the convex hierarchical clustering algorithm to a synthetic graph on 196 nodes. PCA representation of the nodes for **(B)** $\lambda = 0.001$, **(C)** $\lambda = 0.03$, **(D)** $\lambda = 0.2$ and **(E)** $\lambda = 3.0$. 65
3.2 Results of the algorithm averaged over 20 independent trials: (A) efficient rank, (B) homogeneity of the clustering on 28 clusters, (C,D) silhouette scores on respectively 28 and 4 clusters for different values of α. 66
3.4 Results for the Khan Dataset 69

4.1 Summary of the Bayesian Model used in this paper. 77
4.2 Examples of independent sparse, localized loadings over the graphs. 79
4.3 Results for the Bayesian model on the recovery of components on a single graph . . 80
4.4 Comparison accuracy of the recovery 81
4.5 Credible intervals (row-wise and column-wise) 81
4.6 Analysis of the impact of the noise on the recovery of the components 84
4.7 Results for the HNU1 dataset 86
4.8 Comparison of the distributions of the closest neighbor similarity. 87
4.9 Robustness across sessions. 88
4.10 Mean of the first connected components 89
4.11 Mean of Component 3 90
4.12 Mean of Component 4 90
4.13 Mean of Component 5 91

5.1 Hierarchical model for R_0 . . . 94
5.2 Output of simulations showing comparisons of the possible trajectories for contagion models using fixed R_0 vs variable R. Dots indicate the average predicted values, whereas the error bars represent the 98%-confidence interval. . . . 96
5.3 Compartmental SEIR model . . . 99
5.4 Plate model . . . 101
5.5 Parameters (infectious profile and selected groups) chosen for the analysis. . . . 103
5.6 Incidence data for some of our selected groups . . . 104
5.7 Plate model for the real data . . . 106
5.8 Distribution of the recovered spatial reproductive numbers R for the spatial Random-Effects Model. . . . 106
5.9 Transmissibility τ and the recovered daily contact rates $\bar{c}_g$ for the spatial Random-Effects Model. . . . 107
5.10 **Spatial Random-Effects Model.** Predictions (with confidence intervals) for Hubei 0 (first days of the quarantine) and Hubei 1 (last 36 days). Y values are plotted on the log.scale. Green confidence intervals are the one recovered by our Bayesian method, and in pink through the `projection` `R`-package. The blue circles show the observations used for training, and the red triangles are observations from the past six days that we use as validation. . . . 108
5.11 **Spatial Random-Effects Model.** Predictions on the log-scale for a few countries. Training observations shown by the blue circles, validation data by the red triangles. Green confidence intervals are the ones recovered by our Bayesian method, and in pink through the `projection` package. . . . 109
5.12 **Spatial Random-Effects Model.** Predictions (log-scale) for the United States. Green confidence intervals are the one recovered by our Bayesian method, and in pink through the `projection` `R-package`. . . . 110
5.13 Distribution of the recovered spatial reproductive numbers R for the random effects modelling for the full Random-Effects model — fitted on 36 consecutive days from Febrary 9th to March 17th . . . 111
5.14 Transmissibility τ and the average recovered daily contact rates $\bar{c}_g$ for the full Random-Effects model. . . . 112

5.15 **Full Random-Effects Model.** Predictions using random effects on the log-scale for a few countries (training observations shown with the blue rounds, validation data displayed through the red triangles). Pink confidence intervals are the ones recovered by our Bayesian complete random-effect model (new $R_i^{(g)}$ for every time step in each trajectory), green are confidence intervals obtained assuming $R^{(g)}$ is random, but constant through time (one $R^{(g)}$ per trajectory), and in blue, the ones assuming that $R^{(g)}$ is fixed and equal to its mean recovered value 112

5.16 **Full Random-Effects Model.** Predictions (log-scale) for the USA. Training observations shown with the blue round, validation data displayed through the red triangles. Pink confidence intervals are the ones recovered by our Bayesian complete random-effect model (new $R_i^{(g)}$ for every time step in each trajectory), green are confidence intervals obtained assuming $R^{(g)}$ is random, but constant through time (one $R^{(g)}$ per trajectory), and in blue, the ones assuming that $R^{(g)}$ is fixed and equal to its mean recovered value. 113

5.17 **Spatial Random-Effects Model: France.** Comparisons of the outcomes of the different strategies. We compare the estimated likely trajectories in terms of occupied hospital beds using various R: the group's specific and tailored Bayesian R, as well as an overall, general R estimated from the aggregated data. We note the substantial difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R yield. 118

5.18 **Spatial Random-Effects:** France. Comparisons of the outcomes of the different strategies. We compare the estimated likely trajectories in terms of occupied hospital beds using various R: the group's specific and tailored Bayesian R, as well as an overall, general R estimated from the aggregated data. We note the substantial difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R yield. 119

5.19 **Spatial Random-Effects: Italy.** Again, we compare the estimated likely trajectories in terms of occupied hospital beds using various R: the group's specific and tailored Bayesian R, as well as an overall, general R estimated from the aggregated data. We note the substantial difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R yield. 120

5.20 **Spatial Random-Effects: California.** We compare the estimated likely trajectories in terms of occupied hospital beds using various R: the group's specific and tailored Bayesian R, as well as an overall, general R estimated from the aggregated data. We note the substantial difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R yield. 122

5.25 **Spatial Random-Effects:** Time to 1% of the population under hospitalization . . 122

5.21 **Spatial Random-Effects.** United States of America 123
5.22 **Spatial Random-Effects.** Comparison Predictions: this figure further shows for four different groups the estimated impact of a given policy, using different R_0s. This shows the importance of correctly accounting for group-wise heterogeneity in the model. 124
5.23 **Comparison Predictions: United Kingdom** for an Alternating scenario with 5 days of business as usual vs 2 days of 50% lockdown. 124
5.24 **Spatial Random-Effects:** Histograms of the expected Time until Hospitalization Overflow for two groups. 125
5.26 **Full Random-Effects:** Comparison of the Static vs Random R 125
5.27 **Full Random-Effects:** Comparison of the projections of the occupied number of beds. 126
5.28 **Full Random-Effects:** Comparison of the projections of the occupied number of beds in the US under different social distancing scenarios. 126

A1 **Subject F.** Examples of scatterplots for a few bacteria. Colors denote different treatment phases. The affinities between bacteria vary across treatment phases. . . 142
B1 **(A)** Education level. **(B)** Number of cigarettes per day. **(C)** Number of years under dependency. **(D)** Number of years since first use. 144
C1 Visualization of the properties of the dataset 145
D1 Results for the Erdös-Rényi topology. **Top Row:** Comparison of the smooth dynamics (no change point), with 0.1% edges rewired at each time step. **Bottom Row:** Change point detection experiment. 148
D2 Results for the Preferential Attachment topology. **Top Row:** Comparison of the smooth dynamics (no change point), with 5% edges rewired at each time step. **Bottom Row:** Change point detection experiment. 149

E1 Scan Parameters for the HNU1 dataset. 161
E2 A few visualization of the data 162
E3 Structural Graph properties 162
E4 Visualization of the recovered components at various steps of the optimization process. 166

E1 Recovered Credible intervals for the Fraser Model 171
E2 Boxplot of the confidence intervals obtained using the `earlyR`package using the data generated in Algorithm 10. 172
E3 R_0 for the spatial Random-Effects with Dirichlet estimated infectivity profile. . . . 174
E4 $\bar{c}$s for the spatial Random-Effects with Dirichlet estimated infectivity profile. 175
E5 Projection accuracy for a few of the European groups for the spatial Random-Effects with Dirichlet estimated infectivity profile. 176

E6 Projection accuracy for groups in the United States for the spatial Random-Effects with Dirichlet estimated infectivity profile. 177

Chapter 1

Networks, Statistics and Brain Connectomics

From social sciences to biology, scientific communities across a wide number of disciplines have become increasingly interested in the study of networks – that is, graphs in which each entity or data point is assigned to a node, and existing interactions between entities are modeled by edges. If graphs provide a versatile framework for encapsulating structural information in datasets, they also come as an indispensable paradigm in a number of applications where the study of each individual node is either irrelevant or intractable. In these cases, one is more interested in the study of the system as a whole, rather than at an atomic level. Yet, while graphs and networks have been actively researched over the past two decades, uncertainty quantification and inference on graphs remain open fields of research in which much remains to be done. Indeed, most of the literature on network data focuses on studying properties within a given graph at the node or community level (e.g community detection [59, 113], link prediction [159], structural similarity analysis [49] etc.), rather than across graphs. Secondly, many current applications are concerned with the analysis of aligned networks, where nodes are endowed with a given identity and are thus no longer exchangeable — thus making most methods based on the summarization of graphs in terms of exchangeable nodes statistics (degree distribution, diameter, etc.) unfit to the case in point. Examples of such aligned networks arise in numerous settings, ranging from from social sciences — in which one could wish to study snapshots of a single social network observed at different points in time—, to microbiology — in which the focus would be on understanding symbiotic mechanisms between species. Consequently, there appears to be an increasing need for the development of statistical tools and methods tailored to the quantification of the variability and uncertainty across sets of aligned networks.

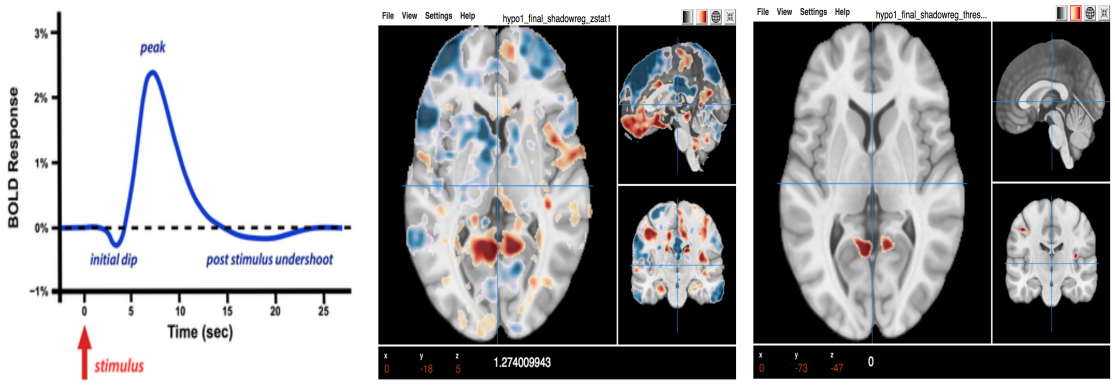
In particular, Brain Connectomics — an area of growing interest in cognitive neuroscience which we use as the guiding motivational thread to this thesis — is a case in point. Let us thus begin

by introducing this field and the various challenges that it presents, as well as the importance of modeling this type of data as a graph.

1.1 Brain Connectomics

Over the recent years, the study of brain connectomics [16, 56, 57] has gained increased interest amidst the neuroscience and cognitive psychology communities. In this framework, the brain is modeled as a graph in which nodes denote voxels or regions of interest (ROIs), while edges represent some notion of functional connectivity, typically inferred from function Magnetic Resonance Imaging (fMRI) scans. Examples of such connectivity measures include Pearson correlations, partial correlations or mutual information between the Blood Oxygenation Level Dependent (BOLD) signals of each region of interest. Brain Connectomics can thus be understood as an extension of the typical analytical pipeline of fMRI images to the analysis of interactions: beyond the detection of node-wise activations of different ROI associated to a given task (motor, speech, etc.), the focus is on capturing how these brain regions interact. The purpose of these studies becomes to relate neuronal coactivation and connectivity patterns to psychiatric illnesses (Alzheimers, dementia, etc.) or cognitive processes. Before delving into the challenges of recovering such interactions, let us begin by introducing this particular data source and the many challenges associated to its analysis — both biological and analytical.

The first important component in the analysis of fMRI data lies in understanding what this imaging modality really measures. From the biological perspective, BOLD signals are an indirect measure of neuronal activity: increased neuronal activity induces changes in regional blood flow, blood volume, and oxygen consumption. Figure 1.1a shows an example of the BOLD response to a single brief stimulus: after a short lag, the BOLD increases for a short period of time (just a few seconds), but long enough to be captured by the MRI. During this phase, regional cerebral blood flow increases to fulfill the activated region's metabolic needs. As a result, the ratio of oxyhemoglobin to deoxyhemoglobin increases, which in turn, creates a distortion in the magnetic field which is captured by the machine. Even with a very brief stimulus, the dominant Hemodynamic Response Function response is slow and delayed, often not occurring until 5-8 seconds later the stimulus. This constitutes one of the main inherent challenges of fMRI analysis: while non invasive, contrary to other quantifying methods such as PET scans, **fMRI is only an indirect method for inferring activity**, and the intensity of the activation can only be understood relative to those in other parts of the brain. This latter notion is important to keep in mind, as it causes fMRI measurements to be vulnerable to many confounding factors (effect of the scanner, magnetic drift, brighter intensities near arteries, etc), which can potentially blur the activation signal that we set out to discover — a crucial soft-spot in many studies which we will further develop in the next subsection.



(a) BOLD Hemodynamic Response Function following a single brief stimulus.

(b) Example of an Unthresholded Z-map (our own NARPS Z-map for hypothesis 1).

(c) Example of a Thresholded Z-map (our own NARPS Z-map for hypothesis 1).

From the purely data analytical perspective, fMRI data consists of a set of time series (one for each of N voxels) $X \in \mathbb{R}^{N \times T}$ capturing the intensity of the magnetic response through time. We thus emphasize, that despite their name, fMRI scans are better described as a matrix of time series, rather than a static image. To infer activation from the raw data, these time series are then heavily pre-processed and either:

- *for task fMRI:* regressed against a given stimulus, yielding coefficients and their associated Z-statistics as indicator of the response of the voxel to the stimulus,
- *for resting-state fMRI,* these series are simply correlated with one another to infer the Default Mode Network (i.e, the alleged background activity of the brain when no task is being performed).

Figures 1.1b and 1.1c show examples of respectively unthresholded and thresholded Z-maps that capture activation patterns arising from the time-series.

The driving hypothesis behind brain connectomics is that the identification of interactions between modules of nodes is crucial to our understanding of cognitive processes and psychological diseases — ranging from depression [97, 107], Attention-deficit/hyperactivity disorder [151], to schizophrenia [58, 123, 147]. Central to the field is the analysis of resting-state functional MRI (rs-fMRI), believed to capture the brain's default activity [55, 70, 118, 119] and the diverse functional interactions between ROIs. Mathematically speaking, the recovery of these interactions boils down to an estimation of the dependency structure between the different brain regions, typically modeled as a graph. These dependencies can be quantified through a variety of measures, whether these correspond to the correlation or precision matrix between the time series, their mutual information, etc.

Yet, because it is only an indirect, relative measure of activity, fMRI data is particularly difficult to analyze and suffers from a high number of potential sources of noise and confounding biases. The

first challenge arises from the intensive pre-processing that fMRI data typically requires: denoising, unringing, motion correction, magnetic drift removal, transformation to a standard template space, etc. No general consensus has been reached in the community, neither with respect to the number nor to the order of these pre-processing steps. This flexibility induces an extremely large number of potential analytical workflows, and thus, in turn, a great variability in the outcome of the results. The variability between scan sessions and individuals further hinders the generalizability of the analysis and complicates the recovery of precise estimates of these interactions. These many challenges show the need for developing robust statistical methods for analyzing this particular type of data, and explain the irreproducibility issues that neuroscience studies have been shown to exhibit, which we exemplify through the results of a reproducibility study that we participated in.

1.2 Illustration: the Neuroanalysis Reproducibility Study

To illustrate the numerous choices and preprocessing steps that have to be made during fMRI analysis, we propose to give as a concrete example a study we participated in. The purpose of this study was to quantify the variability in the results of the analysis of a single dataset led independently by different teams across the world. Collaborating with Dr. Leonardo Tozzi, a postdoctoral fellow in the Department of Psychiatry and Behavioral Sciences at Stanford University, we participated in this challenge as one of the seventy teams performing the analysis. We propose to describe this reproducibility study here since it provides an in-depth illustration of the challenges arising in fMRI analysis, as well as a justification for the development of new, robust statistical methods tailored to this imaging modality. The following is thus based on the paper published in Nature [14].

Our goal as analysts was to pre-process and analyze the fMRI scans of subjects undergoing a mixed-gamble task. The dataset for this study consisted of raw fMRI scans of 108 patients, for which we were asked to assess the significance of activations postulated in nine ex-ante hypotheses. An independent research group coordinating the study then aggregated the results across all analysis teams in order to assess the number of teams concurring and reporting a statistically significant whole-brain corrected result for each hypothesis.

1.2.1 The Data

The data that we were provided consisted in the fMRI scans of 108 healthy participants.

On each of 64 different trials, a mixed gamble was presented to the participant, entailing a 50/50 chance of gaining one amount of money or losing another amount (see Fig 1.2). Each participant was randomly assigned to one of two groups (n=54 in each group in the final experiments). In the equal indifference condition, the matrix of gambles included potential gains twice the range of potential losses [141]; in the equal range condition, the matrix included an equal range of potential gains and losses [40]. All in all, possible losses ranged from 5-20 ILS (in increments of 1 ILS) while gains ranged

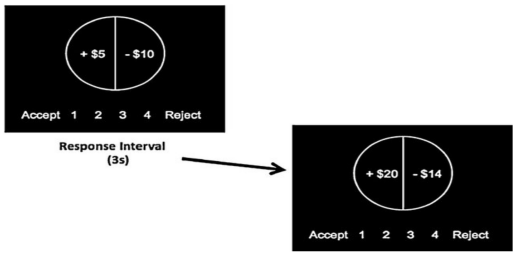


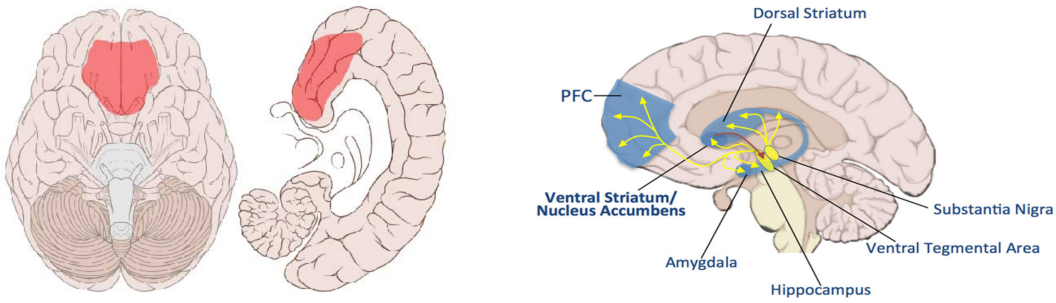
Figure 1.2: Example of a mixed-gamble task.

from 10-40 ILS (in increments of 2) for the equal indifference group or 5-20 ILS for the equal range conditions. All participants were presented all 256 possible combinations of gains and losses across the four runs.

As in [141], participants were asked to evaluate whether or not they would like to play each of the gambles presented to them (strongly accept, weakly accept, weakly reject or strongly reject). They were told that one trial from each of the runs would be selected at random, and if they had accepted that gamble during the task, the outcome would be decided with a coin toss; if they had rejected the gamble, then the gamble would not be played. This incentivizes participants to be consistent in their approach to risk taking and decision making.

Hypotheses. As an analysis team, we were asked to submit yes/no decisions regarding the following anatomical hypotheses for specific contrasts, based on previous results from [141].

- Parametric effect of gain:
  - Positive effect in ventromedial Prefrontal Cortex (PFC) – see Fig. 1.3a for illustration) - for the equal indifference group
  - Positive effect in ventromedial PFC - for the equal range group
  - Positive effect in ventral striatum (see Fig. 1.3b for illustration) - for the equal indifference group
  - Positive effect in ventral striatum - for the equal range group
- Parametric effect of loss:
  - Negative effect in VMPFC - for the equal indifference group
  - Negative effect in VMPFC - for the equal range group
  - Positive effect in amygdala - for the equal indifference group
  - Positive effect in amygdala - for the equal range group
- Equal range vs. equal indifference: Greater positive response to losses in amygdala for equal range condition vs. equal indifference condition.



(a) ventromedial pre-frontal Cortex (b) Striatum and amygdala (figure from [137])

Figure 1.3: Localization of the brain regions tested in the NARPS study.

1.2.2 Our analysis pipeline

Our team decided to use the pre-processed data from the fMRIPrep pipeline [52], which is one of the most established ones and was provided with the data. As per the toolbox's website[1], "fMRIPrep is a functional magnetic resonance imaging (fMRI) data preprocessing pipeline that is designed to provide an easily accessible, state-of-the-art interface that is robust to variations in scan acquisition protocols and that requires minimal user input, while providing easily interpretable and comprehensive error and output reporting."

Consistently with the bulk of the neuroscience analytical literature [114], we model the fMRI response Y using a general linear model (GLM). In this model, the fMRI response $Y \in \mathbb{R}^{T \times N}$ is modeled as a linear combination of temporal task stimuli (in our case, the gambles, which are placed at known times and provided with the data) plus noise:

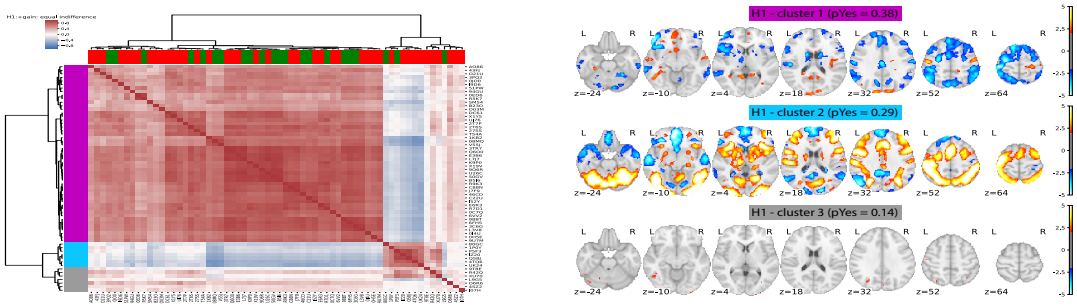
$$Y = X\beta + \epsilon$$

where $\epsilon \sim N(0, V)$, and X is a design matrix containing information about the various signal components. While this model is simple, the main challenge consists in correctly construct an appropriate design matrix X. This is complicated by a number of factors, including the fact that the BOLD response contains low-frequency drifts and artifacts related to head movements, cardiopulmonary-related brain movements, as well as by the variability in hemodynamic brain responses across the brain. To account for these effects, the design matrix X usually consists of both nuisance parameters (corresponding to the drift components and the estimated motion parameters, with variable degrees of freedom) and signal of interest.

We are further interested in combining the results for the individual subjects in order to perform group-level inference, thus calling for a mixed-effect model of our GLM.

Denoting as $Y^{(ij)}$ the j^{th} trial for subject i, the first level model for subject i can be written

[1] `https://fmriprep.readthedocs.io/en/latest/#`



(a) Spearman correlation values between whole-brain unthresholded statistical values for each team

(b) Average statistical maps (thresholded at uncorrected z > 2.0) for each of the three clusters

Figure 1.4: Consistency between team results, as reported in [14]

as: $Y^{(ij)} = X\beta^{(ij)} + \epsilon^{(ij)}$ where $\epsilon^{(ij)} \sim N(0, V^i)$. The second-(group) level can then be written as: $\beta^{(i)} = \beta^{(g_i)} + \eta^{(i)}$, where $\eta^{(i)} \in \mathcal{N}(0, \sigma^2)$, and $g^{(i)}$ in the group to which subject i belongs. For this analysis, we used the simple and widespread approximation of this model done by performing a GLM on each subject (averaging the results across all four trials), and using the resulting activation parameter estimates in a "second-level" group analysis. To perform this analysis, we deployed the neuroscience statistical software FSL on Stanford's Sherlock cluster which allows one to model subject and group effects, and also allows the choice of the basis functions for the hemodynamic response function.

1.2.3 Results of the analysis.

This study revealed the existence of a huge variability in the analysis results and little consensus in the scientific conclusions. Quoting from the resulting paper [14], the flexibility of the preprocessing steps, combined with the flexibility in the data analysis itself "resulted in sizeable variation in hypothesis test results, even for teams whose statistical maps were highly correlated at intermediate stages of their analysis pipeline." In particular, the study identified that the analysis teams' statistical maps could be clustered in three groups (see Fig 1.4a). Interestingly, one of these groups is even anti-correlated with the others. Our team's map lies in the first cluster (51PW). The main authors of the paper continue: " Variation in reported results was related to several aspects of analysis methodology. Importantly, meta-analytic approaches that aggregated information across teams yielded significant consensus in activated regions across teams. [...] The results emphasize the importance of validating and sharing complex analysis workflows, and demonstrate the need for multiple analyses of the same data."

The results shown in the study's papers are reproduced in Table 1.1.

Table 1.1: Variability in the reported results across the NARPS analysis study

	Hypothesis description	Fraction of teams reporting a significant result	Median confidence level	Median similarity estimation
1	Positive parametric effect of gains in the vmPFC (equal indifference group)	0.371	7 (2)	7 (1.5)
2	Positive parametric effect of gains in the vmPFC (equal range group)	0.214	7 (1.5)	7 (1)
3	Positive parametric effect of gains in the ventral striatum (equal indifference group)	0.229	6 (1)	7 (1)
4	Positive parametric effect of gains in the ventral striatum (equal range group)	0.329	6 (1)	7 (1)
5	Negative parametric effect of losses in the vmPFC (equal indifference group)	0.843	8 (1)	8 (1)
6	Negative parametric effect of losses in the vmPFC (equal range group)	0.329	7 (1)	7 (1)
7	Positive parametric effect of losses in the amygdala (equal indifference group)	0.057	7 (1)	8(1)
8	Positive parametric effect of losses in the amygdala (equal range group)	0.057	7 (1)	8 (1)
9	Greater positive response to losses in amygdala for equal range group vs. equal indifference group	0.057	6 (1)	7 (1)

1.2.4 Main take-aways from the NARPS study

This study helped us gain more insight in all the data preprocessing tricks and methods deployed in the analysis of fMRI data. The variability resulting from the pre-processing of fMRI data **highlights the importance of using robust analytical pipelines**, to mitigate the instability in the reported outcomes: in such a noisy setting, it seems imperative to design methods able to account for both the temporal and spatial correlations in the data, while providing robust estimators and reliable confidence intervals. Moreover, as mentioned in the introduction of this chapter, our focus in this thesis is to go beyond voxel activation and to consider interactions and co-activations between brain regions. As such, brain connectomics offers an exceptionally complex playing field for data analysts. Indeed, this requires extending traditional statistical notions (such as the mean of the graph, etc.) to the graph setting — a necessary and crucial block to the statistical analysis of networks and graphs. Moreover, in the context of analyzing brain connectomes, these notions must be robust to noise. Indeed, while interactions might be crucial in getting a more precise understanding of cognitive processes, they come at an increased analytical price. In particular, in most studies, the number of potential edges is significantly greater than the number of time points in the BOLD time series—thus forcing a heavy filtering of the sample correlation matrix to get rid of spurious correlations.

Our neuroscience example thus motivates the following questions, which this thesis attempts to tackle:

- How can we infer robust connectomes from fMRI data?
- How can we compare such connectomes? What is the best measure of similarity that would allow us to account for the inherent noise and variability of this data source?
- Is there a particular scale or coarsening of the different brain regions which would be more appropriate to the study of Brain Connectomics?

1.3 Objectives for this thesis

This PhD thesis focuses on providing some methodological tools for extending statistical inference and uncertainty quantification to graph-structured data—particularly in view of applying these methods to the analysis of fMRI data.

In light of this objective, we begin by studying the case where the graphs that we have to compare are observed and aligned [48]. As such, the first main block of our work — which we describe in detail in Chapter 2 — concentrates on the definition of an appropriate distance for contrasting and comparing aligned networks, thus allowing to characterize the variability of a set of graphs. We then turn in Chapter 3 to the extraction of robust multiscale graph representations. Indeed, in some instances — such as for brain connectomics studies—, the comparison of coarsened representations of graphs can be more informative than the comparison of the original ones. We address this problem here by adapting convex clustering to the analysis of graphs.

The second main part of our work [45] tackles the case where the graphs are unobserved, and need to be simultaneously inferred and contrasted. In particular, Chapter 4 is centered around the extraction of reliable brain connectome networks through the lens of Bayesian Independent Component Analysis — an approach which allows the flexible integration of multiple sources of information while providing Bayesian uncertainty estimates.

Finally, Chapter 5 opens our discussion to the analysis of data that is organized on graphs —rather than an analysis of the graphs themselves. Here, the graph becomes a means or vector guiding the analysis, but not the end-goal of the analysis itself. Indeed, in a number of settings, the underlying organization of a complex system in a graph is crucial in understanding its behavior. In epidemiological studies for instance, social contact networks have been shown to influence the outcome of an epidemic, its propagation speed, or the variability in the transmission rate. In this context, it becomes essential to try to impute and integrate characteristics of the network structure in the analysis — in particular, to account for the heterogeneity of the different components and the impact of such a heterogeneity on the behavior of the system. We focus here on an application

spun from epidemics modeling, in the context of COVID-19 pandemic, still unresolved at the time of writing.

Chapter 2

Inference on Observed Graphs: Distances between sets of aligned networks

We begin by tacking the problem of quantifying the variability in a set of observed, aligned graphs. The previous introductory chapter has highlighted the potential benefits of modeling certain datasets as a graph across a variety of applications and studies, as they offer a powerful framework for describing evolving interactions between agents in complex systems. The focus of such studies typically concerns the system as a whole rather than at an atomic level: brain connectome data represent brain activity by modeling neurons' activation patterns from a network perspective – rather than by recording each individual neuron's activity. Similarly, in microbial ecology, communities of bacteria can be represented by a co-occurrence graph where each bacteria is a node and the edges are a (carefully-selected) function of bacterias' co-abundances. The representation of biological samples as graphs provides a more informative framework than the bacterial counts themselves. Indeed, these graphs can be taken as essential summaries of the data, which can then be used to further investigate the associations highlighted by recent studies between "significant" bacterial communities and various medical conditions, such as obesity [79, 143, 144] or preterm birth [44].

We thus begin by considering the case where these interactions are observed (either through space or through time), and the goal is to understand similarities across different aligned graphs and quantify the variability of the set. The analysis of these networks depends on the selection of the appropriate analytical tools. In particular, a critical step lies in the choice of a distance between graphs capable of reflecting such similarities.

While the literature offers a number of distances that one could a priori choose from, their properties have been little investigated and no guidelines regarding the choice of such a distance have

yet been provided. In particular, most graph distances consider that the nodes are exchangeable and do not take into account node identities. Accounting for the alignment of the graphs enables us to enhance these distances' sensitivity to perturbations in the network and detect important changes in graph dynamics. Thus the selection of an adequate metric is a decisive – yet delicate – practical matter.

In the spirit of Goldenberg, Zheng and Fienberg's seminal 2009 review [69], the purpose of this first methodological chapter is to provide an overview of commonly-used graph distances and an explicit characterization of the structural changes that they are best able to capture. This chapter, based on one of our publications [48], will also serve as our extended literature review for this thesis. To see how the selection of a distance translates in real-life situations, we use as a guiding thread to our discussion the application of these distances to the analysis of both a longitudinal microbiome dataset and a brain fMRI study. We show examples of using permutation tests to detect the effect of covariates on the graphs' variability. Furthermore, synthetic examples provide intuition as to the qualities and drawbacks of the different distances. Above all, we provide some guidance for choosing one distance over another in certain types of applications.

2.1 Motivation: quantifying network similarity

2.1.1 Applications: microbiome and fMRI data

We illustrate our discussion on the analysis of distances between graphs with two main examples, which we present in the two subsequent paragraphs.

The 2011 Relman antibiotics dataset. This longitudinal microbiome study consists of a set of 162 bacterial samples taken from the gut of three distinct subjects (D, E, and F) at different points in time. The subjects were given two courses of antibiotics over ten months, yielding seven distinct treatment phases (pre-treatment, first antibiotic course, week after stopping treatment 1, interim, second course of antibiotics, week after stopping treatment 2, and post-treatment phase). The goal of the study was to assess the antibiotics' effects on microbial communities. Already investigated in the literature [42, 64], we propose to tackle its analysis from a new network perspective. Our analysis provides complementary information to the previous by allowing the analysis of higher-order interactions between bacteria: can we characterize prevalent communities of bacteria for each treatment phase? How do these communities react to the different drugs? The study of co-occurrence

networks in microbiome samples is becoming increasingly popular [5, 67, 102, 117, 150]. A critical step in the analysis is the transformation of the raw bacterial counts into a graph capturing such interactions. For each subject at a given treatment phase, we define a graph in which each node corresponds to a specific bacteria species, and edges $\mathcal{E} = \{(i, j)\}$ capture pairwise "affinities" between bacteria i and j. Intuitively, these affinities capture symbiotic mechanisms between bacteria: do they thrive simultaneously together – or, on the contrary – does one bacteria tend to smother the development of another? While a plethora of methods have been suggested for inferring networks from the abundance matrix (see Appendix A.1 for more references), and because of the particular nature of the data (zero-inflated negative binomial counts), we use a thresholded-Kendall correlation as a measure of the tendency of two bacteria to thrive (or wither) together. Indeed, the Kendall correlation is a ranked-based correlation and is as such less biased by the over-representation of zeros in the data. This produces a set of twenty-one different graphs—one for each of the three subjects during each of the treatment phases— on the 2,582 nodes representing the different species. More details regarding the explicit construction of these graphs are given in Appendix A.1.

Resting-State fMRI data. We will also use different distances to compare brain connectomes, which have radically different properties to the microbiome. The dataset that we study here consists in the resting-states fMRI data of 29 cocaine-dependent patients, and was published as part as a study of the effect of cocaine addiction on functional and structural connectivity[90][1]. In their 2011 article, the authors of the study show the existence of a statistically-significant reduction in the interhemispheric connectivity between cocaine-users and controls, highlighting the existence of an effect of substance-abuse on functional connectivity. In a similar spirit, we apply different network distances to this dataset in order to assess if the number of years under cocaine-dependence correlates with differences between the different connectomes. Each patient's raw fMRI data has been preprocessed through the FSL standard pipeline (more details can be found in Appendix A.2). This procedure yields a total of 116 nodes –each corresponding to a region of interest (ROI) in the brain — with values over 140 time points. To create a (weighted) graph from these filtered time series, we use thresholded Pearson correlation coefficients between nodes. Our approach is akin to [138] who select a threshold by controlling the number of edges in the graphs: the threshold used here is the mean (across subject) of each correlation matrix's 97^{th} quantile. In average, the graphs that we recover are thus sparse, with only 3% edges compared to the fully connected graph.

Given these sets of graphs, the crux of the analysis lies in the choice of a distance capable of identifying similar "graph-states", tailored to the data at hand. In the microbiome example, while the thousands of taxa that constitute the human microbiome allow for a rich variety of potentially different microbial communities, these communities usually involve a very small proportion of the

[1] Data publicly available at the following link: `http://fcon_1000.projects.nitrc.org/indi/ACPI/html/acpi_nyu_1.html`

taxa, and the associated abundance matrices are typically very sparse (here, 17% of the observations are non-zero.) In the brain connectome setting, fMRI data is also usually characterized by a high amount of noise: preprocessing steps, such as the template alignment and Gaussian smoothing performed to realign the brains and account for small head movements during the data retrieval, are known to blur even further the signal-to-noise ratio.

2.1.2 Problem statement and notation

Problem statement. From now on, we assume that the raw data has been transformed into a set of graphs, which we treat as input. In this perspective, the definition of a distance between aligned graphs –that is, graphs defined on the same set of identified nodes– takes on a significant importance. Indeed, distances constitute a stepping stone to any statistical analysis: pairwise comparisons, cluster or variance analysis, or even the definition of a "median", all follow naturally once a distance has been selected –thus making the selection of an appropriate distance a crucial step in the analysis of graph data. While the problem of assessing the distance between two unlabeled graphs is somehow classical and has been well studied[22, 99, 128, 156], our focus is different. Indeed, since in our examples, the nodes have been endowed with a particular identity, permutation-invariant distances would discard potentially relevant information. One might even wish to leverage the information contained in the nodes' labeling to define a distance sensitive to the intensity of changes at key-node locations. While the literature provides us with a number of "off-the-shelf" graph distances – any of which being, in principle, suited for the task–, these distances exhibit in fact distinct properties and capture different types of structural changes. In particular, a decisive parameter lies in the identification of the correct scale for comparing graphs: at which scale is the information contained? Is the analysis interested in capturing local structural changes at an atomic level, or should it focus more broadly on identifying structural similarities at the level of node communities? This notion of scale is at the heart of this review, which we use to classify and investigate the properties of some of the many different distances and similarities between graphs. By organizing each section along the presentation of distances belonging to one of three scale categories (local, meso-scale and global), we hope to provide elements to help the analyst in the selection of an appropriate distance given his/her desiderata.

Starting with structural and spectral distances –respectively local and global distances, two "scale extremes" which constitute the main bulk of metrics proposed in the literature–, we highlight the dynamics and types of structural changes that these two main categories are best able to capture. An improved understanding of these metrics' properties suggests new similarities tailored to specific scenarios. We then present a collection of graph similarities best suited for analyses at the mesoscale. In particular, we introduce a new set of graph similarities based on spectral heat kernels, and argue

that these similarities are optimal in that they combine both local and global structural information. In each case, the performance of the different distances is assessed on both the microbiome and fMRI datasets, as well as on synthetic, controlled experiments. Finally, we extend our illustration by applying them to the analysis of "recipe" networks.

Notation. Throughout this review, we write $G = (\mathcal{V}, \mathcal{E})$ the graph with vertices $\mathcal{V}$ and edges $\mathcal{E}$. We denote as $N = |\mathcal{V}|$ the number of nodes, $|\mathcal{E}|$ the number of edges, and we write $i \sim j$ if nodes i and j are neighbors. Our framework considers undirected binary graphs, with no self-loops (which we extend to the study of weighted graphs in our applications). A refers to the adjacency matrix of the graph, and D to its degree matrix:

$$A_{ij} = \begin{cases} 1 \text{ if } i \sim j \\ 0 \text{ otherwise} \end{cases} \quad \text{and} \quad D = \text{Diag}(d_i)_{i=1\cdots N} \quad \text{s.t.} \quad d_i = \sum_{j=1}^{N} A_{ij}$$

As we are restricting ourselves to undirected graphs, the matrix A is symmetric: $A^T = A$. The Laplacian [12] of the graph is the matrix defined as: $L = D - A$. The Laplacian is symmetric, and we consistently write its (real-valued) eigenvalue decomposition as : $L = U\Lambda U^T$, where U is a unitary matrix, and $\Lambda = \text{Diag}(\lambda_i)$ is the diagonal matrix of the eigenvalues: $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{N-1}$.

2.2 Quantifying local changes via structural distances

Distances between graphs usually fall in one of two general categories, often considered as mutually exclusive: structural vs. spectral distances. The first one captures local changes – that is, changes in the graph structure around each node –, and is thus especially well suited when these local changes can induce radical changes from one graph to the next. By way of illustration, different bonds in molecules can induce radically different properties (toxicity, solubility, etc.), while in brain networks, these distances can capture changes in the behavior of each individual region of interest (ROI). At the other extreme, the second collection of distances assesses the smoothness of the evolution of the overall graph structure by tracking changes in the eigenvalues of the graph Laplacian or its adjacency matrix. As such, these distances are better suited for analyses focusing on overall graph properties – that is, how nodes in the graph are globally organized and interact, rather than each of the nodes' individual function. In brain networks for instance, spectral distances can be employed to assess global changes in connectivity: is the treatment group characterized by a significantly different overall connectivity between areas of the brain compared to the control? We begin our review by analyzing properties of these two most popular types of distances in the two following sections.

2.2.1 The Hamming distance

Definition. The Hamming distance – a special instance of the broader class of Graph-Edit distances – measures the number of edge deletions and insertions necessary to transform one graph into another. While it is widely spread and used in many graph analyses, we will see in this subsection through tests and multidimensional scaling plots that the Hamming distance is nonetheless a blunt tool, and highlight some of its shortcomings.
More formally, let G and $\tilde{G}$ be two graphs on N nodes, as well as A and $\tilde{A}$ their corresponding adjacency matrices, the (normalized) Hamming distance is defined as :

$$d_H(G, \tilde{G}) = \sum_{i,j} \frac{|A_{ij} - \tilde{A}_{ij}|}{N(N-1)} = \frac{1}{N(N-1)} ||A - \tilde{A}||_{1,1} \tag{2.2.1}$$

This defines a metric between graphs, since it is a scaled version of the $L_{1,1}$ norm between the adjacency matrices A and $\tilde{A}$. As such, Eq. 2.2.1 defines a distance bounded between 0 and 1 over all graphs of size N.

Application. Figure 2.1 illustrates the results of the analysis of the microbiome study using the Hamming distance on both the bacterial (top row), as well as the fMRI data (bottom row-right).
General analysis framework. We briefly outline here the framework that we use throughout this chapter to analyze these graphs using each of the dissimilarity measures that we present. For each dataset (microbiome and fMRI) and each distance, we store the pairwise dissimilarities between graphs in a n-by-n dimensional matrix H, where $H_{ij} = d(G_i, G_j)$ and n is the number of graphs ($n = 21$ in the microbiome example and $n = 29$ in the fMRI dataset). We use this matrix H to analyze a set of different factors. In particular, in the microbiome example, each distance is used for the analysis of: (a) the graphs' variability from time frame to time frame, illustrated by plots of distances between consecutive graphs (Figures 2.1B,2.1E) and (b) similarities across subjects and across treatment phases,visualized by both heatmaps or clustermaps of the pairwise distances H between graphs (Figures 2.1A,2.1E) and their low-dimensional projections (multidimensional scaling MDS). Figures 2.1B,2.1D show such two dimensional projections. In the fMRI example, the analysis focuses on the ability of the dissimilarities to capture a relationship between the connectivity graphs and the number of years that subjects spent under cocaine dependency
Results: microbiome data. In the antibiotic study (top row of Figure 2.1), the Hamming distance is computed between the graphs between bacteria communities throughout the treatment course. Similar dynamics across subjects appear, as highlighted by the closely matching shapes of the curves (Figure 2.1C) representing the evolution of the distances between consecutive graphs. The MDS projection (Figure 2.1B) on the first components highlights the existence of a "treatment

gradient": interim phases –located in the bottom right corner of the figure– are closer to the pre-treatment samples and far from the treatment phases (violet and black points at center-left of the figure),consistent with biological interpretation of the treatment effects. While the Hamming distance does not detect stronger similarities between samples belonging to the same individual (no darker blue blocks along the diagonal of the heat map in Figure 2.1A), it is able to identify similar dynamic regimes across subjects – as highlighted by the clustered MDS projections of points corresponding to the same treatment phase. In order to quantify this effect, we run a Friedman-Rafsky test. For a given value of k, we compute the k-nearest-neighbor "metagraph" (or graph of graphs) induced by the pairwise-dissimilarity matrix H.This provides a useful way of extracting information from H by representing it as a graph where each node (corresponding to the co-abundance graph for a patient at a given treatment phase) is itself a graph, and edges reflect the k-strongest similarities between graphs. Henceforward, we refer to this induced k-nearest-neighbor graph of graphs as the k-nn metagraph. Having constructed the k-nn metagraph, we compute the number of its edges which connect graphs of the same class (i.e, in the microbiome dataset, treatment stage or subject). We then permute the labels to generate 50,000 graphs with the same topology, but where the edges randomly connect nodes independently of their class. We compare the original value to this synthetic null permutation distribution to get the associated p-value. This assesses the compatibility of the distance on a given set of labels: if the distance clusters together graphs belonging to the same category, then the p-value should be significantly small. The p-values are reported alongside the plots in Figure 2.1(F), where we have conducted this experiment with $k = 1$ (the nearest-neighbor graph is thus simply the minimum spanning tree). This test fails, in this case, to detect any statistically significant association between the edges in the minimum spanning tree and either the treatment stages or the subject labels. We note that increasing the number of neighbors considered ($k = 2, 3..$) in the metagraph does not uncover any meaningful statistical associations either.

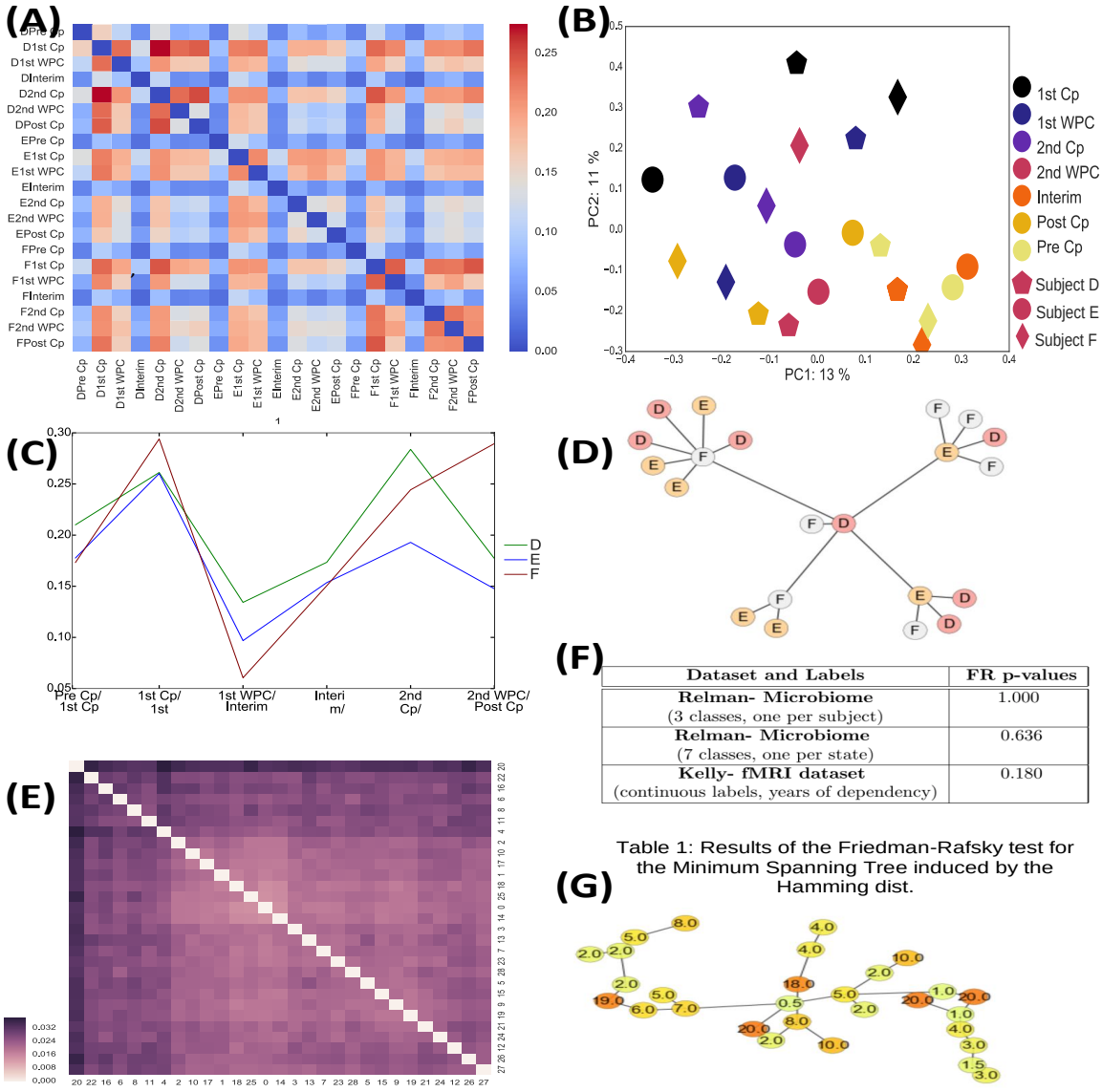
Results: fMRI data. When applied to the resting state-fMRI data, the Hamming distance does not detect any clusters of closely related graphs (as shown by the uniform cluster map in Figure 2.1E and the uniform tSNE projections in Figure 2.1D). We also adapt our previous Friedman-Rafksy test to handle continuous labels instead of discrete classes. This enables use to test the association of the k-nearest neighbor metagraph between patients and the amount of time that they have spent under cocaine dependency. The test statistic is now defined as the sum of the differences between labels (i.e., time under dependence) for all the edges in the graph. In this setting, a small score would indicate that brain networks are more similar to other networks with similar "time under dependence". As shown in both the figures and the plots, the Hamming distance does not detect any significant relationship between the relative distance of the graphs and their labels. We also run an analysis-of-variance type test: we split the graphs into two classes (with roughly the same number of subjects): patients with less than 5 years under cocaine dependency and patients with more than five years. We then compute the ratio $\Delta = \frac{\bar{D}_{12}}{\frac{n_1}{n_1+n_2}\bar{D}_{11}+\frac{n_2}{n_1+n_2}\bar{D}_{22}}$ where $\bar{D}_{ij}$ denotes the average distance

between subjects of class i and j: under the null, this ratio should be centered around 1. We assess the significance of this ratio via a permutation test, which here yields a p-value of 0.62: in this case again, the Hamming distance does not detect any significant difference between the graphs in the two classes.

Analysis. With a cost complexity of the order of $O(N^2)$, the Hamming distance provides a straightforward way of comparing sequences of aligned graphs that only takes into account the number of shared edges. It thus comes as no surprise that this distance has been a long-time favorite in various graph comparison problems. Graph embedding techniques – which provide a vector-valued representation for each graph that captures its geometric properties– are a case in point: in [109], the authors define similarities between subgraphs through their graph-edit distance. Similarly, in [54], the authors introduce the notion of a "median graph" as the minimizer of the sum of pairwise graph-edit distances.

While the Hamming distance is a perfectly valid first candidate graph distance for any type of analysis, it is worth emphasizing that it only reveals some restricted aspect of network similarities.

The first trait to highlight is its uniform treatment of all changes in the graph structure: all additions and deletions are assumed to have similar importance. Changes in the network's core are treated equivalently to changes in the periphery. We will analyze the consequences and limitations of this assumption in section 2.1.3. A second trait is Hamming's sensitivity to the density of the graphs. This yields a limited capacity to recognize similar dynamical processes across graphs with varying sparsity . As an example of the first point, let us consider a dynamic regime in which, at every time point, each edge is randomly flipped independently of the others: it either stays in the graph or disappears with probability p. The total number of disappearances follows a binomial distribution with mean $p|\mathcal{E}|$. For an identical perturbation mechanism, dense graphs are thus placed at higher distances to each other – and are thus considered as more unstable – than sparse graphs. The Hamming distance is unable to recognize that these graphs share in fact the same level of relative variability, which can hinder some aspects of the analysis. Indeed, the random deletion process at hand can be thought of as blurring noise applied to a true underlying graph structure, and is a typical representation our inability to observe all interactions between nodes in a complex system. In this case, it seems more natural to specify the inherent variability of the data in terms of "noise level" rather than "noise quantity", and our analysis should thus recognize similar noise levels independently of the graphs' original sparsity. Similarly, the Hamming distance tends to place nested graphs at a smaller distance to each other than other metrics. Indeed, suppose that graph $\tilde{G}$ comprises 50% of the edges of graph G. The Hamming distance between the two graphs is then simply $d(G, \tilde{G}) = \frac{0.5||A||}{N(N-1)}$, and does not correct for the size of the initial graph. We could nonetheless argue that this distance should be big (or at least close to 0.5), since the structure of the system is radically modified. Our microbiome study is a case in point (Fig. 2.1B): the variety of the microbiota involved in the interim



Dataset and Labels	FR p-values
Relman- Microbiome (3 classes, one per subject)	1.000
Relman- Microbiome (7 classes, one per state)	0.636
Kelly- fMRI dataset (continuous labels, years of dependency)	0.180

Table 1: Results of the Friedman-Rafsky test for the Minimum Spanning Tree induced by the Hamming dist.

Figure 2.1: Hamming distance between bacterial graphs (top rows), and brain graphs (bottom row). Heatmap of the Hamming distances between Kendall-correlation-based bacterial graphs **(A)** and MDS projections on the first two principal components. **(B)**.Colors denote treatment phases, and shapes represent different subjects . Plots of the consecutive distances between bacterial graphs **(C)**. Minimum Spanning tree between bacterial graphs induced by the Hamming distance **(G)**. Friedman-Rafsky test for significance for the different datasets **(F)**. Clustermap of the fMRI graphs **(E)**. Minimum spanning tree between brain connectomes induced by the Hamming distance **(G)**.

phase jumps to almost twice its corresponding value in any of the antibiotics phases (the number of bacteria increases from around 210 bacteria to 420). The distances between the interim phase and the other phases are subsequently smaller than for any of the other phases, with a number of shared taxa.

The Hamming distance is thus a measure of the *amount of change* between two graphs. While this might be adequate for characterizing the evolution of a given system through time, it is nonetheless unfit for finding similarities in broader settings. Tasks such as comparing graph dynamics in the presence of different degree densities or recognizing instances of the same network family (Erdös-Rényi random graphs, preferential attachement graphs, etc.) indubitably require other metrics.

2.2.2 The Jaccard distance

Definition. A potential solution to the aforementioned density-effect problem consists in using the Jaccard distance [105], which includes a normalization with respect to the volume of the union graph:

$$d_{\text{Jaccard}}(G,\tilde{G}) = \frac{|G\cup\tilde{G}|-|G\cap\tilde{G}|}{|G\cup\tilde{G}|} = \frac{\sum_{i,j}|A_{ij}-\tilde{A}_{ij}|}{\sum_{i,j}\max(A_{i,j},\tilde{A}_{ij})} = \frac{||A-\tilde{A}||_{1,1}}{||A+\tilde{A}||_*} \tag{2.2.2}$$

where $||\cdot||_*$ denotes the nuclear norm of a matrix.

Eq. 2.2.2 is known to define a proper distance between the graphs. A straightforward way to see this is to use the Steinhaus Transform, which states that for (X,d) a metric and c a fixed point, the transformation $\delta(x,y) = \frac{2d(x,y)}{d(x,c)+d(y,c)+d(x,y)}$ produces a metric. We apply here this transformation with d the Hamming distance and c the empty graph:

$$\begin{aligned}\delta(G,\tilde{G}) &= \frac{2||A-\tilde{A}||_{1,1}}{||A||_{1,1}+||\tilde{A}||_{1,1}+||A-\tilde{A}||_{1,1}} = \frac{2(|G\cup\tilde{G}|-|G\cap\tilde{G}|)}{2|G\cup\tilde{G}|} \quad (*)\\ &= d_{\text{Jaccard}}(G,\tilde{G}).\end{aligned}$$

In particular, taking for instance $G = G_t$ and $\tilde{G} = G_{t+1}$ the graphs associated to the state of a system at two consecutive time points t and $t+1$ (with $\mathcal{E}_{G_t}$ and $\mathcal{E}_{G_{t+1}}$ their respective set of

undirected edges) and rewriting the left hand side of (*), we have:

$$\begin{aligned} d_{\text{Jaccard}}(G_t, G_{t+1}) &= \frac{d_{\text{Hamming}}(G_t, G_{t+1})}{\frac{|\mathcal{E}_{G_t}|+|\mathcal{E}_{G_{t+1}}|}{2N(N-1)} + \frac{1}{2} d_{\text{Hamming}}(G_t, G_{t+1})} \\ \implies d_{\text{Jaccard}}(G_t, G_{t+1}) &= \frac{\frac{d_{\text{Hamming}}(G_t, G_{t+1})}{\bar{S}}}{1 + \frac{d_{\text{Hamming}}(G_t, G_{t+1})}{2\bar{S}}} \end{aligned} \tag{2.2.3}$$

with $\bar{S} = \frac{|\mathcal{E}_{G_t}|+|\mathcal{E}_{G_{t+1}}|}{2N(N-1)}$ is the average sparsity of the two graphs.

Application. Figure 2.2 shows the result of the analysis carried out using the Jaccard distance. Since the edges in our graphs have been assigned different weights according to the intensity of the interaction between bacteria, we have used the version of the Jaccard distance extended to the weighted graph setting, defined as:

$$d_{\text{Jaccard}}(G, \tilde{G}) = 1 - \frac{\sum_{i,j} \min(A_{ij}, \tilde{A}_{ij})}{\sum_{i,j} \max(A_{ij}, \tilde{A}_{ij})}.$$

This analysis yields somehow different results to the Hamming distance (Figures 2.2A,2.2D, 2.2C,2.2D). We note that the treatment phases express more variability and are far from most on the other samples. The Friedman-Rafsky test for the microbiome data highlighted a significant dependence of the 3-nn metagraph on the subject: with a p-value of 0.0002, this test shows that bacterial graphs corresponding to the same patient are significantly closer than under the random null model. This effect is further confirmed by running a analysis-of-variance type test and computing the statistics $\Delta = \frac{1}{3} \frac{\sum_{i \in \{D,E,F\}} \bar{D}_{i,i^c}}{\sum_{i \in \{D,E,F\}} \frac{n_i}{n_{tot}} \bar{D}_{i,i}}$ where $\bar{D}_{i,i^c}$ denotes the average distance between graphs in class i and graphs in any other class. Under the null, this statistic is centered at 1, and we evaluate its significance through a permutation test. This yields a p-value of 0.0018, highlighting the existence of a significant difference between graphs grouped according to their associated subject Id. This microbiome example is thus a case where the Jaccard distance is a better fit for our analysis: whereas the Hamming fails to uncover any real similarity between bacterial graphs corresponding to the same subject, the Jaccard distance does capture the existence of greater similarities among graphs belonging to the same "block" (i.e, patient), a known effect in microbiome studies. However, when applied to the brain networks, the Jaccard distance displayed an almost uniform distance between all samples and did not recover any significant clustering or grouping of patients (with a p-value associated to the analysis-of-variance test of 0.61).

Discussion. The Jaccard distance adjusts for graph density by including in its normalization the average sparsity of the two graphs. As such, it reflects the *amount of change with respect to the initial graph structure*. To highlight the benefits of this property, let us consider a dynamic regime in which the total number of edges stays fixed, but at each time point, each edge is replugged with

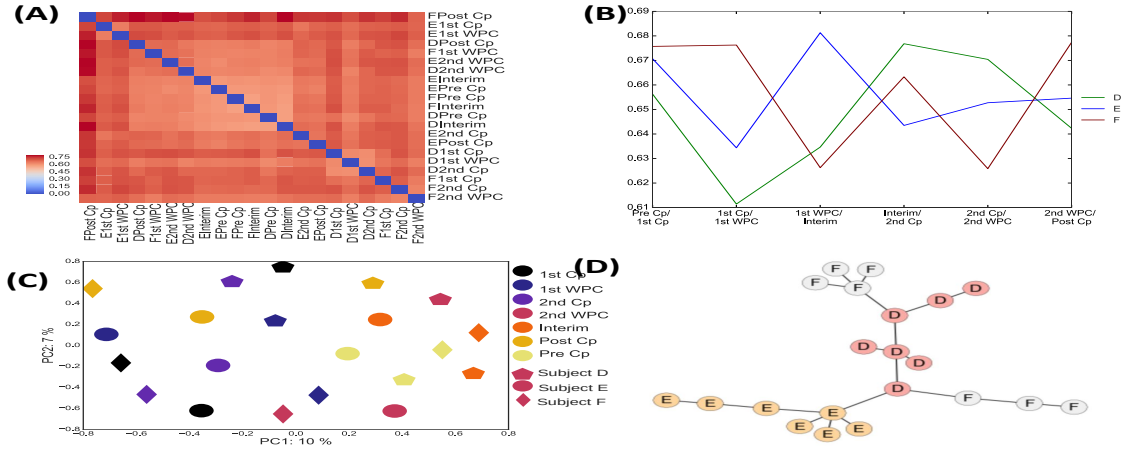


Figure 2.2: Application of the Jaccard distance to the microbiome study. Cluster-map of the Jaccard distances between Kendall-correlation-based bacterial graphs **(A)** . Plots of the consecutive distances between bacterial graphs **(B)**. MDS projection of the bacterial **(C)** graphs on the first two principal axes. Colors denote treatment phases, and shapes represent different subjects. Minimum spanning tree between bacterial graphs **(D)**

.

probability p in a previously vacant connection: the overall number of edges remains identical, but each flipped edge induces an increase in the Hamming distance of size $\frac{4}{N(N-1)}$. Hence, the average Hamming distance between G_t and G_{t+1} admits a closed-form expression of the type:

$$d_{\text{Hamming}}(G_t, G_{t+1}) = \frac{4p|\mathcal{E}|}{N(N-1)} = 4ps$$

where $s = |\mathcal{E}|/(N(N-1))$ is the sparsity of the original graph. By Eq. 2.2.3, the Jaccard distance can be written as: $d_{\text{Jaccard}}(G_t, G_{t+1}) = \frac{4p}{1+2p} = 2[1 - \frac{1}{1+2p}]$, where the later expression is a strictly increasing function of p. The Jaccard distance is thus independent of the sparsity and defines a one-to-one mapping between the rate of change p and the observed distance. In contrast, the effect of p is confounded in the Hamming distance by the influence of the sparsity.

This simple example shows that the Jaccard distance is better suited to comparing different dynamics, where the rate of edge rewiring is the main quantity of interest. Another of its advantages with respect to Hamming is that it provides a more interpretable notion of graph distances. Indeed, the Jaccard distance can be understood as the proportion of edges that have been deleted or added with respect to the total number of edges appearing in either network: a Jaccard distance close to 1 indicates an entire remodeling of the graph structure between time t and $t+1$.In the microbiome study at hand, the Jaccard distance reveals more within-subject variability than Hamming distance, where the blue and red blocks in Figure 2.1(A) highlighted contrasted dissimilarities between graphs: here, while there exists a strong subject effect, on the whole, the almost-uniform clustermap in Figure

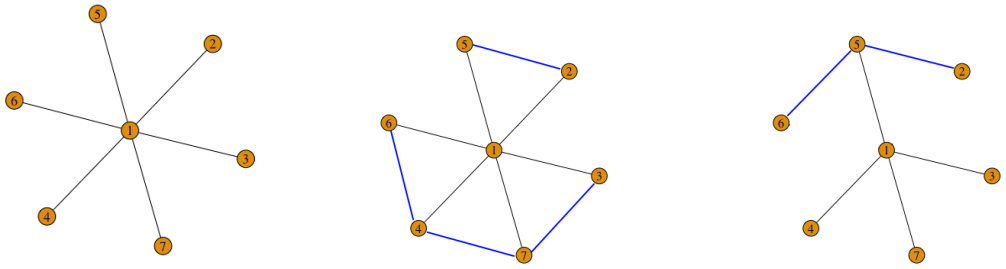
2.2(A) shows that samples within subject are still highly variable.

2.2.3 Shortcomings of local approaches

While the Hamming and Jaccard distances provide straightforward ways of analyzing a graph's dynamics or evolution over time, such measures appear too short-sighted. Indeed, these metrics focus on the direct neighborhood of each node, and fail to capture the "bigger picture" and information on the evolution of the graph as a whole. Figure 2.3 shows an example where a network G_0 undergoes two different dynamic processes, yielding distinct graph structures with similar Hamming distances to the original. In this setup, it is possible to argue that G_1 and G_2 are more similar to each other, since the maximal path length between any two nodes is 2, whereas information percolates less rapidly across the network in the third. Conversely, from another perspective, we could also argue that we should have $d(G_1, G_3) \leq d(G_1, G_2)$, since the two first share a higher number of nodes with identical degree or since they have the same number of spanning trees. This example is meant to show that distances can be adapted to capture specific aspects of a network's properties. In fact, in [100], Koutra and co-authors propose to define a "good" similarity score between graphs as a score satisfying the following set of four characteristics:

1. **Edge-Importance:** modifications of the graph structure yielding disconnected components should be penalized more.
2. **Edge-Submodularity:** a specific change is more important in a graph with a few edges than in a denser graph on the same nodes.
3. **Weight Awareness:** the impact on the similarity measure increases with the weight of the modified edge.
4. **Focus awareness:** random changes in graphs are less important than targeted changes of the same extent.

These serve as guidelines and can be modified and enriched by the data analyst depending on the application at hand. The Jaccard and Hamming distance treat all edges uniformly, irrespective of their status (thus violating criteria 1 and 2 for instance).



(a) Initial Graph G_1

(b) Perturbation of G_1: four edge insertions.

(c) Perturbation of G_1: 2 edge insertions, 2 deletions.

Figure 2.3: Two modifications of the same initial graphs (displayed in figure 2.3a, such that the Hamming distance with the original is $d_H(G_1, G_2) = d_H(G_1, G_3) = \frac{4}{21}$, and the Jaccard distances are $d_J(G_1, G_2) = \frac{2}{5}$ and $d_J(G_1, G_3) = \frac{1}{4}$. The average shortest path length are 1.71 for the initial graph G_1, 1.51 for G_2 and 2 for G_3

2.3 Comparing graph structures: a spectral approach

We now turn to the class of spectral distances. Spectral distances are global measures defined using the eigenvalues of either the adjacency matrix A or of some version of the Laplacian L. Consistent with the notation introduced in section 1, the (combinatorial) Laplacian of the graph is defined as: $L = D - A$, where D is the diagonal matrix such that D_{ii} is the degree of node i. Another popular choice consists in using the normalized Laplacian, defined as $\tilde{L} = I - D^{-1/2}AD^{-1/2}$.

As mentioned in the first section, spectral distances are more suited to analyses where the critical information is contained at the global scale, rather than at the atomic one. For instance, in brain connectome data, such distances are suited to analyses focusing on the overall connectivity patterns between regions of the brain: does cocaine usage decrease the connectivity between different regions? In microbiome studies, they can quantify changes in the overall microbial ecosystem: for instance, does the introduction of a given treatment radically change a microbiome ecosystem by destroying multiple co-existing communities and favoring the rise of a dominant bacteria?

But why should eigenvalues characterize the state of a graph better? Let us first provide some intuition for this spectral approach.The eigenvalues of a graph characterize its topological structure, and in particular the way that energy or information localized at a particular node can be propagated over the graph. As such, they are related to the stability of the complex system that the graph

represents. In quantum chemistry for instance, hydrocarbons are typically represented by graphs, whose adjacency matrix' eigenvalues correspond to energy levels of its electrons. In physics, the eigenvalues of the Laplacian represent the vibrational frequencies of the heat equation. The analysis of the spectral properties of a graph thus provide considerable insight into the dynamics of the system as a whole.

In this section, we brush an overview of various spectral distances that can be used for analyses pertaining global graph properties. Such distances have been well studied and developed in the literature [89, 85, 4]. In [88] for instance, the authors provide an interesting review of several of such spectral distances. This deviates slightly from our original setup: spectral distances are permutation invariant and do not take into account the fact that nodes have been endowed with a particular identity. In fact, such distances can be used to compare any set of graphs, provided that they all share the same number of nodes. Spectral distances are unable to distinguish between isospectral graphs and are in fact pseudo-distances rather than actual distances. However, as the probability of having distinct graphs with identical eigenspectra quickly dwindles as the number of nodes increases, spectral distances are also viable candidates for studying the dynamics of a given complex system through time. While most current algorithms can compute eigenvalue decompositions in $O(N^3)$ steps, some computational tricks bring down this cost to $O(N^2)$ [140]– thus making this spectral approach an appealing, computationally tractable alternative for defining graph similarities [3, 73].

2.3.1 ℓ_p distances on the eigenvalues

Definition. We begin by introducing a general class of versatile spectral distances. A first natural candidate for comparing two graphs based on their eigenvalue decomposition is to choose a representation of the graph (typically its adjacency matrix, combinatorial or normalized Laplacian,etc.) and to simply consider the ℓ_p distance between functions of their eigenspectra.

For any (almost everywhere) differentiable function of the graph's eigenvalues $\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{N-1}$, we can write:

$$d(G, \tilde{G})^p = \sum_{i=0}^{N-1} |f(\lambda_i + \epsilon_i) - f(\lambda_i)|^p \approx \sum_{i=0}^{N-1} |f'(\lambda_i)|^p |\epsilon_i|^p \tag{2.3.1}$$

An important step thus consists in picking an adequate representation of the graph: how to decide between using the graph's adjacency matrix, its Laplacian or normalized Laplacian? These representations and the relationship between their eigenvalues and properties of the graph (average degree, Cheeger constant, etc.) have been well investigated in the literature [35], yet no consensus as to which representation yields more accurate results for comparing graphs has been established. To try and resolve this issue, we suggest the following guidelines:

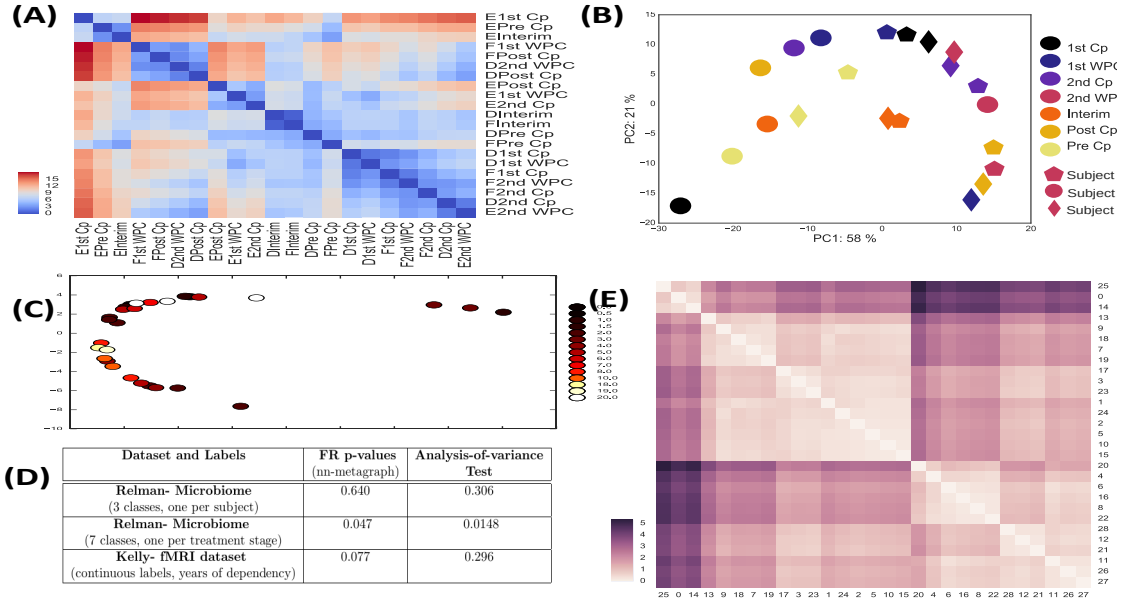
- *leveraging the representation's physical interpretation.* As underlined in the introductory paragraph of this section, both the eigenvalues of the Laplacian and those of the adjacency matrix can be related to physical properties of a system and can thus be considered as characteristics of its states. Whenever such a physical interpretation exists, a good choice thus lies in the selection of the corresponding representation.

- *opting for the most robust alternative.* The adjacency matrix does not down-weight any changes and treats all nodes equivalently. On the other hand, the eigenspectrum of the Laplacian accounts for the degree of the nodes and is known to be robust to most perturbations: a "small" perturbation of the graph –that is, a perturbation that has very little impact on the graph's overall connectivity– will only induce a small change in the eigenvalues [133], thus making them a more attractive alternative for comparing graph structures.

- *choosing a stable representation.* The literature remains divided on which version of the Laplacian to pick. However, the eigenvalues of the normalized Laplacian are bounded between 0 and 2, making it a somehow more stable and preferable representation.

Application. Let's look at the spectral distances on our microbiome and brain data. Figure 2.4 shows the results for the ℓ_2 distance using two different functions of the combinatorial Laplacian eigenspectrum in Eq. 2.3.1 on our datasets: the low-pass filters with randomly chosen parameters $f(\lambda) = e^{-0.1\lambda}$ (Microbiome, top row of Figure 2.4) and $f(\lambda) = e^{-1.2\lambda}$ (rs-fMRI, bottom row of Figure 2.4). The first interesting observation that we make is that these distances produce different results than the previous distances. In particular, we note that for $f(\lambda) = e^{-0.1\lambda}$, the nearest-neighbor metagraph is significantly associated with the treatment stages labels: the p-value associated to the analysis-of-variance test yields a value of 0.025, which is confirmed by the MDS projections (Figure 2.4 B) showing a clear grouping of the graphs per subject. This effect subsides as the scaling value increases while its association with the treatment stage becomes predominant. For $f(\lambda) = e^{-1.2\lambda}$, the analysis-of-variance test (with stages as labels) yields a p-value of 0.015. We also note that this distance is the only one which recovers some meaningful associations between the nn-metagraph of the fMRI data and the age under dependency (Fig.2.4D). We also note that the choice of the representation matters: the fMRI dataset analyzed through the scope of spectral distances based on the adjacency matrix failed to reveal any significant effect. Note that in the microbiome example and the synthetic experiments detailed in section 2.5, the choice of one representation over another was mitigated, with both the Laplacian and the adjacency matrix yielding comparable results. Moreover, as underlined above, the choice of the function itself can lead to the discovery of different effects.

To understand this phenomenon, we build upon the signal processing analogy developed in [129]. In this paper, Shuman and co-authors show that the eigenvalues of the Laplacian can be interpreted as the analog of a signal's frequencies in the temporal domain. In this case, each node's label (or feature representation) can be understood as the value of a signal propagating over the graph at



Dataset and Labels	FR p-values (nn-metagraph)	Analysis-of-variance Test
Relman- Microbiome (3 classes, one per subject)	0.640	0.306
Relman- Microbiome (7 classes, one per treatment stage)	0.047	0.0148
Kelly- fMRI dataset (continuous labels, years of dependency)	0.077	0.296

Figure 2.4: Application of ℓ_2 spectral distances using two functions of the Laplacian eigenspectra in Eq. 2.3.1: low-pass filters $f(\lambda) = \epsilon^{-0.1\lambda}$ on the Microbiome (**top row**) and $f(\lambda) = \epsilon^{-1.2\lambda}$ on the fMRI dataset(**bottom row**). Clustermap of the corresponding distances between bacterial graphs **(A)**/ brain connectomes **(E)** . MDS projection of the bacterial **(B)**/ fMRI **(C)** graphs on the first two principal axes. Colors denote treatment phases/ years of dependency. Pvalue of the FR-test for the 1-nn metagraph across the different datasets, for the low-pass filter $f(\lambda) = \epsilon^{-1.2\lambda}$ **(D)**.

that precise node location. Low eigenvalues and their corresponding eigenvectors are analogous to slowly-varying low-frequency signals over the graph: if two vertices are connected by an edge with a large weight, the values of the signal at those locations are likely to be similar. By contrast, the eigenvectors associated to high eigenvalues vary more rapidly across edges [130, 142]. Hence, "low" eigenvectors encapsulate local information about the structure of the graph (yielding results akin to the Jaccard distance in the microbiome example) while higher values of αs cover a larger portion of the spectrum and allow the incorporation of more global information.

Discussion. We continue upon the signal processing analogy to find an appropriate choice of the function f:

- If the goal of the analysis is to capture *the importance of the changes in the connectivity of the overall graph structure*, the distance should put more emphasis on the first eigenvectors. An adequate choice for f would be thus to select f to act as a low-pass filter: putting more weight on changes occurring in small eigenvalues, and discounting the effect of changes at the higher end of the spectrum. The strength of the modulation of the eigenvalues by the filter depends on the analysis. For instance, taking $f : x \to e^{-\alpha x/\lambda_3}$ ensures associating a weight of

at most $\frac{\alpha}{\lambda_3}\epsilon^{-\alpha}$ in Eq. 2.3.1 on changes in eigenvalues greater or equal to λ_3. In the case where $\lambda_2 << \lambda_3$, this gives more importance to changes occurring in λ_2. In our microbiome study, as previously highlighted, we recover more structure in the dataset by focusing on the lower part of the spectrum (Figure 2.4) and discarding the signal carried by the higher frequency, "noisier" eigenvalues.

- Supposing that one is interested in the *overall change in the "graph's frequencies" at every level* induced by the perturbation, one might actually prefer to take a function that would not discriminate against any value of the eigenfunction, but simply look at the amplitude of the change in eigenvalue. In that case, f could simply be taken to be the identity.

This section has shown the possibility of crafting a distance based on the Laplacian matrix eigenspectrum, tailored to the requirements and objectives of the analysis. However, choosing "an optimal" kernel function for the problem at hand requires domain knowledge or additional insight into the problem – thus requiring more thought than the straightforward Hamming distance.

2.3.2 Spanning tree similarities

Definition. Inspired by A. Kelmans [91, 92], who characterized transformations by their "ability to destroy", we now introduce a similarity which reflects the number of spanning trees that are destroyed or created by the transformation of one graph to another.

The Matrix-Tree theorem provides us with a convenient way of computing the number of spanning trees for a connected graph: denoting $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{N-1}$ the eigenvalues of the Graph Laplacian $L = D - A$, we have:

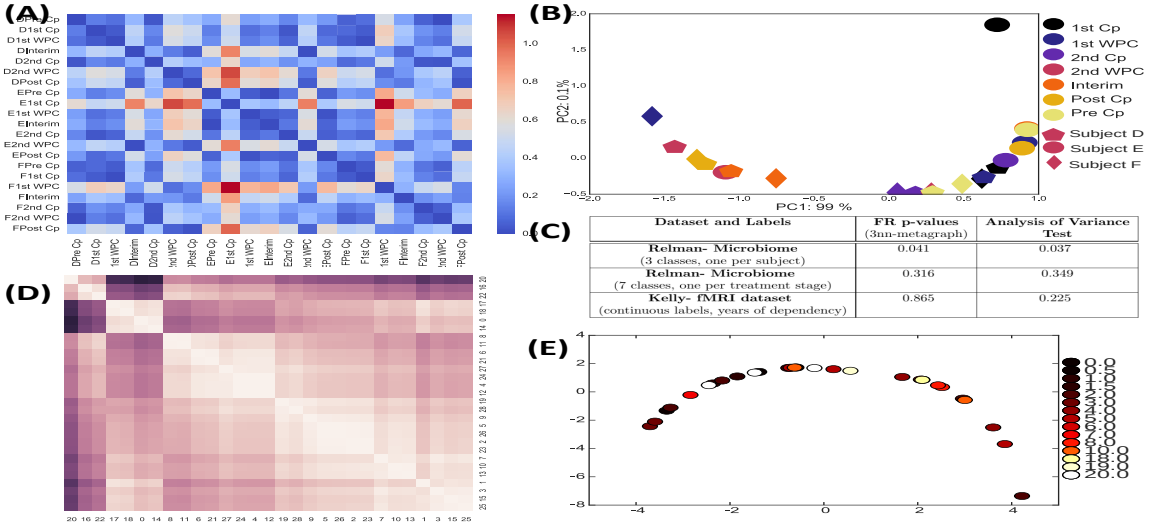
$$\mathcal{T}_G = N_{\text{Spanning tree of } G} = \frac{1}{N}\prod_{i=1}^{N-1}\lambda_i$$

A dissimilarity between two graphs G and $\tilde{G}$ can be defined by comparing the quantities:

$$d_{ST}(G, \tilde{G}) = |\log(\mathcal{T}_G) - \log(\mathcal{T}_{\tilde{G}})| \tag{2.3.2}$$

On an intuitive level, spanning trees are a reflection of the graph's interconnectedness and robustness to change: to draw an analogy with electric current, this amounts to quantifying the effect of one edge deletion on the impedance of the system: how easily does the current still manage to flow?

Application. In this case, the results obtained using the spanning tree dissimilarity (denoted as ST dissimilarity in the rest of the text) are comparable to the results provided by the low-pass filter



Dataset and Labels	FR p-values (3nn-metagraph)	Analysis of Variance Test
Relman- Microbiome (3 classes, one per subject)	0.041	0.037
Relman- Microbiome (7 classes, one per treatment stage)	0.316	0.349
Kelly- fMRI dataset (continuous labels, years of dependency)	0.865	0.225

Figure 2.5: Application of the Spanning tree dissimilarity. **(Top) Microbiome Data/(Bottom) fMRI Data.** Heatmap of the corresponding dissimilarity between Kendall-correlation-based bacterial graphs **(A)**. MDS projection of the bacterial graphs on the first two principal axes **(B)**. Colors denote treatment phases, and shapes represent different subjects. p-values associated to Friedman-Rafsky test of the consistency of the 3-nn metagraph with the labeling of the nodes and analysis-of-variance test **(C)** Clustermap for the fMRI data **(D)**. MDS projections of the brain connectomes on the first two principal components **(E)**.

spectral distance described in the previous subsection. We also observe an interesting phenomenon: the nearest-neighbor metagraph induced by the ST dissimilarity on the microbiome data is fairly consistent with the treatment stages (with an associated FR-pvalue of 0.158), but, as we increase the number of neighbors, this effect becomes rapidly insignificant. However, these later k-nearest neighbor graphs are significantly associated to the subject labels (below the 5% threshold). This effect is confirmed by the analysis-of-variance test described in section 2.2.2, which yields a significant p-value of 0.035.This indicates that the ST dissimilarity does capture both similarities between treatment phases as well as across subjects. This effect can be visualized in Figure 2.5B: the MDS projections of the microbiome graphs along the first 2 principal components follow a curve (which is indicative of a gradient in higher dimensions), along which points belonging to the same subject seem relatively close. Similarly as before, the Spanning tree distance recovers some structure in the fMRI datasets (lighter blocks along the diagonal in Figure 2.5D), although there is no evidence that these clusters are associated with the time under dependency.

Discussion. Suppose that graph G undergoes a “”small” perturbation, yielding a new graph $\tilde{G} = \mathcal{T}(G)$. We know that the eigenvalues of $\tilde{G}$ can be written as a perturbed version of the eigenvalues of G, that is:

$$\forall i, \quad \tilde{\lambda}_i = \lambda_i + \epsilon_i$$

Hence, we can write:

$$\tilde{\mathcal{T}}_G = N_{\text{Spanning tree of } \tilde{G}} = \frac{1}{N}\prod_{i=1}^{N-1}\tilde{\lambda}_i = \mathcal{T}_G \times [1 + \sum_{i=1}^{N-1}\frac{\epsilon_i}{\lambda_i} + \sum_{i,j=1}^{N-1}\frac{\epsilon_i\epsilon_j}{\lambda_i\lambda_j} + \cdots] \tag{2.3.3}$$

Combining 2.3.3 and 2.3.2 yields:

$$d_{ST}(G, \tilde{G}) = |\log(1 + \sum_{i=1}^{N-1}\frac{\epsilon_i}{\lambda_i} + \sum_{i,j=1}^{N-1}\frac{\epsilon_i\epsilon_j}{\lambda_i\lambda_j} + \cdots)| \tag{2.3.4}$$

The impact of the change is thus inversely proportional to the value of the eigenvalues. This is an attractive property for weakly connected graphs (i.e, that have small λ_1), where the addition or deletion of a critical edge can have a huge impact on the graph's overall connectivity. Conversely, changes in larger eigenvalues have less impact: to continue with the temporal frequency analogy drawn in the previous section, this similarity automatically discounts changes that are related to noise, and accentuates the impact of changes on low eigenvalues which are considered to be more reflective of the graph's structure. We emphasize that, once again, this defines a pseudo-distance between eigenspectra (or dissimilarity score between graphs), rather than a distance on the graphs themselves. This effect is shown in Figure 2.5, which exhibits results close to the low-pass filter approach developed in the previous subsection. The advantage of the ST dissimilarity is that it does not require the specification of a particular ad-hoc low-pass kernel on the eigenspectrum. However, this does come at an increased price in terms of the variability of the results: because the effect of perturbations is measured with respect to the inverse of the eigenvalues (Eq. 2.3.4), this distance is less stable than the low-pass filter spectral distance. We will study this in more depth in our synthetic experiments in section 2.5.

2.3.3 Distances based on the eigenspectrum distributions

General framework

Rather than focusing on the graph's eigenspectra, another alternative proposed by [88, 73] considers *continuous spectral distributions*. The continuous spectral distribution is obtained from each graph by computing the graph's eigenvalues and considering a kernelized version of its eigenvalue distribution. For a Gaussian kernel, the spectral distribution is defined as:

$$\rho_G(x) = \frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\lambda_i)^2}{2\sigma^2}}$$

A pseudo-distance between graphs is based on the distance between spectrum distributions, which, in the case of the ℓ_1-distance used in [73], yields the following expression:

$$d(G, \tilde{G}) = \int |\rho_G(x) - \rho_{G'}(x)| dx.$$

In their 2016 article [73], Gu and co-authors show that, in the limit of an infinite number of nodes, these distances have the added benefit of distinguishing between different types of graphs (Erdös-Rényi vs Preferential Attachment, etc). As such, these distances are able to recognize important geometrical information in the overall graph structure. We now investigate a variant of such a class of distances. Proposed in [82], it has been shown to exhibit interesting properties [87]: the Ipsen-Mikhailov distance.

Definition of the IM distance

Definition. First introduced by Ipsen[82] for graph reconstruction and later extended to the broader "graph-comparison" problem by Jurman and co-authors [86, 87, 88], the Ipsen-Mikhailov distance is a spectral measure which relates a network on N nodes to a system with N molecules connected by elastic strings. The connections are dictated by the graph's adjacency matrix A and the system can thus be described by a set of N equilibrium equations:

$$\frac{\partial^2 x_i}{\partial t^2} + \sum_{j \neq i} A_{ij}(x_i - x_j) = 0$$

In this setting, the eigenvalues of the Laplacian matrix of the network are interpreted as the squares of the vibrational frequencies ω_i of the system: $\lambda_i = \omega_i^2$ with $\lambda_0 = \omega_0 = 0$.

The Ipsen-Mikhailov distance characterizes the difference between two graphs by comparing *their spectral densities*, rather than the raw eigenvalues themselves. The spectral density of a graph is defined as the sum of Lorenz distributions:

$$\rho(\omega, \gamma) = K \sum_{i=1}^{N-1} \frac{\gamma}{\gamma^2 + (\omega - \omega_i)^2}$$

where γ is a parameter common to all vibrational frequencies that we will have to determine, and K is the normalization constant defined such that: $\int_0^\infty \rho(\omega, \gamma) d\omega = 1$. This spectral distance between

two graphs A and B is defined as:

$$\epsilon_\gamma(A,B) = \sqrt{\int_0^\infty [\rho_A(\omega,\gamma) - \rho_B(\omega,\gamma)]^2 d\omega} \tag{2.3.5}$$

The latter expression depends on the choice of the scale parameter γ. Jurman and co-authors [88] set $\gamma = \bar{\gamma}$ as the unique solution of:

$$\epsilon_{\bar{\gamma}}(\mathcal{E}_N, \mathcal{F}_N) = 1$$

So the IM distance is bounded between 0 and 1 and its upper bound is attained only for $\{A,B\} = \{\mathcal{E}_N, \mathcal{F}_N\}$ where $\mathcal{E}_N$ denotes the empty graph and $\mathcal{F}_N$ the complete graph on N nodes. In Appendix A.5, we investigate a closed form formula for these parameters.

The Hamming-Ipsen-Mikhailov distance

Definition. So far, none of these spectral distances have used the fact that particular nodes can be matched. There is no way of discriminating changes (that is, emphasizing changes in areas of the graph deemed important to the analyst), or of accounting for rare - but existing- isospectral graphs. To bridge the two approaches, Jurman and al [87] propose a distance that is a weighted linear combination of the Ipsen-Mikhailov and the normalized Hamming.

$$d_{HIM}^{\xi} = \frac{1}{\sqrt{1+\xi}} \sqrt{IM^2 + \xi H^2}$$

Application. The results of the microbiome analysis carried out with this distance are displayed in Figure 8. We note that the improvement with respect to the Ipsen-Mikahilov distance is only marginal.

Discussion. This distance benefits from the advantages of both the Hamming and the Ipsen-Mikhailov distances by combining local and global information. Note that, since it is a linear combination of a distance with a non-negative quantity, this defines a proper distance between graphs. The parameter ξ provides additional flexibility to the metric by allowing to favor one type of information over another. However, empirically, we have observed this distance to be computationally expensive, and thus difficult to apply to the study of large graphs and/or large datasets.

Application. Figure 2.6 shows the results of the analysis using the HIM distance on our microbiome study. The MDS projection (Figure 2.6A) seem to highlight a similarity between graphs corresponding to the same treatment. The Friedman-Rafsky test on the minimum spanning tree with

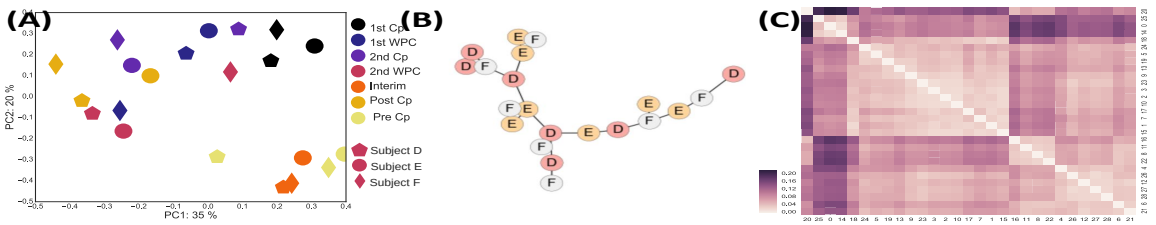


Figure 2.6: Application of the Hamming-Ipsen-Mikhailov distance. MDS projection of the bacterial graphs on the first two principal axes**(A)** . Colors denote treatment phases, and shapes represent different subjects. Minimum Spanning Tree induced on the bacterial graphs by the HIM distance **(B)**. Application of the HIM distance to the fMRI data set: clustermap of the different distances between connectomes **(C)**.

the treatment phases as labels is significant, with a p-value of 0.00048. This is further confirmed by the analysis of variance test described in section 2.2.2 with the stages as labels, yielding a pvalue of $p = 10^{-5}$. As also shown by Figure 2.6C, the HIM is able to make the best of both the Hamming and spectral distances, and is thus able to spot more structure in the datasets. Overall, because these spectral distances are "unlocalized" and make no use of the nodes' identities, they are suited to the comparison of graphs' overall structure without any prior on where "critical" changes occur in the spectrum. On an aside note, both the IM and HIM distances were the lengthiest to compute – perhaps restricting their scope of use to the comparison of small sets of reasonably-sized graphs.

2.3.4 The Polynomial Approach

Definition and motivation. The previous spectral distances all shared a common problem: they require an explicit computation of the graph's eigenvalues – which, computational tricks aside, still generally has complexity $O(N^3)$ – and is sensitive to global properties of the graph (as captured by the eigenvalues). Structural distances (Hamming and Jaccard) however, were too short-sighted and concentrated on changes in each node's direct neighborhood. Another interesting type of distances would thus be at an intermediate scale, and compare changes in local neighborhoods. For instance, changes in sparse regions of the graphs might be more informative than perturbations in very dense ones. Following Koutra and co-authors's [100] proposed guidelines for distance selection, a "good" similarity score should be able to capture such nuances and attribute more weight to changes in areas of the graphs deemed more critical by the data analyst.

In this new setup, a possible solution is to work directly with the powers of the graphs' adjacency matrix A^k. Indeed, the powers of the adjacency matrix relate directly to a graph's local topology

through the coefficients A^k_{ij}, which corresponds to the number of paths (possibly with cycles) that start at i and arrive at j in k hops. Hence, by design, these powers are inherently local. The coefficients A^k_{ij} can be thought of as a characterization of the connectivity between two nodes with respect to the k-hop neighborhoods: nodes i and j at distance greater than k hops have connectivity index $A^k_{ij} = 0$, whereas nodes within each other's k -hop neighborhood will typically have high connectivity index A^k_{ij} if the neighborhood is dense, and lower A^k_{ij} if the region is sparse. As such, the powers of the adjacency matrix seem to offer an attractive starting point to quantify changes on the mesoscale.

Typically, for each neighborhood (centered around a node a), perturbations should be assigned weights that are monotonically decreasing functions of the distance: a perturbation has higher impact in the local neighborhood if it is closer to the center than the periphery. In this spirit, denoting as $A = Q\Lambda_A Q^T$ the eigenvalue decomposition of the adjacency matrix A of a given graph, a proposed similarity score is defined with a polynomial $P(x) = x + \frac{1}{(N-1)^\alpha}x^2 + \cdots + \frac{1}{(N-1)^{\alpha(K-1)}}x^K$ of the adjacency:

$$P(A) = QWQ^T$$

where $W = \Lambda_A + \frac{1}{(N-1)^\alpha}\Lambda_A^2 + \cdots + \frac{1}{(N-1)^{\alpha(K-1)}}\Lambda_A^K$.

The distance between two graphs G_1 and G_2 can simply be computed by comparing the polynomials of their associated adjacency matrices A_1 and A_2:

$$d_{\text{pol1}}(G_1, G_2) = \frac{1}{N^2}||P(A_1) - P(A_2)||_{2,2} \tag{2.3.6}$$

In a way, this distance is a straightforward extension of the Hamming distance to the mesoscale: rather than looking at perturbations at the atomic level – counting the number of removed and inserted edges without assessing the effect of the perturbation on the overall structure, this polynomial distance compares neighborhoods of larger sizes and thus attempt to capture the effect of perturbation at an intermediate scale. The weighting factor α is a way of discounting "peripheral" changes in neighborhoods of larger sizes with respect to neighborhoods of smaller size. We note that Eq. 2.3.6 is just a proposed class of polynomial distances, but this set of distance can be more broadly customized to a specific problem at hand, including domain knowledge to choose the size of the neighborhood, etc.

Application. Figure 2.7 shows the application of the polynomial distance to the microbiome data. Similar to the Hamming distance, the polynomial distance does detect significant similar dynamics across subjects (closely matching curves in Figures 2.7C). However, in this case, the polynomial approach seems a weak compromise between structural and spectral distances, and does not benefit from any of their advantages: the polynomial distance is neither significantly associated to states or subjects (as per the associated Friedman-Rafsky and analysis-of-variance type tests).

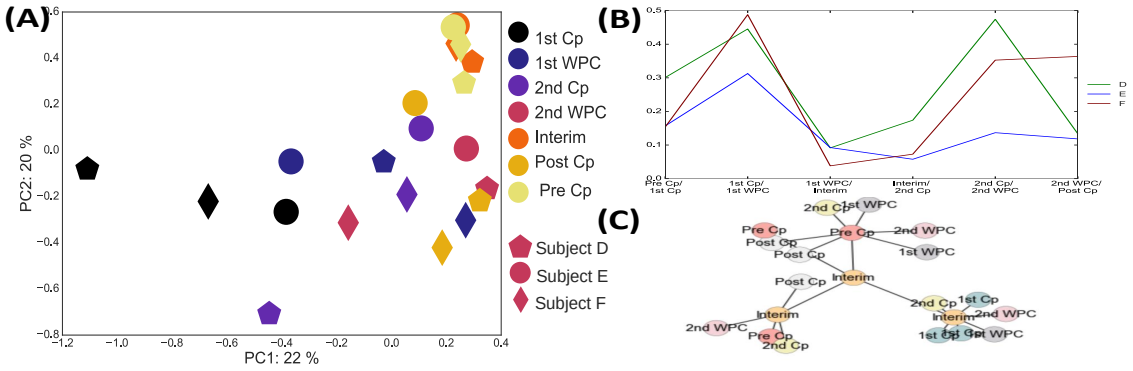


Figure 2.7: Application of the polynomial dissimilarity to the microbiome bacterial graphs, for $K = 3, \alpha = 0.9$. Heatmap of the corresponding dissimilarity **(A)** . MDS projection of the bacterial graphs **(B)** on the first two principal axes. Colors denote treatment phases, and shapes represent different subjects. Plots of the consecutive distances between bacterial graphs **(C)**.

Comparison of polynomial, spectral and structural distances. The main advantage of the polynomial distances over the Hamming and Jaccard distances is that the former takes into account the properties of each node – past each node's immediate neighborhood. Indeed, it is possible to show that the effect of a perturbation (that is, the addition of one edge) on the graph can be directly related to properties of the graph at a higher order than simply the one-hop neighborhood. By construction, polynomials of order k reflect the effect of the perturbation on k-hop neighborhoods. They can be expressed in terms of polynomials of the degree of the nodes of the added edge, as well as the size of the intersection of their neighborhoods (up to size k). Hence, while its form (powers of the Laplacian) make it an intrinsically local distance, the polynomial distance is a first step towards bridging purely structural and spectral distances, by extending the Hamming distance to neighborhoods of greater depth. However, this might come at an increased price: the real-application studies have shown this distance to be blurred by both the perturbation of the organizational structure of the microbiome from one phase to the next and the variability of the bacteria across subjects.

2.4 Quantifying change at the mesoscale

Most of the distances described in the previous sections can be considered as extremes: structural distances have proven to be too "local" and agnostic to perturbations' effects on a given complex system's organization as a whole. Spectral distances are global and fail to use the information captured in the nodes' identities. However, in several cases, the information of interest to the analyst

lies somewhere in between. In brain connectome data for instance, while the individual neuron's activities might be too noisy to yield any significant results, the neighborhoods (in this case, larger regions of the brain) might however shed some light on the effect of such or such drug on the nervous system. Similarly, in microbiome networks, analyses focusing on small modules of bacteria might alleviate some of the noise due to"reading errors". Both of these examples are case in point where an analysis of the graph at the neighborhood level (or "meso-scale") is desirable.

On the other hand, polynomial distances – which we had originally proposed as a scalable version of eigenvalue-based distances– quantify changes with respect to the k-hop neighborhoods, and have shown promising properties in both real and synthetic experiments: in the case where nodes' identities hold some insightful information, this extension of standard structural metrics seem to have brought a solution, trading off between the locality of the changes and their impact on the organization of the system as a whole. This indicates that considerable insight can be gained by comparing graphs at this intermediate "neighborhood" scale. This approach thus calls for the need for characterizing topological properties of these neighborhoods. In this section, we investigate graph comparison through a "glocal" lense (borrowing an expression from [87]), extending the class of mesoscale polynomial distances introduced in section 2.3.4 by suggesting two alternative characterizations of neighborhoods' topological properties.

2.4.1 Quantifying interactions: connectivity-based distances

We begin with a simple intuitive distance based on some measure of nodes' pairwise interactions. Indeed, as previously underlined, we want a distance that: (a) preserves information about each node's identity and (b) incorporates information characterizing nodes by their relationship to the whole graphs, rather than uniquely with respect to their direct neighbors. A general framework is to consider the set of graph dissimilarities defined as:

$$d_{\text{centrality}}(G_t, G_{t+1}) = \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} (s_{ij}^{(t+1)} - s_{ij}^{(t)})^p \Big)^{1/p} \tag{2.4.1}$$

where s_{ij}^t is some measure of the interaction or affinity between nodes i and j in graph G_t. This dissimilarity metric thus quantifies how much the different interactions have changed from one graph to the other. This approach satisfies our constraints: it is both local and respects nodes' identities while accounting for the whole graph structure by summing over all pairwise "interaction" scores.

In the simplest, most intuitive case, we can simplify this expression by using centrality measures. Indeed, centrality measures (betweenness, harmonic, etc.) can typically be used to characterize them as either belonging to part of the core or the periphery of the graph, and thus encode global topological information on the status of node within the graph. These metrics are thus natural

candidates to characterize 'mesoscopic' changes. More formally, in this setting, denoting $c_i^{(t)}$ as the betweenness-centrality of each node i in the graph at time t, one defines a distance between two graphs G_t and G_{t+1} as:

$$d_{\text{centrality}}(G_t, G_{t+1}) = \sqrt{\sum_{i=1}^{n} (c_i^{(t+1)} - c_i^{(t)})^2} \tag{2.4.2}$$

One of the positive aspects of this metric is that centrality measures are "integrated" quantities; measuring the number of paths that typically pass through a given node. As such, similarly to eigenvalues, these metrics are more robust to small perturbations in the graph structure than the Hamming distance. Moreover, each "drift" measure in Eq. 2.4.2 is interpretable: a change in centrality can be understood as a drift of the node away from (or towards) the core of the network. However, the problems associated with this approach are two-fold. First of all, one has to choose a single "good" centrality measure (harmonic, betweenness, etc.), which may require domain-knowledge, since this captures a specific aspect of the network's evolution. Moreover, the computation of betweenness centrality on unweighted graphs typically requires algorithms with complexity $O(|\mathcal{V}||\mathcal{E}|)$. This approach is thus unfortunately, like the IM distance, difficult to extend to larger graphs.

In order to make this approach more tractable, recent work has proposed using approximation algorithms to compute alternative interaction metrics in Eq.2.4.1. For instance, in [115] Papadimitriou and co-authors suggest five different scalable similarities. In [100], Koutrai and co-authors propose a low-dimensional approximation of these scores based on loopy belief propagation algorithm – yielding a method (DeltaCon), able to approximate Eq.2.4.1 with a computational complexity linear on the number of edges in the graphs.

2.4.2 Heat spectral wavelets

Another alternative is to derive characterizations of each node's topological properties through a signal processing approach: the values of the nodes constitute a signal over the graph, which can be filtered by modulating the graph's spectrum. This yields a different similarity than in the eigenvalue-based setting: whereas in the previous section, eigenvalue distances simply computed a distance between the modulation of two graphs' eigenvalues, here, the eigenvalues are modulated and combined with their respective eigenvectors to yield a "filtered" representation of the graph's signal. Such an approach could follow work initially done by Monning and co-authors, who, in their recent 2016 paper [111], build upon the DeltaCon similarity to create a (proper) distance between graphs: they introduce the *Resistance Perturbation* index, a metric based on the eigenvalues and eigenvectors of a modified version of the graph Laplacian. In this section, we focus on a closer analogy

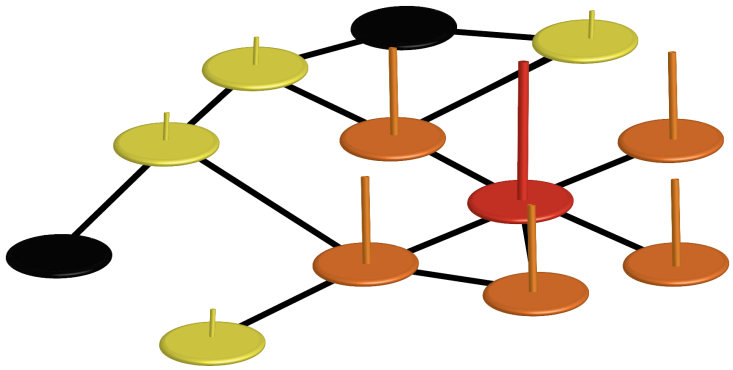


Figure 2.8: Representation of the heat kernel: the source node (red) has the highest temperature (interpreted here as a "signal value") – illustrated by the high red bar. The heat wave diffuses over the neighborhood (illustrated by the varying intensities of the nodes' colors and height of the bars representing the dwindling temperature/signal strengths.)

to signal processing and use recent work in the graph signal processing literature to derive such characterizations. In this subsection, we focus on an approach inspired by [49] for the purpose of structural role identification. In that paper, inspired by the emerging field of graph signal processing [129], the authors suggest using heat spectral wavelets to characterize each node's local topology for the purpose of structural role identification. To use a concrete analogy, this method operates in a way similar to sonar detection: each node probes the network by diffusing a heat wavelet, and the way that the network responds to each of these probes – that is, the different heat prints that are obtained for each node – is taken as a signature for each of the nodes' topological neighborhoods.

More formally, denoting $L = U\Lambda U^T$ as the Laplacian's eigenvalue decomposition, where $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{N-1}$, the heat scaling-wavelet [130] $\Psi^{(\tau)}_{\cdot,a}$ centered at node a with scale τ is defined as the column vector of the matrix $Ue^{-\tau\Lambda}U^T$:

$$\Psi^{(\tau)}_{\cdot,a} = Ue^{-\tau\Lambda}U^T\delta_a \implies \forall m, \quad \Psi^{(\tau)}_{m,a} = \delta_m^T Ue^{-\tau\Lambda}U^T\delta_a = \sum_{j=0}^{N-1} e^{-s\lambda_j}U_{aj}U_{mj}$$

where δ_m is the indicator vector associated to node m, and $e^{-\tau\Lambda}$ is the diagonal matrix $\text{Diag}(e^{-\tau\lambda_0}, e^{-\tau\lambda_1}, \cdots, e^{-\tau\lambda_{N-1}})$. In [49], the authors propose to define a structural signature for each node as the unordered set of coefficients:

$$\chi_a = \{\Psi^{(\tau)}_{m,a}\}. \tag{2.4.3}$$

By comparing these distributions, one captures information on the connectedness and centrality of each node within the network, thereby providing a way to encompass in a Euclidean vector all the necessary information to characterize nodes' topological status within the graph. Moreover, in order to compute these wavelets in a tractable fashion that extends to large graphs, Hammond and co-authors [75] suggest the use of Chebychev polynomial approximations.The cost of computing the

wavelet transforms becomes simply $O(K|\mathcal{E}|)$–making spectral wavelets an attractive approach for characterizing structural roles. While these signatures were initially devised to detect structural similarities across a network, they can also be employed to characterize similarities across a set of aligned graphs. In this setting, network similarity between graphs G_t and G_s is defined by comparing each node's topological signature in G_t with its counterpart in G_s. Here, a large dissimilarity between graphs indicates either an important "volume" of change (as in the Hamming distance) or that some nodes have undergone important topological changes. It thus captures changes at both the fine and intermediary scales.

More formally, this dissimilarity between graphs amounts to an ℓ_2 distance between each node's structural embedding:

$$\begin{aligned} d(G_t, G_{t+1}) &= \frac{1}{N}\sum_{a\in\mathcal{V}} ||r_a^{(t)} - r_a^{(t+1)}||_2 = \frac{1}{N}\sum_{a\in\mathcal{V}} \delta_a \Delta^T \Delta \delta_a \\ \implies d(G_t, G_{t+1}) &= \frac{1}{N}\mathrm{Tr}[\Delta^T \Delta] \end{aligned} \tag{2.4.4}$$

where $\Delta = U_t e^{-\tau\Lambda_t} U_t^T - \tilde{U_{t+1}} e^{-\tau\Lambda_{t+1}} U_{t+1}^T$.

To formalize the link with the previous subsection, we argue that these wavelet coefficients are in fact robust integrated centrality scores. Indeed heat kernels can be understood as a robust *page rank score* at each node [30]. By design, the heat kernel integrates over the neighborhoods (the size of which depends on the scale of the kernel) and is thus less sensitive to small perturbations. Hence, these wavelet coefficients provide a tractable alternative to the centrality measures proposed in section 2.4.1. As such, they benefit from these measures' interpretability, while being generalizable to larger networks.

Discussion. We summarize the advantages of this method as follows.

- **tractability**: as already noted, the cost of computing the wavelets via a polynomial approximation is linear in the number of edges, making it a suitable approach for large sparse graphs.

- **granularity**: since this metric compares each node's status in the two graphs, this approach benefits from granular information which allows the possibility of identifying the nodes that have undergone the most drastic changes. This is particularly useful in a number of applications where the identification of the area of the graph which changed the most is also of interest (what bacteria radically changed, which neurons adopted a completely different role in the graph, etc.).

- **inclusion of 'mesoscopic' information**: the wavelets allow us to compare neighborhoods of the node at different scales automatically. This enables a less short-sighted representation of the overall graph structure than standard structural distances.

- **inclusion of 'multiscale' information**: the topological signatures that we obtain for each node can be further enriched to contain multiscale information: in [49], a multiscale topological signature associated to scales $\{s_1, \cdots s_j\}$ is defined as the concatenation of the representations : $\chi_a = [\Psi_a^{s_1}, \Psi_a^{s_2}, \cdots \Psi_a^{s_j}]$. In this case, the heat-distance between two graphs is simply computed by replacing r_a in Equation 2.4.4 by this new value χ_a.This allows for a more robust representation of the topological role assumed by each node. An application of distances based on this multiscale signature on the recipes network is presented in section 2.6.

These observations also hold for the connectivity-based distances introduced in section 2.4.1.

2.4.3 Application to the microbiome and fMRI study

Figure 2.9 shows the results of the analysis of the microbiome study using a heat based distance with $\tau = 1.2$. Interestingly, this distance is one of the few showing a significant link between the years under dependency and the graphs in the fMRI datasets (Friedman-Rafsky for the 5-nn metagraph has a p-value below the 0.05 threshold, Figure 2.9B). However, in the microbiome dataset, this distance is dominated by a clear subject effect (Figures 2.9B,D and E). This is further confirmed by the analysis-of-variance test with the subjects as labels described in section 2.2.2 yields a p-value below 10^{-4}. We also note that this distance clearly indicates similar dynamics across subjects (Fig. 2.9C).

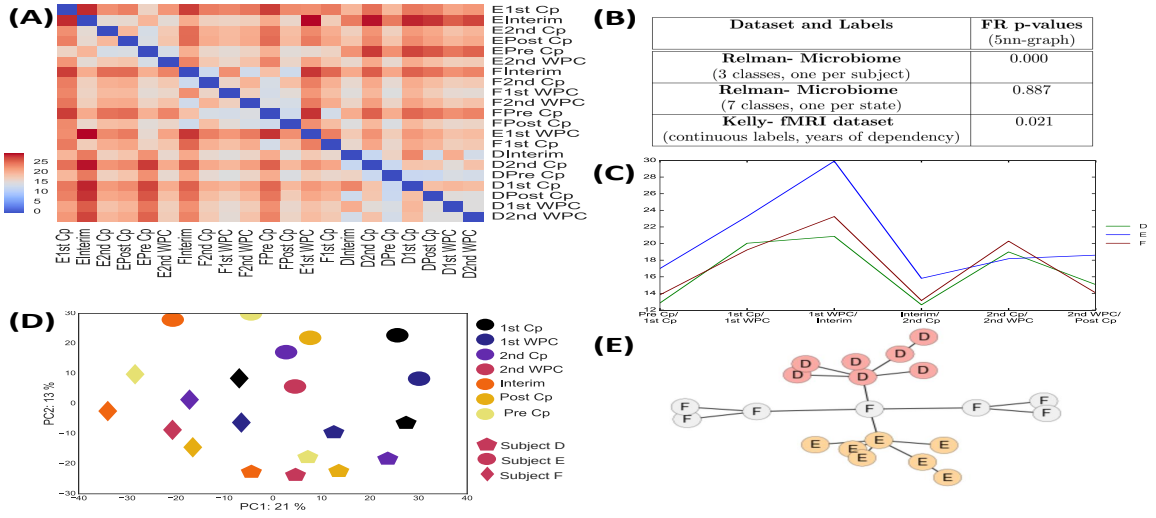


Figure 2.9: Application of the heat wavelet characteristic distance to the microbiome bacterial graphs, for $\tau = 1.2$. Heatmap of the corresponding dissimilarity **(A)** . P-values of the Freidman-Rafksy test on the 5-nearest-neighbor graphs induced by the heat distance, on each dataset. Plots of the consecutive distances between bacterial graphs **(C)**. MDS projection of the bacterial graphs **(D)** on the first two principal axes. Colors denote treatment phases, and shapes represent different subjects. Minimum Spanning tree induced on the bacterial graphs **(E)**.

2.5 Synthetic Experiments

We use several synthetic experiments on toy graphs to enhance our understanding of the different distances' behavior and relative advantages. Synthetic experiments [2] have the benefit of offering a controlled environment for testing the different distances' sensitivity relative to:

- **the graph topology**: we tested the different distances on 5 types of graphs:
  - The Erdös-Rényi model (with $N = 81$ nodes, and a probability of connection $p = 0.1$)
  - a Preferential Attachment (PA) graph on $N = 81$ nodes (with $\alpha = 1.$)
  - a Stochastic Block Model (SBM) graph, with 3 equally sampled communities and connection matrix: $C = \begin{pmatrix} 0.4 & 0.1 & .001 \\ 0.1 & 0.2 & 0.01 \\ 0.001 & 0.01 & 0.5 \end{pmatrix}$

  These sets of network families present different global and local degree densities, and will allow us to assess the impact of the topology on the analysis of network dynamics. The Erdös-Rényi graphs are denser than the preferential attachment graphs, which have an almost star-like structure with a few hubs. The SBM model is somewhere between the two: there are three relatively dense cliques with only a few edges connecting them.

- **the perturbation mechanism**: in our first set of experiments, three initial graphs are generated according to a given topology– ER, preferential attachment, and SBM. We simulate network dynamics as follows: at each time step, for each graph, η % of the edges are removed and re-connected elsewhere (at random, following a preferential attachment model). In an attempt to replicate real life situations, we add to this procedure a "background" depletion/thickening process: edges are deleted with probability 0.015 and "formerly absent" edges are added with probability 0.015. How do the distances behave in this simple setting? We expect the curves for the distances depending on topological properties to be stable for denser graphs. In this case, modifications rarely impact the structure of the graph, but structural distances are large, because many edges are being moved. On the other hand, we might observe more instability in the plots for sparser graph structures, where the deletion of a critical edge can have a much stronger impact on the overall connectivity of the graph.

- **changes in the intensity of the perturbation mechanism**: in our second set of experiments, we want to assess metrics' sensitivity to changes in the dynamical process. At time T=0,

[2] The code for all synthetic and real experiments developed in this review is public and available at: `https://github.com/donnate/TrackingNetworksChanges`

we generate on initial graph according to one of our three proposed topologies and simulate a dynamical mechanism as before in which, at each time step, 8.5% of edges are randomly re-wired, and edges are randomly deleted or added with probability 0.015. At $T = 6$, the perturbation mechanism increases its rewiring probability to 34% , and its random deletion/ addition probabilities to 6%. At T=13, the process reverts back to its original characteristics. This induces three distinct time blocks in the time series. The aim here is to see which distances show the existence of a change-point in the graphs' dynamics.

In order to analyze the results, we:

1. quantify different distances' ability to cluster graphs belonging to the same time series: in the first set of experiments, for a low level of noise η, the distance should recognize graphs belonging to the same time series. To quantify this effect, we use agglomerative clustering on the distance matrix to recover 3 different clusters. We then compute the homogeneity and completeness of these clusters.

2. assess the consistency of the ordering of the graph induced by each distance with the time series: each graph at time t should be closer to its "parent" graph at time $t-1$ and "child" graph at time $t+1$.

3. estimate the ability of each distance to spot changes in the dynamic regime. To this effect, we visualize the heatmaps of the different distances, and compute the ratios $r_1 = \frac{\bar{D}_{12}}{\sqrt{D_{11}D_{22}}}$ and $r_2 = \frac{\bar{D}_{32}}{\sqrt{D_{33}D_{22}}}$, where $\bar{D}_{ij}$ denotes the average "between" (or "within", if $i = j$) distance between graphs in time chunk i and graphs in time chunk j.

The legend in each of the subsequent figures indicate the correspondence between curves and distances.

We note that in the SDM case, the eigenvalue-based distances, both the Laplacian and adjacency-based representation yield comparable results.

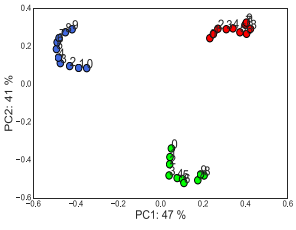
(I) Clustering Performance experiment (3 graphs, noise level σ=0.1)

A. Clustering by graph (3 labels)

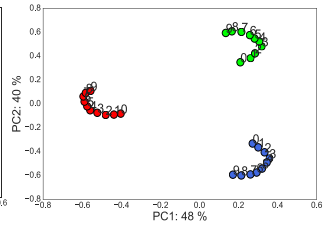
Dissimilarity	FR p-values (nn-graph)	Homogeneity	Completeness
Hamming	0.000	1.000	1.000
Jaccard	0.000	1.000	1.000
HIM	0.000	1.000	1.000
Spanning Tree	0.854	0.128	0.173
Polynomial $K=3, \alpha=0.5$	0.000	1.000	1.000
Heat, $\tau=0.1$	0.541	0.090	0.191
Eigenvalue-based $f(\lambda)=e^{-1.2\lambda}$	0.674	6.73e-02	7.62e-02

A.i. Clustering performance and FR test

A.ii. Homogeneity as the proportion of rewired edges increases



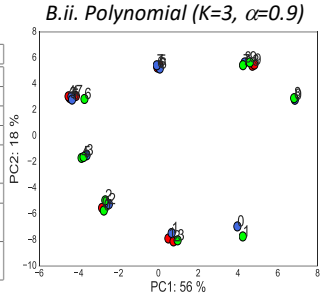
B.i. Hamming Distance



B.ii. Polynomial (K=3, α=0.9)

Dissimilarity	FR p-values
Hamming	0.000
Jaccard	0.000
HIM	0.000
Spanning Tree	0.000
Polynomial $K=3, \alpha=0.5$	0.000
Heat, $\tau=0.1$	0.011
Eigenvalue-based $f(\lambda)=e^{-3\lambda}$	0.267
Eigenvalue-based $f(\lambda)=1_{\lambda<2}\lambda$	1.10^{-4}

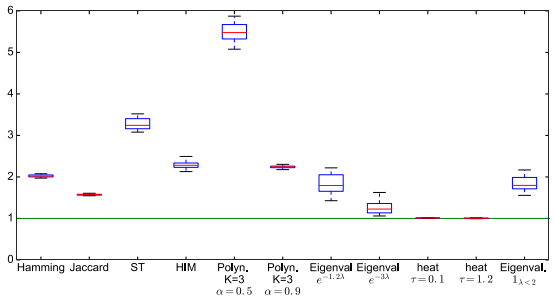
B.iv. FR-test of the consistency between the MST and time



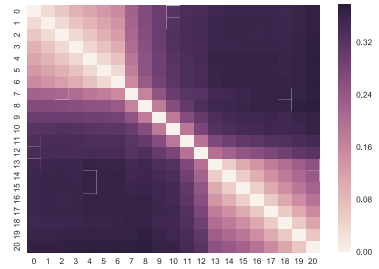
B.iii. Eigen. Distance $f(\lambda)=\lambda 1_{\lambda<2}$

B. MDS projection on the first two principal components

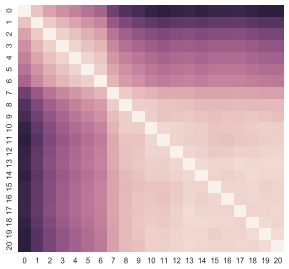
(II). Change-Point Detection experiment (T=7 and T=13)



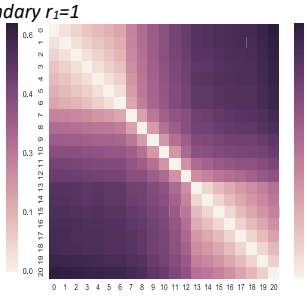
C.i. Ratio $r_1=D_{12}/(D_{11}D_{22})^{1/2}$. Green line indicates the boundary $r_1=1$



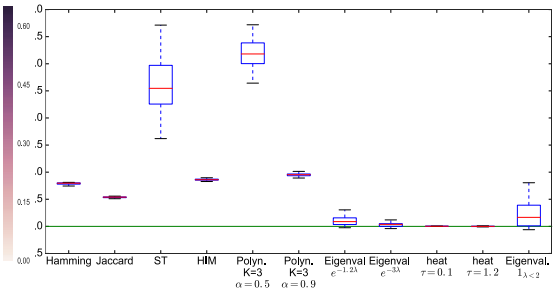
C.ii. Heatmap: Hamming distance



C.iii. Heat distance (τ=0.1)



C.iv. Polynomial distance (K=3, α=0.9)



C.v. Ratio $r_2=D_{23}/(D_{33}D_{22})^{1/2}$. Green line indicates the boundary $r_2=1$

Figure 2.10: Results for the Stochastic Block Model topology. **Top Row:** Comparison of the smooth dynamics (no change point), with 0.05% edges rewired at each time step. **Bottom Row:** Change point detection experiment.

We summarize our findings as follows:

- **Smooth dynamical process**: We observe that the performance of the different distances is not consistent across topologies. In particular, the dynamic regime associated to the Preferential Attachment graph seems to yield more instability, and none of the distances are able to cluster graphs belonging to the same time series correctly. Nor do they yield an ordering of the graphs consistent with their ordering in the time series. However, in the case of the ER and SBM graphs, all structural distances (Hamming, Jaccard, polynomial), as well as some of the eigenvalue based ones (IM) exhibit perfect recovery of the different clusters and the temporal ordering. This is probably due to the higher graph degree densities. Thus, the addition of an edge perturbs the overall structure less. In this case, it seems that the heat distance as well as the low pass filter emphasize similarities across graphs at a similar "state in time": the MDS projection shows a curve with closely clustered points representing graphs at identical times.
- **Change-point detection problem**: the structural distances exhibit better behavior (as shown by the clear blocks along the diagonal). We note that the ST dissimilarity and the polynomial distance have high r_1 and r_2 ratios, making them perfect candidates for detecting a change in regime.

2.6 Case study for spatial dynamics: worldwide recipe networks

In this final section we extend the scope of our analyses from temporal to spatial dynamics through the study of a concrete example: a worldwide recipe network.

In this example, each cuisine is modeled by a graph in which nodes represent ingredients, and edges measure their co-occurrence frequency in various recipes. The motivating intuition behind this graph-based co-occurrence is that cuisines can be better characterized by typical associations of ingredients. For instance, the Japanese cuisine might be characterized by a higher associativity of ingredients such as “rice” and “nori” than Greek cuisine. In our analysis, the graphs were obtained by processing the 57,691 recipes scraped from three different American culinary websites (*allrecipes*, *epicurious*, and *menupan.com*) in [1] as part as a study on food-pairing associations, and counting the co-occurrences of 1,530 different ingredients for 49 different cuisines (Chinese, American, French, etc.)[3]. Each cuisine is then characterized by its own co-occurrence network. The weight on the edge is the frequency of co-occurrence of the two ingredients in a given cuisine. The final graph for a given cuisine thus consists in a collection of disconnected nodes (ingredients that never appear in a single

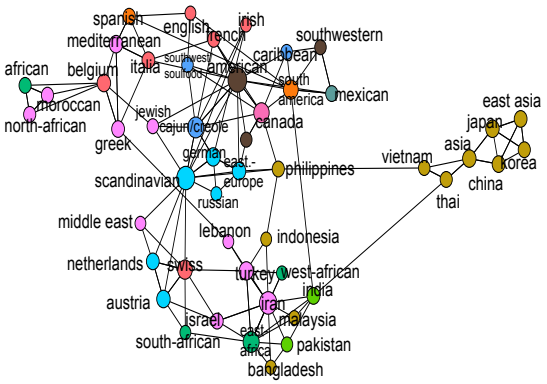
[3]The data can be downloaded at the following link `http://yongyeol.com/pub/`

recipe for that cuisine) and a weighted connected component. The construction of these graphs is further discussed in Appendix A.3.

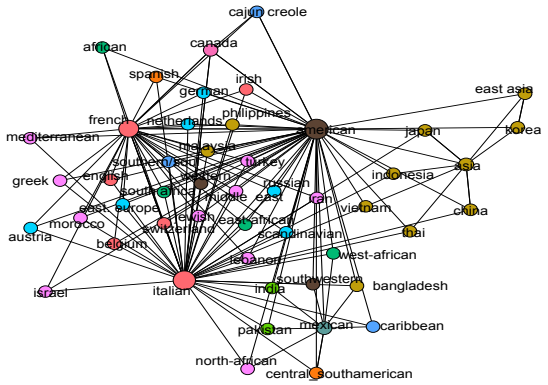
The goal of this analysis is to show which meaningful similarities can be captured by our different distances, and to highlight which is distances are better suited to the comparison of the different cuisine-graphs in this very sparse and unbalanced setting. In this case, natural groupings of cuisines are intuitive, and the results are thus easy to benchmark. Here, we use our inferred pairwise distance matrix between ingredient co-occurrence networks and evaluate our results by plotting both the heat maps of the pairwise distances and constructing "3-nearest-cuisine metagraphs" for each type of distance. In this graph, each node corresponds to a given cuisine c. The neighbors of cuisine c correspond to its three-nearest neighbors with respect to a given pairwise similarity matrix. This yields a directed graph of order 3, which we treat here as undirected – hence the degree of each node can be greater than 3 if the node is among the three nearest-neighbors of several cuisines. This provides a way of filtering the information contained in the distance matrix, and quickly visualizing whether the similarities recovered by the distance make intuitive sense.

Structural distances.

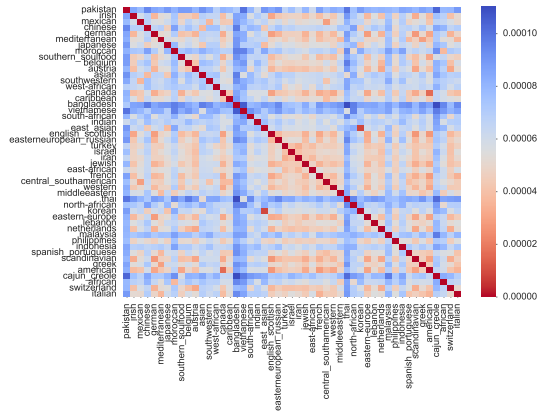
We begin by analyzing the similarities captured by the Hamming and Jaccard distances. We note that these two structural distances yield very different results. The Jaccard 3-nearest-cuisine summary graph exhibits an interesting tri-cephalic structure (Figure 2.11b): almost every node in the graph is connected to three main hubs (American, French and Italian). This shows that the Jaccard similarity mostly captures the proportion of shared co-occurrences (as opposed to other network properties). Indeed, here, the American, Italian and French cuisines have the largest connected components (Figure C1c), hence the overlap with the other cuisines' connected components is greater. As such, the Jaccard distance fails to recover more subtle structure in the food network. At the other extreme, the Hamming distance recovers more structure than the Jaccard distances: it manages to recover clusters corresponding to East Asian and East European cuisines. We note here that the similarities are linked with the number of shared ingredients between two cuisine. In particular, the Bangladesh cuisine – whose connected component comprises only 22 ingredients) is uniformly far from the other graphs (Figure 2.11c). The Hamming distance only reflects the overlap in connected components without accounting for the components relative sizes.



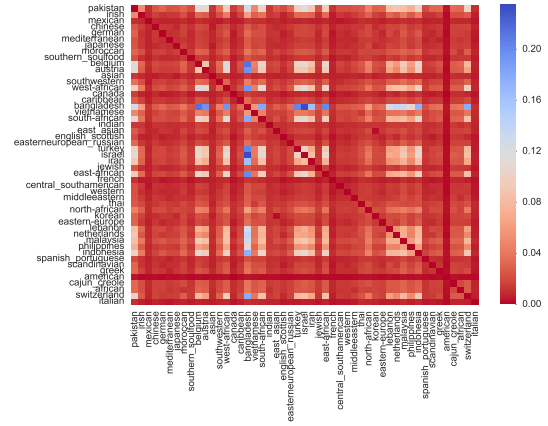
(a) 3-nearest-neighbor Proximity Graph between cuisines (Hamming Distance)



(b) 3-nearest-neighbor Proximity Graph between cuisines (Jaccard distance)



(c) Pairwise distances between cuisines (Hamming Distance)



(d) Pairwise distances between cuisines (Jaccard Distance)

Figure 2.11: Comparison of the pairwise distances and three-nearest cuisine summary graphs (**Left column:** Hamming distance. **Right column:** Jaccard). The three-nearest cuisine summary graph is constructed by representing each co-occurrence network by a node and linking it to its three nearest neighbors according to a given pairwise similarity matrix.

Spectral distances. Spectral distances also struggle to recover similarities between cuisines. In this example, the IM and HIM yielded the same 3-nearest-neighbor graphs (Figure 2.12). We note that this graph consists of two connected components, and does not follow the expected clustering of cuisines: the Mediterranean cuisines for instance (pink nodes in Figure 2.12) are scattered in each of the two clusters. This might be due to the fact that the IM distance fails in the presence of disconnected graphs, where the eigenvalue 0 has a high order of multiplicity for every graph: in particular, the Bangladesh graph, where 0 has multiplicity 1,508, is very distant from the others. We note that Bangladesh sits unusually far from America, where 0 has order of multiplicity 1,189. For the polynomial distances (Section 2.3.4), we have taken parameters $\alpha = 0.9$ and $K = 5$ (a study of $\alpha = 0.5$ and $K = 3$ has achieved the same results). We have also computed a eigenspectrum-based distance

(Section 2.3.1) with $f(x) = e^{-0.9x}$ (using the adjacency matrix of the graph). The polynomial distance seems to recover clusters that are almost consistent with geographical proximities of the different cuisines. However, the lack of structure (no block elements or pronounced groupings) apparent from the heat maps (Figures 2.13a and 2.13d) highlights the fact that spectral distances struggle to find definite patterns in this dataset. We thus conclude that a distance based on eigenvalues seems to achieve very unconvincing results for the study of graphs with many disconnected components: in this case, comparing the structure of the graph is insufficient, and we need to include information contained in the nodes' labels.

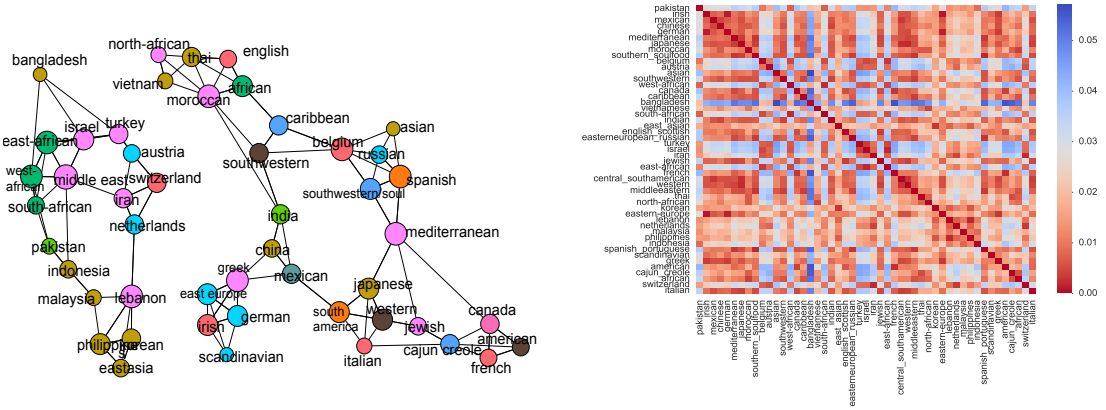
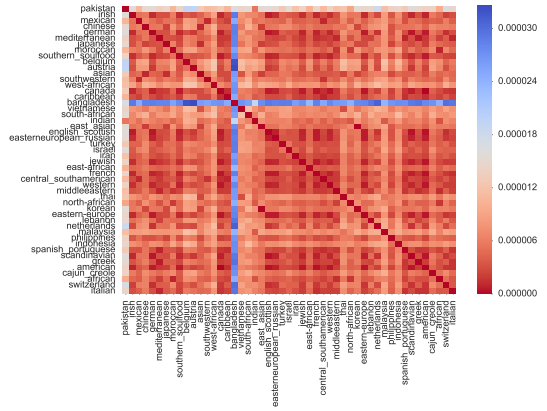
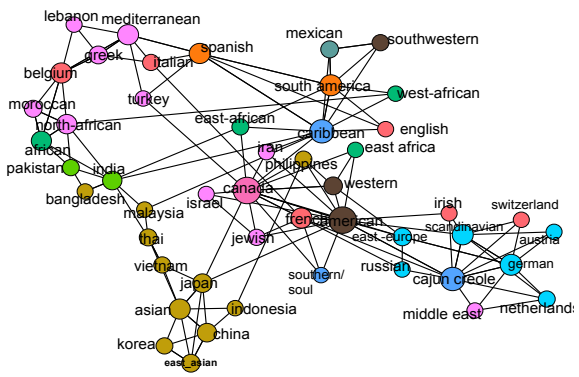


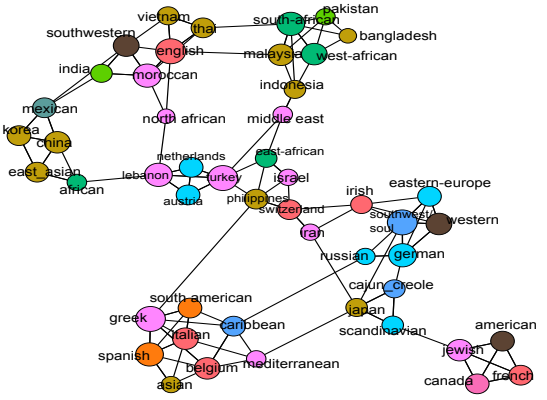
Figure 2.12: **Ipsen-Mikhailov distance. (right)** 3-nearest-neighbor Proximity Graph between cuisines **(left)** Pairwise distances between cuisines.



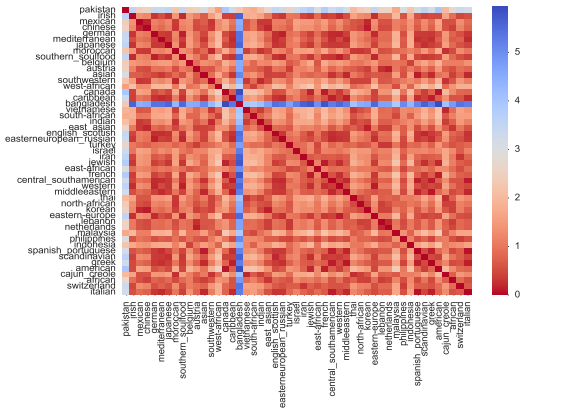
(a) Pairwise distances between cuisines (polynomial distance)



(b) 3-nearest-neighbor Proximity Graph between cuisines (polynomial distance)



(c) Pairwise distances between cuisines (Eigen Distance)



(d) Pairwise distances between cuisines (Eigen Distance)

Figure 2.13: Pairwise distances between cuisines for various spectral distances. **Top row:** Ipsen-Mikhailov distance. 2^{nd} **row:** Polynomial distance (section 2.3.4), with $\alpha = 0.9$ and $K = 5$. **Bottom row:** Eigenspectrum-based distance (section 2.3.1) with $f(x) = e^{-0.9x}$.

Wavelet distances. In this case, we have computed the heat wavelet signatures for each node according to their multiscale version described in section 2.4. The scale s was chosen to take values in $\{1, 2, \dots 29\}$.

Figure 2.14 shows the 3-nearest cuisine metagraph that this distance yields. We see that the graph that we are able to recover is consistent with geographical proximities would expect. We note for instance the clusters of Scandinavian cuisines and south-western European cuisines, as well as a high proximity of Mediterranean cuisines and Asian cuisines. It is interesting to note that this approach puts Bangladesh cuisine with high centrality. This is due to the limited number of recipes that we have for Bangladesh cuisine yielding more homogeneous and higher edges weights, and thus seemingly closer distance to the other graphs.

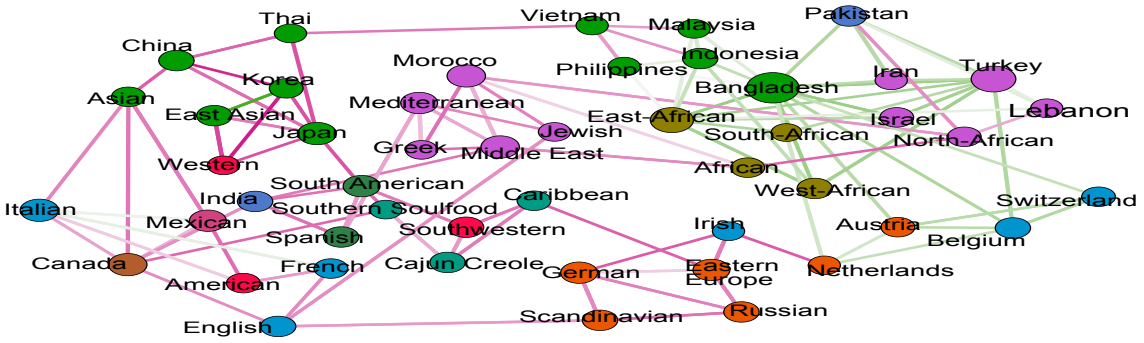


Figure 2.14: Proximity Graph between cuisines (heat-wavelet based distance)

2.7 Conclusion: which distance should we use?

In summary, in this chapter, we have given an overview of different metrics and similarity measures for comparing graphs for which we have node labels. Graphs created in real applications rarely have exchangeable nodes and our main focus here has been to supplement the ample literature on permutation invariant graph distances with more refined ones that account for these node identities. Our main focus has been to reflect on the types of changes in dynamic and spatial scenarios that these distances are best suited for. We have provided highlights of these performances on both synthetic and real-data graph analyses.

We have illustrated the use of these metrics for doing statistics on graph-objects in much the same way we did for binary rooted trees in [20]. Finding the right distance for the problem at hand can enable us to further our analyses by constructing the Fréchet mean graph or decomposing the sums of squares of distances between graphs enabling the type of analysis-of-variance type tests that we illustrated in this article. Based on the results of our set of experiments as well as the real-life results that were shown in the previous sections, we now conclude with a discussion of the strengths and relative advantages of the different distances.

Overall, structural distances seem particularly fitted for tracking temporal evolutions through time when the nodes' Ids are well defined and hold a particular importance in the network. Indeed, these distances focus on some measure of the volume of edges that change from one graph to the other, and as such, are especially able to recognize graphs belonging to same trajectory. This has proven useful in the microbiome case, where the Jaccard distance was able to recognize a strong "subject effect". The Hamming distance can be further enriched by taking into consideration higher order information and comparing larger neighborhoods with the polynomial distances. However, in real datasets, this distance appears to suffer from the same drawbacks: all changes are treated

equivalently across nodes, and the distance is blurred by the numerous changes, hindering its ability to correctly capture subtle similarities between graphs. On the other hand, as shown by the synthetic experiments [2.10,D1,D2], global distances (or meso-scale distance with large neighborhood scopes (e.g., large scaling factor τ in the heat distance) however seem less fit for the task of recognizing graph trajectories: these distances are more focused on comparing graph properties at a higher level.

However, the real-life experiments have also shown the advantages of using global spectral information. This is especially useful in real-life "noisy" setting where the correspondence between nodes from one graph to another is only approximative. In the fMRI dataset typically, the role played by one node in a graph can in fact be assumed by its neighbor in another. In that case, while these graphs will be classified as dissimilar by structural distances, spectral and meso-scale distances are able to recognize the similarity. This seems to explain the results obtained by the spectral and heat distances on the fMRI dataset, where some interesting associations between the one-nearest neighbor metagraph and the number of years under dependency are detected. The heat-based distance seems to offer a promising way of achieving "glocality" – that is finding an appropriate middle ground between local structural distance and the global spectral ones. Indeed, the scaling factor controls for the propensity of the distance to take into account local information. In order to provide more intuition to this phenomenon, let us come back to the signal processing analogy of the Laplacian's eigenvalues: the small eigenvalues of Laplacian are related to low frequency signals over the graph's vertices, inducing neighboring nodes to share similar signal values, while large eigenvalues correspond to fast varying signal across the graph's edges and can be compared to high-frequency noise. In this setting, the heat-kernel distances act by filtering out these noisy "high frequencies" and keeping only its low-frequency components– that is, information about neighborhoods. When applied to the microbiome example for instance, heat-distances with lower scaling factors were able to recover the strong subject effect in the data. However, as the scaling factor τ increases, more global information is taken into account and, similar to the eigenvalue-based distances, the heat-kernel distance is then able to recover meaningful treatment stage effects.

Figure 2.15 summarizes the different distances exposed throughout this article, highlighting their relative advantages and drawbacks.

Distances are useful in assessing many sources of variability in a dataset and as we have shown, can even detect the existence of change-points in dynamics of complex systems such as microbial communities. Pairwise dissimilarity matrices can be used to draw heatmaps (and visualizing the existence – or lack-there-of– of structure in a dataset). Multidimensional scaling embeddings of graphs in Euclidean space allow us to detect latent clusters or gradients.

However, much remains to be done to construct a complete framework for quantifying differences between graphs. For instance, we have not used subgraphs and motif counts that could also be useful in quantifying such similarities as was suggested in [13]. Another possible perspective which we have not covered here focuses on the use of graph kernels [148] to define similarities between graphs.

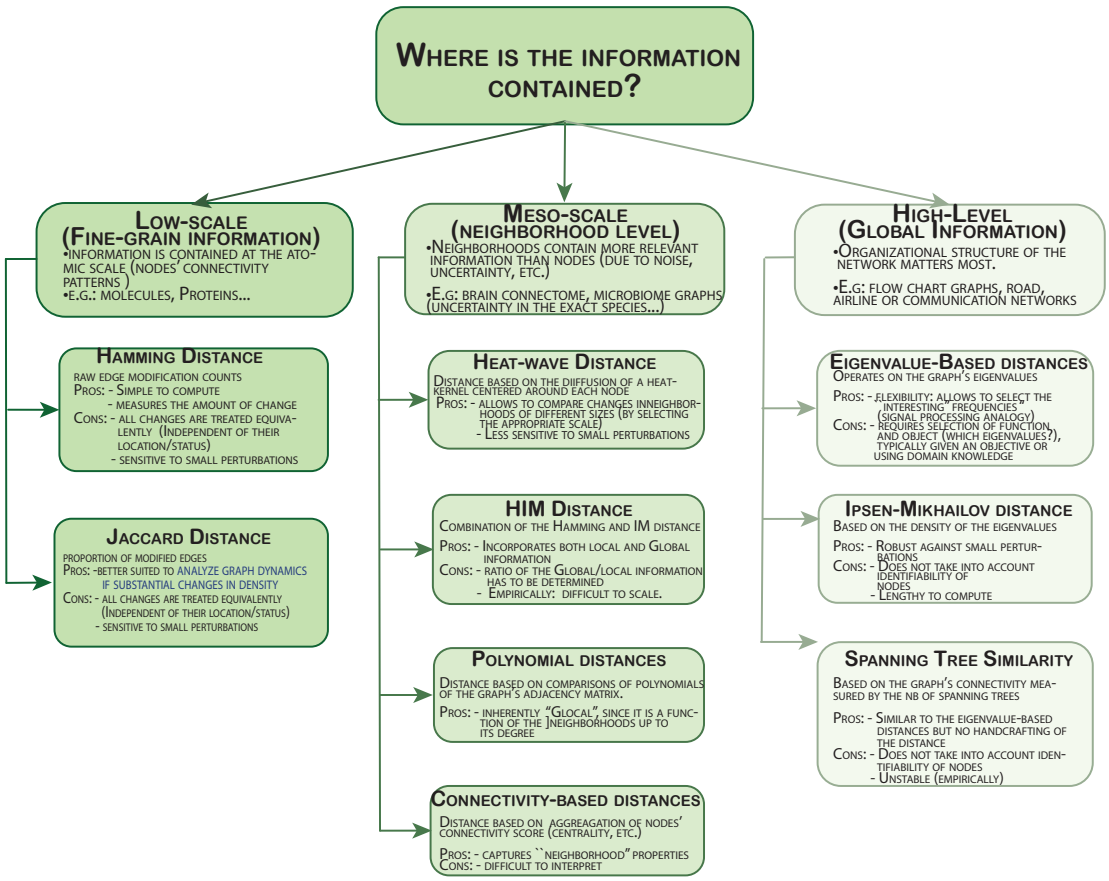


Figure 2.15: Summary of the distances detailed in this chapter.

Finally, we have only taken into account simple node identifiers, whereas incorporating more information about node covariates and edge lengths would enable higher resolution studies and enhanced change-point detection.

Chapter 3

Comparing Multiscale Representations of Observed Graph via Convex Clustering

Chapter 2 investigated several distances used to compare sets of aligned networks. But one of the main strengths of graphs consists in the hierarchical organization that they are able to capture. By progressively coarsening the nodes, we are able to get higher-level representations of the organization of the data, and recover long-range dependencies between graph regions. In some cases, such coarser representations of the graph might even be more informative than the original ones. Taking for instance the case of brain connectomics, one could imagine that the very fine grain connections between voxels are inherently noisy and subject to substantial variation from individual to individual. On the other hand, we do expect some consistency across patients as to the broader connections between brain regions. In this case, the analysis of the coarsened weaving of the brain connections might thus be more appropriate to the discovery of consistent patterns and structure across individuals. This amounts to comparing hierarchical representations of the data. Given the amount of inherent noise in our dataset, it becomes important to be able to extract hierarchical representations of aligned graphs that are both robust to noise and amenable to the comparison of aligned graphs at multiple scales. We propose to focus in this chapter on one such representation, based on an adaptation of Convex Clustering and presented in one of our publications [46].

Convex clustering [26] is a recent stable alternative to hierarchical clustering. It formulates the recovery of progressively coalescing clusters as a regularized convex problem. While convex clustering was originally designed for handling Euclidean distances between data points, our goal is to extend it here to the case where the data is directly characterized by a similarity matrix or weighted graph.

Having defined an appropriate convex objective, the crux of this adaptation lies in our ability to provide: (a) an efficient recovery of the regularization path and (b) an empirical demonstration of the use of our method. We address the first challenge through a proximal dual algorithm, for which we characterize both the theoretical efficiency as well as the empirical performance on a set of experiments. Finally, we highlight the potential of our method by showing its application to several real-life datasets — thus going beyond its potential brain connectomics application—, thus providing a natural extension to the current scope of applications of convex clustering.

3.1 Motivation: why do we need robust multiscale representations for graphs?

From gene sequencing to biomedical studies [27, 33, 68, 127], hierarchical clustering [39, 84] is currently one of the most widely-used procedures for data analysis. The resulting dendrogram yields a complete summary of the data and bypasses the need to prespecify an adequate number of clusters. The visualization of these clusters' progressive coalescence provides a comprehensive and intuitive view of their similarities. However, hierarchical clustering is an inherently greedy procedure, which typically constructs the clusters' fusion path by iteratively aggregating (or splitting) clusters. The recovered coalescence path is dependent on the choice of the linkage function, and has also been shown to be highly sensitive to outliers and perturbations of the dataset— thus allowing the formation of spurious clusters and consequently hindering the generalizability of the analysis [24]. This is particularly problematic in applications where these multiscale representations of the data are compared, contrasted and analyzed. Such is the case in brain connectomics, where a topic of interest is the comparison of the multiscale representations of the network woven by white matter tracts across different people or groups. In such noisy regimes, the definition of a robust and optimal hierarchical clustering takes on a particular importance.

Convex clustering. To overcome these issues, convex clustering [24, 77, 116] is a recent alternative formulation of hierarchical clustering as the solution of a convex optimization problem with a regularization penalty. In its original form, denoting each observation i by its corresponding vector $X_i \in \mathbb{R}^d$(with $i = 1 \cdots N$), and introducing $U_i \in \mathbb{R}^d$ the centroid of the cluster associated to point i, convex clustering solves the following objective:

$$\underset{U \in \mathbb{R}^{d \times N}}{\operatorname{argmin}} \sum_{i=1}^{N} ||X_i - U_i||^2 + \lambda \sum_{i,j=1}^{N} W_{ij} \mathrm{Pen}(U_i - U_j) \tag{3.1.1}$$

where W_{ij} are coupling weights (typically chosen as the k-nearest neighbors of each observation). In the previous expression, Pen is a penalty function (typically the ℓ_q-norm, with $q \geq 1$) which encourages coupled observations to share the same centroid. The solution path $U^{(\lambda)}$ is comparable

to the coalescence path recovered by hierarchical clustering, and the regularization parameter λ, to the different levels in the hierarchical clustering dendrogram[24, 77]:

- for $\lambda = 0$, the solution of Eq. 3.1.1 is $U^{(0)} = X$, and each point belongs its own cluster.
- as λ increases, the penalty term induces the centroids U_i to fuse, until in the limit, these centroids reach a consensus value, thus forming a single cluster $U^{(\infty)} : \forall i, j \in \{1, N\}, \quad U_i^{(\infty)} = U_j^{(\infty)}$.

The strict convexity of the objective function of Problem 3.1.1 guarantees the existence of a globally optimal solution, as well as its robustness against perturbations [24], thus making this convex formulation an extremely appealing alternative to hierarchical clustering. We refer the reader to [136] for a thorough review of the properties of convex clustering as well as a formal analysis of its parallel with hierarchical clustering.

Contributions and related work. One of the main drawbacks of convex clustering is that the optimization procedure associated to Problem 3.1.1 is computationally more involved than the greedy optimizations performed by standard hierarchical clustering. While some work has already been put into the design of efficient solutions [26], to the best of our knowledge, the derivation of algorithms for convex clustering has been restricted to the setting where data are Euclidean: observations are represented by vectors in $\mathbb{R}^d$, and similarities are simply characterized through pairwise Euclidean distances. However, in an increasing number of applications, such representations are difficult to obtain and the data come more readily as a graph in which nodes represent observations, and edges reflect some function of similarities between data points. In connectomics for instance, correlations between brain regions are summarized by a weighted graph, which provides a more amenable support to the study of functional connectivity [90, 138]. Similarly, in social sciences, relationships between individuals are readily modeled by a graph, where edges denote interactions between users. In many of these graph-structured datasets, hierarchical clustering is an indispensable tool since it allows the analysis of the data at different scales. The derivation of a convex multiscale summary of the data with the same global optimum and robustness guarantees as its Euclidean counterpart thus represents an impactful problem with many applications – a challenge which we propose to tackle in this paper. Our contributions consist in (a) the adaptation of the convex objective posed in Problem 3.1.1 to the graph setting and (b) the derivation of two provably efficient solutions. We analyze and validate our method through a set of synthetic experiments, and show its application on several real-world datasets.

3.2 Mathematical Problem Statement

Throughout this chapter, we assume that the data comes under the form of a weighted similarity matrix K between N elements (i.e, for instance, the adjacency matrix or diffusion map associated to

a graph), which we assume to be sparse. We adopt the standard convention of referring to the i^{th} column of any given matrix M as M_i.

Positive Definite Symmetric Input Matrices. We begin by studying the case where the similarity matrix K is symmetric and Positive Definite. By direct application of the spectral lemma, we can re-write K as a dot-product in a higher-dimensional space: $K = \Phi^T\Phi$ i.e. $\forall i, j, \quad K_{ij} = \Phi(X_i)^T\Phi(X_j)$. This provides an amenable setting for the generalization of convex clustering, where the goal becomes to recover the centroids U_i associated to each implicit high-dimensional vector $\Phi(X_i)$. Since each centroid lies in the convex hull of its corresponding vectors, we require U to have the form:

$$U = \Phi(X)\pi, \quad \text{where} \quad \pi\mathbf{1} = \mathbf{1}, \quad \mathbf{1}^T\pi = \mathbf{1}^T \text{ and } \pi \geq 0 \tag{3.2.1}$$

In this setting, the doubly-stochastic matrix π benefits from a bi-dimensional interpretation: the columns correspond to the centroids' representation using the original observations as dictionary, while the rows can be interpreted as soft membership assignments of observations to clusters. However, we highlight that this constraint further adds to the algorithm complexity of the original convex clustering algorithm, and is non-trivial to implement.

Using the kernel trick, Eq. 3.1.1 can be adapted here to:

$$\underset{\pi \in \Delta_N}{\operatorname{argmin}} \ \mathrm{Tr}[\pi^T K\pi - 2K\pi] + \lambda \sum_{i,j} K_{ij}\mathrm{Pen}(\pi_i - \pi_j) \tag{3.2.2}$$

where $\Delta_N = \left\{\pi \in \mathbb{R}^{N\times N} : \pi\mathbf{1} = \mathbf{1}, \mathbf{1}^T\pi = \mathbf{1}^T, \pi \geq 0\right\}$ is the set of doubly stochastic matrices.

Proof. We begin by showing how, in the case where the kernel matrix K is assumed to be positive definite, the adaptation of convex clustering proposed in Eq. 3.2.2 naturally follows. To begin with, we remind the reader that, by Mercer's theorem, we can simply write a high-dimensional equivalent formulation $\hat{\mathcal{P}}$ of convex clustering as:

$$\hat{\mathcal{P}} = \min_{\pi \in \Delta_N} \sum_{i=1}^{N} ||\Phi(X_i) - \sum_j \Phi(X_j)\pi_{ji}||^2 + \lambda \sum_{i,j} W_{ij} ||\sum_k \pi_{ki}\Phi(X_k) - \sum_k \pi_{k'j}\Phi(X_{k'})||$$

This induces the following set of equivalents:

$$\begin{aligned}
\hat{\mathcal{P}} &\iff \arg\min_{\pi\in\mathcal{S}} \sum_{i=1}^{n} \Big(||\Phi(X_i)||^2 + ||\Phi(X)\pi_{\cdot,i}||^2 - 2\Phi(X_i)^T(\Phi(X)\pi_{\cdot,i}) \Big) + \lambda \sum_{i,j} W_{ij} \Big(||U_i - U_j|| \Big) \\
&\iff \arg\min_{\pi\in\mathcal{S}} Tr[\pi^T\Phi(X)^T\Phi(X)\pi] - 2Tr\Big(\Phi(X)^T(\Phi(X)\pi)\Big) + \lambda \sum_{i,j} W_{ij} \Big(||U_i - U_j|| \Big) \\
&\iff \arg\min_{\pi\in\mathcal{S}} Tr[\pi^T\Phi(X)^T\Phi(X)\pi] - 2Tr\Big(\Phi(X)^T(\Phi(X)\pi)\Big) + \lambda \sum_{i,j} W_{ij} \Big(\underbrace{||\Phi(X)[\pi_{\cdot i} - \pi_{\cdot j}]||}_{\leq L||\pi_{\cdot i} - \pi_{\cdot j}||} \Big) \\
&\implies \arg\min_{\pi\in\mathcal{S}} Tr[\pi^T K\pi] - 2Tr\Big[K\pi\Big] + \lambda \sum_{i,j} W_{ij} \Big(||\pi_{\cdot i} - \pi_{\cdot j}|| \Big)
\end{aligned} \tag{3.2.3}$$

where, in the last line, we have used the fact that the columns of U are in fact the coordinated of the centroids in the dictionary of the original observations – hence, penalizing the pairwise differences between euclidean representation is equivalent to penalizing the dictionary coordinates.

□

The next important step consists in choosing the coupling penalty, which we take here to be a mixed total-variation penalty:

$$\text{Pen}(\pi_{\cdot i} - \pi_{\cdot j}) = \alpha||\pi_i - \pi_j||_{2,1} + (1-\alpha)||\pi_i - \pi_j||_1.$$

This choice is motivated by the fact that the ℓ_1-penalty is known to provide nested sequences of clusters [77], while the ℓ_{21}-penalty allows the recovery of a more stable solution. Total variation distances have also been shown to encourage the recovery of piecewise linear functions and to provide solutions with sharp edge contrasts [21] — a desirable property in our setting, since this amounts to "clamping" the centroids together as they progressively coalesce.

To ease notation, for any square matrix M, we denote as $\delta^{(M)} \in \mathbb{R}^{N\times N^2}$ the $N \times N^2$-dimensional matrix of pairwise differences such that: $\forall i, j, k \leq N, \quad \delta^{(M)}_{k,(i,j)} = (\mathbf{e}_{ki} - \mathbf{e}_{kj})M_{ij}$, where $\mathbf{e}_{ki} = \mathbb{1}_{k=i}$ is the i^{th} cartesian column-basis vector. The final constrained minimization problem can thus be compactly written as:

$$\underset{\pi\in\Delta_N}{\text{argmin}} \Big\{ \text{Trace}[\pi^T K\pi - 2K\pi] + \lambda\Big(\alpha||\pi\delta^{(K)}||_{2,1} + (1-\alpha)||\pi\delta^{(K)}||_1\Big) \Big\} \tag{3.2.4}$$

As for its Euclidean counterpart, the solution of Eq. 3.2.4 is consistent with its interpretation as a cluster coalescence path:

- when $\lambda = 0$, the solution of the previous equation is the identity: $\pi^{(0)} = I_N$. It is easy to check that I_N is a solution to $\text{argmin}_{\pi\in[0,1]^N} \text{Tr}[\pi^T K\pi - 2K\pi]$. Since I_N is doubly-stochastic, by strict convexity of the objective in Eq. 3.2.4, we deduce that it is the solution for $\lambda = 0$.

- when $\lambda = \infty$, on the other hand, the solution of Eq. 3.2.4 must be such that $||\pi\delta^{(K)}||= 0$. This is given by the consensus matrix $\pi^{(\infty)} = \frac{1}{N}\mathbf{1}\mathbf{1}^T$, which is the intersection of the set $\{A \in \mathbb{R}^{N\times N} : \forall i, j, \leq N \quad A_i = A_j, A \geq 0\}$ with the set of doubly-stochastic matrices.

Discussion. The assumption that K is positive definite is by no means restrictive from the modeling perspective. Indeed, in many applications (brain connectomes, etc.), the kernel K corresponds to some transformation of a positive definite similarity (typically, to some thresholded-measure of the correlation between vertices). Even if this is not the case, Positive-Definiteness can be achieved by regularizing the kernel: $\hat{K} = K+\gamma I_n$. As for many clustering algorithms (choice of the most adequate distance metric in k-means, bandwidth in spectral clustering, etc.), the choice of the appropriate transformation should depend on the analysis.

We also highlight that, in contrast to hierarchical clustering, only for the choice $\alpha = 0$ is the algorithm proven to output a nested sequence of clusters[77]. However, we emphasize that the goal of this chapter is to extract robust multiscale representations (rather than strictly nested ones): the regularization path allows the recovery of progressively coarser and coarser representations of the data. The strict convexity of the objective in Eq. 3.2.4 ensures its global optimality, which in this case, is potentially a more desirable quality than strict nestedness.

3.3 Algorithm: a dual FISTA approach

The main challenge consists in devising an efficient frithm for solving the previous optimization problem. While this problem is strongly convex, exact solvers are extremely slow, making the computation of the full regularization path almost intractable. In this chapter, we propose two main methods. The first is based on an adaptation of the Fast Iterative Shrinkage and Thresholding Algorithm [7], a method originally proposed by Beck and Teboulle for image deblurring in 2009 [6] and which we have selected for both its theoretical efficiency and its empirical performance. Our contribution here lies in the adaptation of this method to the evermore-challenging setting of Eq. 3.2.4, in which the optimization has to be done on the set of doubly stochastic matrices — a much more constrained and complicated setting than for image processing. We also provide a gradient descent-based implementation, based on a linearization of the objective and more suitable to the analysis of larger graphs, as well as an ADMM-based implementation [15] for the sake of comparison. For the sake of clarity, we outlay in the main text the derivation of FISTA, and leave the derivation of the alternative approaches to B.1 and B.2.

Algorithm. Broadly speaking, FISTA [7] is an algorithm for efficiently solving optimization problems of the form: $\min_x f(x) + g(x)$, where g is proper convex (but not necessarily smooth, as typically for ℓ_1 penalties and indicator set functions) and the subgradients of f are Lipschitz. One of the most appealing characteristics of FISTA lies in (a) the absence of any user-defined parameters—making it a completely parameter-free method—and (b) a $1/k^2$-accelerated convergence rate. In the spirit of the

algorithm proposed by Beck and Teboulle [6] for image denoising and deblurring under total-variation penalty, we propose to solve our similarity-based convex problem 3.2.4 using FISTA on the dual. The additional challenges that our approach faces with respect to the original method are two-fold: (a) our set of constraints is given by the graph adjacency matrix K and is thus more general than the regular 2D-grid in [6], and (b) we are optimizing over the set of doubly stochastic matrices, thus requiring an efficient projection algorithm.

From primal to dual. As in the previous section, we begin by supposing that the similarity matrix K is positive semi definite. K factorizes as: $K = \Phi^T\Phi$, where, by writing $K = U\Lambda U^T$ the spectral decomposition of K, we have: $\Phi = \Lambda^{1/2}U^T$. We emphasize that, while we introduce this (potentially computationally expensive) decomposition to highlight the parallel with image deblurring, we will never have to explicitly compute it. Eq. 3.2.4 can thus be equivalently re-written as:

$$\text{Minimize}_{\pi\in\mathbb{R}^{N\times N}}\frac{1}{2}||\Phi\pi-\Phi||_F^2+\mathbb{1}_{\pi\in\Delta_N}+\lambda\Big(\alpha||\pi\delta^{(K)}||_{2,1}+(1-\alpha)||\pi\delta^{(K)}||_1\Big)$$

This is akin to an image deblurring problem, with π playing the role of the true image, and Φ the observed image and blurring process. Similarly to Beck and Teboulle, we thus propose to start with the associated image denoising problem, and will generalize to the original deblurring problem in a subsequent step:

$$\underset{\pi\in\Delta_N}{\text{Minimize}}\frac{1}{2}||\pi-\Phi||_F^2+\lambda\Big(\alpha||\pi\delta^{(K)}||_{2,1}+(1-\alpha)||\pi\delta^{(K)}||_1\Big) \tag{3.3.1}$$

Proposition 1. *The dual of Eq. 3.3.1 is given by:*

$$\max_{p\in\mathcal{P},q\in\mathcal{Q}}||\Pi_{(\Delta_N)^C}\Big(\Phi-\lambda(\alpha p\delta_K^T+(1-\alpha)q\delta_K^T)\Big)||_F^2-||\Phi-\lambda(\alpha p\delta_K^T+(1-\alpha)q\delta_K^T||_F^2 \tag{3.3.2}$$

where we denote as Π_{Δ_N} the orthogonal projection operator on the set Δ_N and $\Pi_{\Delta_N^C} = I - \Pi_{\Delta_N}$ the projection onto its complement, and where the sets $\mathcal{P}$ and $\mathcal{Q}$ are respectively the ℓ_2- sphere and the unit cube in $\mathbb{R}^N$:

$$\mathcal{P} = \{p\in\mathbb{R}^{N\times N^2} : \forall i,j\in[1,N]^2, \quad ||p_{\cdot,ij}||_2\leq 1\}$$
$$\textit{and } \mathcal{Q} = \{q\in\mathbb{R}^{N\times N^2} : \forall i,j\in[1,N]^2, \quad ||q_{\cdot,ij}||_\infty\leq 1\}$$

The subgradients associated to this objective are Lipschitz with constant $L = 16\lambda^2\max_i||K_i^2||_2$.

Proof. We begin by observing that:

$$\max_{p\in\mathbb{R}^N:||p||_2\leq 1}p^Tx = \sqrt{\sum_{i=1}^n x_i^2} \text{ and } \max_{q\in\mathbb{R}^N:||q||_\infty\leq 1}q^Tx = ||x||_1.$$

Proof. We here provide a brief proof of the statement in Eq. 3.3:

$$\max_{p\in\mathbb{R}^N:||p||_2\leq 1} p^T x = \sqrt{\sum_{i=1}^{n} x_i^2} \text{ and } \max_{q\in\mathbb{R}^N:||q||_\infty\leq 1} q^T x = ||x||_1$$

To see this, let us first consider the equality on p, and introduce the Lagrangian corresponding to the constraint:

$$\mathcal{L}(p,\lambda) = -p^T x + \lambda(p^T p - 1), \quad \lambda \geq 0$$

where the primal is $\min_p \max_{\lambda\in\mathbb{R}^+} \mathcal{L}(p,\lambda)$, and the dual can be written as: $\max_{\lambda\in\mathbb{R}^+} \min_p \mathcal{L}(p,\lambda)$. The latter inner minimization with respect to p is achieved for:

$$\nabla_p \mathcal{L}(p,\lambda) = -x + 2\lambda p = 0 \iff p = \frac{1}{2\lambda}x,$$

and the dual problem reduces to:

$$\max_\lambda -\frac{||x||^2}{2\lambda} - \lambda(\frac{1}{4\lambda^2}||x||^2 - 1) = \max_\lambda -\frac{||x||^2}{4\lambda} + \lambda$$

The latter is achieved for $\lambda = \frac{||x||}{2}$, and thus: $p = \frac{1}{||x||}x$. Hence, $\max_{p\in\mathbb{R}^n, p^T p\leq 1}[p^T x] = ||x||_2$, which concludes the proof.
Similarly for q, it is easy to check that:

$$||x||_1 = \max_{s:\quad s_i\in\{-1,1\}} s^T x.$$

By relaxing the constraint on s, we have: $||x||_1 = \max_{s:\quad s_i\in[-1,1]} s^T x$, which concludes the proof. □

This allows Eq.3.2.4 to be re-written as:

$$\min_{\pi\in\Delta_N} ||\pi - \Phi||_F^2 + 2\lambda \max_{p\in\mathcal{P}, q\in\mathcal{Q}} \text{Trace}\Big(\alpha p^T \pi \delta_K + (1-\alpha) q^T \pi \delta_K\Big)$$

The corresponding dual problem $h(p,q)$ is thus given by:

$$\begin{aligned}
&\max_{p\in\mathcal{P}, q\in\mathcal{Q}} \min_{\pi\in\Delta_N} ||\pi - \Phi||_F^2 + 2\lambda \text{Trace}\Big(\alpha\delta_K p^T \pi + (1-\alpha)\delta_K q^T \pi\Big) \\
&= \max_{p\in\mathcal{P}, q\in\mathcal{Q}} \min_{\pi\in\Delta_N} ||\pi - \Big(\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha) q\delta_K^T)\Big)||_F^2 - ||\Phi - \lambda\Big(\alpha p\delta_K^T + (1-\alpha) q\delta_K^T\Big)||_F^2
\end{aligned}$$

The inner expression here is minimized by the projection of $\Phi - \lambda\Big(\alpha p\delta_K^T + (1-\alpha) q\delta_K^T\Big)$ onto the set Δ_N, allowing an explicit formulation of the dual as $\max_{p\in\mathcal{P}, q\in\mathcal{Q}} h(p,q)$, with:

$$h(p,q) = ||\Pi_{\Delta_N^C}\Big(\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha) q\delta_K^T)\Big)||_F^2 - ||\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha) q\delta_K^T)||_F^2$$

which concludes the first part of the proof.

We now have to prove that the subgradients of the dual $h(p,q)$ are Lipschitz. Taking derivatives with respect to p and q, we can show that h is Lipschitz with constant:

$$L(h) = 16\lambda^2 \max[\alpha^2, (1-\alpha)^2] \ \times (\max_i ||K_i||_2^2)$$

Proof. We here show that the dual problem is Lipschitz with respect to each of the variables p and q. We have:

$$h(p,q) = ||\Pi_{\Delta_N^C}\Big(\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha)q\delta_K^T)\Big)||_F^2 - ||\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha)q\delta_K^T)||_F^2$$

Thus:

$$\begin{aligned}\nabla_p h(p,q) &= -\lambda\alpha\Big(2\Pi_{\Delta_N^C}[\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha)q\delta_K^T)] - 2(\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha)q\delta_K^T))\Big)\delta_K \quad (*)\\ &= 2\lambda\alpha\Pi_{\Delta_N}[\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha)q\delta_K^T)]\delta_K\end{aligned}$$

where $(*)$ follows from the fact that $\frac{\partial||\Pi_{\Delta_N^C}[x]||_F^2}{\partial x} = \frac{\partial||\Pi_{\Delta_N^C}[x]||_F^2}{\partial \Pi_{\Delta_N^C}[x]}\frac{\partial \Pi_{\Delta_N^C}[x]}{\partial x} = 2\Pi_{\Delta_N^C}[x]$.
Similarly:

$$\nabla_q h(p,q) = 2\lambda(1-\alpha)\Pi_{\Delta_N}[\Phi - \lambda(\alpha p\delta_K^T + (1-\alpha)q\delta_K^T)]\delta_K$$

We note that, by definition of δ_K:

$$\forall M \in \mathbb{R}^{N\times N^2}, \quad ||M\delta_K||_{ij}^2 = K_{ij}^2||M_i - M_j||_2^2$$

Hence:

$$\begin{aligned}&\forall M \in \mathbb{R}^{N\times N^2}, ||M\delta_K||_F^2\\ &= \sum_{ij} K_{ij}^2||M_i - M_j||^2 \leq \sum_{ij} 2K_{ij}^2(||M_i||^2 + ||M_j||^2)\\ &\leq \sum_{ij}(2K_{ij}^2||M_i||^2 + 2K_{ji}^2||M_j||^2) \ \text{ by symmetry of } K\\ &\leq 4\sum_i ||K_{i,.}||_2^2||M_i||^2 \leq 4\max_i\{||K_{i.}||_2^2\} \times ||M||_F^2\end{aligned} \tag{3.3.3}$$

Hence, using the non-expensiveness property of the orthogonal projection operator, we can show

that the subgradients of h are Lipschitz, since:

$$\begin{aligned}
&||\nabla h(p_1,q_1) - \nabla h(p_2,q_2)||_F^2 \\
&= ||\nabla_p h(p_1,q_1) - \nabla_p h(p_2,q_2)||_F^2 + ||\nabla_q h(p_1,q_1) - \nabla_q h(p_2,q_2)||_F^2 \\
\leq & 8\lambda^2 \max[\alpha^2,(1-\alpha)^2] \times 4 \times \max_i ||K_{i\cdot}||_2^2 \\
& \times ||\Pi_{\Delta_N}\Big(\Phi - \lambda(\alpha p_1 \delta_K^T + (1-\alpha) q_1 \delta_K^T)\Big) - \Pi_{\Delta_N}\Big(\Phi - \lambda(\alpha p_2 \delta_K^T + (1-\alpha) q_2 \delta_K^T)\Big)||_F^2 \\
\leq & 32\lambda^4 \max[\alpha^2,(1-\alpha)^2] \ \times (\max_i ||K_{i\cdot}||_2^2) \times ||\Big(\alpha(p_1-p_2) + (1-\alpha)(q_1-q_2)\Big)\delta_K^T||_F^2
\end{aligned} \tag{3.3.4}$$

$$\begin{aligned}
&||\nabla h(p_1,q_1) - \nabla h(p_2,q_2)||_F^2 \\
\leq & 128\lambda^4 \max[\alpha^2,(1-\alpha)^2] \ \times (\max_i ||K_{i\cdot}||_2^2) \times ||(\alpha(p_1-p_2) + (1-\alpha)(q_1-q_2))||_F^2 \\
\leq & 128\lambda^4 \max[\alpha^4,(1-\alpha)^4] \times (\max_i ||K_{i\cdot}||_2^2) \times 2(||p_1-p_2||_F^2 + ||q_1-q_2))||_F^2) \\
\leq & 256\lambda^4 \max[\alpha^4,(1-\alpha)^4] (\max_i ||K_{i\cdot}||_2^2) \times ||(p_1,q_1) - (p_2,q_2)||_F^2
\end{aligned} \tag{3.3.5}$$

Thus:

$$||\nabla h(p_1,q_1) - \nabla h(p_2,q_2)||_F \leq 16\lambda^2 \times \max[\alpha^2,(1-\alpha)^2] \ \times (\max_i ||K_{i\cdot}||_2) ||(p_1,q_1) - (p_2,q_2)||_F$$

□

■

This proposition lays the grounds for using accelerated ascent algorithms such as FISTA on the dual: given that we have shown that the subgradients of dual are Lipschitz, FISTA ensures to solve the objective with a convergence rate in $O(\frac{1}{k^2})$, where k denotes the number of iterations [7]. However, as for image deblurring, our setting is further complicated by the presence of the "blurring" matrix Φ. While we have assumed here K to be positive definite and could potentially solve exactly the projection update(*), this update would in particular require the inversion of the operator $\Phi^T\Phi = K$ — a costly operation that does not transfer well in the case where K is nearly singular. Instead, we adopt the approximate strategy of Beck and Teboulle, and view it as a rough equivalent to their deblurring problem. Denoting the solution of the denoising problem by $D(\Phi, \lambda)$, the authors show the optimal solution of the deblurring problem can be obtained by iteratively solving:

$$D(Y - \frac{2}{L}\Phi^T(\Phi\pi - \Phi)), \frac{2\lambda}{L}) = D(Y - \frac{2}{L}(K\pi - K)), \frac{2\lambda}{L})$$

Empirical results (section 3.5) validate our approach.

The FISTA updates of the dual variables are described in Algorithm 1.

Input: (fixed) variables π^0, K
Output: Denoising problem output
Initialization: $(p, q) = (s_0, r_0) = (\mathbf{0}_{N\times N^2}, \mathbf{0}_{N\times N^2})$
while not converged **do**
$(p_k, q_k) = \Pi_{\mathcal{P},\mathcal{Q}}\Big[r_k + \frac{2\lambda(\alpha,1-\alpha)}{L(h)}\pi_k\delta_K\Big]$
Or equivalently:
$p_k = \Pi_{\mathcal{P}}\Big[r_k + \frac{\alpha}{8\lambda \max[\alpha^2,(1-\alpha)^2]\ \times(\max_i||K_i||_2^2)}\pi_k\delta_K\Big]$
$q_k = \Pi_{\mathcal{Q}}\Big[s_k + \frac{1-\alpha}{8\lambda \max[\alpha^2,(1-\alpha)^2]\ \times(\max_i||K_i||_2^2)}\pi_k\delta_K\Big]$
$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
$(r_{k+1}, s_{k+1}) = (p_k, q_k) + \frac{t_k-1}{t_{k+1}}(p_k - p_{k-1}, q_k - q_{k-1})$
$\pi_{k+1} = \Pi_{\Delta_N}[\pi^k - \frac{2}{L}(K\pi^k - K) - (\alpha\lambda r_k\delta_K^T + (1-\alpha)\lambda s_k\delta_K^T)]\delta_K\Big]$
end while

Algorithm 1: Update for π

Proposition 2. *The projection operators onto the sets* $\mathcal{P}, \mathcal{Q}$ *are given by:*

- $\Pi_{\mathcal{P}}[p] = \frac{p_k}{\max[1,||p||_2]}$
- $\Pi_{\mathcal{Q}}[q] = \frac{q_k}{\max[1,|q_k|]}$

In particular, the previous two-step procedure has the advantage of bypassing the need to explicitly compute and invert Φ. In order to efficiently perform the updates on the set of doubly stochastic matrices, we use the scalable iterative scheme proposed in [108], which we detail in Algorithm 2.

The procedure is summarized in Algorithms 1 and 2, and a Python implementation is provided on Github[1].

Ymatrix to project onto Δ_N
$P^* = \arg\min_{D\in\Delta_N}||Y - D||_F^2$
Initialization: $P \leftarrow Y$*;*
while not converged **do**
$P \leftarrow P + (\frac{1}{n}I + \frac{\mathbf{1}^T P\mathbf{1}}{n^2}I - \frac{1}{n}P)\mathbf{1}\mathbf{1}^T - \frac{1}{n}1 1^T P$
$P \leftarrow \frac{P+|P|}{2}$
end while
Return π^* such that $\Phi\pi^* = \Pi_{\Phi\Delta_N}[\Phi - \lambda\mathcal{L}(r_k, s_k)]]$

Algorithm 2: Projection onto Δ_N

[1] https://github.com/donnate/HC_dev

3.4 Performance Analysis

Computational cost analysis. An inspection of the updates in Algorithm 1 reveals that only a subset of the coordinates of p_{ij} have to be updated at each iteration— that is, whenever $K_{ij} = 0$, p_{ij} remains identically zero. Denoting $|\mathcal{E}|$ as the number of non-zero entries in K, the memory required to store both p and q is thus $O(N|\mathcal{E}|)$. At each step, the algorithm thus relies on either (a) element-wise operations on matrices of size $O(N|\mathcal{E}|)$ or (b) matrix multiplications with cost at most $O(|\mathcal{E}|N^2)$. As such, the overall complexity and memory storage of the algorithm grows linearly with the number of edges, but quadratically with the number of nodes. We have not as of yet optimized the design of an algorithm capable of efficiently storing a representation of K and leave this to future work (see discussion in the conclusion).

Validation of the empirical efficiency. We begin by assessing the efficiency of our algorithm through a set of synthetic experiments. We generate a synthetic random graph with 3-level fractal structure: the coarsest level corresponds to an Erdős-Rényi graph on 4 “meta nodes". Each of these meta nodes can be further divided in a set of communities on 7 “super nodes", each corresponding to a dense clique on 7 nodes. This graph generation process yields a graph with 2 levels of clustering (i.e, levels of resolution): a coarse one at the meta level (4 clusters) and a fine-grain one at the super-node level (28 clusters). Figure 5.1(A) illustrates the generation process. Our convex clustering objective in Eq. 3.2.2 makes it particularly amenable to classification: the columns of the recovered matrix $\pi(\lambda)$ provide a representation of the centroids using the observations as dictionary – allowing to use any off-the-shelf machine learning algorithm to analyze these representations. We run our hierarchical method on the regularized 2-hop adjacency matrix of the induced graph ($K = \mathrm{D}^{-1/2}A^2\mathrm{D}^{-1/2}$ with $D = \mathrm{Diag}(A^2\mathbf{1})$) and assess the results both visually through the associated PCA plots (Figure 5.1) and quantitatively by running k-means on the recovered centroids $\pi(\lambda)$.

Performance Metrics. We quantify the amount of structure recovered at each regularization level λ through:

- **the effective rank $er(\lambda)$ [122] of the similarity between centroids:** letting D_π be the distance matrix between observations (i.e., $D_\pi[i,j] = \pi_i^T K \pi_j$) and $\{\sigma^{(D_\pi)}\}_j$ its eigenvalues, the effective rank is defined as:

$$er(\pi) = \exp\{-\sum_{k=1}^{N} \frac{\sigma_k^{(D_\pi)}}{\sum_{j=1}^{N}\sigma_j^{(D_\pi)}} \log\left(\frac{\sigma_k^{(D_\pi)}}{\sum_{j=1}^{N}\sigma_j^{(D_\pi)}}\right)\}.$$

  This measures the entropy of the eigenvalue distribution of the similarities between centroids, and should progressively decrease from N to 1 as λ increases.

- **the k-means cluster accuracy and silhouette score.** We run k-means on the centroids for respectively 4 and 28 clusters, and assess the accuracy of the recovered multilevel clustering: a

high accuracy and silhouette score indicate that the convex clustering algorithm has successfully recovered the multi scale structure of the data.

All the results that we show here are averaged over 20 different random graphs. Figure 5.1 shows the progressive coalescence of the centroids as λ increases. We note the consistency of the recovered cluster path with hierarchical clustering: for very small values of λ, each cluster contains only one node, and the centroids representations progressively merge as λ increases. This can be further quantified by computing the effective rank (Fig. 3.2B), which progressively dwindles with the increase of the regularization penalty. Note that, as indicated above, this decline is not strictly monotonic. This behavior is further quantified in Table 3.1, where it becomes apparent that the different coarsened graph representations induced by λ recover different levels of resolution: the accuracy both at the meta-community (4 clusters) and super-node (28 clusters) level is extremely high. Yet, as λ increases and the centroids progressively fuse, the silhouette score decreases. In particular, Fig. 3.2 (C,D) show that the optimal silhouette score for clustering at the fine and coarse levels shifts from $\lambda^*_{\text{fine}} \approx 1$ to $\lambda^*_{\text{coarse}} \approx e^6$ (peak of the curve).

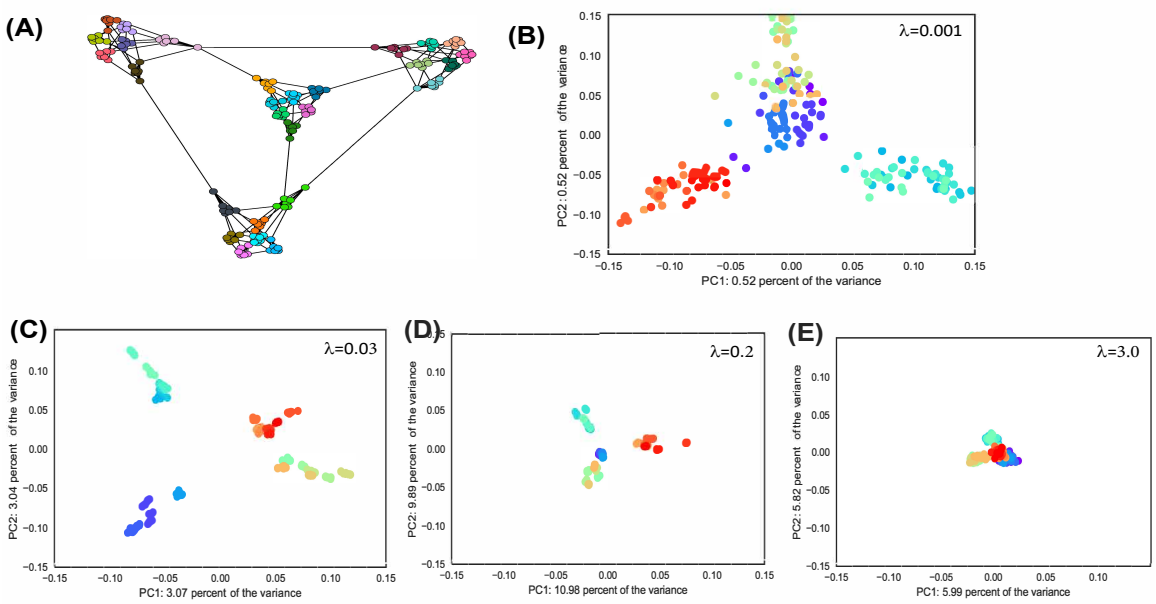


Figure 3.1: Application of the convex hierarchical clustering algorithm to a synthetic graph on 196 nodes. PCA representation of the nodes for **(B)** $\lambda = 0.001$, **(C)** $\lambda = 0.03$, **(D)** $\lambda = 0.2$ and **(E)** $\lambda = 3.0$.

Impact of the choice of α. A comparison of the results for different values of α is presented in Fig. 3.2. We observe that the behavior of the results is roughly similar, although smaller values of λ seem to encourage faster clamping of the centroids and a steeper convergence towards the consensus matrix $\pi^\infty = \frac{1}{N}\mathbf{1}\mathbf{1}^T$.

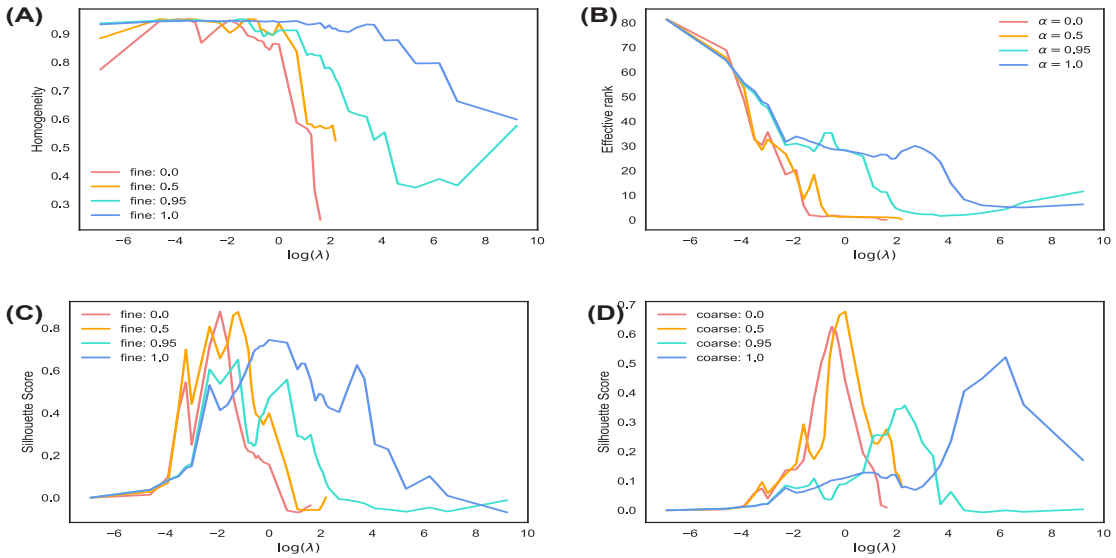


Figure 3.2: Results of the algorithm averaged over 20 independent trials: (A) efficient rank, (B) homogeneity of the clustering on 28 clusters, (C,D) silhouette scores on respectively 28 and 4 clusters for different values of α.

		28 classes			**4 classes**		
λ	$er(\lambda)$	Accuracy	Completeness	Silhouette	Accuracy	Completeness	Silhouette
0.001	81.3	0.86	0.94	$8.1e^{-4}$	0.95	0.94	$1.3e^{-4}$
0.1	30.4	0.89	0.94	$6.1e^{-1}$	0.95	0.95	$8.5e^{-2}$
0.5	35.2	0.70	0.91	$2.6e^{-1}$	0.95	0.94	$3.8e^{-2}$
1	28.45	0.65	0.91	$4.7e^{-1}$	0.95	0.95	$8.9e^{-2}$
40	1.64	0.29	0.53	$-2.6e^{-2}$	0.87	0.74	$2.1e^{-2}$

Table 3.1: Performance of kmeans clustering on the raw embeddings, with respectively 4 or 28 classes as ground truth labels, for $\alpha = 0.95$.

3.5 Validation with Real-life Experiments

Our method was driven by its application to multi-resolution graph analysis. In this setting, a typical goal is to obtain a coarser and coarser approximation of the similarity matrix (progressively fusing the clusters together) in order to capture the underlying structure of the graph at multiple scales. With this objective in mind, we now provide a few examples of the performance of our method on two real datasets.

Connectomics. In this application, we wish to compare the structural connectomes of healthy individuals, undergoing a longitudinal test-retest Reliability and Dynamical Resting-State fMRI study. The data is a subset of the HNU1 cohort [160]. In particular, we focus on the structural

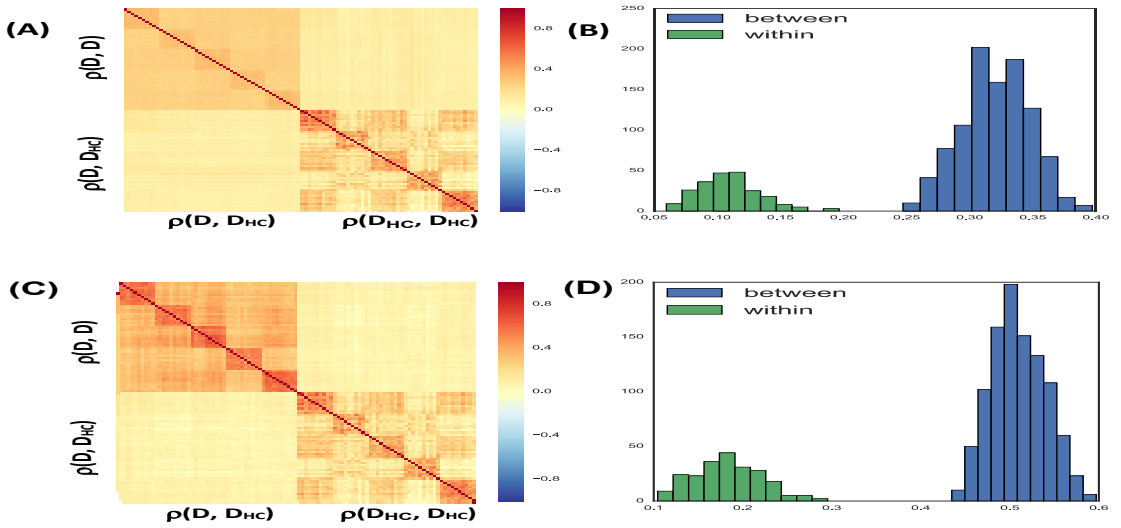
connectomes of 5 subjects obtained over the course of 10 distinct scan sessions (three days apart from another)[2]. For each connectome, we compute its convex clustering representation for various values of the parameter λ. This yields a multiscale representation from the raw connectomes, which we then compare. The goal is to assess whether the multiscale representations obtained via Convex Clustering are more consistent and robust across subjects and scans than the ones obtained via traditional single linkage Hierarchical Clustering (HC). To compare the output of our convex clustering procedure (i.e a set of centroids) and the dendrogram obtained via single linkage HC, we compare the distance matrices between centroids that these output induce (in particular, we use the cophenetic distance [131] to convert the HC dendrogram into a distance matrix).

Fig. 3.3a shows, on the left side, the Kendall rank correlation between these similarity matrices for two values of λ, as well as their correlation with the cophenetic distance induced by single linkage HC. Interestingly, both HC and Convex clustering recovered multiscale representations with a strong subject effect, as highlighted by the red blocks along the diagonal: representations corresponding to different scans of the same subject are more alike than scans across subjects. This is highlighted by the column on the right side of Fig. 3.3a, which shows a clear separation in the distances between scans belonging to the same subject ("within distances") and scans across different subjects ("between"). This effect fades away as the regularization increases. This is consistent with our expectation that the overall organization of the brain is globally the same across subjects, while differences between individuals are more salient at the fine-grain scale.

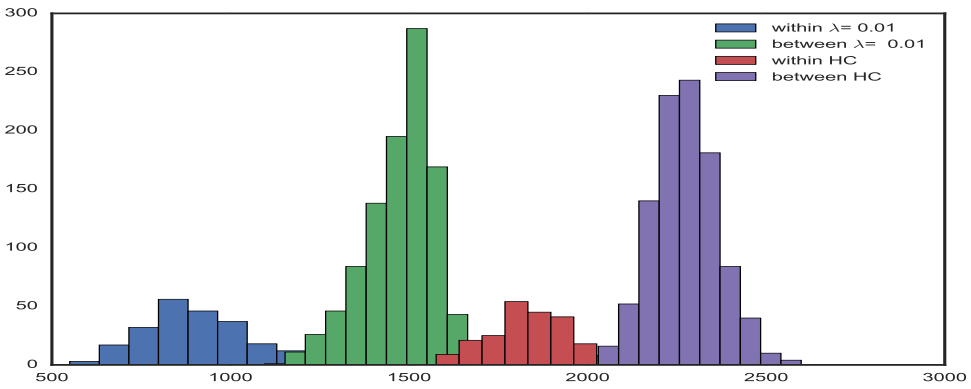
To quantify the relative performance of our algorithm with standard HC, we estimate the variability of its output: for each scan and each value of λ, we compute the 5 nearest-neighbor graphs that the coarsened similarity matrices induce. We then compute the distances between these 5 nearest-neighbor graphs (using the Hamming distance, that is, the raw ℓ_2-distance between adjacency matrices). We observe that the variability in these graph is smaller for convex clustering than for graphs obtained using single linkage HC: the distribution for two values of λ are plotted on Fig. 3.3b, and we observe that these differences are significantly inferior than to the ones of HC. This indicates that the 5-nearest neighbor graphs recovered by our convex clustering procedure induce more robust and consistent multiscale representations of the connectomes across subjects and scans.

Khan gene expression. We now demonstrate the robustness of our method by applying it to the Khan dataset [94]. This dataset consists of gene expression profiles of four types of cell tumors of childhood. In this case, we want to show that the clusters recovered by our procedure are more robust than those recovered by standard hierarchical clustering, in that the multiscale representation of the similarities between genes that they capture are more reproducible: we split the dataset between training and testing, and assess the similarity between the multi scale representations that we extract out of those. In this case, a set of 64 arrays and 306 gene expression values are used for training, and 25 arrays for testing. We apply hierarchical clustering on the similarity matrix induced by the

[2]The preprocessed structural connectomes are readily available at `https://neurodata.io/mri-cloud/`.



(a) Results for the connectomics study using different values of the regularization (A, B: $\lambda = 0$ and C,D: $\lambda = 0.01$). Left-column: pairwise Kendall rank correlation between coarsened brain networks induced by convex clustering $\rho(D, D)$ and HC's associated cophenetic distance $\rho(D_{HC}, D_{HC})$ as well as Kendall cross-correlation between representations $\rho(D, D_{HC})$ (matrix sketch on the left of the picture). Right column: distribution of the distances between coarsened DTI scan representations for scans belonging to the same subject ("within") and across different subjects ("between").

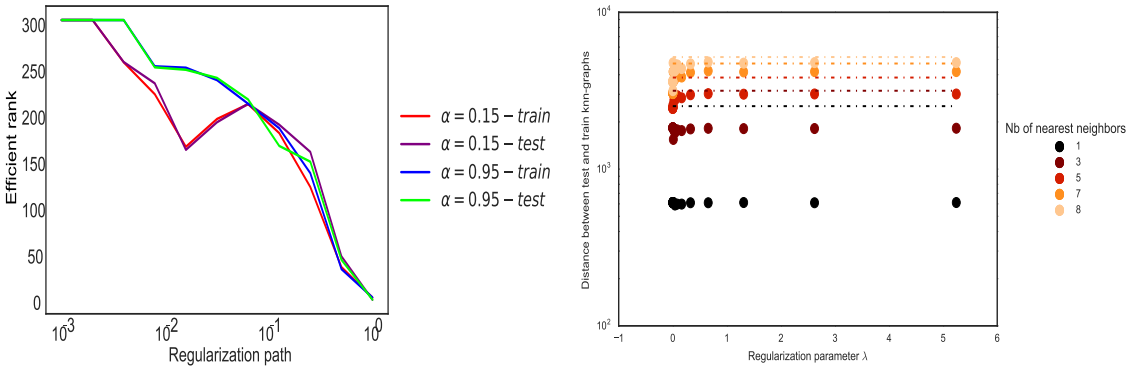


(b) Comparison of the "within" and "between" subject distances between coarsened representations of DWI scans for various levels of λ.

data's 10 nearest-neighbor graph and assess the stability of the induced hierarchy: at each level, we aggregate the centroids in both training and testing based on their efficient rank and compute the clusters' homogeneity score using the training labels as ground truth. We compare this against standard agglomerative clustering. Interestingly, the results (displayed in the table in Fig. 3.4) indicate a better homogeneity of our method with respect to the greedy one for intermediary values of the regularization, indicating that the more unstable clusters of standard HC's clustering are at the intermediary levels. Fig. 3.4b also shows that the distances between train and test k-nn graphs

(as in the connectome study) is consistently smaller than for the k-nn graphs induced using HC's cophenetic distance. This indicates greater consistency between test and train results for convex clustering.

Figure 3.4: Results for the Khan Dataset



(a) Efficient rank for recovered along the regularization path, on the test and train sets for two values of α.

(b) Distances between test and train k-nearest neighbor graphs as the regularization λ increases (colored by k). Dashed lines indicate the HC-cophenetic baseline for the number of neighbors considered.

λ	**Effective rk** $er(\pi)$	**Homogeneity FISTA** ($\alpha = 0.95$)	**Homogeneity standard HC**
0.032	242	**0.937**	0.935
0.256	141	**0.774**	0.721
0.512	37	**0.446**	0.292
1.024	7	**0.100**	0.087

(c) Table: Homogeneity score between test and train predictions at different points of the regularization path.

3.6 Conclusion: how can we compare hierarchical structures?

In conclusion, we have proposed an adaptation of FISTA on the dual for solving convex hierarchical clustering in the case where the data are directly a graph or a similarity matrix. We have shown the performance of our method on both synthetic and real datasets, highlighting its ability to recover different important scales with better consistency and robustness than standard Hierarchical Clustering. To begin working on scaling up this method, we also devised a gradient descent-based implementation, based on a linearization of the objective and more suitable to the analysis of larger graphs, as well as an ADMM version [15] for the sake of comparison. One intrinsic limit to the scalability of our method lies in its requirement to store a matrix of size N^2 – an aspect that we could attempt solving through progressive matrix factorizations of $\pi = A^T A$, which we leave for future work.

Chapter 4

Inference on Latent Graphs: a Bayesian Independent Component Analysis framework for Brain Connectomics

The two previous chapters were devoted to the analysis of observed networks. We now propose to extend the discussion to the case where the graphs are latent, and have to be inferred and compared. In particular, we gear the treatment of this topic here to the inference of subnetworks for Brain Connectomics, where the high-dimensional, low-sample, and noisy regimes that typically characterize fMRI data, makes the recovery of such interactions an ongoing challenge: how can we discover patterns of co-activity between brain regions that could then be associated to cognitive processes or psychiatric disorders? In this chapter, we investigate a constrained Bayesian ICA approach which, in comparison to current methods, simultaneously allows (a) the flexible integration of multiple sources of information (fMRI, DWI, anatomical, etc.), (b) an automatic and parameter-free selection of the appropriate sparsity level and number of connected submodules and (c) the provision of estimates on the uncertainty of the recovered interactions. Our experiments, both on synthetic and real-life data, validate the flexibility of our method and highlight the benefits of integrating anatomical information for connectome inference.

4.1 Motivation: extracting relevant subnetworks in Brain Connectomics data

Motivation. Over the recent years, the study of brain connectomics [16, 56, 57] has gained increased interest amidst the neuroscience and cognitive psychology communities. In this framework, the brain is modeled as a graph in which nodes denote voxels or regions of interest (ROIs), while edges represent some notion of functional connectivity (partial correlations, mutual information, etc.), typically inferred from fMRI scans. As explained in Chapter 1, the driving hypothesis behind brain connectomics is that the identification of interactions between modules of nodes is crucial to our understanding of cognitive processes and psychological diseases. Central to the field is the analysis of resting-state functional MRI (rs-fMRI), believed to capture the brain's default activity [55, 70, 118, 119] and the diverse functional interactions between ROIs. Mathematically speaking, the recovery of these interactions boils down to an estimation of the correlation (or precision) matrix between brain regions. Yet, in addition to the intensive pre-processing that fMRI data typically requires (denoising, unringing, motion correction, transformation to a standard template space, etc), in most studies, the number of potential edges is significantly greater than the number of time points in the BOLD time series—thus compelling a heavy filtering of the sample correlation matrix to get rid of spurious correlations. The variability between sessions and individuals further hinders the generalizability of the analysis and complicates the recovery of precise estimates of these interactions.

Meanwhile, structural connectivity has been a subject of increased interest in the brain connectomics literature over the past few years. Structural connectivity reflects the physical wiring of the brain by inferring white matter tracts from Diffusion Weighted Images (DWI). Since structural connectivity is widely believed to restrict functional connectivity [37, 78, 114, 134, 145], the incorporation of DWI data as supplementary information seems like an appealing way of making the estimation of functional connectivity more robust. However, structural data does not come without its fair share of caveats—thus perhaps explaining why more emphasis has been given to structural connectivity to guide (rather than strictly constrain) functional analysis. In particular, several studies have shown that, while structural connectomes typically did not contain many false negative edges [78, 124]—in other words, we can assume that all the long-range anatomical connections are recovered,—they might however contain many phantom tracts. Moreover, by design of the recovery process, longer-range connection weights are typically inflated compared to their shorter counterparts, thus yielding additional uncertainty in the actual significance of the strength of the recovered structural connections. In spite of these drawbacks, the joint use of different imaging modalities could nonetheless allow an increased prediction accuracy in the study of clinical or behavioral outcomes, as well as a deeper understanding of the interplay between neurophysiological signals and cognitive processes. This new multimodal setting calls for the development of new statistical methods [114] tailored to the integration of these sources of information.

Prior Work: Single-modality methods. There is an extremely rich literature on single modality methods for the estimation of functional connectivity and subsequent network analyses. In this setting, only fMRI BOLD time series are considered, and the inference boils down to a filtering of the sample correlation matrix to detect significant interactions between brain regions and limit the number of false discoveries — a setting which thus transcends the field of brain connectomics. In this broader setting, in order to "filter" the correlation matrix, practioners have typically chosen one of three mutually exclusive paths.

Shrinkage-based methods. When it comes to correlation estimation, the first of these paths consists in using shrinkage methods [25, 34, 51, 98, 103, 125] and Random Matrix Theory [2, 155] to correct for the spectrum deviations due to the data's high dimensional regime. However, these approaches are typically completely agnostic to either anatomical constraints or sparsity assumptions — perhaps explaining why such approaches are seldom used in fMRI studies.

Threshold-based methods. Another popular branch of analysis uses thresholding as a way to constrain and recover estimates of the inverse covariance matrix [9, 10, 18, 63, 126, 157]. Such estimates are typically more organically aligned with our prior assumptions on structural connectivity, as we only expect a subset of regions to be actively involved in a given cognitive process. Yet, while the regularization and inclusion of a prior on the sparsity in the estimation of the connectome is a useful — if not necessary— starting point, all of these methods remain agnostic to the underlying anatomy of the brain. When applied to brain connectomics, such threshold-based approaches exhibit several deficiencies:

1. They rely on an arbitrary definition of the cutoff threshold. This is usually achieved by controlling for the level of sparsity of the recovered graph, which has itself to be selected—thus introducing additional sources of variability in the analysis.
2. They provide no way of accommodating for the uncertainty of the edge weights.
3. Mainly, these methods do not allow the inclusion of complementary sources of information, such as DWI data.

Factor Analysis The last popular pipeline for filtering the sample correlation matrix in spite of the high-dimensionality of the data consists in using a flavor of factor analysis. Broadly speaking, factor analysis has been suggested as a useful way to "clean up" correlation matrices in a number of of applications ranging from neuroscience to empirical finance, where it has been shown to accurately and efficiently recover precision matrix [17]. However, in the case of brain connectomics, we have no a priori model for the factors, and K must thus be inferred from the data. In neuroscience in particular, Independent Component Analysis (ICA) has been a long-time favorite amongst factor methods, as it allows the discovery of sparse interacting regions of the brain [38]. In its most widely adopted form, "Vanilla" ICA [80, 81, 110] is a matrix decomposition technique which utilizes the non-Gaussianity of the distribution of the multivariate time series $Y \in \mathbb{R}^{T\times N}$ to decompose it as a

linear combination of non-Gaussian sources:

$$Y = SA \tag{4.1.1}$$

where $S \in \mathbb{R}^{T \times K}$ denotes the time series associated to each independent component (assumed to be centered and non-Gaussian, such that $\mathbb{E}[\frac{S^T S}{T}] = I$) and $A \in \mathbb{R}^{K \times N}$, their corresponding loadings. Each loading $A_{k\cdot}$ can thus be associated to a subnetwork of co-activated nodes. In other words, this assumes that the time series observed at each node i is a linear combination of independent cognitive processes (each associated to a given source). The coefficient $|A_{ki}|$ thus quantify how much node i responds to the activation of source k, so that when $|A_{ki}|$ is high, we say that node i is activated by source k. The study of the components thus allows to identify regions of the graph that are "co-activated" by the same source and that thus potentially contribute to the same cognitive process. Bayesian models for ICA have been proposed in the past, but to the best of our knowledge [23, 29, 146], such models have seldom been used in conjunction with structural information to analyze resting-state fMRI data [8, 71, 72, 96, 106].

Prior work: Multimodal methods. Recently, there has been a surge of "multi-modal" methods [114, 154] for analyzing fMRI data. Most of these methods define probability models that allow the integration of the anatomical information to guide connectome inference. Another set of methods favors joint ICA [60, 114, 135], where a traditional ICA model is fit on horizontally concatenated structural and functional data. However, to the best of our knowledge, these methods make at most one assumption for the prior of the data (i.e, selects from either structural or sparsity constraints), and none of these methods enforce the recovery of connected components.

Objectives of the paper. We propose a Bayesian Independent Component Analysis-based model (Eq. 4.1.1) of the covariance matrix that leads to the inference robust functional connectomes. Using DWI structural information to build the prior on the ICA components, we devise a sparse Bayesian hierarchical model that permits the automatic selection of the appropriate number of components and sparsity level through the use of an Automatic Relevance Determination (ARD) prior. In contrast to existing methods, our approach simultaneously leverages the following natural assumptions on the data:

- **Hypothesis 1:** The loadings are assumed to be sparse: $||A||_0 \leq s$, where s denotes a sparsity level. From a biological perspective, this is equivalent to assuming that only a few processes over a subset of brain regions are involved in the activity captured by each fMRI scan.
- **Hypothesis 2:** Each loading (or component) $A_{k\cdot}$ is assumed to correspond to an anatomically-connected subset of nodes.

Under these constraints, our formulation of Bayesian ICA simultaneously achieves (a) the flexible integration of multiple sources of information (fMRI, DWI, anatomical, etc.), (b) an automatic and

parameter-free selection of the appropriate sparsity level and number of connected submodules and (c) the provision of estimates on the uncertainty of the recovered interactions.

4.2 A Bayesian ICA model

To fulfill these objectives, we build upon the standard Bayesian ICA model [8, 28, 121]. We assume that Y comes from a noisy mixture of independent components, that is:

$$Y = SA + \epsilon \quad \text{with} \quad \mathbb{E}[Y] = 0, \quad \mathbb{V}\text{ar}[Y] = 1 \tag{4.2.1}$$

Note that we assume here that, following standard ICA procedure, the data Y has been centered and scaled. This is also consistent with the bulk of the connectome literature, which typically focuses on studying correlations — rather than covariances — between Regions of Interest (ROIs).

Eq. 4.1.1 induces the following correlation structure on the time series:

$$\begin{aligned} \epsilon &\sim N(\mathbf{0}, \text{Diag}(\gamma)) \\ \frac{1}{T}\mathbb{E}[Y^T Y] = \frac{1}{T}\mathbb{E}[A^T S^T S A] + \frac{1}{T}\text{Diag}(T\gamma) &= \mathbb{E}[\mathbb{E}[A^T \frac{S^T S}{T} A | A]] + \text{Diag}(\gamma) \\ &= \mathbb{E}[A^T A] + \text{Diag}(\gamma), \qquad \text{assuming } \frac{1}{T}\mathbb{E}[S^T S] = 1 \end{aligned} \tag{4.2.2}$$

Let us now turn to the connectedness and sparsity assumptions which constitute the bulk of our contribution. These additional modelling assumptions are closely aligned with the biology and guide the selection of our priors on A:

1. **Sparsity**: each component A_k should be non-zero on a small subset of nodes. Indeed, as explained in the introduction, $A_{k\cdot}$ can be interpreted as a "cognitive" subnetwork, as the magnitude of the coefficients in $A_{k\cdot}$ indicate which node is activated by source k. From a biological perspective, we expect specific cognitive processes to involve only a fraction of the brain. From a mathematical viewpoint, ICA relies on the non-Gaussianity of the timeseries Y. Since the number of components K here can be arbitrarily large and since each timepoint Y_{ti} is a mixture of coefficients with finite fourth moment, by virtue of the central limit theorem, our non-Gaussian mixture can only involve a few components.

2. **Connectedness:** each component A_k should span a subset of anatomically connected nodes. The connectedness assumption also stems from biogical considerations: the activation of the different ROIs originates from neurophysiological signals which travel along the white-matter tracts in the brain. As such, it is natural to assume that cognitive processes activate pathways, rather than disconnected sets of nodes.

3. **Non-negativity:** The components A are assumed to be non-negative. From a biological viewpoint, this assumes that we are only considering positive excitation mechanisms (and discarding inhibitory ones). From a modeling perspective, this also alleviates some of identifiability issue of the traditional ICA components, which are only determined up to a sign flip.

To meet criteria (1-3), we model the components A_k as independent multivariate Gaussians with a carefully selected precision matrix. Let $L = D - W$ be the graph's combinatorial Laplacian (where D is a diagonal matrix in which each diagonal entry D_{ii} is the degree of node i, and W is the graph's adjacency matrix). We define the regularized Laplacian L_α as: $L_\alpha = L + \alpha I$. Here, α is a small regularization coefficient which ensures L_α to be positive semi definite. With these notations, we propose the following generative mechanism for the components:

$$\forall k \leq K, \quad A_{k\cdot} \sim \left| \mathcal{N}\Big(0, \Sigma_\alpha\Big) \right| \tag{4.2.3}$$

where $\Sigma_\alpha^{-1} = \Omega_\alpha = L_\alpha$. Under these assumptions, the negative log-likelihood of our model thus includes a penalty on a term of the form:

$$\begin{aligned} A_{k\cdot} L_\alpha A_{k\cdot}^T = \frac{1}{2} \sum_i \sum_{j \sim i} \omega_{ij} \Big(A_{ki} - A_{kj} \Big)^2 \\ + \alpha \sum_i A_{ki}^2. \end{aligned} \tag{4.2.4}$$

.

This effectively pushes the values of the activations of neighboring nodes to be close and thus enforces the connectivity constraint of assumption (2). While the underlying motivation behind this model is clearly explained with the graph combinatorial Laplacian, we also note in passing that, under this model, any version of the Laplacian (normalized, etc.) can be used in lieu of L, with similar effects: the only difference lies in the value of the penalty ω_{ij} in Eq. 4.2.4.

The multivariate normal model creates components that are smooth and dense over the graph. To enforce localization, we convolve these components with binary mask D, such that:

$$A = \tilde{A} \odot D \tag{4.2.5}$$

where $\tilde{A} \sim \left| \mathcal{N}\Big(0, \Sigma\Big) \right|$ are our multivariate Gaussian components and D is roughly $\{0, 1\}$-valued and effectively creates a localization effect.

Let us now turn to the value of the coefficients D and the selection of the number of components. The latter can be inferred using automatic relevance determination if we further decomposing the loadings A as:

$$A = \Lambda \tilde{A} \qquad \text{with } \Lambda = \text{Diag}((\lambda_i)_{i=1}^K) \tag{4.2.6}$$

where the rows of $\tilde{A}$ are the convolved multivariate Gaussians of Eq. 4.2 and each λ_k roughly comes

from a mixture with point mass at 0 so as to effectively "switch on" or "off" the different components. In this model, the component A_k is activated if λ_k is significantly greater than 0.

Before going any further in our selection of a prior for D, let us consider the identifiability of our Bayesian model. As underlined in [28], the Bayesian ICA model in Eq. 4.2.1 is not identifiable as such, as it is invariant to row and column permutations, as well as rotations. To get rid of this permutation invariability, we order the λ_ks by increasing order. The additional sparsity assumption on the components A also alleviates some of the rotational identifiability issues of our model. In our current formulation of the problem, there remains the problem of the coupling of D and Λ: a scenario in which λ_k is almost 0 amounts the same model if we had instead $\lambda_k > \delta$ but $\forall i, D_{ki} \approx 0$. We thus constrain the columns of the mask D to have fixed norm equal to 1 in order to avoid any such scaling issues. This motivates modelling $D^2_{\cdot i}$ as a Dirichlet distribution:

$$\forall i \in [1, N], \quad D^2_{\cdot i} \sim \text{Dirichlet}[\frac{1}{K}] \in \mathbb{R}^K.$$

Indeed, from the modelling perspective, D_{ki} can be regarded as an indicator variable which selects which component node i is responsive to: $D_{ki} \neq 0$ if node i is in the component, and 0 everywhere else. By independently switching "on" and "off" the coefficients A_{ki} within each component k, we allow the component to be localized on the graph. Since we are considering voxels' scaled time series, by design, we must have:

$$\frac{1}{T}\text{diag}(Y^T Y) = 1$$

This implies that the variance of each node must be such that:

$$\begin{aligned} \forall i \in [1, N], \quad 1 = \frac{1}{T} Y_{\cdot i}^T Y_{\cdot i} &= \sum_{k=1}^{K} \lambda_k^2 D_{ki}^2 A_{ki}^2 \frac{(S^T S)_{kk}}{T} + \sigma^2 \\ &\approx \sum_{k=1}^{K} \lambda_k^2 D_{ki}^2 \Sigma_{ii}^{(A)} + \sigma^2 \end{aligned} \tag{4.2.7}$$

Each entry $\lambda_k^2 D_{ki}^2$ can thus be understood as a measure of the proportion of the variance attributed to component k in node i.

The full model is summarized in the plate model in Figure 4.1.

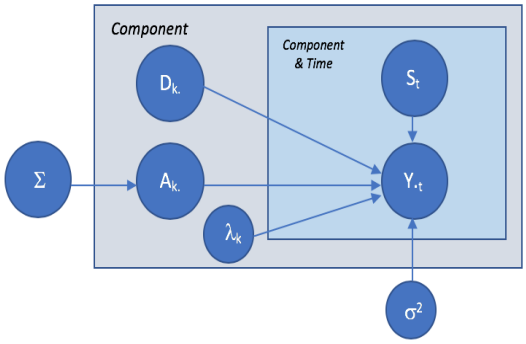
Solving the problem. In order to solve model 4.1, we propose two different methods:

- *Exact inference:* We use the Stan programming suite to perform MCMC sampling and get both modes and posterior confidence intervals for the distributions [66][1]. Stan uses Hamiltonian Monte Carlo to solve for the different components. This is particularly suited to our problem,

[1]The code is publicly available at `https://github.com/donnate/ConstrainedBayesianICA`

$$
\begin{aligned}
S_t &\sim \text{Laplace}(0, \frac{1}{\sqrt{2}}) \\
\sigma^2 &\sim 1/\Gamma_{1,1} \\
\forall k, A_k &\sim \left|\mathcal{N}(\mathbf{0}, \Sigma)\right| \\
\Sigma &: \text{from structural} \\
&\quad \text{information.} \\
\lambda_k &\sim \text{Mixture}_{\pi_\Lambda}(\Gamma_{1,1}, \Gamma_{10,10}) \\
D_k^2 &\sim \text{Dirichlet}(\frac{1}{K})
\end{aligned}
$$

(a) Explicit Bayesian model



(b) Full plate model

Figure 4.1: Summary of the Bayesian Model used in this paper.

which, with a total of roughly $K \times (2 \times N + T + 1)$ parameters to be estimated, is high-dimensional and thus difficult to solve using standard Monte Carlo methods. Despite its high-dimensionality, we underline that Bayesian ICA still represents an efficient compression of the original time series as long as $K << T$ and $K << N$.

- *Numerical Approximation:* The MCMC-based procedure is unfortunately difficult to scale up to more than a few hundred nodes. For larger-scale problems, we thus resort to a numerical approximation to Model 4.1. This approximation solves directly for the subnetworks A and is well-suited for fast approximations of numerical point estimates consistent with model 4.1. The full derivation of this method is provided in Appendix C.2.

Finally, further stability and faster convergence of the algorithm are also achieved by warm-starting it in a region of plausible solutions, obtained via numerical optimization.

4.3 Testing: Synthetic Experiments

As a proof of concept, we begin by checking the ability of our method to recover signals over synthetic graphs: the goal of this subsection is to show (a) the ability of our method to recover **sparse, localized, connected** components, and (b) to show that the recovered components are indeed more reliable than for any other method.

In order to mimic our underlying assumptions on cognitive brain mechanisms, we generate the data according to the following process:

1. **Step 1: Backbone graph generation:** we begin by creating the structural graph which will dictate the partial correlations between the different parts of the brain. To do so:

(a) We generate 5 random structural graphs (i.e, white matter pathways), each representing a separate component (i.,e, a set of co-activated nodes).

(b) We then fill the edges of these subgraphs matrix by uniformly sampling in $[0.3, 0.4]$. This allows the creation of an invertible precision matrix– that is, the partial correlations – corresponding to each submodule.

(c) The full graph is created by patching together the adjacency matrices of these subgraphs (some nodes are shared across components, in order to ensure that the final graph is truly fully connected).

2. **Step 2: Creating the loadings:** We sample each component from a folded multivariate normal distribution, where the precision matrix is the partial correlation of each of the submodules from Step 1.

3. **Step 3: Sampling the sources:** we then sample S from a Laplace distribution, with scale parameter $\frac{1}{\sqrt{2}}$ (each entry is identically distributed, there is no temporal auto-correlation in the data).

An illustration of one of such synthetic graph and its corresponding five different components is shown in Fig. 4.2. Note that in this setting, the components can overlap and some nodes can be involved in several processes.

Model Analysis. We begin by testing the ability of our model to recover five different components which are consistent with the desired specifications (hypotheses 1-3): sparse, localized and connected. Fig. 4.4 displays the graphs that are recovered for a low-noise level ($\sigma = 0.01$), which we use to start our discussion.

.

We see that the model correctly identifies 5 distinct components (Fig.4.3c). The norm for the different components (blue dots in Fig. 4.3d) show indeed that only 5 of those are above 0. By contrast, for the regular ICA model (green diamonds in Fig. 4.3d) show many of these components being distinct from 0, thus highlighting the difficulty of assessing the correct number of components in the traditional ICA setting. In particular, the correlation with the Bayesian components with the ground truth components is almost perfect (Fig.4.4b), while substantially more diffuse for the traditional ICA model (Fig.4.4a). We also see that our model's recovered coefficients are both sparse and localized, as most of the coefficients of the loadings matrix $\Lambda A \odot D$ are around 0. (Fig. 4.3c,4.3d, Fig.4.5). Finally, the correlation between the Bayesian ICA components and the ground-truth (Fig 4.4b) is extremely high, highlighting an accurate recover of the components. By comparison, the standard ICA model (Fig 4.4a) fails to capture the appropriate number of components and to accurately recover them.

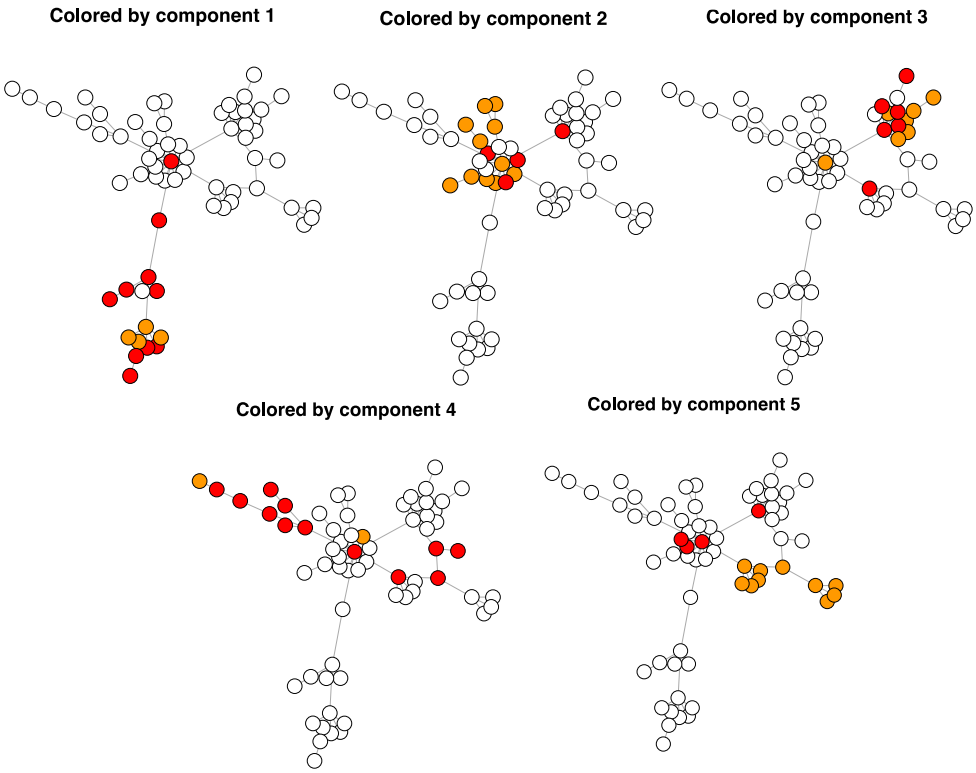
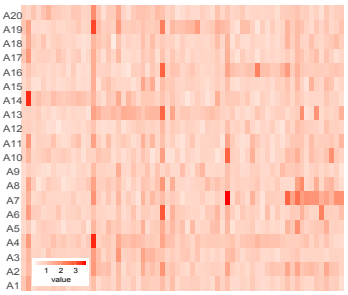
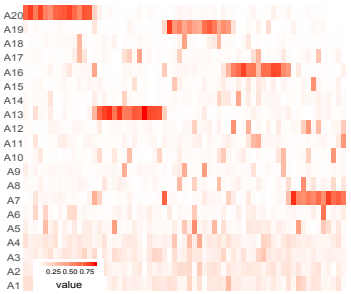


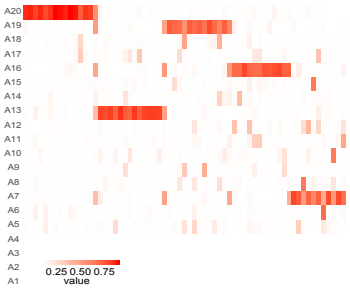
Figure 4.2: Examples of independent sparse, localized loadings over the graphs.



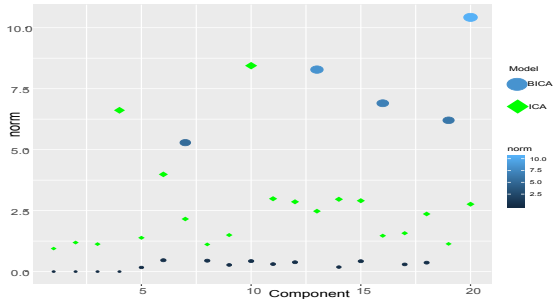
(a) Components A



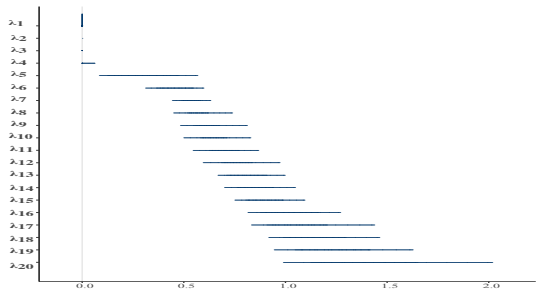
(b) Components D



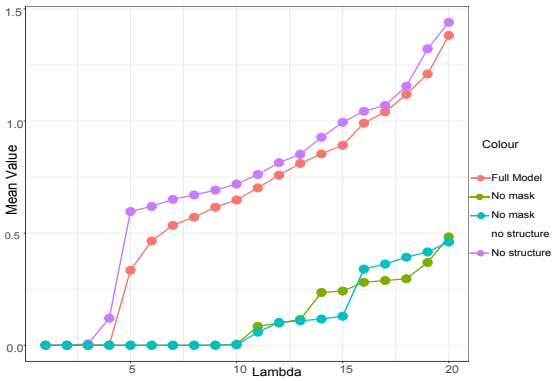
(c) Full Loadings



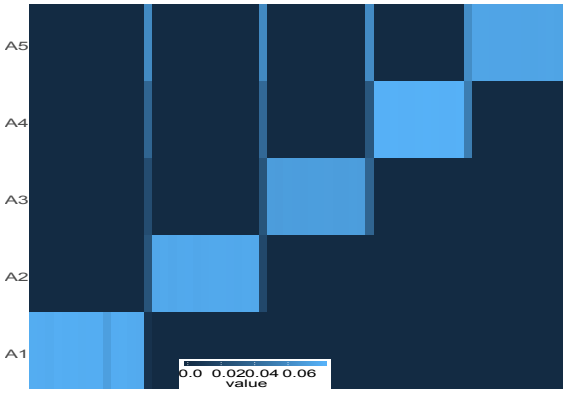
(d) Norm of the recovered coefficients



(e) Posterior distribution for the λ_k with median and 80% intervals.

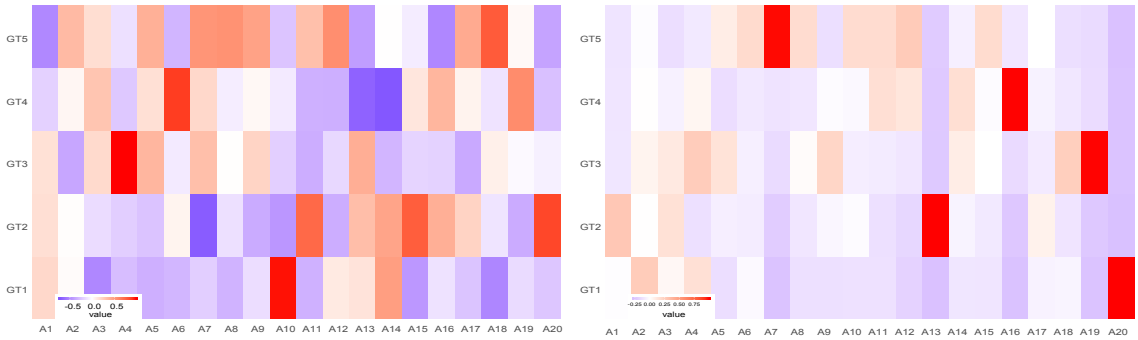


(f) Recovered λs



(g) Ground truth components

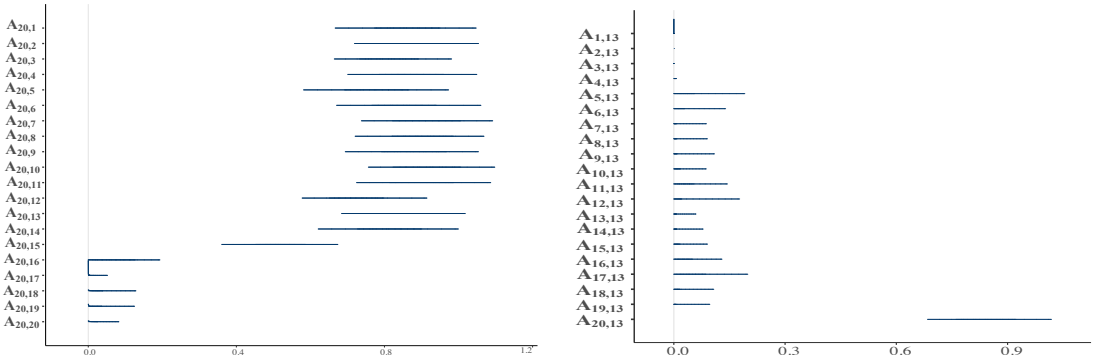
Figure 4.3: Results for the Bayesian model on the recovery of components on a single graph



(a) Correlation between ground-truth components and vanilla ICA components.

(b) Correlation between ground-truth components and Bayesian recovery.

Figure 4.4: Comparison accuracy of the recovery



(a) Credible intervals for Node 20.

(b) Credible intervals for Component 20.

Figure 4.5: Credible intervals (row-wise and column-wise)

Assessing the impact of the different assumptions. Building upon the previous visual inspection of the results, we quantify the performance of the method by running this experiment multiple times with the same setting. We also quantify how much our recovered coefficients abide with our structural assumptions:

- *spatial localization:* to quantify the spatial localization of the different subgraphs, we evaluate two quantities:
  - The average spread of the components, defined as $s = \frac{1}{K}\text{Trace}[A_k L A_k^T]$. The smaller the contrast, the smoother the component.
  - A "localization" coefficient, defined a the ratio of the within norms of the outside and inside activations associated to each ground truth component. That is, letting $\mathcal{I}$ be the set of all nodes and $\mathcal{I}_k$ be the indices of the nodes in the ground-truth component that is the most correlated with A_k, $\rho = \sum_{k=1}^{K} \frac{||A_{\mathcal{I}\setminus\mathcal{I}_k}||^2}{||A_{\mathcal{I}_k}||^2}$. Thus, the smaller the coefficient, the more localized the component.
- *sparsity:* We measure the sparsity of the recovered components by assessing:
  - the ℓ_1-norm of the components, $s = ||A||_1$
  - the number of recovered components, which we define as the number of components that have correlation above 0.8 with the ground-truth components. For perfect recovery, we expect this number to be 5. Higher number indicate redundancy in the components that are recovered, whereas lower number indicate that too few components were actually recovered.
- *accuracy:* finally, we assess the accuracy of the recovered components by comparing the correlation of the ground-truth loadings with the recovered ones. We quantify this accuracy by providing the mean of the top-5 correlations: the higher the mean, the more accurately each component is recovered.

We display the results in Tables 4.1 and 4.2, by providing the average and standard deviation of these metrics over 55 independent trials. In particular, these tables highlight the performance of the Bayesian algorithm: a higher number of factors are recovered compared to the vanilla counterpart.

Algorithm	**Model**	**Sparsity** ($\times 10^{-2}$)	**Spread** s	**Localization** ρ
ICA	Vanilla	14.11 ± 0.66	34.46 ± 1.94	16.6 ± 4.45
BICA	Full	5.04 ± 0.39	22.68 ± 2.69	2.51 ± 1.73
	Vanilla	7.11 ± 0.71	22.36 ± 2.57	3.04 ± 1.60
	No localization	6.97 ± 0.70	22.03 ± 2.66	2.75 ± 1.40
	No structure	5.07 ± 0.36	22.83 ± 2.54	2.45 ± 1.46

Table 4.1: Comparison of the different algorithms' results on our synthetic problem

Algorithm	Model	Nb Recovered Factors	Mean top-5 λs	Accuracy
ICA	Vanilla	2.02 ± 1.13	NA	0.76 ± 0.094
BICA	Full	5.00 ± 0.00	1.20 ± 0.09	1.00 ± 0.001
	Vanilla	5.04 ± 0.19	0.37 ± 0.03	0.99 ± 0.002
	No localization	5.04 ± 0.19	0.35 ± 0.04	0.99 ± 0.002
	No structure	5.07 ± 0.13	1.28 ± 0.09	1.00 ± 0.001

Table 4.2: Comparison of the different algorithms on synthetic problem: Evaluation of the different components

We observe that the accuracy of the recovery is perfect for the Bayesian ICA model (both in the number of recovered factors and in their accuracy). On the other hand, the recovery of these factors is much noisier with traditional ICA, with an average best top-5 correlation hovering around 0.76. Similarly, the BICA components are 2.5 times more sparse and 1.56 more spread than ICA, thus varying more smoothly over the graph.

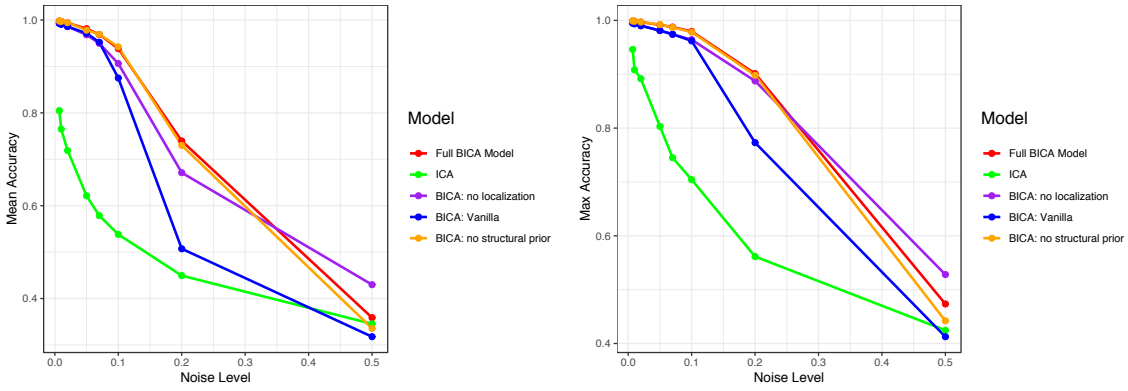
Robustness to Noise. We also evaluate the performance of the different models under varying levels of noise (the ϵ in Eq. 4.2.1). The results are displayed in Figure 4.6. In particular, this setting, we see that, up to reasonable amounts of noise, the Bayesian ICA model is far more accurate that the Vanilla ICA: the top-5 Recovered/GT pairs have a correlation that is up to 20% bigger than the Vanilla ICA model, indicating an extremely accurate detection of the underlying factors. We also see that the inclusion of structural priors (red and purple curves) yield components that are overall more accurate under varying levels of noise. The inclusion of both localization and structural priors (red curve) yields increased mean and max accuracies.

4.4 Validation: Real-life experiments

Our method having shown good promise on controlled synthetic examples, the goal of this section is to assess its behavior on real data, and if it can be used to process and produce interesting results for time series on graphs. We focus here on one particular test/retest fMRI study: the The Hangzhou Normal University (HNU1) dataset [160].

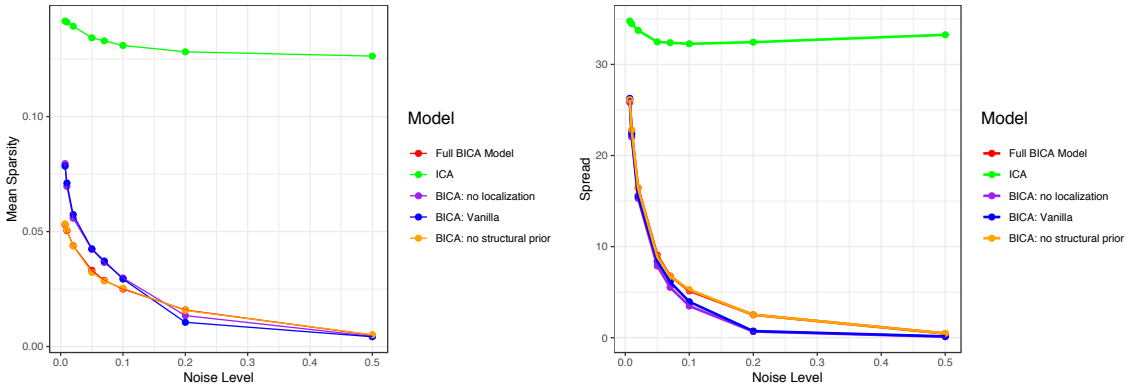
The Hangzhou Normal University dataset.[160] This dataset was gathered as part of a one-month Test-Retest Reliability and Dynamical Resting-State study. It consists of the fMRI of 30 healthy patients taken every 3 days across one month (each patient thus has roughly a total of ten scans). Five modalities (EPI/ASL/T1/DTI/T2) of brain images were acquired for all subjects. As per the dataset's website[2], "during functional scanning, subjects were presented with a fixation cross and were instructed to keep their eyes open, relax and move as little as possible while observing the fixation cross. Subjects were also instructed not to engage in breath counting or meditation.". For

[2] http://fcon_1000.projects.nitrc.org/indi/CoRR/html/hnu_1.html



(a) Mean of the top-5 pairwise correlations between recovered components and ground-truth factors.

(b) Max correlation between recovered component/ground-truth factors pairs.



(c) Sparsity (ℓ_1 norm).

(d) Spread

Figure 4.6: Analysis of the impact of the noise on the recovery of the components

the purpose of this study, we chose to use the Craddock 200-node atlas, which provides a manageable, yet relatively fine-scale representation of the brain. More details on the properties of this dataset, the study, the parameters of the scan as well as the fMRI pre-processing can be found in Appendix A. The structural matrices were obtained using pre-processed connectomes using NeuroData's MR graphs package, **NDMG** [**?**, 95][3]. More details can also be found in Appendix A. This dataset provides an extremely convenient framework to test the validity of our model: not only do we have functional and structural information for each scan, we can test for subject effects and robustness across subjetcs and sessions.

First results. Fig. 4.7 shows an example of the components that can be extracted from a given scan (session 1 for the subject 24527, the first of the cohort). As can be seen on Fig. 4.7a and 4.7g , our Bayesian model manages to capture components that are localized: only a few coefficients are non-zero within each component. By way of comparison, the Vanilla ICA components are much more scattered over the brain (Fig.4.7e). The Bayesian model also allows us to quantify the uncertainty associated to each coefficient, as displayed in Fig.4.7b: we see that the right Superior Frontal Gyrus for Subject 24527 is significantly involved in 8 components. In this particular case, the BICA model allows the incorporation of up to 18 components. Our method also allows a characterization of the uncertainty estimates for each node (Fig 4.7b): the Right Superior Frontal Gyrus for instance is only significantly activated in 8 components. All in all, this first visual inspection allows us to check that our method does indeed recover components that are aligned with our original hypotheses: sparse, connected and localized.

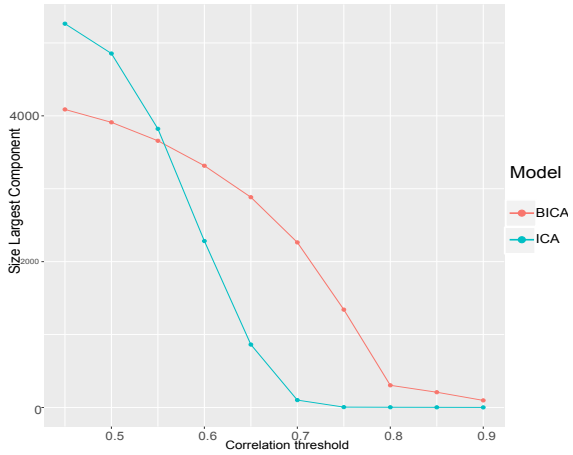
Robustness of the model. We further quantify the model's performance by evaluating its capacity to identify robust components across scans. To do so, we concatenate all components (as ordered by the λs) extracted from all 270 scans–thus yielding a total matrix of size 5400–, and we compute their cross-correlations.

Subject effects. The first step is to establish the existence (or lackthereof) of subject effects: it seems natural to assume that scans from the same subject should yield components that are generally closer together than across subjects. With this aim in mind, for each component, we find at its k-closest neighbors (where k is taken to vary from 1 to 10). We denote as $\mathcal{E}_k$ the set of all such selected pairwise correlations. To create a test statistic, we then compute the cumulative sum of the correlations over all pairs $(i, j) \in \mathcal{E}_k$ belonging to the same subjects:

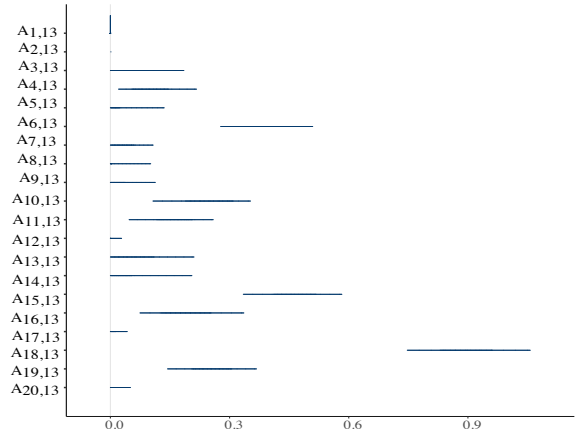
$$s_k = \sum_{(i,j)\in\mathcal{E}_k} \rho_{ij} \mathbb{1}_{G_i=G_j}$$

where G_i is the grouping induced by subject i. To compare this against a null distribution, we propose a Friedman-Rafsky test: we permute the labels, and compute the test statistic $s_k^{(\pi)}$ for the permuted data. In this scenario, a high value for s_k would be indicative of a strong subject effect,

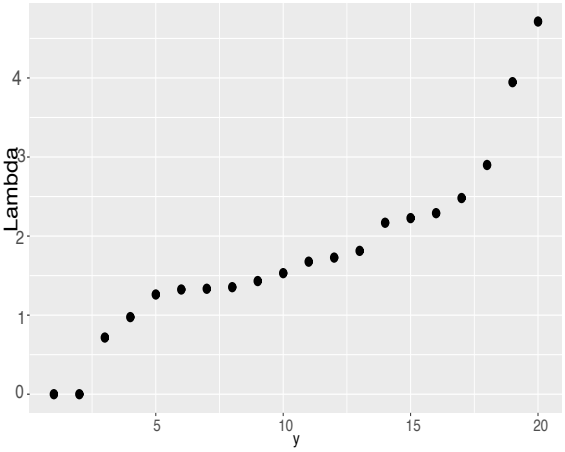
[3]The data and more information are available on NeuroData's website: `https://neurodata.io/mri/`



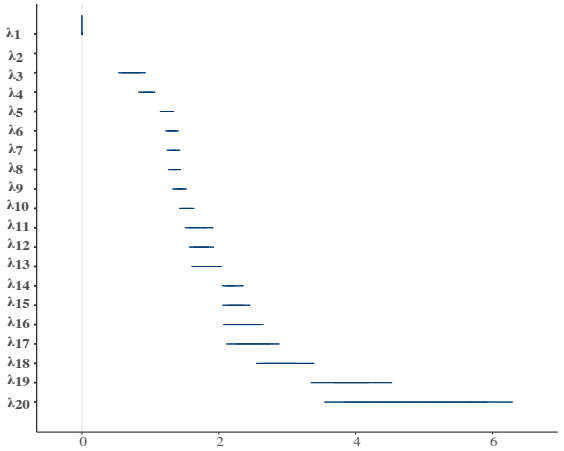
(a) Sizes of the connected components



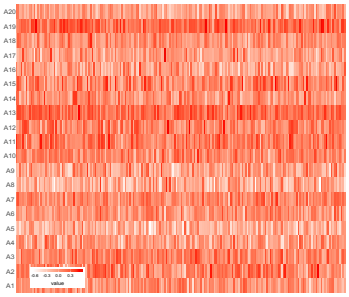
(b) Posterior Distribution of the activations of Node 13 per component (Right Superior Frontal Gyrus), with medians and 80% intervals.



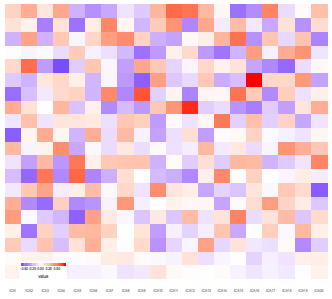
(c) Recovered Lambdas



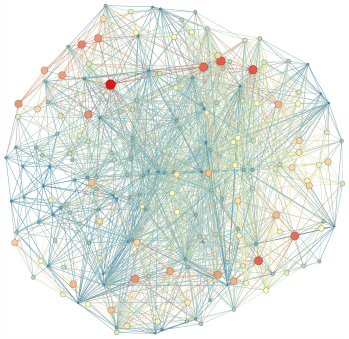
(d) Credible intervals for the Λs, with medians and 80% intervals.



(e) "Vanilla ICA" components



(f) Correlation between BICA and Vanilla ICA



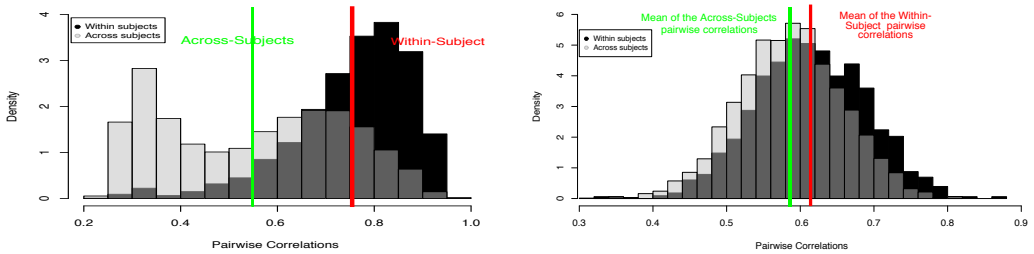
(g) Example of a given component (comp 20)

Figure 4.7: Results for the HNU1 dataset

since s_k increases with both the strength of the correlation and the inherent number of such pairwise interactions. For $k = 1$ (nearest-neighbor test), we observe that:

- **There exists a strong subject effect**: with $s_k = 1072.31$ and an average correlation of 0.78 between nearest neighbor pairs belonging to the same subject (compared to 0.68 for pairs corresponding to different subject), the Friedman-Rafsky test is overwhelmingly significant: nearest neighbor-pairs are much closer if they belong to the same subject than across subject (Fig.4.8a. We note that this effect is also present–even if less overwhelming—for the Vanilla ICA model (average correlation of 0.60 for within-subject pairs, and 0.56 across pairs, Fig. 4.8b). Note that the number itself of within-subject correlation edges does not come out as significant in either model.

- **The most correlated components are from a different session:** given a set of "same subject" pairwise correlations, we want to establish of these correlations are from the same scan or if they come from different sessions. The later indeed indicates that the model is able to identify robust subject-specific components. We thus permute the labels of the scans for each of these "within-subject" pairwise correlations. Interestingly, we note that the test is significant for our Bayesian ICA model, but not for the Vanilla ICA.

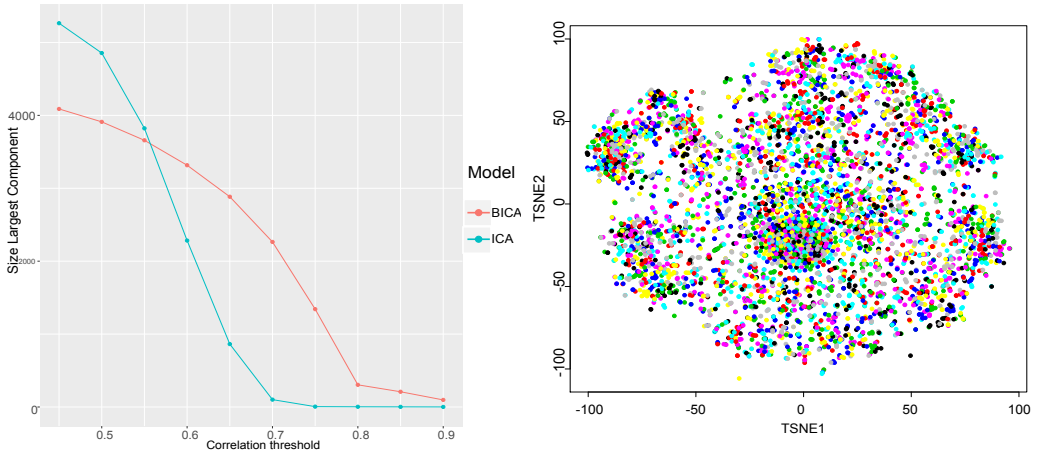


(a) Strongest Pairwise Correlations (Bayesian ICA) (b) Strongest Pairwise Correlations (Vanilla ICA)

Figure 4.8: Comparison of the distributions of the closest neighbor similarity.

Robustness across scans. Finally, we wish to evaluate the model's ability to detect components that are robust across scans and across individuals. To do so, we consider the cross-correlations across components. For a given set of thresholds $\eta \in [0.45, 0.95]$, we create a binary similarity matrix between components by setting all correlations above η to 1, and below to 0. We investigate the size of the connected components that this binary similarity yields. We first draw a TSNE plot of the different loadings (Fig. 4.9b), where each color denotes a subject. We do not observe a clustering by subject, which is reassuring, as the goal of the method is to extract components that generalize across subjects and sessions. The results are displayed in Fig. 4.9. Figure 4.9a shows in particular that the largest connected component for very high correlations is big (304 components at threshold

0.8, vs 3 for the Vanilla ICA). This shows that the Bayesian model has successfully identified robust components across scans and subjects. We represent each of these connected communities of loadings by its mean: the average top-5 connected components at threshold 0.8 are displayed in Fig.4.10-4.13. Interestingly, we notice that these average most robust loadings are extremely localized: the first one is located in the pre-frontal cortex, the second in the sensory motor cortex. Interestingly, the fifth component defines a network that appears to be in line with the robust connectome findings in [**?**].



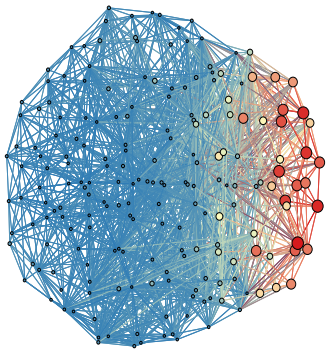
(a) Size of the largest connected component

(b) TSNE plot of the recovered loadings, colored by subjects.

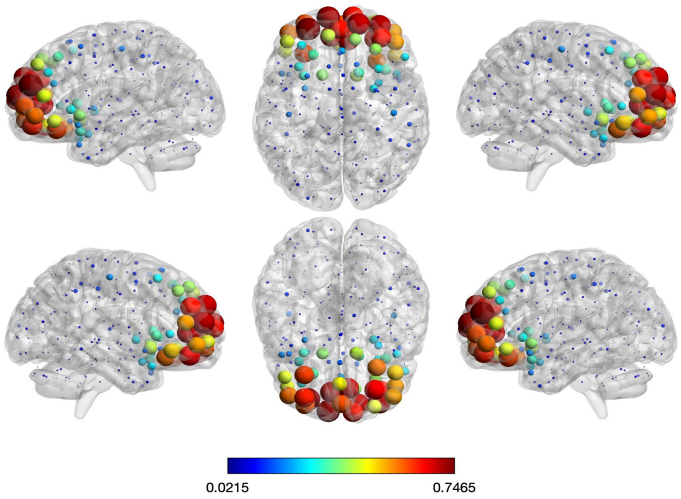
Figure 4.9: Robustness across sessions.

4.5 Conclusion: advantages, disadvantages and further research directions

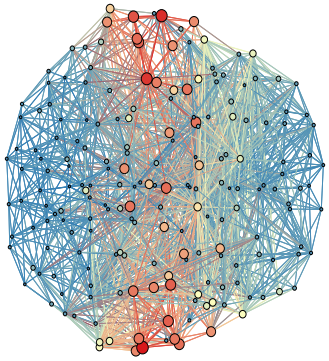
We have proposed here a version of Bayesian ICA that allows the recovery of biologically interpretable, connected subcomponents from time series that lie on a graph, with automatic self-tuning and minimal user input while providing uncertainty estimates on the recovered loadings. We have also provided a fast numerical approximation of point estimates. Our experiments on synthetic graphs validate the theoretical framework and show that submodules of co-activated nodes can be accurately recovered. The application of our method to real-life data seems to indicate that such an approach does indeed promote reliability and reproducibility among different connectomes, as well as biologically connected components– a necessary ingredient in brain connectome studies, where the main objective is to relate such subnetworks to cognitive processes.



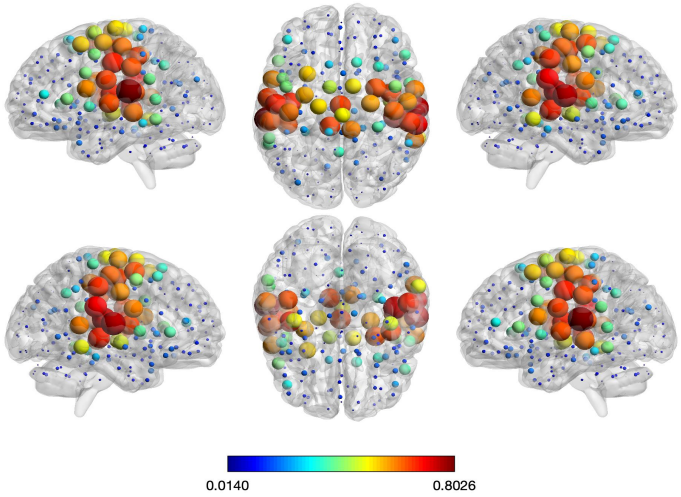
(a) Largest Connected Component



(b) Largest Connected Component (3D view)

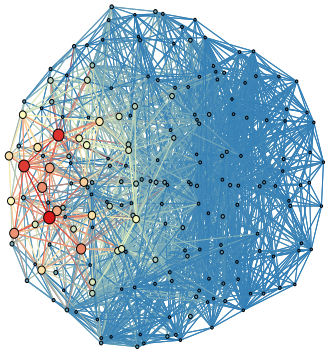


(c) 2^{nd} Largest Connected Component

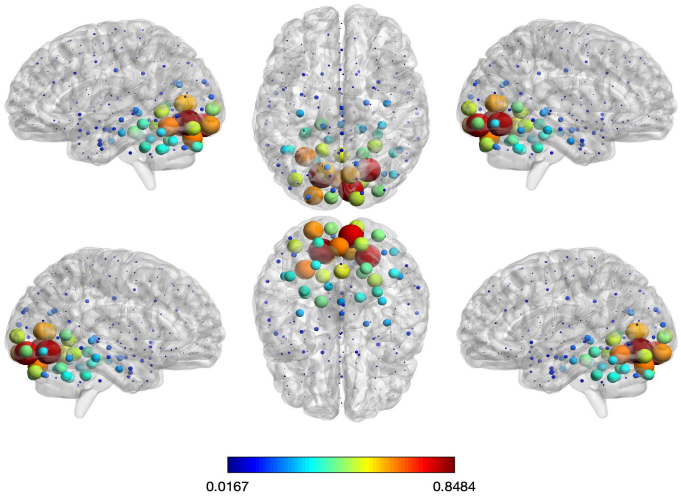


(d) 2^{nd} Largest Connected Component (3D view)

Figure 4.10: Mean of the first connected components

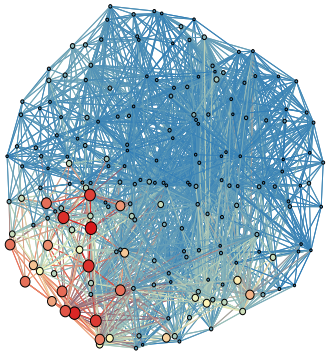


(a) 3^{rd} Largest Connected Component

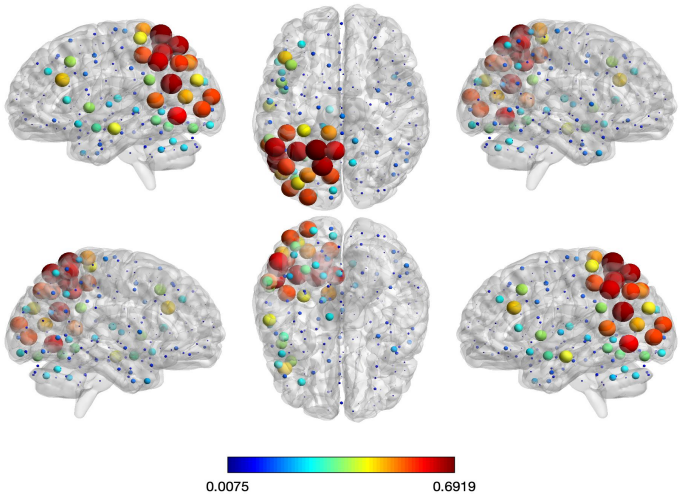


(b) 3^{rd} Largest Connected Component (3D view)

Figure 4.11: Mean of Component 3

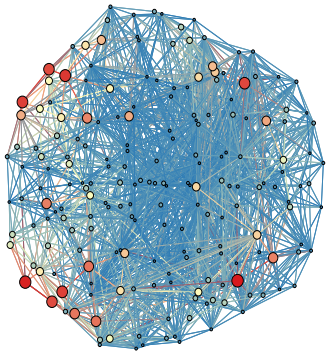


(a) 4^{th} Largest Connected Component

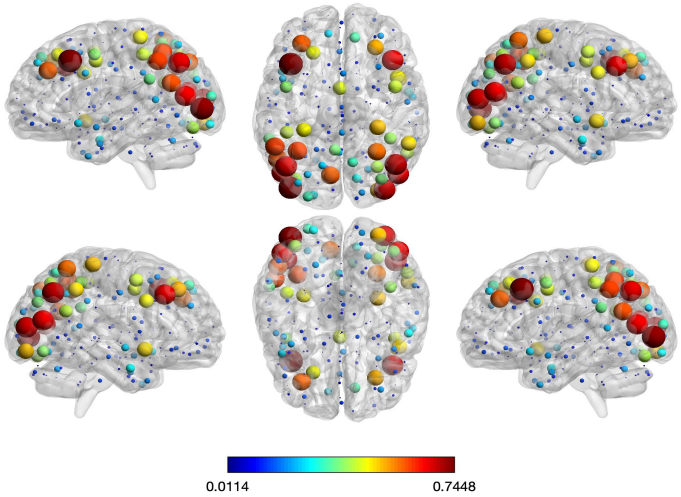


(b) 4^{th} Largest Connected Component (3D view)

Figure 4.12: Mean of Component 4



(a) 5^{th} Largest Connected Component



(b) 5^{th} Largest Connected Component (3D view)

Figure 4.13: Mean of Component 5

Chapter 5

Inference with Latent Networks: a COVID-19 case study

The previous chapters tackled the problem of uncertainty quantification and variability estimation on sets of aligned networks, whether these graphs were observed or latent. We propose to open the discussion from inference *for* graphs to inference *with* graphs. That is, in a number of applications, the data that we observe lives on a graph, whose (potentially unobserved) structure in crucial in understanding the behavior of the system. The practical motivating application for this new facet of inference for graph-structured data is to understand the link between the inherent contact network between individuals in a society and the mechanisms of disease propagation — especially in light of the COVID-19 epidemic, still unresolved at the time of writing.

Indeed, the Coronavirus' reproductive number R, which characterizes the average number of secondary cases generated by each primary case, takes on a significant importance in the quantification of the potential scope of the pandemic. Yet, in most models, R is assumed to be a universal constant for the virus across outbreak clusters and populations — thus neglecting the inherent variability of the transmission process due to varying population densities, demographics, temporal factors, etc. — and in particular, due to the inherent variability of the degree distribution in the social contact network. In fact, it can be shown that this degree distribution is skewed, thus potentially leading to a highly variable reproductive number. Foregoing this inherent variability in R and considering its expected value in contagion models thus leads to biased or loose results in the reported predictive scenarios, especially as these are tailored to a given country or region.

The goal of this chapter is to thus lay the foundations to the examination of (a) the impact of the reproductive number R's variability due to the heterogeneity of the contact network on important output metrics, and (b) the effect of percolation of this variability in projected scenarios so as to

provide uncertainty quantification. In this perspective, in the absence of contact network estimates, we propose a Bayesian approach to model this variability: instead of considering a single R, we consider a distribution of reproductive numbers R and devise a simple Bayesian hierarchical model that builds upon current methods for estimating the R to integrate its heterogeneity. We then simulate the spread of the epidemic, and the impact of different social distancing strategies using a probabilistic framework that models hospital occupancy. This shows the strong impact of this added variability on the reported results. We emphasize that our goal is not to replace benchmark methods for estimating the basic reproductive numbers, nor to devise accurate predictive scenarios, but rather to discuss the importance of the impact of R's heterogeneity on uncertainty quantification for the current COVID-19 pandemic.

5.1 Motivation: contagion, graphs and heterogeneity

First detected in Wuhan (Hubei Province, China) in December 2019, the current COVID-19 pandemic has thrown the entire world in a state of turmoil, as governments closely monitor the spread of the virus and have taken unprecedented measures to contain local contagion outbreaks. In order to adequately inform public policy makers, experts in epidemiology are currently trying to assess the potential scope of this global pandemic and to draw predictive scenarios. Standard epidemiological research uses the basic reproductive number R_0 as the key parameter in almost all contagion models — whether these scenarios are drawn using variants of the Susceptible-Exposed-Infected-Removed (SEIR) deterministic equations [76, 93, 152, 120] or of exponential growth models [158].

By definition, the basic reproductive number R_0 characterizes the expected number of secondary cases directly produced by one single typical infectious case in a population of completely susceptible individuals. To give more intuition on the underlying transmission mechanisms that it captures, R_0 can be decomposed as the product of three terms [36]:

$$R_0 = \tau \bar{c} D_I \tag{5.1.1}$$

where τ is the transmissibility (i.e., probability of infection given contact between a susceptible and infected individual), $\bar{c}$ is the average number of contact per day between susceptible and infected individuals, and D_I is the duration of infectiousness — that is, the number of days during which an infected patient can be expected to contaminate others. R_0 thus serves as an epidemiological metric to describe the contagiousness or transmissibility of infectious agents: the outbreak is expected to continue if R_0 is greater than 1, or to naturally subside if R_0 is strictly less than 1. As recently highlighted by Delameter et al. [41], this coefficient inherently depends on some local characteristics of the population. In particular, going back to the decomposition provided in Eq. 5.1.1, R_0 is intrinsically tied to temporal and spatially-varying factors, such as population age demographics,

political or environmental variables, cultural or social dynamics, or the density of the population — all favoring or diminishing the rate of contacts $\bar{c}$ between individuals. R_0 can thus be naturally modeled using a hierarchical framework, which accounts for the reproductive number's heterogeneity by decomposing it according to different strata. For instance, the reproductive number R could be hierarchically broken down according to countries (or regions), age groups, and, at the most granular level, across subjects. The expected number of secondary cases is indeed contingent on each primary cases' socio-economic status, age, etc., and perhaps even time — as one could imagine the contact rates varying between weekends and weekdays. A very fine-grain analysis of the R_0's heterogeneity would thus model R_0 as a distribution over cases and time in a given population. Figure 5.1 shows one such potential hierarchical stratification.

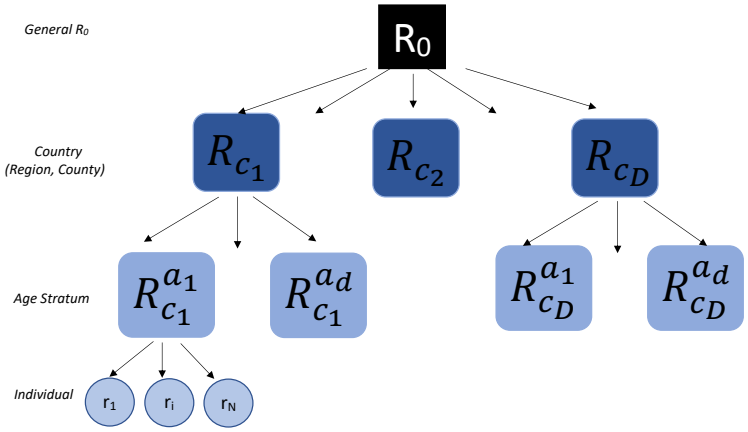


Figure 5.1: Hierarchical model for R_0

The "universal" R_0 used in epidemiological models to characterize the disease might thus be thought of as a general summary statistic, averaged over individuals and populations — thus discarding any form of local variability. The underlying assumption is that the dynamics of the pandemic are similarly described by the trajectory estimated using the average R_0, or the average of the epidemic's trajectories with varying R_0. Yet, because the number of new incident cases each day depends exponentially on the history of the trajectory, this averaging approximation might come at a huge accuracy cost in prediction models. To give a clearer picture of the potential effects of this additional randomness on the model, let us consider two naive experiments.

First Experiment: Inherent effect of the randomness on the model. In the first experiment, we consider a simplification of the exponential growth model for an epidemic. In this model, for a given reproductive number R, each new infectious case generates a Poisson(R) number of new cases the following day. This amounts to considering an instantaneous incubation period and that each

primary case is only contagious for day. At each time t, the number of new cases is thus generated as:

$$X_{t+1} = \text{Poisson}(RX_t)$$

We assume that exactly 1% of these incident cases will not recover from the disease, so that the cumulative number of deaths at time t can be written as $D_t = 0.01 \sum_{s=1}^{t} X_s$). Using an Anscombe transform, we consider $Y_t = 2\sqrt{X_t}$, and we know that the generation mechanism of Y_t can be approximated by a normal distribution:

$$\forall t, \quad Y_{t+1} = 2\sqrt{X_{t+1}} \approx \text{Normal}(2\sqrt{RX_t}, 1)$$

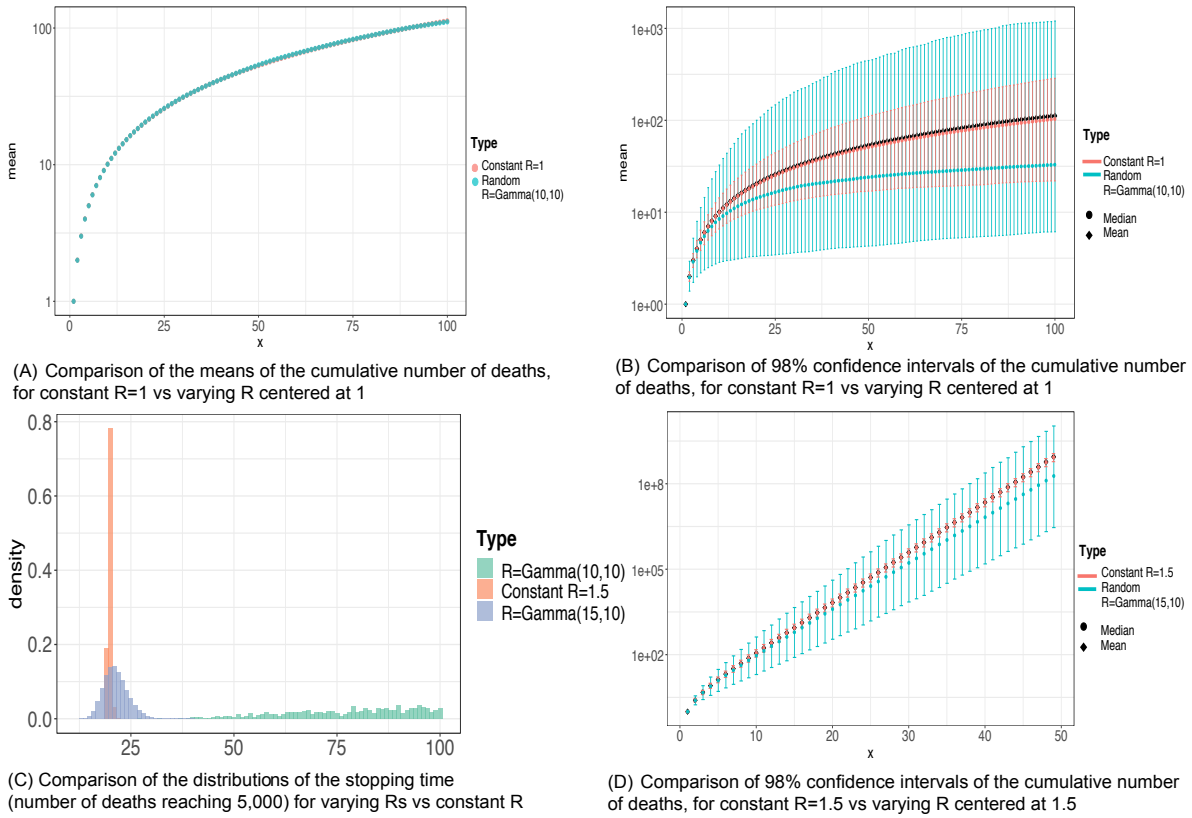
Assume now that R is a random variable, with expected value $R_0 = \mathbb{E}[R]$ and finite variance. Using Taylor's expansion around the mean R_0, we can thus write the probability of the number of new incident cases exceeding a given threshold η (or rather, its monotonous transformation to a domain spanning $\mathbb{R}$) as:

$$\begin{aligned}
f(\eta, R) &= \log(\frac{\mathbb{P}(X_{t+1} < \eta|R)}{\mathbb{P}(X_{t+1} > \eta|R)}) \\
&= \log(\mathbb{P}(z - 2\sqrt{RX_t} < 2\sqrt{\eta}\Big|R)) - \log(\mathbb{P}(z - 2\sqrt{RX_t} > 2\sqrt{\eta}\Big|R)) \\
&= \log(\Phi\Big(2(\sqrt{\eta} + \sqrt{RX_t})\Big)) - \log(1 - \Phi\Big(2(\sqrt{\eta} + \sqrt{RX_t})\Big)) \\
&= f(\eta, R_0) + \frac{\partial f}{\partial R}(\eta, R_0)(R - R_0) + \frac{\partial^2 f}{\partial^2 R}(\eta, R_0)(R - R_0])^2 + o((R - R_0)^2)
\end{aligned}$$

Thus, integrating with respect to R_0 yields:

$$f(\eta) = \log(\frac{\mathbb{P}(X_{t+1} < \eta)}{\mathbb{P}(X_{t+1} > \eta)}) = f(\eta, R_0) + \frac{\partial^2 f}{\partial^2 R}(\eta, R_0)\text{Var}(R) + o(\text{Var}(R)) \tag{5.1.2}$$

where $\frac{\partial^2 f}{\partial^2 R}(\eta, R_0) = -\sqrt{\frac{X_t}{R_0}} \frac{\Phi'(A_t)}{\Phi(A_t)(1-\Phi(A_t))} \Big[\frac{1}{2R_0} + \frac{\Phi'(A_t)(1-2\Phi(A_t))}{\Phi(A_t)(1-\Phi(A_t))} A_t\Big] > 0$ and $A_t = 2(\sqrt{\eta} + \sqrt{R_0 X_t})$. Note that the latter is simply a constant that only depend on R_0 and the threshold η considered. In particular, this factor is equal to 0 for $A_t = 0$ and is an increasing function of η. This Taylor expansion allows us to conclude that by considering R to be constant and equal to its expectation $R_0 = \mathbb{E}[R]$, for high values of η, our approximation will be off by a term of the order of the variance of R_0. This is potentially quite problematic: this error is exponentially propagated throughout the projected trajectory for the epidemic, thus potentially leading to (a) an increased variability of the number of deaths $D_t = 0.01 \sum_{s=1}^{t} X_s$ and (b) the introduction of new "worst cases scenarios". Consider for instance the stopping time corresponding to the total number of deaths reaching 5,000: $\tau = \min\{t \in \mathbb{N} : D_t \geq 5,000\}$. Since the model is build sequentially, it is difficult to get close form formulas and exact probabilistic results. We thus resort to simulating 40,000 contagion events over



(A) Comparison of the means of the cumulative number of deaths, for constant R=1 vs varying R centered at 1

(B) Comparison of 98% confidence intervals of the cumulative number of deaths, for constant R=1 vs varying R centered at 1

(C) Comparison of the distributions of the stopping time (number of deaths reaching 5,000) for varying Rs vs constant R

(D) Comparison of 98% confidence intervals of the cumulative number of deaths, for constant R=1.5 vs varying R centered at 1.5

Figure 5.2: Output of simulations showing comparisons of the possible trajectories for contagion models using fixed R_0 vs variable R. Dots indicate the average predicted values, whereas the error bars represent the 98%-confidence interval.

100 days to quantify the impact of this added variability. In one case, we simulate the propagation of the epidemic using a fixed, constant R_0. In the other, we simulate the propagation assuming that at each time step, R is sampled from a gamma distribution centered around R_0. We compare the effect of the two models on the evolution of the number of deaths and the stopping time τ. The results are reported in Figure 5.2.

Based on those simulations, we make two observations: (a) while the mean number of deaths is roughly the same for both scenarios (Fig 5.2A), the distributions are substantially different (Fig 5.2B,D). In particular, the worse-case scenario (99th quantile) is bigger by orders of magnitude when considering a variable R, with respect to its constant counterpart. This is an important observation: average predictions for the fixed and variable R models are seemingly the same, yet their associated uncertainty estimates and catastrophic scenarios are radically different. Moreover, for constant

$R_0 = 1$, the stopping time $\tau = \min\{t \in \mathbb{N} : D_t \geq 5,000\}$ is never reached. It is nonetheless reached in 0.19% of cases using a varying R (Fig 5.2C), thus making it a non-zero probability event and enlarging the space of possible events. The variable-R model thus presents a wider scope of worst-case scenarios than the ones predicted using a constant, average R_0 — a fact that is potentially crucial for policy makers to make informed decisions.

Second Experiment: effect of the randomness on the estimation procedure. We have shown that a constant R_0 might not be satisfactory from the model's perspective – we now also assess how the error induced by the averaging is also reverberated in the estimation procedure. In this second experiment, we simulate an exponential growth of the number of incident cases over the course of 20 days using a Gamma distributed R with shape 1.2 and scale 1. This is in fact mimicking a scenario under which R varies every day, thus accounting for some temporal effects (weekend vs week days), subject-effects across newly infected cases, etc. Let us now try to recover the R_0 using the Exponential Growth model in the `R0` `R`-package. The average difference between the recovered and true mean R_0 over 1,000 simulations is 2.94 (with only 8.5 % coverage by the recovered confidence intervals)— that is, more than twice the 1.35 mean error that we obtain by simulating data under the same setting using a constant $R_0 = 1.2$. This brings to light two new observations: (a) standard R_0 estimation procedures seem to perform even less well with variable R_0, and (b) the confidence intervals usually provided are too narrow, and do not correctly emphasize the high uncertainty of the predicted R_0 value.

In light of these synthetic experiments, assuming the reproductive number R_0 to be constant thus comes at a huge cost in terms of accuracy of the reported predictive scenarios. In particular, the worst-case scenarios associated to these predictions could be either (i) too optimistic without appropriately characterizing their uncertainty, (ii) unable to account for the existence of "super-spreaders" in the general population, and (iii) fail to allow certain rare events leading to the formation of outbreaks — thus potentially misleading policy makers and begging the question: for the analysis of our real data, how much variability do we need to account for in the modeling of R_0? Is the aggregated version sufficient to provide informative scenarios, or is a hierarchical model preferable?

From a statistical viewpoint, the issue of accounting for the R_0's variability can be tackled from two perspectives: (a) a "spatial random-effects" approach, splitting the data into groups in a hierarchical fashion and assuming the R_0 to be constant within groups — the underlying assumption being that the within-group variance is significantly smaller than the global variance, and that this stratification thus reduces the error in the estimation of the uncertainty — or (b) a "full random-effects" model, incorporating both spatial and temporal variability. The latter builds upon the spatial random effects model, but adds a further layer of variability by assuming a random daily effect. In other words, whereas in the "spatial random-effects" approach, the R is assumed to vary across groups

but is held constant throughout the trajectory, in the "full random-effects" model, R varies at each time step. This assumes that the inherent daily variations of the reproductive number Rs are too substantial to be neglected. In both cases, a more granular estimation of the R_0 using geographical, weekday, weather, and other sources of information could make day-to-day variations in the R_0 provide more realistic epidemiological predictions of the outbreak propagation speed, as well as the expected times before hospitals reach capacity — both crucial quantities for informing policy makers as they arbitrate between different courses of action, especially as drastic public health measures typically come at significant social and economical costs. We emphasize that our goal here is not to come up with a new model or definition for the R_0, nor to pretend to a better predictive model than experts in epidemiology. Rather, our focus is simply to assess – as statisticians – the effect of this added variability in predictive scenarios, in order to grasp a little better how this variability is propagated in downstream analyses. One of the hypotheses that we would thus like to test is if the heterogeneity of the R_0 coefficient can severely impact predictive scenarios for the outbreaks: how certain are we about the predictions that we are making? In light of the observed heterogeneity of the R_0's, how confident are we of the transferability of a given policy in one country to another?

In this paper, we deal with stochasticity and limited/missing data using a Bayesian perspective. We begin by describing the Bayesian hierarchical model that we use to estimate the varying reproductive number R. This approach provides a more natural framework for uncertainty quantification through the provision of credible intervals. We show the impact of this variability on the predictive scenarios and the effect of public policy measures (e.g. social distancing or alternating lockdown days) that can be drawn using these models. All of our experiments here are deployed on the current COVID-19 pandemic. The code and data used for this analysis are openly available on the authors' Github[1].

5.2 Model and Theory: towards a heterogeneous R

Our evaluation of the effect of heterogeneity in predictive scenarios comprises of two steps, whose details are provided separately in the two following sections. The first consists in estimating the number of new cases (or incidence cases) per day, using a model for the heterogeneity of the reproductive number R. The second step models the hospital and ICU occupancy as a consequence of the surge in incident cases, which we detail in section 5.3. We propose to account for the heterogeneity of the reproductive number by dividing the data into geographical groups (Level 1 of the hierarchical framework presented in Figure 5.1), and compare the results for the spatial random-effects model with the ones obtained for the full random-effects model.

Modeling an heterogeneous incidence proportion. We base the first step of our analysis on a simple Bayesian hierarchical extension to models currently deployed throughout the literature to

[1]Code and data at: `https://github.com/donnate/heterogeneity_R0`

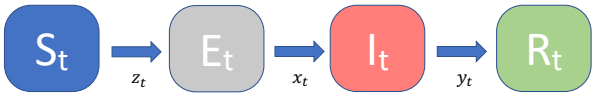


Figure 5.3: Compartmental SEIR model

compute the R_0. The Bayesian formalism is indeed particularly amenable to uncertainty quantification and modeling with limited information (as we track the epidemic at various stages of progression across locations) through the use of priors and posterior credible intervals.

Throughout the remainder of this analysis, let us adopt the following formalism. Let G be the number of groups that we want to analyze (these could either be localized virus outbreak clusters , regions or countries). Let N_g denote the population of each of these groups, initially assumed to be completely susceptible. For the sake of simplicity, we neglect the number of births, natural deaths, and incoming/outgoing flux between groups.

Literature Review. The classical model for predicting the spread of an epidemic within each group is the Susceptible-Exposed-Infected-Removed (SEIR) compartmental model, which was used to estimate COVID19's R_0 in its early days [152]. In this setting, each group of size N_g is split in one of four different compartments (see Figure 5.3): people are either susceptible, exposed, infected (which we understand as exhibiting symptoms) or removed (including recoveries and deaths). The SEIR is thus a deterministic model, in which the evolution of the populations in each compartment is modeled through a set of differential equations:

$$\begin{aligned}
\frac{dS_k(t)}{dt} &= -\frac{S_k(t)}{N_k}\frac{R_0^{(k)}}{D_I}I_k(t) \\
\frac{dE_k(t)}{dt} &= -\frac{S_k(t)}{N_k}\frac{R_0^{(k)}}{D_I}I_k(t) - \frac{E_k(t)}{D_E} \\
\frac{dI_k(t)}{dt} &= \frac{E_k(t)}{D_E} - \frac{I_k(t)}{D_I}
\end{aligned} \tag{5.2.1}$$

where:

- $S_k(t), E_k(t), I_k(t)$, and $R_k(t)$ are the number of susceptible, latent, infectious, and removed individuals at time t in group k;
- D_E and D_I are the mean latent (assumed to be the same as incubation) and infectious period (equal to the serial interval minus the mean latent period);
- $R_0^{(k)}$ is the basic reproductive number is population k.

The main issue with this deterministic set of equations lies in the fact that it does not provide any natural uncertainty quantification for estimates of R_0, nor any Maximum Likelihood Estimate formulation of the R_0 coefficient — and thus, provides no natural notion of uncertainty, especially given that all the parameters that are fed into these equations are (informed) guesses, that come with their own level of uncertainties. Some studies have introduced stochastic components in SEIR models, for instance in the study of Ebola [104]. However, it is not standard to take into account the heterogeneity of the basic reproductive numbers R_0 — thus potentially hindering the realism of their predictive scenario.

Model. Here, we build upon a simplification of the compartmental model. The heterogeneity of R_0 is modeled using a Bayesian hierarchical workflow. Our model is based on the non-parametric model by Fraser [61], also used for estimating R_0 in Cori et al [32]. A version of this model was implemented in the `R`-package `EarlyR` [139], which has been used in recent studies[158] to infer COVID 19's R_0. Instead of explicitly modeling the exposed and infected periods separately –based solely on the number of new infections– this model foregoes the modeling of latent cases and relies solely on inferring the number of new cases from previous observations using an "infectivity profile" [32]. In this setting, each infected case is expected to contaminate on average of R_0 patients (by definition) — but the distribution of this number of new infections is given by a probability distribution which only depends on the time s elapsed since infection: one could indeed imagine a patient becoming increasingly contagious over the first few days of the infection as the viral load builds up, and decreasingly so after the peak of the illness. This infectious profile is typically modeled as a Gamma distribution. Since this quantity is generally unknown and hard to estimate from available data, Cori et al [32] propose the use of the parameters of the serial interval (for which we typically have much more substantial observational data and means of estimation) as a good proxy. The only drawback of this model compared to the SEIR compartmental one is that the exponential growth phase is only valid for the first stages of the epidemic (on shorter timelines), and will thus not yield informative scenarios in longer time horizons (several years) — but it does provide a valid estimate of R_0 that we can then plug in as parameter in any deterministic model. Our goal is to assess the toll of hospital load that a rapidly propagating pandemic can induce. As such, we emphasize that we focus on **short-term estimation**, and the study of the uncertainty for time frames of a few weeks, rather than months — since it could also be argued that due to the dynamic nature of the problem and rapidly changing policies, scenarios for months or years ahead are extremely hard to devise.

Moreover, in our setting, we focus solely on the uncertainty on R, which we assume to have a distribution over space and time. That is, we assume the parameters of this serial interval to have been correctly estimated and thus, the coefficients w_s to be known. We discuss in the appendix how to add some uncertainty to these parameters.

We call X the number of incident (new infectious) cases each day. The incidence on day t

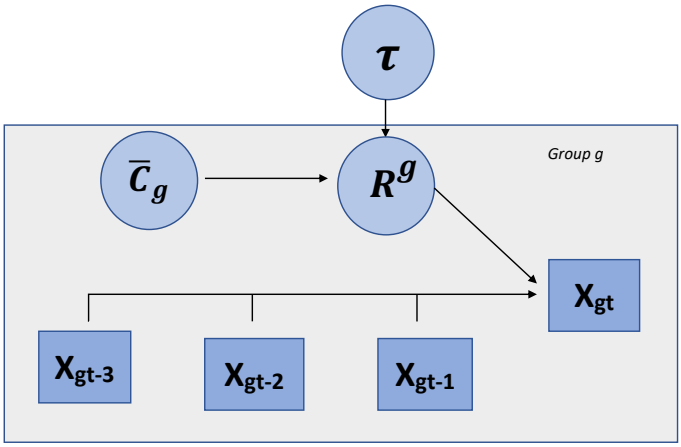


Figure 5.4: Plate model

conditioned on the previous incidences can be modeled by a Poisson distribution of the form:

$$\forall t \leq T, \quad X_t \sim \text{Poisson}(R_0 \sum_{s=1}^{t-1} w_s X_{t-s}) \quad (*)$$

where $w_s = \mathbb{P}[\Gamma_{\alpha,\beta} \in (s, s+1)]$.

Here, we assume a hierarchical structure following Eq. 5.1.1 for R, which takes into account the inherent spatial variability and decomposes it as the product of the transmissibility τ, the daily contact $\bar{c}_g$ and the duration of individual infections D_i — which we assume to be known. We propose to assume here the transmissibility to be common across all groups, and the rate of contact $\bar{c}_g$ to be group-specific, as it is intrinsically tied to locally varying parameters (age demographics, social and cultural habits and current local policies), etc. This decomposition also provides a convenient way to model the effect of the measures deployed by governments to try to control the spread, which effectively targets diminishing the reproductive number R by reducing the daily contact rate $\bar{c}$. Here, we propose to model the transmission rate as being drawn from a Beta distribution, while we assume a Gamma distribution for the $\bar{c}_g$s. The generating mechanism is summarized in the plate model provided in Figure 5.4.

One of the issues with this model is the one of identifiability: the product $\bar{c}\tau$ is invariant by rescaling of the two factors. To get rid of this identifiability issue, we propose to adopt a similar strategy as in classical logistic regression examples: we fix the first group's daily contact rate $\bar{c}_1$ to a fixed value — we pick it here to be 1. Intuitively, this assumes that any infected person in population 1 to be in average in contact with one susceptible person per day. All other values of $\bar{c}$ can thus be understood as relative measures with respect to this benchmark group — thus a $\bar{c}_2$ with value 2 would indicate that, on a daily basis, an infected individual in population 2 has twice as many

contacts in the susceptible group than in population 1. This benchmark value could be either an arbitrary benchmark value (which should allow the potential R to vary within reasonable ranges), or an informed measure of social interactions — for instance, a daily contact of one person per day might seem like an appropriate value for a population in complete lockdown, such as seen in Wuhan as of January 22^{nd}.

The model is summarized below:

$$\begin{aligned} \forall t \leq T, \forall g \leq G, \quad X_{t,g} &\sim \text{Poisson}(R^{(g)} \sum_{s=1}^{t-1} w_s X_{g,t-s}) \\ \forall g = 2 \cdots G, \quad \bar{c}_g &\sim \Gamma(2,1) \\ \tau &\sim \beta(1,39) \\ R^{(g)} &= \bar{c}_g \tau D_I \end{aligned} \tag{5.2.2}$$

The full random-effect version of this model follows the same framework, except, instead of considering a fixed $R^{(g)}$, at each time step, the effective reproductive rate $R_t^{(g)}$ is sampled from a gamma distribution:

$$R_t^{(g)} \sim \Gamma(R_0^{(g)} * 10, 10)$$

We provide in Appendix D.2 a formal comparison of this hierarchical framework with the results provided by the methodology detailed in Cori et al [32] when R_0 is fitted on the whole data, or respectively independently on each group.

Model Fitting and Validation. To fit this hierarchical model, we use the `RStan` programming suite[19]. This uses Hamiltonian Monte Carlo to generate samples and estimate the different parameters of the model. We use a total of 8 chains, with 5,000 warmup iterations and 1,000 sampling steps. All the associated code and data are provided on Github[2]. Appendix D.1 provides a set of synthetic experiments that we use to benchmark the accuracy of our method.

Analysis of the COVID-19 Data

We now deploy our approach on the 2020 COVID-19 pandemic dataset openly provided by the Center for Systems Science and Engineering at Johns Hopkins University[3]. The goal of this subsection is to deploy our approach to the analysis of large geographical groups, where we expect social and environmental factors to vary substantially — and thus, the R_0 to exhibit a high amount of variance. In most countries outside of China, the past few weeks have seen the critical surge of the pandemic. In contrast, current reports have shown that the epidemic is slowing down in mainland China. We

[2] `https://github.com/donnate/heterogeneity_R0`

[3] Data openly available: `https://github.com/CSSEGISandData/COVID-19`

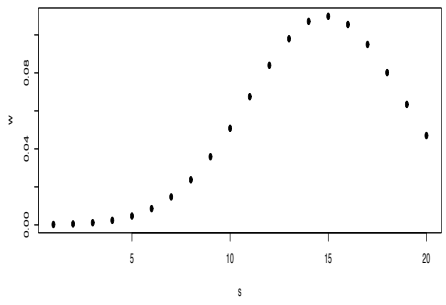
thus propose to focus on the analysis of the period from February 9th to March 17th (where the epidemic has steadily grown in the world without – for the most part — massive social distancing or public policy measures). We use the next five days (March 18th to 22nd) for validation .

We consider a total of 19 distinct geographical groups, spread around the world in order to gauge the amount of variability shown in the reproductive number R:

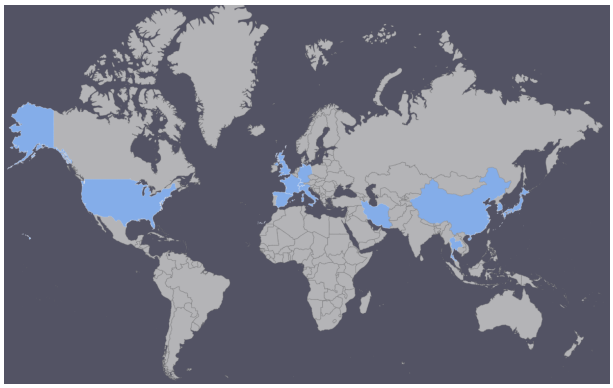
- the six countries reporting the highest numbers for the epidemic in Europe (Italy, Spain, France, Germany, the United Kingdom and Switzerland),
- seven groups in Asia (Hong Kong, the Chinese provinces of Guizhou and Hubei, Singapore, Thailand as well as Japan and South Korea),
- Iran,
- and finally, the United States as a whole, as well as the states of California, Washington and New York.

These groups are highlighted on the map in Figure 5.5b.

For most of these groups, as shown in the plots in Figure 5.6 , the epidemic still seems to be in its early stages and growing exponentially. To contrast it with later stages in the epidemic, we also fit the model separately to the data from the provinces of Hubei for first 36 days after the beginning of the quarantine (starting January $22^{nd}, 2020$) — a group that we refer to as "Hubei 0". This group is taken to be our reference group (we assume that in this case, the number of daily contacts is 1).



(a) Infectious Profile (w_s)

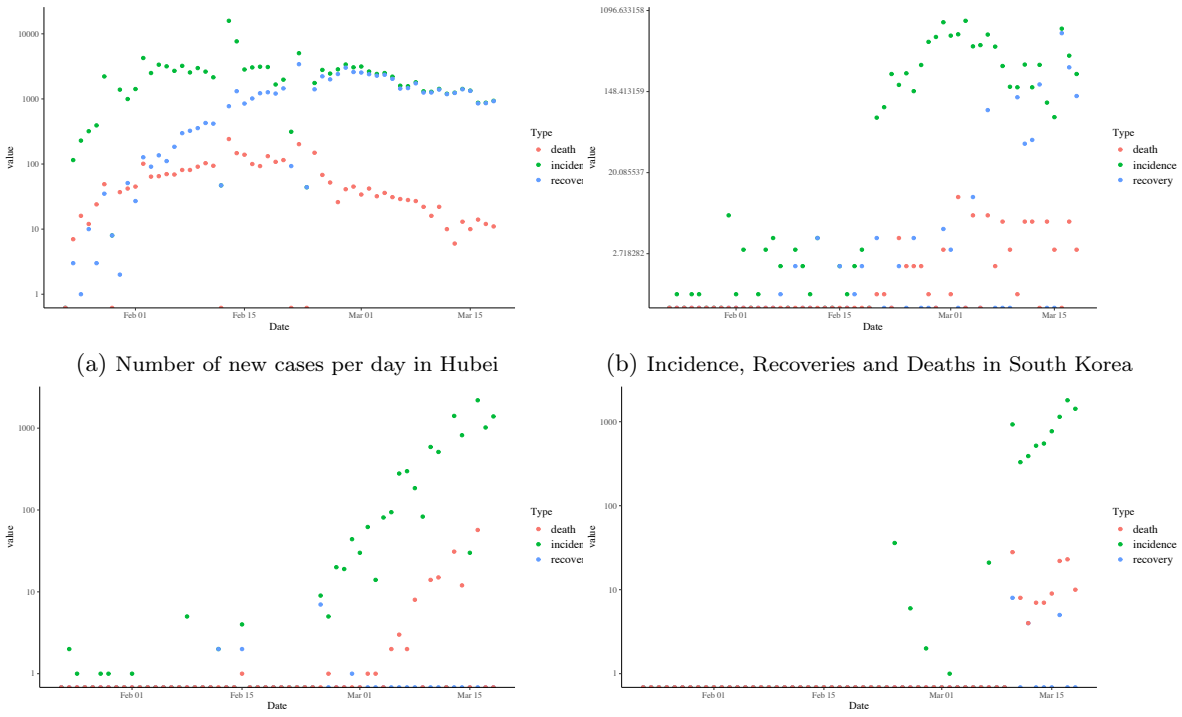


(b) Localization of the different groups selected for the analysis.

Figure 5.5: Parameters (infectious profile and selected groups) chosen for the analysis.

Preliminary Exploratory Data Analysis.

The plots shown in Figure 5.6 show the time series for a few of our selected epidemic groups, and highlight the need to make our analysis more robust. Indeed, the exact date of the onset of the



(a) Number of new cases per day in Hubei

(b) Incidence, Recoveries and Deaths in South Korea

(c) Incidence, Recoveries and Deaths in France

(d) Incidence, Recoveries and Deaths in the US

Figure 5.6: Incidence data for some of our selected groups

epidemic in each group is uncertain, as is shown on those plots by the fact that the different countries are currently at different stages of the epidemic. In particular, we note the substantial lag of the United States with respect to Europe — though the sudden spike in number of reported cases seems to indicate an under-reporting of previous cases, potentially due to lack of testing. To account for this under-reporting of the number of early cases, we introduce a random variable $\mu_0^{(g)}$ modeling these phantom, unreported cases. We also observe the existence of what seems to be (a) different times of onset of the epidemic across the various groups and (b) erroneous reporting– as the increments in the number of deaths are sometimes negative. We correct the latter by thresholding the increments at 0 — that is, we assume that these negative increments are due to an error in the reporting. For the first, we propose a slight adaptation of the Poisson Model proposed in the previous subsection. To model the different onsets of the epidemics, we introduce switching dynamics through a variable $\theta_g \in [1...36]$, indicating the time of the onset of the exponential growth in the corresponding group.

We use a variance stabilizing transform: instead of directly modeling Poisson counts, we perform an Anscombe transform of the data:

$$Y_{gt} = 2\sqrt{X_{gt} + \frac{3}{8}}$$

which has the effect of ensuring:

$$Y_{gt} \approx \mathcal{N}(2\sqrt{R^{(g)}\sum_{s=1}^{K} w_s X_{g,t-s} + \frac{3}{8}}, 1)$$

The generative model is summarized below and in Fig. 5.7:

$$\begin{aligned}
\forall \theta_g \leq t \leq T, \forall g \leq G, \quad Y_{t,g} &\sim \mathcal{N}(2\sqrt{R^{(g)}\sum_{s=1}^{K} w_s X_{t-s,g} + \frac{3}{8}}, 1) \\
\forall g = 2 \cdots G, \quad \bar{c}_g &\sim \Gamma(2,1) \\
\tau &\sim \beta(1,29) \\
\forall k, \mu_{g,k}^{(0)} &\sim \text{Gamma}(50,1) \\
R^{(g)} &= \bar{c}_g \tau D_I \\
\theta_g &\sim \text{Unif}(1,36)
\end{aligned} \tag{5.2.3}$$

where $X_{g,-k} = \mu_{g,k}^{(0)}, \quad \forall k \in [1 \cdots K]$

Let us now consider the problem of choosing the inherent parameters of the model – that is, satisfactory values for w_s and D_I. We use an infectivity profile following a normal distribution with parameters $\mu = 5.2$, and standard deviation $sd = 3.7$, as reported in [50]. Indeed, in this study, the authors show that the serial interval for COVID-19 is actually closer to a normal distribution, rather than the gamma distribution that is traditionally used in such cases. Recent studies have also shown that viral shedding could last up to 20 days. We thus take w to be a 20 dimensional vector, and $D_i = 20$. Figure 5.5a shows the distribution of the values of the coefficients of w_s.

Spatial Random-Effects Model: Fitting. As for the synthetic experiments, we fit the model in Eq. 5.2.3 using `Rstan` [19], 8 chains with 5,000 warmup iterations and 1,000 sampling steps. Figures 5.8 and 5.9 show the posterior credible intervals for the daily contact rates, transmissibility, as well as each group's spatial reproductive number R_g itself. As a first sanity check to the performance of the model, we compare the values of the recovered R for Hubei 0 and 1 – two different stages of the epidemic in Hubei, with a tightening of the lockdown. It is interesting and reassuring to see in Figure 5.8 that the reproductive number for Hubei does seem to have gone down, from a value of roughly 0.91 in the first days of the quarantine to 0.82 over the past 36 days — whether this decrease is due to the even tighter lockdown, or the decrease in the susceptible population, the order between these two coefficients is in the order we'd expect . Moreover, these values recovered by our model are consistent with both numbers that have been found in other studies [158, 152], as well as aligned with the R_0 recovered by using the `earlyR` [139] programming suite (see (see Fig.5.11b, 5.11f,5.12d) — yielding in most cases similar predictions, as shown by the overlapping green (Bayesian) and pink (`earlyR`)

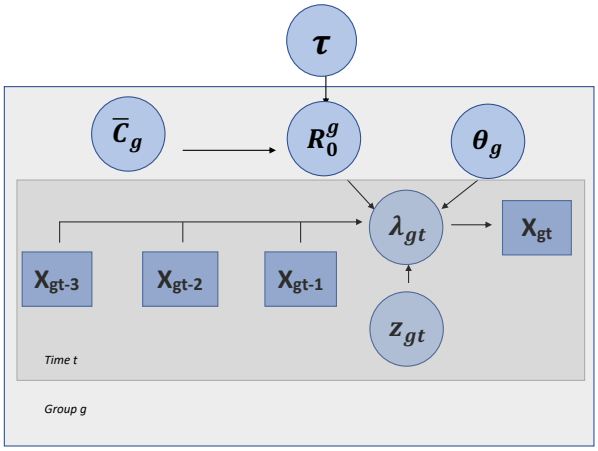


Figure 5.7: Plate model for the real data

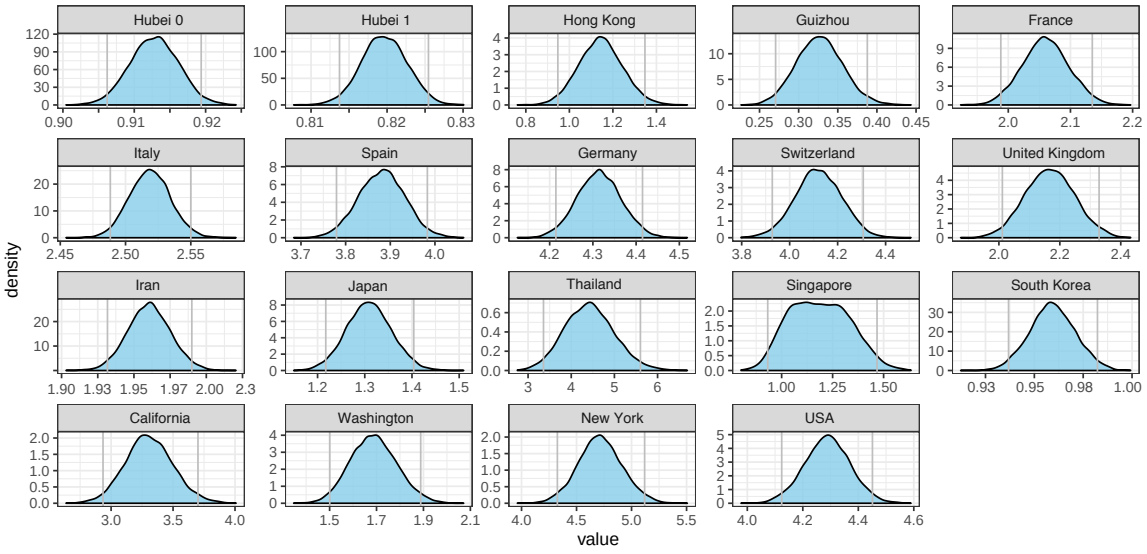


Figure 5.8: Distribution of the recovered spatial reproductive numbers R for the spatial Random-Effects Model.

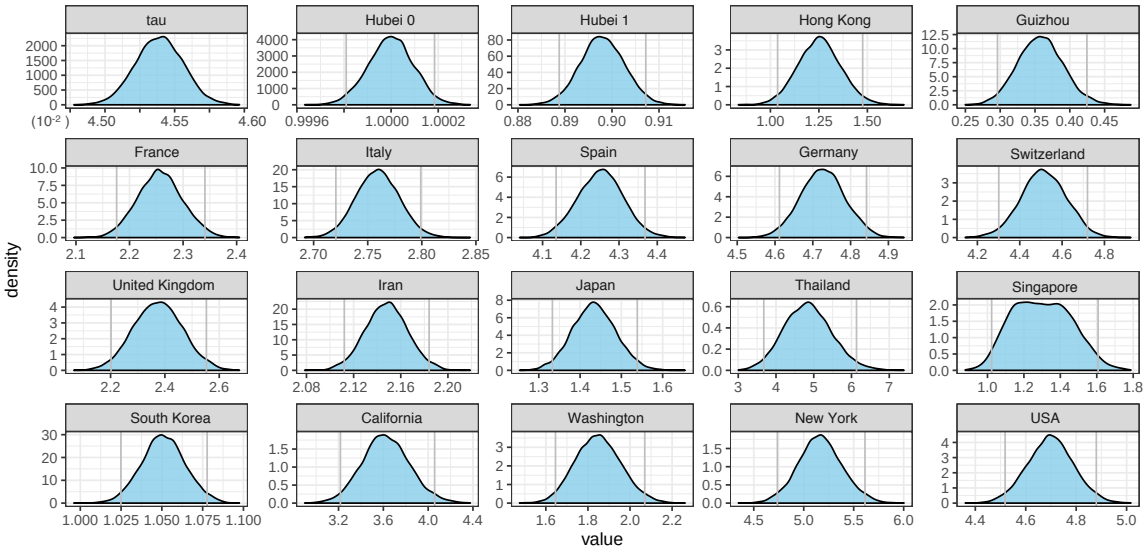


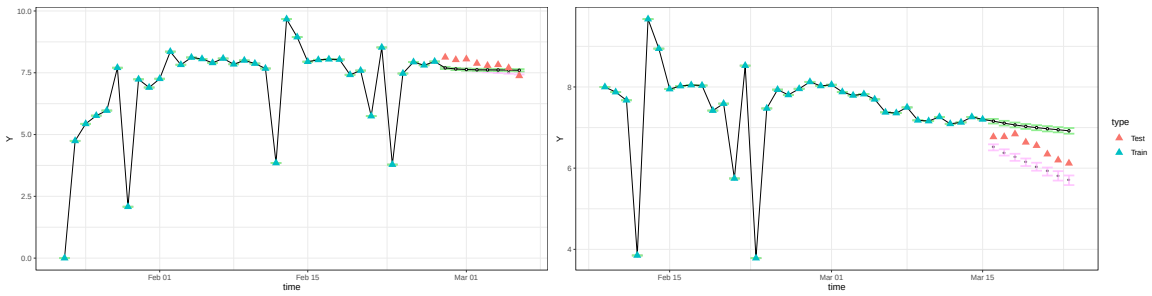
Figure 5.9: Transmissibility τ and the recovered daily contact rates $\bar{c}_g$ for the spatial Random-Effects Model.

confidence intervals. This serves as a benchmark test to our model: the Bayesian model that we use here is consistent with other statistical methods for estimating the R_0. Again, our purpose here is not to "beat" established methods, but to focus on the problem of uncertainty quantification. While in some cases, the Bayesian framework seems to be more amenable to missing and/or incomplete data (Fig.5.11g), our method is by no ways a substitution for this software. Rather, it provides a more amenable framework for the analysis of the variability in the predictions, and we base the following discussion on the credible intervals that our Bayesian framework provides.

Spatial Random-Effect Model: Discussion. Looking at the results, we note substantial heterogeneity in the reproductive numbers associated to the different states in America, as the reported reproductive number R varies from 3.5 in California to 1.8 in Washington — thus highlighting the importance of tailoring the estimation of the reproductive number R to a given population. The reported between-group variance is substantial compared to the within-group variance, thus emphasizing the need to adapt the estimation of the R_0 to a given geographical group. This thus begs the question: in today's general discussion about the epidemics, which R_0 is actually used? Indeed, the spatial heterogeneity alone severely impacts the accuracy of the reported incident cases. This is in particular reflected in Table 5.1, which shows the prediction performance associated to predicting the evolution of the disease using group-specific coefficients fitted using our Bayesian framework and `earlyR`, contrasted against the predictions spanned from using the R_0 computed on the aggregated dataset. For the Bayesian method, we compare predictions that make use of the spatial heterogeneity of the method, against a set of predictions based on the fixed average group

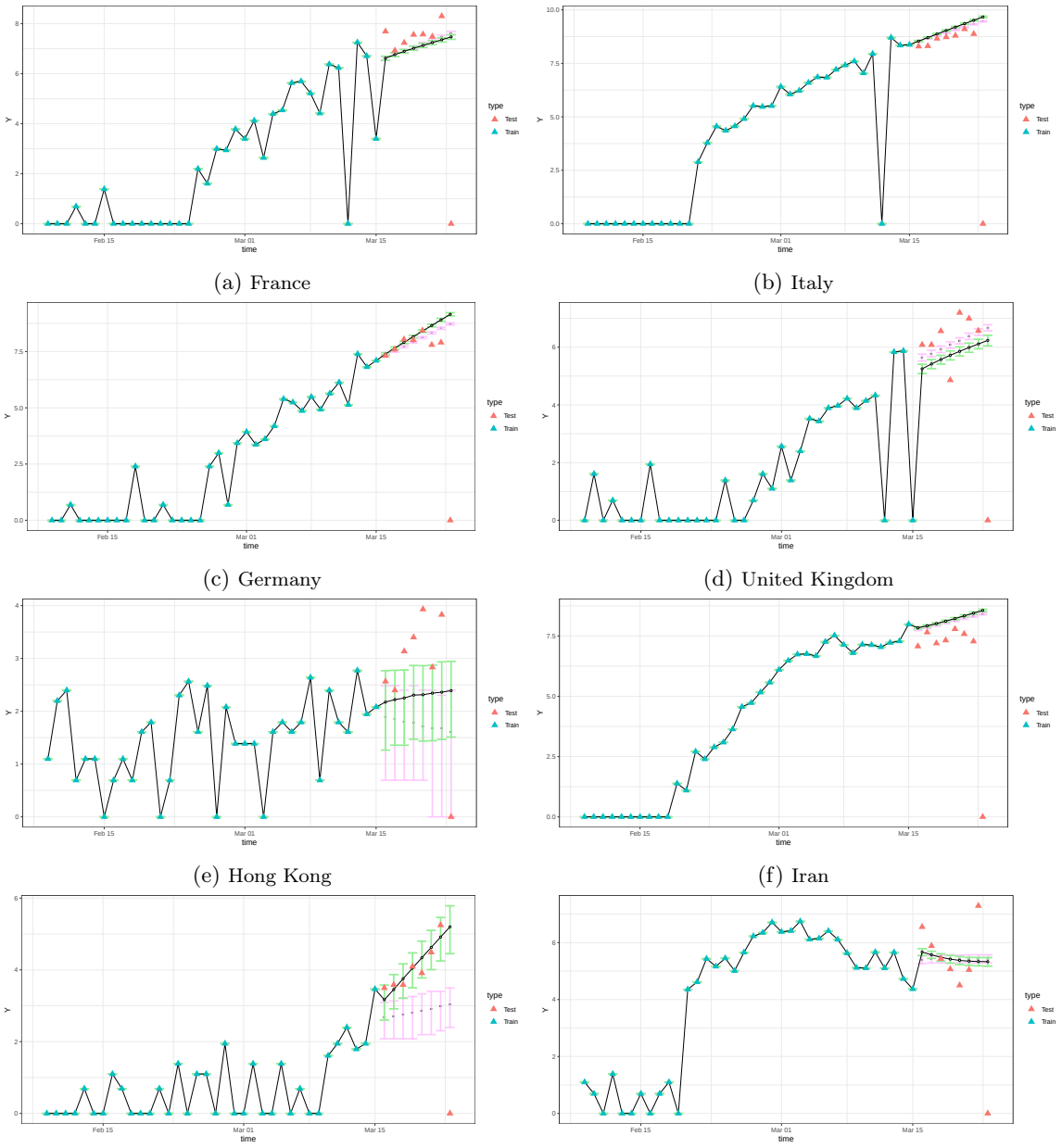
value, and a third set that uses the fixed mean R_0 over all groups to make the predictions. In each case, the R_0 was fitted on 36 days of data (starting on February 9th), and performance evaluated on the time span from March 18th to March 22nd. We focus on the comparison of the three following metrics: (a) the length of the confidence interval provided by the method (in terms of the number of reported cases), (b) the coverage (that is, the percentage of times the observed value in the test data is contained in the interval provided by the model) and (c) the accuracy, expressed as the average difference between the predicted mean value and the actual observation. In particular, the error between the average predicted values and the actual observations using a group-specific R_0 is 30% lower than the one obtained using an average R_0 over the aggregated data (for both the Bayesian model and the `earlyR` estimates). We also note that the group-specific R_0 obtained through our Bayesian procedure exhibits similar accuracy to the one inferred using `earlyR` (which do not benefit from the hierarchy), while yielding larger confidence intervals and substantially better coverage.

Moreover, we observe a significant variability in the confidence intervals associated to each of these Rs. For instance, the width of the confidence interval for the reproductive number in Spain is roughly of length 0.2, while it is two or three times as big in other groups (e.g, California, Switzerland). This local variability thus also needs to be taken into account when running predictive scenarios, since this uncertainty is then exponentially reverberated in downstream predictive scenarios.



(a) Hubei (b) Hubei 1

Figure 5.10: **Spatial Random-Effects Model.** Predictions (with confidence intervals) for Hubei 0 (first days of the quarantine) and Hubei 1 (last 36 days). Y values are plotted on the log.scale. Green confidence intervals are the one recovered by our Bayesian method, and in pink through the `projection` `R`-package. The blue circles show the observations used for training, and the red triangles are observations from the past six days that we use as validation.



(a) France (b) Italy

(c) Germany (d) United Kingdom

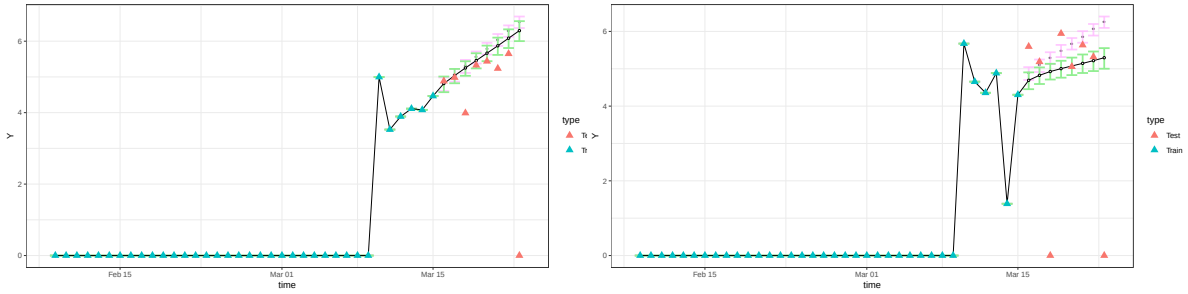
(e) Hong Kong (f) Iran

(g) Thailand (h) South Korea

Figure 5.11: **Spatial Random-Effects Model.** Predictions on the log-scale for a few countries. Training observations shown by the blue circles, validation data by the red triangles. Green confidence intervals are the ones recovered by our Bayesian method, and in pink through the `projection` package.

Method	Length CI	Percentage coverage	Mean Error
Bayesian Method	214	63.9	861
Bayesian Method (Fixed, average group R)	137	63.2	861
Bayesian Method (Overall average R)	139	59.4	1227
`EarlyR` (Fitted per group)	159	51.9	769
`EarlyR` (Overall aggregated R_0)	73	21.1	1034

Table 5.1: **Spatial Random-Effects Model:** Comparison of performances of the different prediction (over 5 days).For the Bayesian method, we compare predictions that make full use of the spatial heterogeneity of the method (1st row), vs one that uses the fixed average group value (2nd row), and a third that uses the mean R_0 over all groups (3rd row) to make the predictions. We compare these results to the coefficients obtained using `earlyR` fitted independently on each cluster (4th row) or over the aggregated data (5th row).



(a) California (b) Washington State

(c) New York (d) United States (whole)

Figure 5.12: **Spatial Random-Effects Model.** Predictions (log-scale) for the United States. Green confidence intervals are the one recovered by our Bayesian method, and in pink through the `projection R-package`.

Table 5.1 and the predictive plots in Figures 5.11 and 5.12 thus emphasize the importance of tailoring the R coefficient to the given group: note in particular the variability in the length of the confidence intervals from group to group.

These results thus highlight the need to integrate the heterogeneity of the spatial distribution of R in the model to yield informative models. Note however that this model is only valid for the first weeks of the epidemic, as this assumes an exponential growth model.

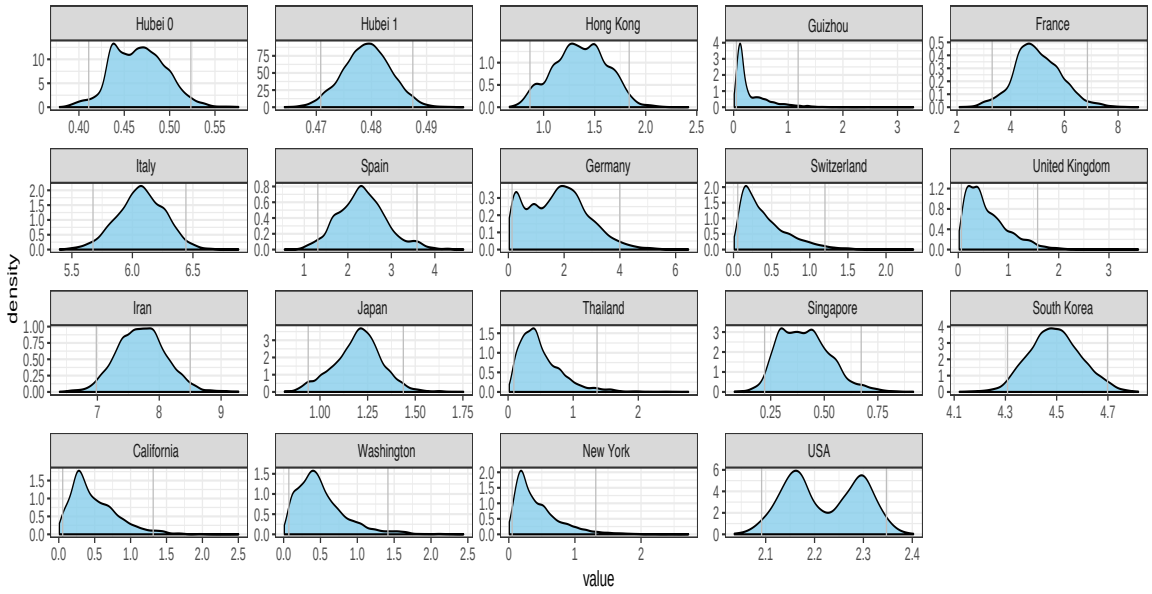


Figure 5.13: Distribution of the recovered spatial reproductive numbers R for the random effects modelling for the full Random-Effects model — fitted on 36 consecutive days from Febrary 9th to March 17th

Full Random-Effects Model: Fitting. We now assess the impact of adding further variability in the group $R^{(g)}$ reproductive coefficient, modeling it itself as a random variable which is re-sampled at every time step. Similarly to the previous setting, we fit the random-effect model using `RStan` with 8 chains, 5,000 warmup iterations and 1,000 sampling steps. We discard two of the chains that had not mixed with the others.

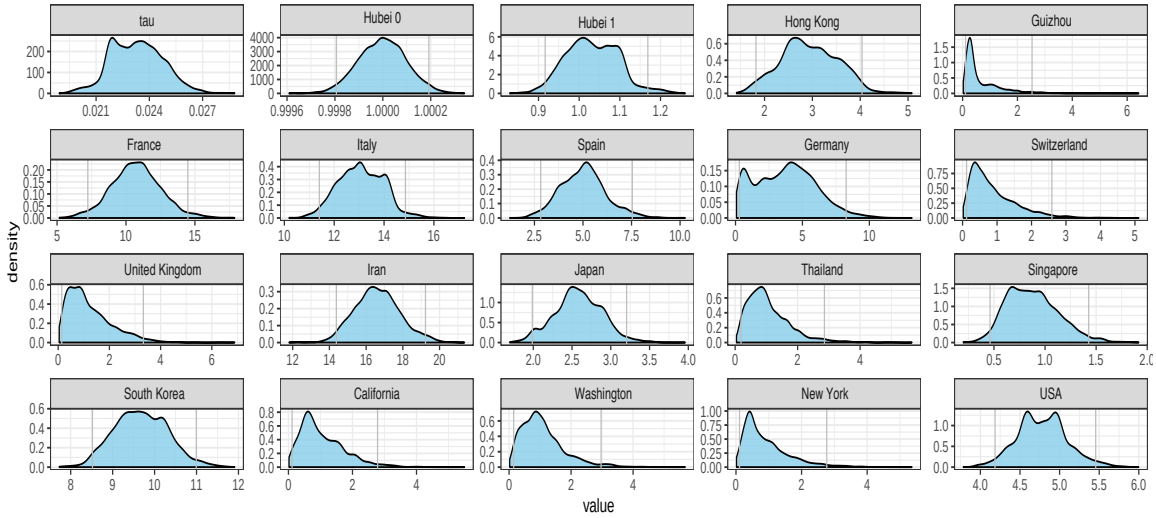
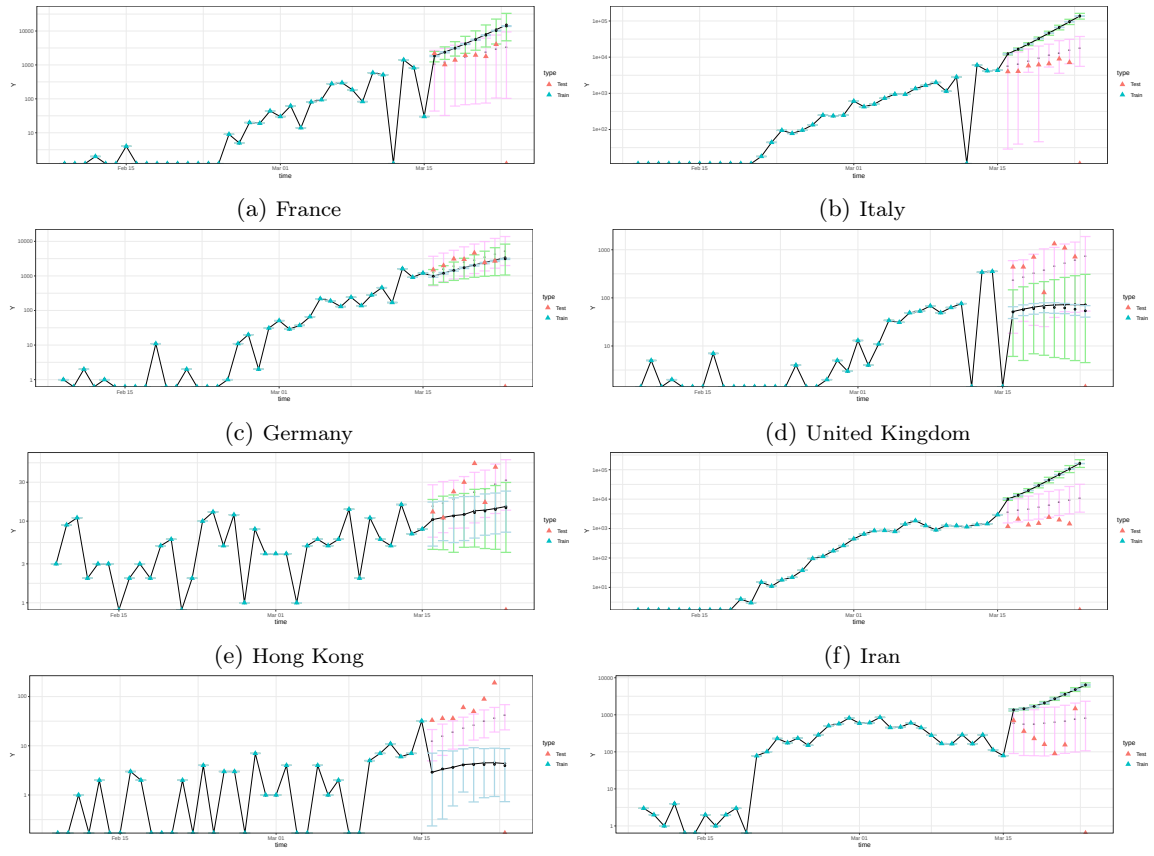


Figure 5.14: Transmissibility τ and the average recovered daily contact rates $\bar{c}_g$ for the full Random-Effects model.



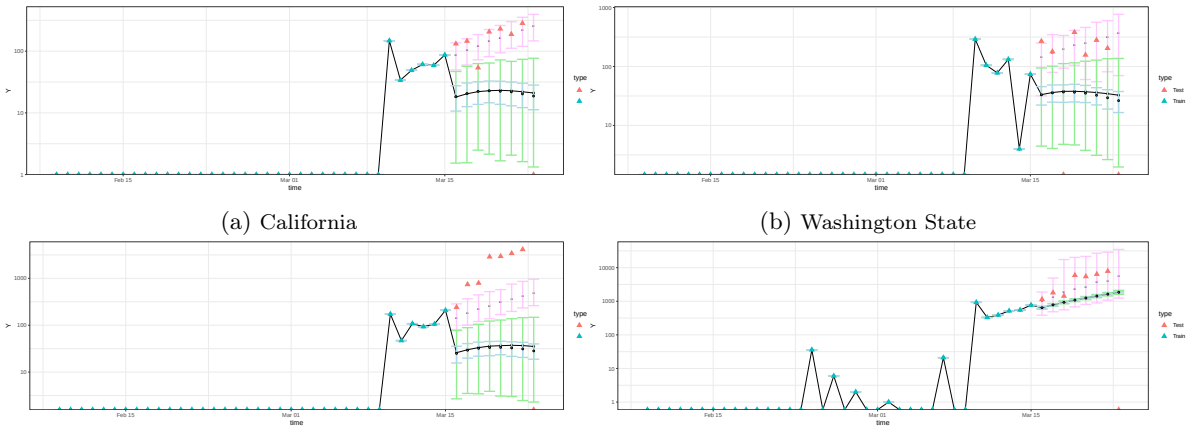
(a) France (b) Italy

(c) Germany (d) United Kingdom

(e) Hong Kong (f) Iran

(g) Thailand (h) South Korea

Figure 5.15: **Full Random-Effects Model.** Predictions using random effects on the log-scale for a few countries (training observations shown with the blue rounds, validation data displayed through the red triangles). Pink confidence intervals are the ones recovered by our Bayesian complete random-effect model (new $R_i^{(g)}$ for every time step in each trajectory), green are confidence intervals obtained assuming $R^{(g)}$ is random, but constant through time (one $R^{(g)}$ per trajectory), and in blue, the ones assuming that $R^{(g)}$ is fixed and equal to its mean recovered value



(a) California (b) Washington State

(c) New York (d) United States (whole)

Figure 5.16: **Full Random-Effects Model.** Predictions (log-scale) for the USA. Training observations shown with the blue round, validation data displayed through the red triangles. Pink confidence intervals are the ones recovered by our Bayesian complete random-effect model (new $R_i^{(g)}$ for every time step in each trajectory), green are confidence intervals obtained assuming $R^{(g)}$ is random, but constant through time (one $R^{(g)}$ per trajectory), and in blue, the ones assuming that $R^{(g)}$ is fixed and equal to its mean recovered value.

Discussion: Full Random-Effects Model. The results for the full random-effects model are displayed in Figures 5.13 and 5.14. Note that the shapes of the distribution of the R_g are radically different that in the Spatial Random-Effect Model. In particular, the credible intervals are substantially larger. We also note that the distribution for the R_0 corresponding to the United States exhibits a slight bimodal behavior: this could be due to the heterogeneity of the United States data, which consists of the raw counts of the epidemic aggregated over different states at various stages of the epidemic.

Figures 5.15 and 5.16 illustrate the advantage of adding temporal variability to the R_0. Indeed, for this fitted model, the confidence intervals for the R are much larger than in the Spatial Random-Effects model. The predictions are, in average, more accurate using the fully random model rather that using the average group R_0 (the average group value fitted in this version of the model). For instance, using the average R_0 recovered in this setting, the average difference between the mean trajectory and the actual observations is around 5,000. The error is however cut by 40% by using the fully Bayesian model. The confidence intervals — albeit far greater for the fully random model than for the the spatial random-effects model — nonetheless provide a better coverage of the actual observations (87% for the pink fully random credible intervals, to be compared against 49% for the green partially random ones, and 40% for the blue ones corresponding to the constant R_0), thus pointing to a more realistic quantification of the uncertainty.

Partial Conclusion: Effect of the Variability on Uncertainty Quantification. From our simulations and experiments on real data, it seems that taking into account the complete variability

of the R_0 is important to plan ahead for worst-case scenarios. Indeed, while the mean error is twice as big for the full random-effects model compared to the spatial one and the average-case scenarios thus seem to be better predicted by the less variable model, the true difference lies in the extreme cases. In particular, the coverage by the provided credible intervals is less optimal for the spatial random-effects model compared to the fully Bayesian one (64% vs 87%). In either of these models, the predictions obtained by holding R_0 constant, aggregated over the whole data, yields confidence intervals that are too narrow, and predictions that are completely off. Additional temporal and spatial variability seem therefore necessary to draw a more complete portrait of the outcomes of the epidemic (and especially the extreme), as well as to correctly quantify its uncertainty.

5.3 Evaluating the impact of adding heterogeneity in predictive scenarios.

The second stage of our analysis is to use our fitted model for the heterogeneous R and local group demographics to predict the impact of different strategies on the outcome of the epidemic. Indeed, policy makers are currently faced with the difficult task of implementing efficient policies to limit the spread of the virus, while arbitrating between societal and economical costs. An inspection of the decomposition of the reproductive number provided in Eq. 5.1.1 exhibits why a policy geared towards a lowering of the daily contact rate $\bar{c}$ should efficiently limit the spread of the virus. The goal of this section is thus to quantify the effect of governmental measures on the "flattening of the curve". Again, we emphasize that our study does not aspire to provide state-of-the-art prediction models, but rather to assess how informative our exponential growth model truly is when used in the context of drawing predictive scenarios under such huge uncertainty. As such, we use our fitted reproductive number to generate new predictions for the next 200 days. The results that are presented in this section are inferred from running 800 simulations.

Note that we do not want to assume here that a given policy can manage to bring the R_0 to a given value (e.g. 1) – in other words, that the effect of the policy is absolute. Rather, because different countries have different governmental structures, population (and household) dynamics, it seems more sensible to talk about a spectrum of social distancing measures and to discuss the effect of a policy in relative terms. We thus consider policies that divide the daily contact rate by a certain factor, rather than in absolute value. On an aside note, one could envision assessing a given policy's effective slashing of the contact rate using population census data (for the structure of the household) as well as general mobility data. Historical data could indeed be used estimate baseline contact rates, while current mobility data would reflect the effective contact rate associated to a given policy. Thus, we characterize policies not as categorical variables (e.g, "Total Freedom", "Shelter-In-Place", "Lockdown", "R=1", and so on), but as continuous variables (i.e, reduction of $\bar{c}$ to 20% of its original values, etc.).

Since one of the main issues when handling the pandemic consists in the access to healthcare, we use the predicted epidemics trajectories to model hospital bed occupancy. Studies have indeed shown that hospitals are able to deliver the best care up to 85% of capacity — threshold after which the quality of the care decreases This can lead to a potential increase in fatality rates for COVID-19, as well as unwanted additional comorbidities, as people suffering from other ailments do not receive proper access to treatment. Thus, this part will focus on modeling the number of beds (general hospitalized and Intensive Care Unit (ICU)) required by patients suffering from COVID-19 at any moment in time (rather than the incident cases themselves), as well as the cumulative death toll.

The model that we adopt here is the following. For each day:

1. We generate the number of new incident cases based on the Bayesian model detailed and fitted in the previous section.

2. We then generate the number of people among these incidence cases that will require hospitalization. This number is generated by a binomial distribution, with a hospitalization rate that is contingent on the geographic localization and takes into account the age demographic layout of each cluster:
$$\pi^{(g)}_{\text{Hosp}} \sim 0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{Hosp}}, 1)$$
where α_g is the proportion of each age group in location g (divided in 4 groups: from “0-19" years-old, “20-54", “54-65", and “65+”), and π_α is the hospitalization rate per group (expressed in percentages, and assumed to be universal across all contagion groups).

3. Once the number of newly hospitalized people has been selected, we choose among them using a binomial distribution the people directly admitted into an Intensive Care Unit (ICU). The parameter for the binomial is also contingent on the demographics:
$$\pi^{(g)}_{\text{ICU|Hosp}} \sim \frac{0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{ICU}}, 1)}{\pi^{(g)}_{\text{Hosp}}}$$
where π_α^{ICU} is the ICU rate per group (also expressed in percentages, and assumed to be universal across all contagion groups).

4. Finally, the fatalities are chosen among the people placed in the ICU, and sampled from a binomial distribution with probability:
$$\pi^{(g)}_{\text{death|ICU}} \sim \frac{0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{death}}, 1)}{\pi^{(g)}_{\text{ICU}}}$$

5. All the hospitalizations, ICU and number of deaths being selected, we assign a time of death of each patient and of departure from the hospital/ICU by sampling from a normal distribution,

whose mean and standard deviation have been selected based on numbers based on recent studies.

The scenarios are thus sampled as follows:

$$
\begin{aligned}
\tau &\sim \text{Posterior}(\tau) \\
\forall g, \quad \bar{c}_g &\sim \text{Posterior}(\bar{c}_g) \\
X_{t,g} &\sim \frac{1}{2}\Big(\mathcal{N}\Big(2\sqrt{R_0 \sum_{s=1}^{K} w_s X_{t-s} + \frac{3}{8}}, 1\Big)\Big)^2 - \frac{3}{8} \\
\pi^{(g)}_{\text{Hosp}} &\sim 0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{Hosp}}, 1) \\
\pi^{(g)}_{\text{ICU|Hosp}} &\sim \frac{0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{ICU}}, 1)}{\pi^{(g)}_{\text{Hosp}}} \\
\pi^{(g)}_{\text{death|ICU}} &\sim \frac{0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{death}}, 1)}{\pi^{(g)}_{\text{ICU}}} \\
\text{Hosp}_{t,g} &\sim \text{Binomial}(X_{t,g}, \pi^{(g)}_{\text{Hosp}}) \\
\text{ICU}_{t,g} &\sim \text{Binomial}(\text{Hosp}_{t,g}, \pi^{(g)}_{\text{ICU|Hosp}}) \\
\text{Deaths}_{t,g} &\sim \text{Binomial}(\text{ICU}_{t,g}, \pi^{(g)}_{\text{death|ICU}}) \\
\forall i \in [1...\text{Deaths}_{t,g}], \quad T_i^{\text{Deaths}_{t,g}} &\sim N(\mu_d, \sigma_d) \\
\forall i \in [1...\text{ICU}_{t,g}] \quad T_i^{\text{ICU}_{t,g}} &\sim N(\mu_{ICU}, \sigma_{ICU}) \\
\forall i \in [1...\text{Hosp}_{t,g}] \quad T_j^{\text{Hosp }_{t,g}} &\sim N(\mu_h, \sigma_h)
\end{aligned}
\tag{5.3.1}
$$

We emphasize again that our predictive model for the number of incident cases is based on some version of the exponential growth model. As such, it is only valid for the first stages of the epidemic but not for long-term predictions of the disease, where traditional SEIR models are typically better suited to the task. Our goal here is to assess the speed at which hospitals can be swamped with patients at the beginning of the epidemic, and to understand how the variability impacts our selection of a course of action mitigating the social and economical costs associated to a complete lockdown of a geographical region. We also want to assess the efficient of alternation-based scenarios, in which the population would remain under lockdown for x days (that is, a state of activity in which the normal number of daily contacts is divided by a factor y), and pursue almost regular activities for the remainder of the week.

Spatial Random-Effects Model: Discussion. We focus on the analysis of the results for the projected epidemic trajectories for a few of our groups. Note that the results here should be interpreted as the trajectories, based on the data obtained on March 18th, if the governments had immediately implemented a given public policy. As such, there might be a few discrepancies with the current observed numbers — but again, the focus of this paper is to assess the impact of the

added variability on the uncertainty of the projected scenarios rather than on the accuracy of the scenarios themselves, so as to answer the following question: how informative truly are our predictive scenarios?

Case Study 1: France Panel 5.17 shows the evolution of hospital occupancy and death rate in France using different reproductive numbers Rs as input and using preventive measures to stall it. In particular, Panel 5.17(A) shows the predictive results for the number of beds occupied by COVID-19 patients, ICU units, and daily death toll for strategies using an average R_0 (fitted on the aggregated data for the 19 groups that we have been considering in this paper). These have to be contrasted against the predictions we can obtain in Panel 5.17(B), where the group-specific R for France is used to draw the predictive scenarios. Note in particular the lack of consensus between the outputs of these scenarios: for the general mean R_0, slashing daily contacts by 60% is sufficient to obtain prevent hospitals from overflowing within the next two months. This is barely sufficient for the group-specific R (with much longer resolution horizon), and scenarios, such as alternating 3 days of lockdown with an 80% reduction in the daily contact rates with four days of business as usual is no longer sufficient to resolve the healthcare overflow. This highlights the importance of using the group-specific reproductive rate.

Figure 5.18(D) shows the hospital capacity as predicted by the different R_0s — that is, for models where the R_0 is either drawn from the distribution of another country (to test for transferability), from the group-specific distribution, or is a fixed R_0 estimated from the aggregated data. Not only do the scenarios vastly differ, we note that some of the alternating strategies will not be valid under an inappropriate R_0, as they will not be to contain the growth of the epidemic and the overflow of the hospital system.

Case Study: Italy We proceed to a similar study in the case of Italy (Panel5.19). We note that the impact of the pandemic on Italy is an order of magnitude greater than in France. Again, Panel 5.19 shows the divergence of the scenarios that are produced when using an “averaged" version of the reproductive number. Note as well that in the case of Italy, an alternating strategy with 5 days of quarantine and 2 days of regular activity is barely sufficient to keep the Hospital occupancy at 10,000.

Case Study: The United States We also show the same results for California (Fig. 5.20) and the United States (Fig. 5.21). Due to the country’s large R_0, the model seems to hint towards a complete saturation of the hospitals within three weeks. We also note that in this case, a slashing of $\bar{c}$ to 40% of its original value is no longer sufficient to prevent the explosion of the hospital occupancy. Finally, the difference between the predictive scenarios for California and the United States as a whole seem to highlight the need to perform the analysis at a very fine grain level.

Figure 5.22 shows the different scenarios produced for four different groups and comparing the

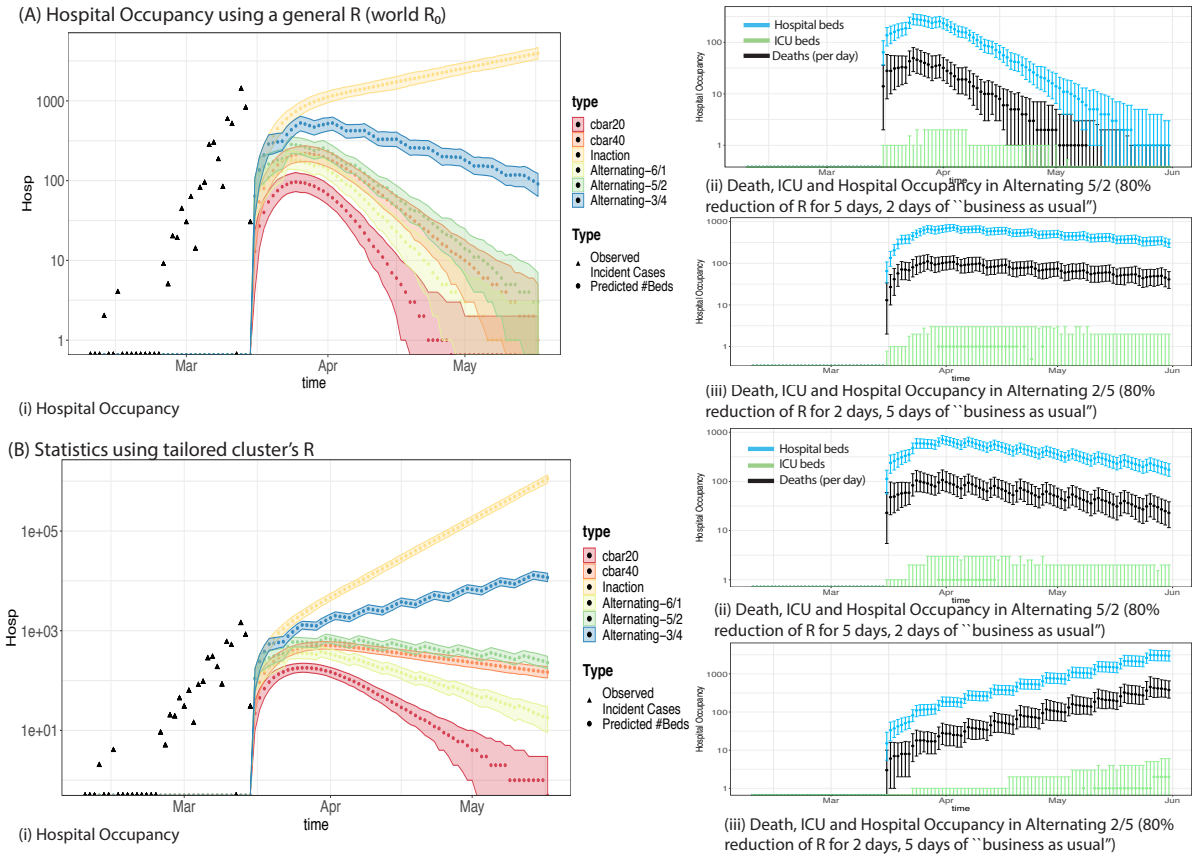
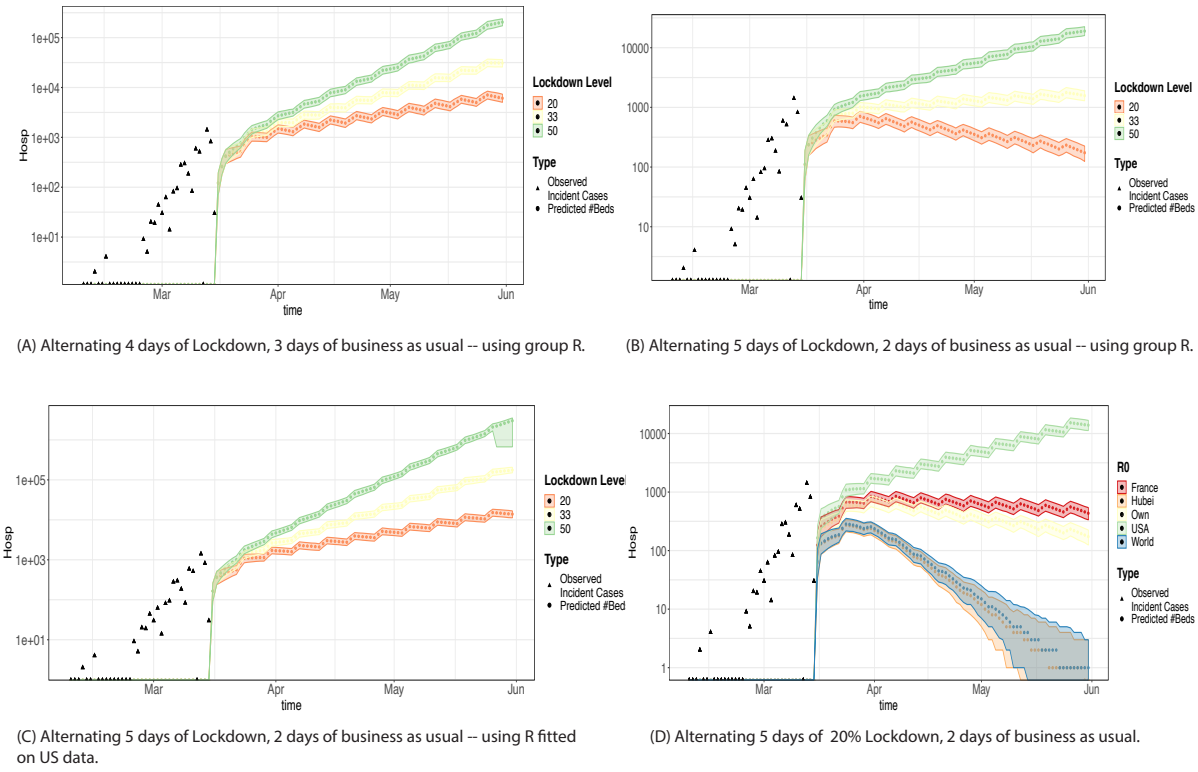


Figure 5.17: **Spatial Random-Effects Model: France.** Comparisons of the outcomes of the different strategies. We compare the estimated likely trajectories in terms of occupied hospital beds using various R: the group's specific and tailored Bayesian R, as well as an overall, general R estimated from the aggregated data. We note the substantial difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R yield.

effect of using different fitted values for the R_0, compared to their own. This highlights the impact of fitting the appropriate R_0, and the lack of transferability between the different clusters.

On the other hand, we note that sampling from each posterior interval for each R_g obtains similar prediction bands that by running the same model using the mean R_g. Fig. 5.23 shows that the confidence bands for our predictions of the number of new incident cases per day are not significantly narrower than our projected scenario using the mean R_0 of the distribution. This is because the confidence intervals recovered by the Bayesian model (as per Fig. 5.8) are for the most part quite narrow. A more realistic model for the R_0 would be to model it using a heavier-tailed distribution, so as to accurately capture the existence of super-spreaders.

Our study cases and Fig. 5.22 thus show the impact of selecting the right R_0: not all policies yield the desired flattening of the curve, thus highlighting the need to perform a fine grain analysis of



(A) Alternating 4 days of Lockdown, 3 days of business as usual -- using group R.

(B) Alternating 5 days of Lockdown, 2 days of business as usual -- using group R.

(C) Alternating 5 days of Lockdown, 2 days of business as usual -- using R fitted on US data.

(D) Alternating 5 days of 20% Lockdown, 2 days of business as usual.

Figure 5.18: **Spatial Random-Effects:** France. Comparisons of the outcomes of the different strategies. We compare the estimated likely trajectories in terms of occupied hospital beds using various R: the group's specific and tailored Bayesian R, as well as an overall, general R estimated from the aggregated data. We note the substantial difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R yield.

each cluster to draw informative scenarios.

Stopping times. To quantify the impact of the variability on the R_0, we now look at the time expected until 1% of the population is hospitalized, using different policies and R_0 in the fitting procedure. Results are reported in Table 5.2 and in Fig. 5.24 and 5.25 , which show the distribution of the stopping time obtained for a few groups. In particular, results indicate that an alternating lockdown (5 days of 50% lockdown, 2 days of business as usual) results in an explosion of the number of occupied hospital bbeds in 87% of time,while it occurs with 100% of the times using other less stringent strategies. This explosion would occur roughly in 8 months (237 day — a horizon too far into the future for our model to be able to gauge it accurately, but an indicator in the efficiency of delaying this peak from 3 months to 8. On the other hand, a sustained, continuous 50% lockdown would allow the hospitalization mass to remain manageable in all cases). Similar observations follow

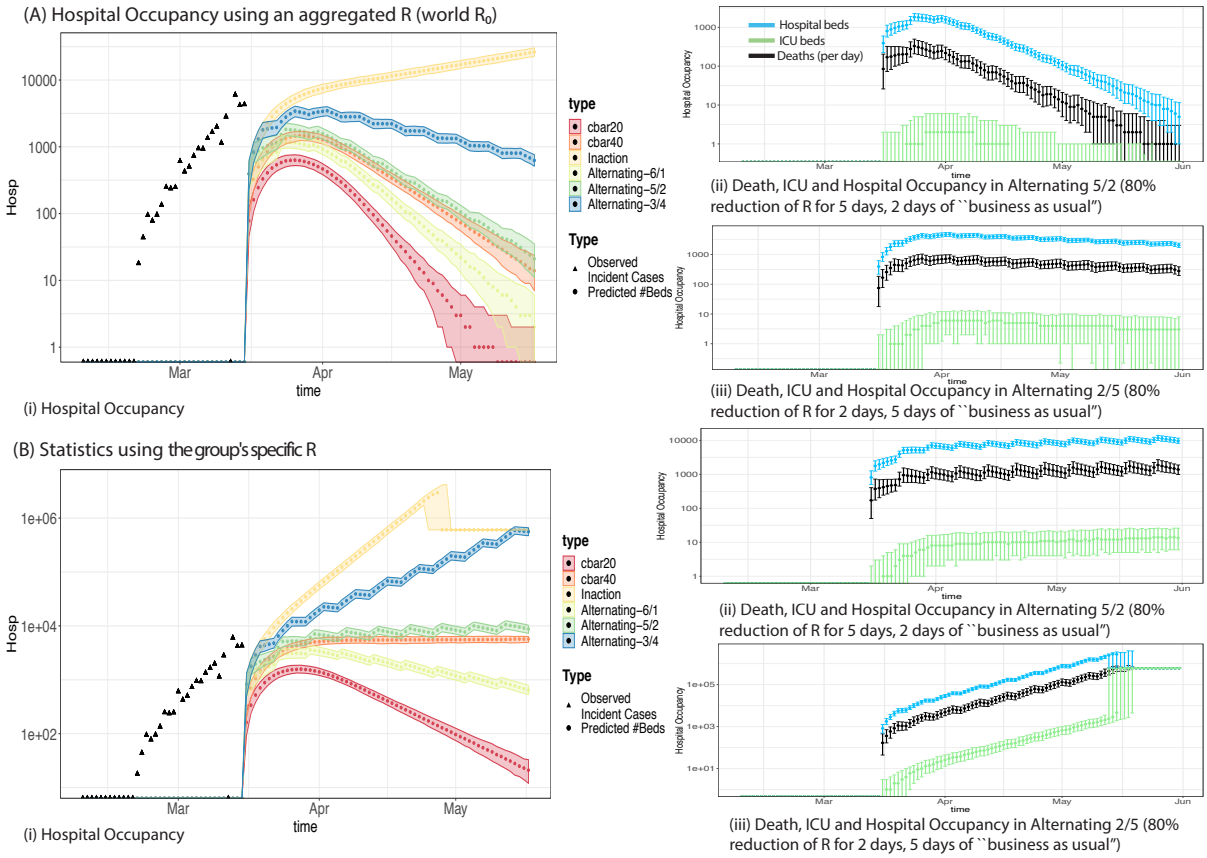


Figure 5.19: **Spatial Random-Effects: Italy.** Again, we compare the estimated likely trajectories in terms of occupied hospital beds using various R: the group's specific and tailored Bayesian R, as well as an overall, general R estimated from the aggregated data. We note the substantial difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R yield.

the same line for the other groups.

Stopping Time before Hospital Overflow						
Country	R_0 used	Strategy	Percentage	Mean τ	q97.5 τ	q2.25 τ
France	Specific	Inaction	100	110.5	107	113
		$\bar{c}^* = 0.5\bar{c}^0_{France}$	0.00	NA	NA	NA
		Alternating 4(50%)/3	100	180	186	173
		Alternating 5(50%)/2	0	NA	NA	NA
	USA	Inaction	100	80.5	78	83
		$\bar{c}^* = 0.5\bar{c}^0_{France}$	100	106	103	109
		Alternating 4(50%)/3	100	81	80	81
		Alternating 5(50%)/2	100	86	88	81
California	Specific	Inaction	100	86.1	84	88
		$\bar{c}^* = 0.20\bar{c}_{Cal}$	100	170.0	165	176
		Alternating 4(50%)/3	100	109.0	108	113
		Alternating 5(50%)/2	100	121.3	113	123
	USA	Inaction	100	75	73	77
		$\bar{c}^* = 0.5\bar{c}_{Cal}$	100	120	117	124
		Alternating 4(50%)/3	100	89	87	93
		Alternating 5(50%)/2	100	96	94	100
USA	Specific	Inaction	100	74.7	73	76
		$\bar{c}^* = 0.5\bar{c}_{US}$	100	119.9	117	123
		Alternating 4(50%)/3	100	88.4	87	92
		Alternating 5(50%)/2	100	95.0	94	99
	France	Inaction	100	125	122	128
		$\bar{c}^* = 0.50\bar{c}_{US}$	0	NA	NA	NA
		Alternating 4(50%)/3	0	NA	NA	NA
		Alternating 5(50%)/2	0	NA	NA	NA

Table 5.2: **Spatial Random-Effects:**Comparison of the stopping times associated to the scenarios drawn using different R (group-specific or another group's). The N/A values indicate that the stopping time has not been reached in any of our 800 simulations.

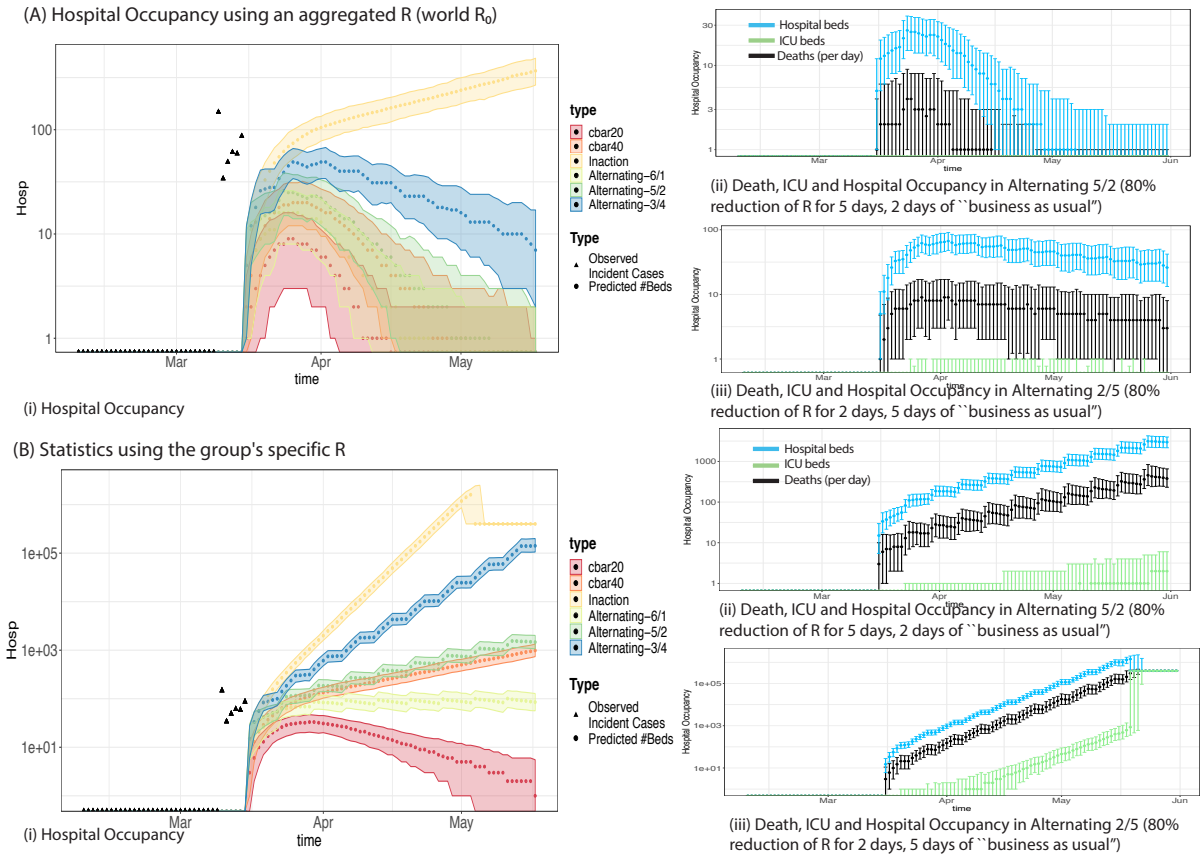


Figure 5.20: **Spatial Random-Effects: California.** We compare the estimated likely trajectories in terms of occupied hospital beds using various R: the group's specific and tailored Bayesian R, as well as an overall, general R estimated from the aggregated data. We note the substantial difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R yield.

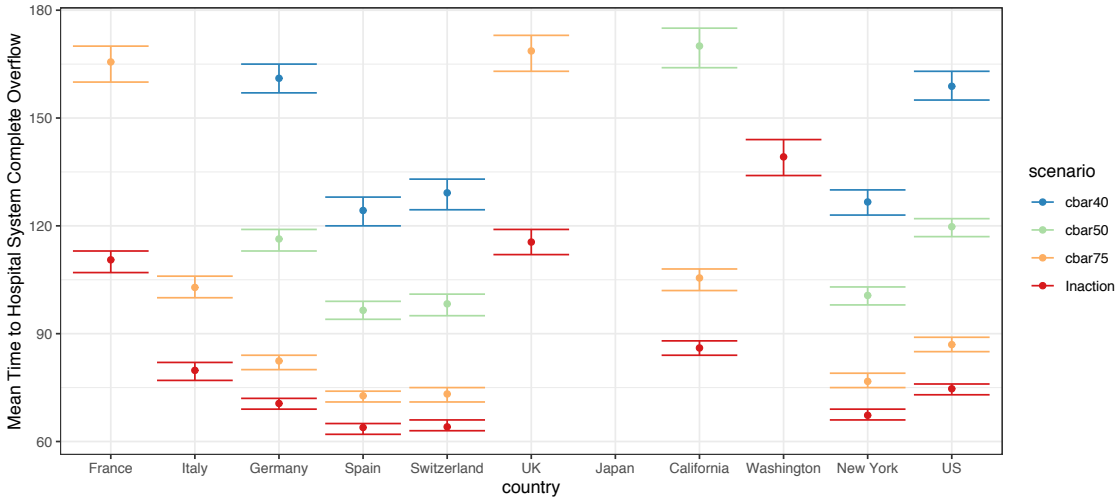


Figure 5.25: **Spatial Random-Effects:** Time to 1% of the population under hospitalization

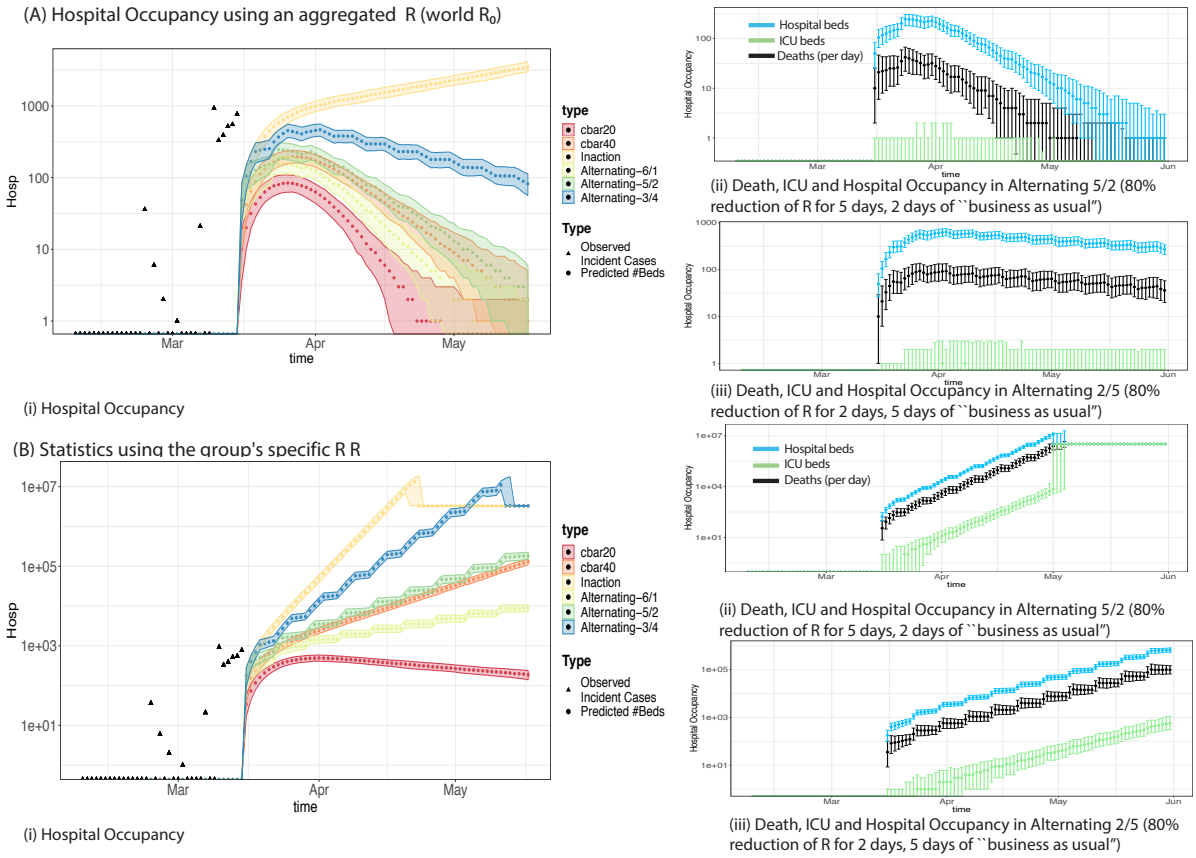


Figure 5.21: **Spatial Random-Effects.** United States of America

Full Random-Effects Model. We now compare the results that we obtain using a full random-effects model, as opposed to the spatial random-effect model that we have been considering so far. As a reminder, the difference between the two is that in the full random-effects model, the R_0 is considered random and sampled from a gamma distribution at each time step. On the other hand, for the spatial random-effect model, R_0 is also random, but sampled once at the beginning of the trajectory and held constant across time steps.

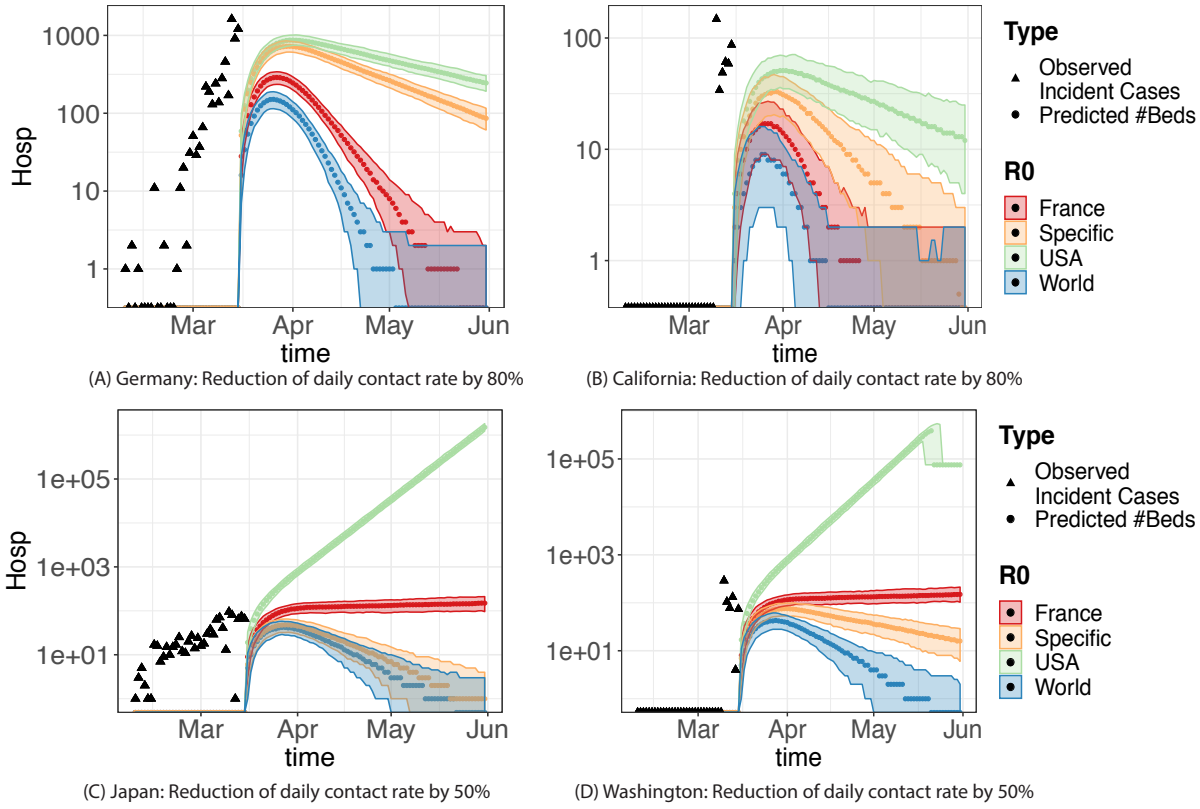


Figure 5.22: **Spatial Random-Effects.** Comparison Predictions: this figure further shows for four different groups the estimated impact of a given policy, using different R_0s. This shows the importance of correctly accounting for group-wise heterogeneity in the model.

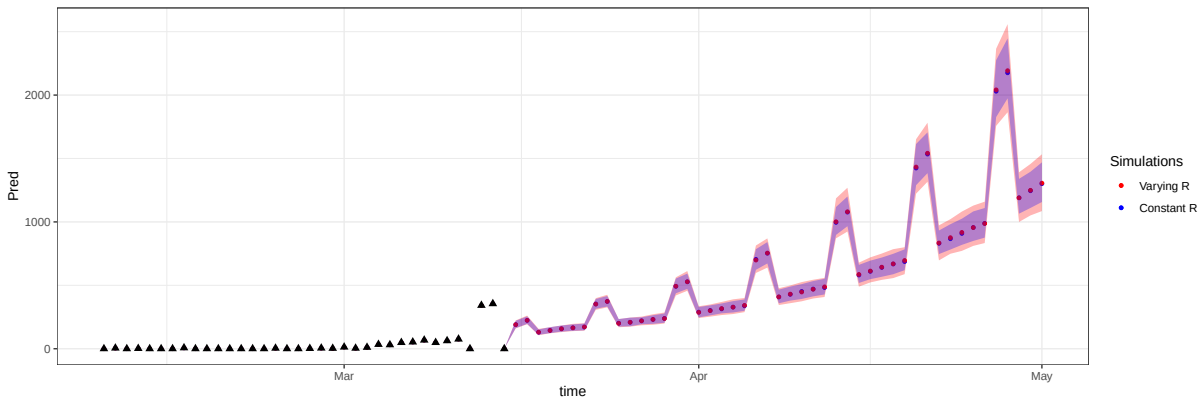
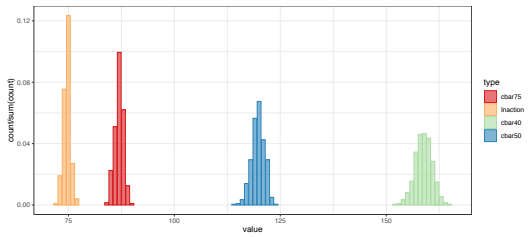
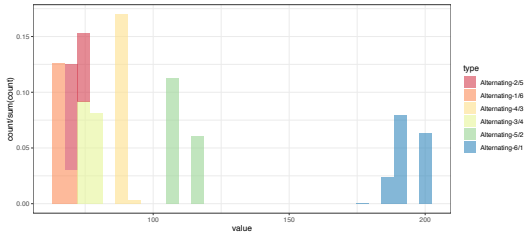


Figure 5.23: **Comparison Predictions: United Kingdom** for an Alternating scenario with 5 days of business as usual vs 2 days of 50% lockdown.

[H]

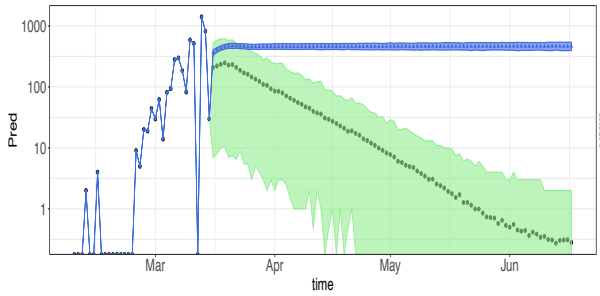


(a) USA: Time to 1% of the population under hospitalization

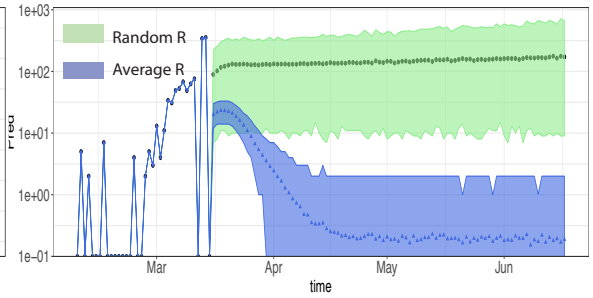


(b) Spain : Time to 1% of the population under hospitalization for different scenarios

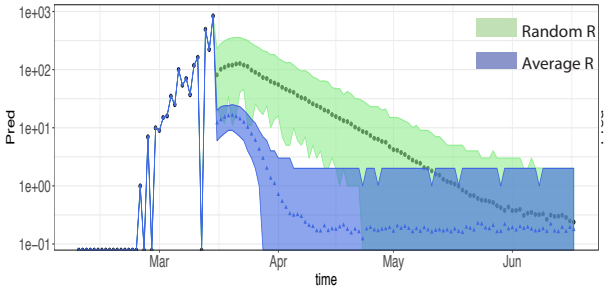
Figure 5.24: **Spatial Random-Effects:** Histograms of the expected Time until Hospitalization Overflow for two groups.



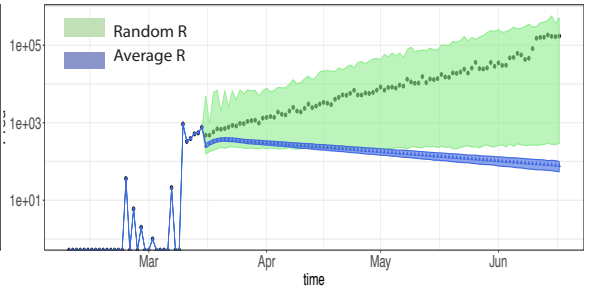
(A) France: Comparison of the effect of dividing the daily contact rate by 5 on the predicted average number of new indident cases (and 95% confidence bands), using an average, constant R (blue) vs a variable one (green).



(B) United Kingdom: Comparison of the effect of dividing the daily contact rate by 60% on the predicted average number of new indident cases (and 95% CI), using an average, constant R (blue) vs a variable one (green).



(C) California: Comparison of the effect of dividing the daily contact rate by 5 on the predicted average number of new indident cases (and 95% confidence bands), using an average, constant R (blue) vs a variable one (green).



(D) United Sates: Comparison of the effect of slashingthe daily contact rate by 60% on the predicted average number of new indident cases (and 95% confidence bands), using an average, constant R (blue) vs a variable one (green).

Figure 5.26: **Full Random-Effects:** Comparison of the Static vs Random R

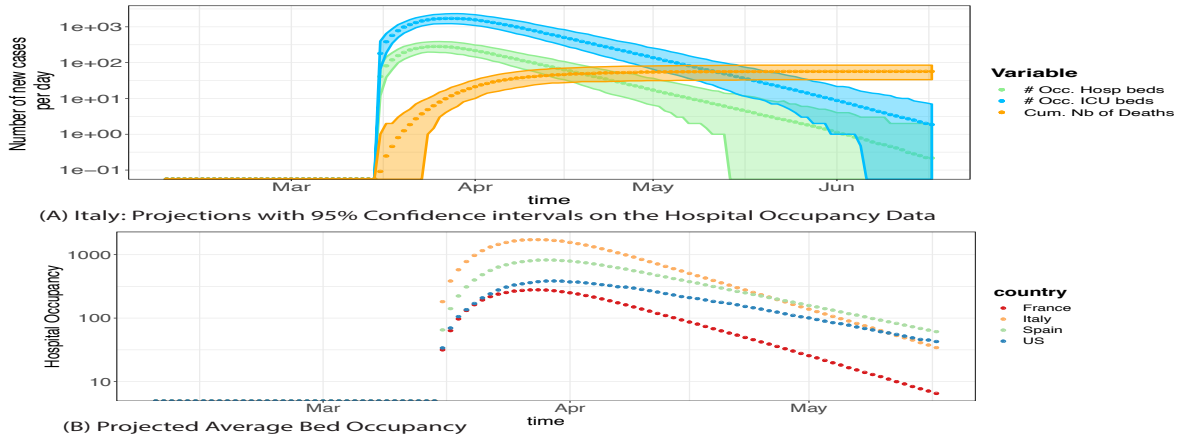


Figure 5.27: **Full Random-Effects:** Comparison of the projections of the occupied number of beds.

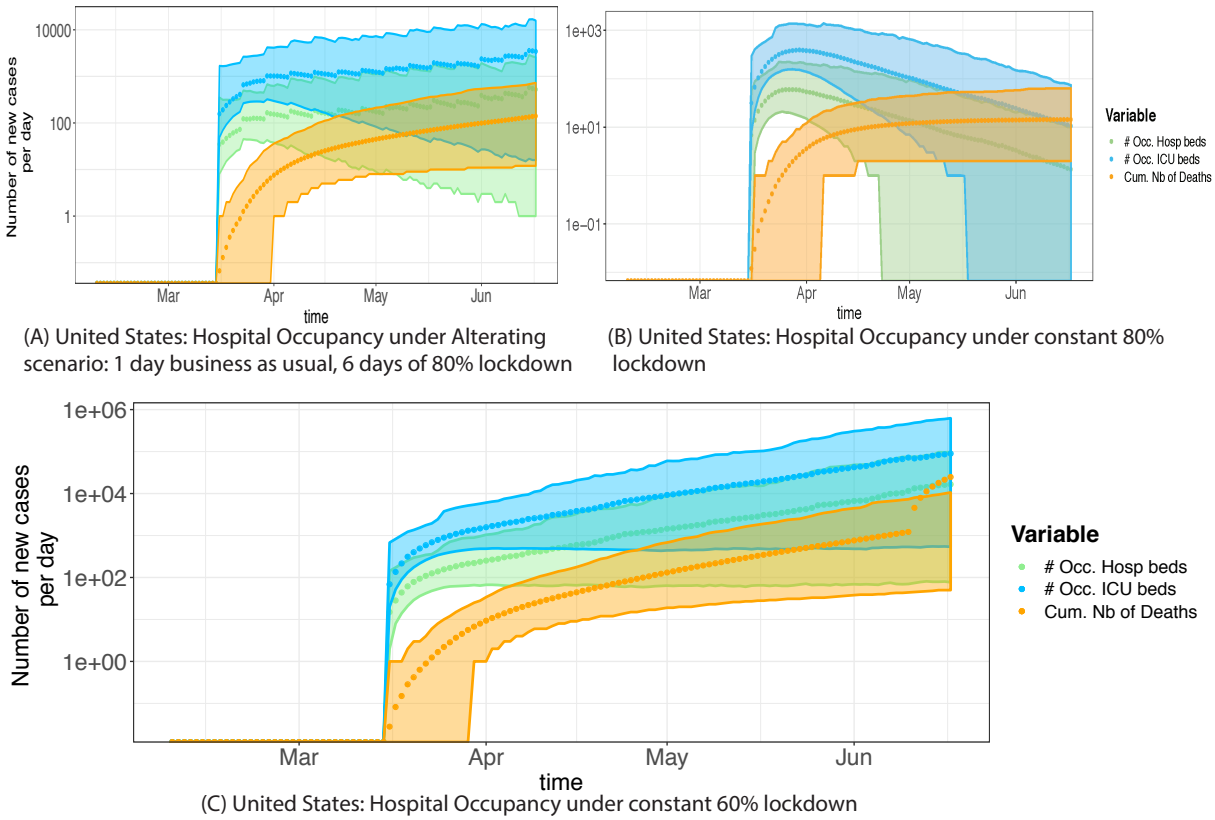


Figure 5.28: **Full Random-Effects:** Comparison of the projections of the occupied number of beds in the US under different social distancing scenarios.

Fig. 5.26 shows the comparison between the projected epidemic trajectories using a constant average mean R_g (blue dots and 95% confidence bands), vs a variable one (green dots and 95% confidence bands). We observe that the two approaches yield extremely different results. In particular, the confidence bands obtained for the average R_0 are typically too narrow, and yield in Fig. 5.26(B-D) predictions that are more optimistic than the ones obtained with the added variability in the R_0: the resolution of the conflict seems in general much longer using the fully random R (depicted in green in Fig. 5.26), and can be even less likely than what is projected using the group average R_g (Fig. 5.26(D)). The high difference here is due to the large credible intervals for each mean R_g that we had obtained. Thus the fixed average and fully variable scenarios rapidly diverge.

Figures 5.27 and 5.28 show more estimates of the projected average hospital occupancy across different countries. In particular, Figure 5.28 compares three different policies: continuous and sustained slashing of social distance by 60 and 80% (Fig 5.28 (C,B)), and an alternating policy of 6 days of severe lockdown (slashing of the social distance by 80%), and one of business as usual. It is interesting to see that the one day of difference between Fig.5.28(A) and (B) makes a significance difference: the continuous slashing is allows to prevent an overflow of the hospital with probability 1, whereas an hospital overflow happens with non-zero probability in Fig. 5.28(A). As already observed in the introduction, the additional variability in the full-random model extends the domain of plausible events — a crucial fact for policy-makers to correctly assess worst-case scenarios.

5.4 Conclusion:

In conclusion, we have presented here an analysis targeted at assessing the level of granularity in terms of the variability necessary in the modeling of the reproductive number R to draw informative scenarios. In particular, we have shown that the modeling of this heterogeneity is crucial to correctly model extreme scenarios and characterize their uncertainty. Indeed, using a spatial and temporal random-effects model, we have shown that the added variability is necessary to (a) provide better coverage of the confidence intervals, and thus, more appropriately quantify the uncertainty associated to a certain prediction or the effect of a given policy and (b) explain rare events and understand the formation of outbreaks — which averaged models would not allow and which are nonetheless crucial elements to take into account when weighting different scenarios.

Our analysis of the real data has also shown that the reproductive number vastly varies depending on the group considered — as such, it seems that an informative model would at least try to take into account the spatial heterogeneity, if not the full one. We emphasize again that our study does not aspire to draw predictive scenarios, but rather to understand how models and predictive scenarios are truly impacted by the choice and inherent variability of the R – and the great variability that we have imputed seems to highlight the need for a fine -grain analysis.

Further Discussion on the Variability of R. In our data analysis for the spatial random-effects model, we have assumed $\bar{c}$ to follow a well-behaved γ distribution. We have also tried changing this to a Cauchy distribution (which has much fatter tails, and thus, could add more variability). The results we obtained with this new prior were similar to ones obtained using the Gamma distribution, thus highlighting the fact that our spatial random-effects model does not seem to be extremely sensitive to the choice of the prior.

The purpose of this chapter was to assess how we could start integrating some heterogeneity — in part due to the skewed degree distribution of the contact network — into standard models for building predictive scenarios. Since we have no estimate of this underlying contact network, we have built in this heterogeneity using a Bayesian model. To continue upon this analysis, it would be interesting to add more structure to the model of the variability by adding proxies for the degree distribution and the underlying network structure — which we leave as future work.

Chapter 6

Conclusion

Summary of the Contributions. In conclusion, throughout this Ph.D thesis, we have investigated multiple facets of uncertainty quantification for graphs and networks. In particular, we strived to motivate this study by its concrete, practical applications to real-life data —- and, to especially, to brain connectomics analysis. Each chapter thus is thus geared towards one such application.

Consequently, this thesis has allowed us to gain a brief overview of the different challenges arising along the multiple steps of the analysis graph-structured data, which we summarize as follows:

- **Inferring Graphs.** In most biomedical applications, the data does not spontaneously arise under the form of a graph. Rather, these graphs have to be inferred from the raw data — thus begging the question: what measure of interaction/correlation should we use? How robust should our graph estimates be? This thesis has tackled this problem — tailoring it to Brain Connectomics — through a Bayesian ICA approach in Chapter 4 . This has in particular allowed to achieve more robust subnetworks estimates by (a) fusing different data types and (b) quantifying the uncertainty of each recovered subnetwork through the provision of credible intervals. In particular, we were able to successful deploy this method on a real rs-fMRI data.

- **Comparing Graphs.** Once the different graphs have been inferred, they need to be contrasted and analyzed. The question thus tackled in Chapters 2 and 3 consisted in the definition of an appropriate measure of similarity between aligned networks. To this end, Chapter 2 has aimed to provide a thorough review of different metrics, as well as more guidance as to which distance to choose depending on the analysis at hand. Chapter 3 was geared towards extending this review by providing a way to compare multiscale representations of these graphs.

- **Analysis of signal on Graphs** Finally, Chapter 5 extended our discussion to the case where the graphs are latent, but their structure impacts the behavior of the system as a whole – and in particular, the mechanisms of contagion over graphs. It has thus allowed us to broaden our exposure to graphs and their use in data analysis.

Bibliography

[1] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási. Flavor network and the principles of food pairing. *Scientific reports*, 1:196, 2011.

[2] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[3] A. Banerjee. *The spectrum of the graph Laplacian as a tool for analyzing structure and evolution of networks*. PhD thesis, University of Leipzig, 2008.

[4] A. Banerjee and J. Jost. Spectral plot properties: Towards a qualitative classification of networks. *NHM*, 3(2):395–411, 2008.

[5] A. Barberán, S. T. Bates, E. O. Casamayor, and N. Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal*, 6(2):343, 2012.

[6] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[7] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[8] C. F. Beckmann and S. M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, 23(2):137–152, 2004.

[9] P. J. Bickel, E. Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[10] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[11] S. Biswas, M. McDonald, D. S. Lundberg, J. L. Dangl, and V. Jojic. Learning microbial interaction networks from metagenomic count data. In *International Conference on Research in Computational Molecular Biology*, pages 32–43. Springer, 2015.

[12] B. Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 2013.

[13] A. Bonato, D. F. Gleich, M. Kim, D. Mitsche, P. Pralat, Y. Tian, and S. J. Young. Dimensionality of social networks using motifs and eigenvalues. *PloS one*, 9(9):e106052, 2014.

[14] R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, pages 1–7, 2020.

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[16] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186, 2009.

[17] J. Bun, J.-P. Bouchaud, and M. Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

[18] T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

[19] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[20] J. Chakerian and S. Holmes. Computational tools for evaluating phylogenetic and hierarchical clustering trees. *Journal of Computational and Graphical Statistics*, 21(3):581–599, 2012.

[21] A. Chambolle, V. Duval, G. Peyré, and C. Poon. Geometric properties of solutions to the total variation denoising problem. *Inverse Problems*, 33(1):015002, 2016.

[22] P.-A. Champin and C. Solnon. Measuring the similarity of labeled graphs. In *International Conference on Case-Based Reasoning*, pages 80–95. Springer, 2003.

[23] K. Chan, T.-W. Lee, and T. J. Sejnowski. Variational bayesian learning of ica with missing data. *Neural Computation*, 15(8):1991–2011, 2003.

[24] G. K. Chen, E. C. Chi, J. M. O. Ranola, and K. Lange. Convex clustering: An attractive alternative to hierarchical clustering. *PLoS computational biology*, 11(5):e1004228, 2015.

[25] Y. Chen, A. Wiesel, and A. O. Hero. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107, 2011.

[26] E. C. Chi, G. I. Allen, and R. G. Baraniuk. Convex biclustering. *Biometrics*, 73(1):10–19, 2017.

[27] H. Chipman and R. Tibshirani. Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, 7(2):286–301, 2005.

[28] R. Choudrey and S. Roberts. Variational bayesian independent component analysis with flexible sources. *University of Oxford, Tech. Rep*, 2001.

[29] R. A. Choudrey and S. J. Roberts. Variational mixture of bayesian independent component analyzers. *Neural computation*, 15(1):213–252, 2003.

[30] F. Chung. The heat kernel as the PageRank of a graph. *PNAS*, 104(50):19735–19740, 2007.

[31] N. Connor, A. Barberán, and A. Clauset. Using null models to infer microbial co-occurrence networks. *PloS one*, 12(5):e0176751, 2017.

[32] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9):1505–1512, 2013.

[33] F. Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, 16(22):10881–10890, 1988.

[34] R. Couillet and M. McKay. Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, 131:99–120, 2014.

[35] D. Cvetković. Spectral recognition of graphs. *Yugoslav Journal of Operations Research*, 22(2):145–161, 2012.

[36] D. J. Daley and J. Gani. *Epidemic modelling: an introduction*, volume 15. Cambridge University Press, 2001.

[37] J. S. Damoiseaux and M. D. Greicius. Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity. *Brain structure and function*, 213(6):525–533, 2009.

[38] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'ardenne, W. Richter, J. D. Cohen, and J. Haxby. Independent component analysis for brain fmri does not select for independence. *Proceedings of the National Academy of Sciences*, 106(26):10415–10422, 2009.

[39] W. H. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[40] B. De Martino, C. F. Camerer, and R. Adolphs. Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–3792, 2010.

[41] P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang, and K. H. Jacobsen. Complexity of the basic reproduction number (r0). *Emerging infectious diseases*, 25(1):1, 2019.

[42] L. Dethlefsen and D. A. Relman. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4554–4561, 2011.

[43] S. Diamond and S. Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

[44] D. B. DiGiulio, B. J. Callahan, P. J. McMurdie, E. K. Costello, D. J. Lyell, A. Robaczewska, C. L. Sun, D. S. Goltsman, R. J. Wong, G. Shaw, et al. Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences*, 112(35):11060–11065, 2015.

[45] C. Donnat and S. Holmes. A Constrained Bayesian ICA model for connectome inference. *arXiv preprint*, 2019.

[46] C. Donnat and S. Holmes. Convex Clustering for Graph Data. In *Proceedings of the IEEE Asilomar conference*. IEEE, 2019.

[47] C. Donnat and S. Holmes. Modeling the heterogeneity in covid-19's reproductive number and its impact on predictive scenarios. *arXiv preprint arXiv:2004.05272*, 2020.

[48] C. Donnat, S. Holmes, et al. Tracking network dynamics: A survey using graph distances. *The Annals of Applied Statistics*, 12(2):971–1012, 2018.

[49] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec. Spectral graph wavelets for structural role similarity in networks. *arXiv preprint arXiv:1710.10321*, 2017.

[50] Z. Du, X. Xu, Y. Wu, L. Wang, B. Cowling, and L. Meyers. Serial interval of covid-19 among publicly reported confirmed cases. *Emerging infectious diseases*, 26(6), 2020.

[51] B. Efron, C. Morris, et al. Multivariate empirical bayes and estimation of covariance matrices. *The Annals of Statistics*, 4(1):22–32, 1976.

[52] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111, 2019.

[53] K. Faust and J. Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538, 2012.

[54] M. Ferrer, I. Bardají, E. Valveny, D. Karatzas, and H. Bunke. Median graph computation by means of graph embedding into vector spaces. In *Graph Embedding for Pattern Analysis*, pages 45–71. Springer, 2013.

[55] A. Fornito and E. T. Bullmore. What can spontaneous fluctuations of the blood oxygenation-level-dependent signal tell us about psychiatric disorders? *Current opinion in psychiatry*, 23(3):239–249, 2010.

[56] A. Fornito and E. T. Bullmore. Connectomics: a new paradigm for understanding brain disease. *European Neuropsychopharmacology*, 25(5):733–748, 2015.

[57] A. Fornito, A. Zalesky, and E. Bullmore. *Fundamentals of brain network analysis*. Academic Press, 2016.

[58] A. Fornito, A. Zalesky, C. Pantelis, and E. T. Bullmore. Schizophrenia, neuroimaging and connectomics. *Neuroimage*, 62(4):2296–2314, 2012.

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

[60] A. R. Franco, J. Ling, A. Caprihan, V. D. Calhoun, R. E. Jung, G. L. Heileman, and A. R. Mayer. Multimodal and multi-tissue measures of connectivity revealed by joint independent component analysis. *IEEE journal of selected topics in signal processing*, 2(6):986–997, 2008.

[61] C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PloS one*, 2(8), 2007.

[62] J. Friedman and E. J. Alm. Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9):e1002687, 2012.

[63] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[64] J. Fukuyama, P. J. McMurdie, L. Dethlefsen, D. A. Relman, and S. Holmes. Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Biocomputing 2012*, pages 213–224. World Scientific, 2012.

[65] E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams, and I. Nimmo-Smith. Quickbundles, a method for tractography simplification. *Frontiers in neuroscience*, 6:175, 2012.

[66] A. Gelman, D. Lee, and J. Guo. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543, 2015.

[67] G. K. Gerber. The dynamic microbiome. *FEBS letters*, 588(22):4131–4139, 2014.

[68] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

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

[70] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, 2003.

[71] A. R. Groves. *Bayesian learning methods for modelling functional MRI*. PhD thesis, Oxford University, UK, 2009.

[72] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich. Linked independent component analysis for multimodal data fusion. *Neuroimage*, 54(3):2198–2217, 2011.

[73] J. Gu, J. Jost, S. Liu, and P. F. Stadler. Spectral classes of regular, random, and empirical graphs. *Linear algebra and its applications*, 489:30–49, 2016.

[74] E. Gülden, F. S. Wong, and L. Wen. The gut microbiota and type 1 diabetes. *Clinical Immunology*, 159(2):143–153, 2015.

[75] D. Hammond, P. Vandergheynst, and R. Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[76] H. W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[77] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, page 1, 2011.

[78] C. Honey, O. Sporns, L. Cammoun, X. Gigandet, J.-P. Thiran, R. Meuli, and P. Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009.

[79] M. A. Hullar and J. W. Lampe. The gut microbiome and obesity. In *Obesity Treatment and Prevention: New Directions*, volume 73, pages 67–79. Karger Publishers, 2012.

[80] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

[81] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[82] M. Ipsen and A. S. Mikhailov. Evolutionary reconstruction of networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 66(4):6–9, 2002.

[83] J. Jalanka-Tuovinen, J. Salojärvi, A. Salonen, O. Immonen, K. Garsed, F. M. Kelly, A. Zaitoun, A. Palva, R. C. Spiller, and W. M. de Vos. Faecal microbiota composition and host–microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome. *Gut*, pages gutjnl–2013, 2013.

[84] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[85] J. Jost and M. P. Joy. Evolving networks with distance preferences. *Physical Review E*, 66(3):036126, 2002.

[86] G. Jurman, M. Filosi, S. Riccadonna, R. Visintainer, and C. Furlanello. Differential network analysis and graph classification: a glocal approach. pages 1–13, 2016.

[87] G. Jurman, R. Visintainer, M. Filosi, S. Riccadonna, and C. Furlanello. The HIM glocal metric and kernel for network comparison and classification. *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, 7(1):46109, 2015.

[88] G. Jurman, R. Visintainer, and C. Furlanello. An introduction to spectral distances in networks. *Frontiers in Artificial Intelligence and Applications*, 226:227–234, 2011.

[89] G. Jurman, R. Visintainer, S. Riccadonna, M. Filosi, and C. Furlanello. A glocal distance for network comparison. *arXiv preprint arXiv:1201.2931*, 2012.

[90] C. Kelly, X.-N. Zuo, K. Gotimer, C. L. Cox, L. Lynch, D. Brock, D. Imperati, H. Garavan, J. Rotrosen, F. X. Castellanos, et al. Reduced interhemispheric resting state functional connectivity in cocaine addiction. *Biological psychiatry*, 69(7):684–692, 2011.

[91] A. K. Kelmans. Comparison of graphs by their number of spanning trees. *Discrete Mathematics*, 16(3):241–261, 1976.

[92] A. K. Kelmans. Transformations of a Graph Increasing its Laplacian Polynomial and Number of Spanning Trees. 18:35–48, 1997.

[93] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[94] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673, 2001.

[95] G. Kiar, E. W. Bridgeford, V. Chandrashekhar, D. Mhembere, R. Burns, W. R. G. Roncal, and J. T. Vogelstein. A comprehensive cloud framework for accurate and reliable human connectome estimation and meganalysis. *bioRxiv*, page 188706, 2017.

[96] D. Kim, J. Burge, T. Lane, G. D. Pearlson, K. A. Kiehl, and V. D. Calhoun. Hybrid ica–bayesian network approach reveals distinct effective connectivity differences in schizophrenia. *Neuroimage*, 42(4):1560–1568, 2008.

[97] M. S. Korgaonkar, A. Fornito, L. M. Williams, and S. M. Grieve. Abnormal structural networks characterize major depressive disorder: a connectome analysis. *Biological psychiatry*, 76(7):567–574, 2014.

[98] A. Kourtis, G. Dotsis, and R. N. Markellos. Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix. *Journal of Banking & Finance*, 36(9):2522–2531, 2012.

[99] D. Koutra, A. Parikh, A. Ramdas, and J. Xiang. Algorithms for graph similarity and subgraph matching. In *Proc. Ecol. Inference Conf.*, 2011.

[100] D. Koutra, N. Shah, J. T. Vogelstein, B. Gallagher, and C. Faloutsos. Deltacon: principled massive-graph similarity function with attribution. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(3):28, 2016.

[101] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5):e1004226, 2015.

[102] M. Layeghifard, D. M. Hwang, and D. S. Guttman. Disentangling interactions in the microbiome: a network perspective. *Trends in microbiology*, 25(3):217–228, 2017.

[103] O. Ledoit, M. Wolf, et al. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

[104] P. E. Lekone and B. F. Finkenstädt. Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177, 2006.

[105] M. Levandowsky and D. Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.

[106] R. Li, K. Chen, A. S. Fleisher, E. M. Reiman, L. Yao, and X. Wu. Large-scale directional connections among multi resting-state neural networks in human brain: a functional mri and bayesian network modeling study. *Neuroimage*, 56(3):1035–1042, 2011.

[107] W. Liao, J. Li, X. Duan, Q. Cui, H. Chen, and H. Chen. Static and dynamic connectomics differentiate between depressed patients with and without suicidal ideation. *Human brain mapping*, 39(10):4105–4118, 2018.

[108] Y. Lu, K. Huang, and C.-L. Liu. A fast projected fixed-point algorithm for large graph matching. *Pattern Recognition*, 60:971–982, 2016.

[109] M. M. Luqman, J.-Y. Ramel, and J. Lladós. Multilevel analysis of attributed graphs for explicit graph embedding in vector spaces. In *Graph Embedding for Pattern Analysis*, pages 1–26. Springer, 2013.

[110] D. J. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Technical report, Citeseer, 1996.

[111] N. D. Monnig and F. G. Meyer. The resistance perturbation distance: A metric for the analysis of dynamic networks. *arXiv preprint arXiv:1605.01091*, 2016.

[112] B. Naul and J. Taylor. Sparse steinian covariance estimation. *Journal of Computational and Graphical Statistics*, 26(2):355–366, 2017.

[113] M. E. Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

[114] H. Ombao, M. Lindquist, W. Thompson, and J. Aston. *Handbook of Neuroimaging Data Analysis*. Chapman and Hall/CRC, 2016.

[115] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina. Web graph similarity for anomaly detection. *Journal of Internet Services and Applications*, 1(1):19–30, 2010.

[116] K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor. Convex clustering shrinkage. In *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*, 2005.

[117] S. R. Proulx, D. E. Promislow, and P. C. Phillips. Network thinking in ecology and evolution. *Trends in ecology & evolution*, 20(6):345–353, 2005.

[118] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001.

[119] M. E. Raichle and A. Z. Snyder. A default mode of brain function: a brief history of an evolving idea. *Neuroimage*, 37(4):1083–1090, 2007.

[120] J. M. Read, J. R. Bridgen, D. A. Cummings, A. Ho, and C. P. Jewell. Novel coronavirus 2019-ncov: early estimation of epidemiological parameters and epidemic predictions. *medRxiv*, 2020.

[121] S. Roberts and R. Choudrey. Bayesian independent component analysis with prior constraints: An application in biosignal analysis. In *International Workshop on Deterministic and Statistical Methods in Machine Learning*, pages 159–179. Springer, 2004.

[122] O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. In *Signal Processing Conference, 2007 15th European*, pages 606–610. IEEE, 2007.

[123] M. Rubinov and D. S. Bassett. Emerging evidence of connectomic abnormalities in schizophrenia. *Journal of Neuroscience*, 31(17):6263–6265, 2011.

[124] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[125] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[126] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Advances in neural information processing systems*, pages 2101–2109, 2010.

[127] E. Segal and D. Koller. Probabilistic hierarchical clustering for biological data. In *Proceedings of the sixth annual international conference on Computational biology*, pages 273–280. ACM, 2002.

[128] Y. Shimada, Y. Hirata, T. Ikeguchi, and K. Aihara. Graph distance for complex networks. *Scientific reports*, 6:34944, 2016.

[129] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

[130] D. Shuman, B. Ricaud, and P. Vandergheynst. Vertex-frequency analysis on graphs. *Applied and Computational Harmonic Analysis*, 40(2):260–291, 2016.

[131] P. H. Sneath, R. R. Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

[132] M. O. Sommer, G. M. Church, and G. Dantas. The human microbiome harbors a diverse reservoir of antibiotic resistance genes. *Virulence*, 1(4):299–303, 2010.

[133] D. A. Spielman. Spectral graph theory and its applications. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 29–38. IEEE, 2007.

[134] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun. A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of neuroscience methods*, 204(1):68–81, 2012.

[135] J. Sui, G. Pearlson, A. Caprihan, T. Adali, K. A. Kiehl, J. Liu, J. Yamamoto, and V. D. Calhoun. Discriminating schizophrenia and bipolar disorder by fusing fmri and dti in a multimodal cca+ joint ica model. *Neuroimage*, 57(3):839–855, 2011.

[136] K. M. Tan and D. Witten. Statistical properties of convex clustering. *Electronic journal of statistics*, 9(2):2324, 2015.

[137] E. H. Telzer. Dopaminergic reward sensitivity can promote adolescent health: A new perspective on the mechanism of ventral striatum activation. *Developmental cognitive neuroscience*, 17:57–67, 2016.

[138] P. Tétreault, A. Mansour, E. Vachon-Presseau, T. J. Schnitzer, A. V. Apkarian, and M. N. Baliki. Brain connectivity predicts placebo response across chronic pain clinical trials. *PLoS biology*, 14(10):e1002570, 2016.

[139] P. N. Thibaut Jombart, Anne Cori. *earlyR: Estimation of Transmissibility in the Early Stages of a Disease Outbreak*. https://CRAN.R-project.org/package=earlyR, 2017.

[140] M. Thüne. *Eigenvalues of matrices and graphs*. PhD thesis, University of Leipzig, 2012.

[141] S. M. Tom, C. R. Fox, C. Trepel, and R. A. Poldrack. The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518, 2007.

[142] N. Tremblay et al. Graph wavelets for multiscale community mining. *IEEE TSP*, 62(20):5227–5239, 2014.

[143] P. J. Turnbaugh, F. Bäckhed, L. Fulton, and J. I. Gordon. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell host & microbe*, 3(4):213–223, 2008.

[144] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *nature*, 444(7122):1027–131, 2006.

[145] K. Uludağ and A. Roebroeck. General overview on the merits of multimodal neuroimaging data fusion. *Neuroimage*, 102:3–10, 2014.

[146] H. Valpola and P. Pajunen. Fast algorithms for bayesian independent component analysis. In *Proceedings of the second international workshop on independent component analysis and blind signal separation, ICA'00*, pages 233–238, 2000.

[147] M. P. van den Heuvel and A. Fornito. Brain networks in schizophrenia. *Neuropsychology review*, 24(1):32–48, 2014.

[148] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(Apr):1201–1242, 2010.

[149] B. Wahlberg, S. Boyd, M. Annergren, and Y. Wang. An admm algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, 45(16):83–88, 2012.

[150] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 10(7):1669, 2016.

[151] S. Whitfield-Gabrieli, S. Ghosh, A. Nieto-Castanon, Z. Saygin, O. Doehrmann, X. Chai, G. Reynolds, S. Hofmann, M. Pollack, and J. Gabrieli. Brain connectomics predict response to treatment in social anxiety disorder. *Molecular psychiatry*, 21(5):680–685, 2016.

[152] J. T. Wu, K. Leung, and G. M. Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet*, 2020.

[153] L. C. Xia, D. Ai, J. Cram, J. A. Fuhrman, and F. Sun. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics*, 29(2):230–237, 2012.

[154] W. Xue, F. D. Bowman, A. V. Pileggi, and A. R. Mayer. A multimodal approach for determining brain networks by jointly modeling functional and structural connectivity. *Frontiers in computational neuroscience*, 9:22, 2015.

[155] J. Yao, S. Zheng, and Z. Bai. *Sample covariance matrices and high-dimensional data analysis*, volume 2. Cambridge University Press Cambridge, 2015.

[156] L. A. Zager and G. C. Verghese. Graph similarity scoring and matching. *Applied mathematics letters*, 21(1):86–94, 2008.

[157] R. Y. Zhang, S. Fattahi, and S. Sojoudi. Large-scale sparse inverse covariance estimation via thresholding and max-det matrix completion. *arXiv preprint arXiv:1802.04911*, 2018.

[158] S. Zhao, Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*, 2020.

[159] D. Zhou, B. Schölkopf, and T. Hofmann. Semi-supervised learning on directed graphs. In *NIPS*, volume 5, pages 1633–1640, 2004.

[160] X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. C. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1:140049, 2014.

Appendices

Appendix A

Complements for the distances between networks

A.1 The 2011 Relman microbiome study : a network perspective

In this appendix, we present more details about the 2011 Relman microbiome study that serves as one of the guiding thread examples to our discussion.

The data. As briefly mentioned in the introduction, this longitudinal study consists of a set of 162 bacterial samples taken from the gut of three distinct subjects (D, E, and F) at different points in time, with roughly the same number of samples (52 to 57) per subject. The subjects were given two courses of antibiotics over a ten month period, yielding seven distinct treatment phases (pre-treatment, first antibiotic course, week after stopping treatment 1, interim, second course of antibiotics, week after stopping treatment 2, and post-treatment phase).

The Graph Paradigm. While this dataset has been thoroughly analyzed over the past years [64, 42], we view it through a novel network-based angle. Representation of microbial interactions as a graph has become standard practice[117, 5, 150, 102, 67]. Indeed, bacteria live in symbiosis, feed and proliferate within each body site, yielding potentially complex higher-order interactions. Recent developments in ecology has shown the synergy between bacteria and how it explains various pathologies, such as drug resistant infections [132], inflammatory-bowel disease [83], or diabetes [74]. A deeper understanding of these interactions is thus a necessary step towards more effective translational medicine. In this framework, networks come as a natural tool. They allow to ask a

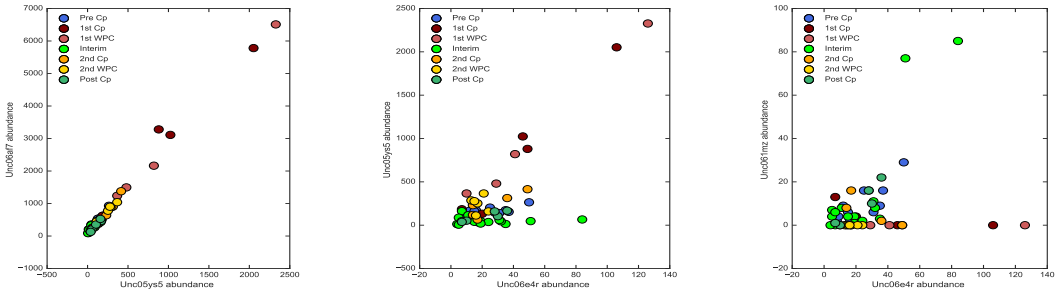
variety of questions such as: are there any significant microbial synergies? Can these be associated to a given pathology? Network structure can also yield insight in the response of the microbial community to perturbations [31, 53].

Network Inference. The pre-processing of the data to infer graphs plays a crucial step in the analysis. Microbiome samples are particularly challenging to analyze: generally modeled as zero-inflated negative binomial data, microbiome samples typically exhibit a high number of zero counts. Methods have to take into account the specificity of these data, and a plethora of different methods have been suggested for finding associations between bacteria [62, 11, 101, 153]. We refer the reader to [150] and [102] for a review and comparison of the different methods for inferring networks from co-occurrence data.

In our setting, we do have a small number of samples per treatment phase, and as such, the estimated correlations are inherently noisy. Keeping in mind this noise level, we construct a set of Bacterial "symbiosis" graphs as follows:

• For each subject at a given treatment phase, we define a graph in which each node corresponds to a specific species of bacteria, and edges $\mathcal{E} = \{(i, j)\}$ capture pairwise "affinities" (as measured by the correlation of the abundances through time within each of the different treatment phases) between bacteria i and j. Due to the large number of zeros in the data, we use the rank-based correlation metric, Kendall's ρ, as a measure of correlation.

• We fix a threshold for each graph: in our case, the threshold that we selected was 0.5, and ensured reasonable sparsity of the network (from 0.02 to 0.3). We emphasize that in this setting, the edges can be "spurious' in the graph, and should not be interpreted as "statistically significant interactions" between bacteria. In fact, these edges are simply a reflection of the co-occurrence between bacteria within different subjects at each different time phase. Figures A1a and A1b show the dynamics in a few cases, highlighting the existence of different synergies for each of the treatment phases.



(a) Unc05ys5 vs Unc06af7 (b) Unc06e4r vs Unc05ys5 (c) Unc06e4r vs Unc061mz

Figure A1: **Subject F.** Examples of scatterplots for a few bacteria. Colors denote different treatment phases. The affinities between bacteria vary across treatment phases.

A.2 From fMRI data to brain connectomics

In this section, we present more details on the fMRI dataset we used as one of our examples.

The dataset that we have selected consists in the resting-state fMRI of 29 patients under cocaine-dependency. This dataset was gathered by Kelly and co-authors[90] as part of an NIDA-funded grant (R03DA024775) on the impact of cocaine addiction on structural and functional connectivity [1]. This dataset consists in 6-minute resting-state fMRI of patients, pre-processed as per the AFNI and FSL standard pipelines. In particular, the preprocessing included:

- various corrections to account for small head movements and heterogeneity of the fMRI images. These corrections included (as per [90]): slice time correction; 3-D motion correction; temporal despiking; spatial smoothing (FWHM=6mm); mean-based intensity normalization; temporal bandpass filtering (0.009–0.1Hz); linear and quadratic detrending;
- nuisance signal removal (white matter, CSF, global signal, motion parameters) via multiple regression.
- linear registration of functional to structural images (with intermediate registration to a low-resolution image and b0 unwarping)
- nonlinear registration of structural images to the MNI152 template: this allows to align the brain to a common "template" brain.

This yields a total of 116 time-series corresponding to the MNI152 template's nodes, with 140 points each. Consistent with [90], we use these filtered time-series to define a graph based on these series' pairwise Pearson correlation. In order to filter each subject's correlation matrix, we followed an approach akin to [138] and opted for a threshold controlling the graphs' sparsity. The threshold used here is the mean (across subject) of the 97^{th} quantile of each patient' correlation matrix (taking only the off-diagonal coefficients). On average, the graphs that we recover have around 3% sparsity, a level in accordance with typical analyses in the field [138] .

This dataset also contains covariates for each subject, with a total 14 variables including gender, age, number of years since first use, smoker, and number of years under dependency. Figure B1 shows the distribution of some of these features.

In their 2011 article [90], the authors use these covariates as evidence of a decreased functional connectivity for patients under cocaine dependence (with respect to healthy control). In this paper, we use our different graph distances to assess whether patients with similar "years of dependency" are more similar.

[1] The data is publicly available at the following link `http://fcon_1000.projects.nitrc.org/indi/ACPI/html/acpi_nyu_1.html`.

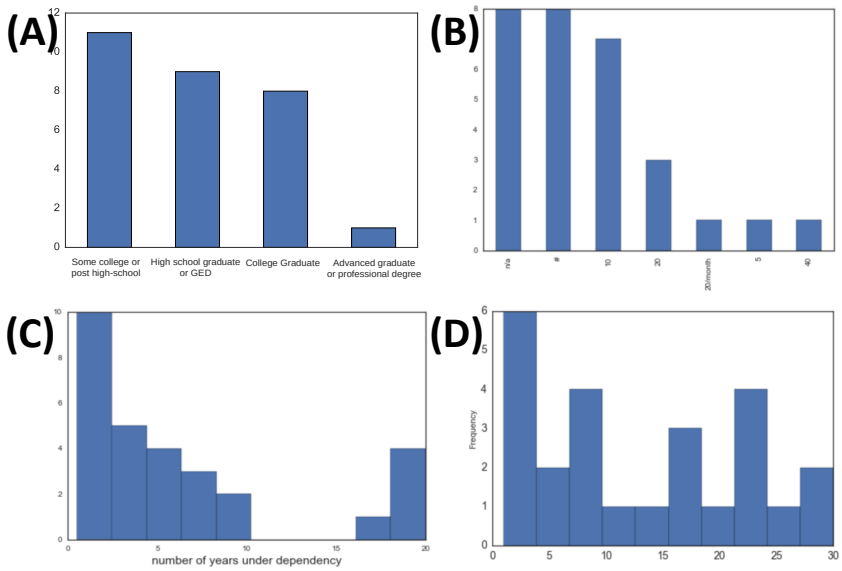


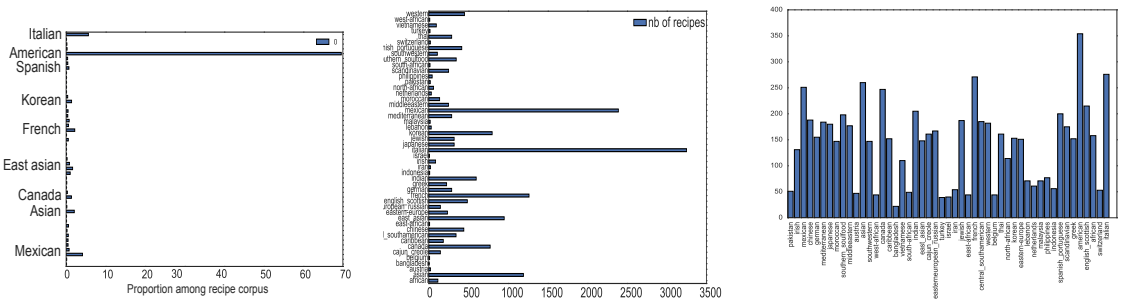
Figure B1: **(A)** Education level. **(B)** Number of cigarettes per day. **(C)** Number of years under dependency. **(D)** Number of years since first use.

A.3 The recipes network

In this appendix, we present in greater details the recipes dataset analyzed in section 2.6. The original dataset consists in 57,691 recipes scraped from three different American culinary websites (*allrecipes*, *epicurious*, and *menupan.com*) that were gathered in [1] as part as a study on food pairing associations. Each recipe is labeled by corresponding cuisine (French, American, Greek, etc...). This yields a total of 49 different labels. In this review, we analyzed this dataset from a network perspective by constructing graphs from the co-occurrences of the 1,530 different ingredients in each cuisine. Each of 1,530 ingredients constitutes a node in the graph and each of the 49 cuisine is assigned to a weighted graph. The weight on the edge is the frequency of co-occurrence of the two ingredients for that particular cuisine. We note that in this case, each graph includes a collection of disconnected nodes (ingredients that never appear in a single recipe) and a weighted connected component.

Before beginning the analysis of the different graphs, let us quickly highlight the potential challenges of this particular dataset:

- the representation of the different cuisines is highly imbalanced (Figures C1a and C1b). While the American cuisine is extremely well represented (with a little over 40,000 recipes, or 70% of the recipes), conversely, other cuisines are underrepresented: the Bangladesh cuisine, for



(a) Percentage of recipes from a given cuisine in the corpus. We note the predominance of American cuisine, with more than $2/3^{rd}$ of the corpus.

(b) Number of recipes per cuisine (aside from the American cuisine, for visualization purposes). The minimum is achieved by the Bangladesh cuisine, with only 4 recipes)

(c) Number of ingredients in the connected component. Maximum is achieved by the American cuisine (354 connected ingredients) while minimum is achieved by the Bangladesh cuisine (22 ingredients).

Figure C1: Visualization of the properties of the dataset

instance, only appears 4 times in this corpus. It follows that the number of the ingredients appearing in the connected component of the co-occurrence graphs also varies substantially (Figure C1c).

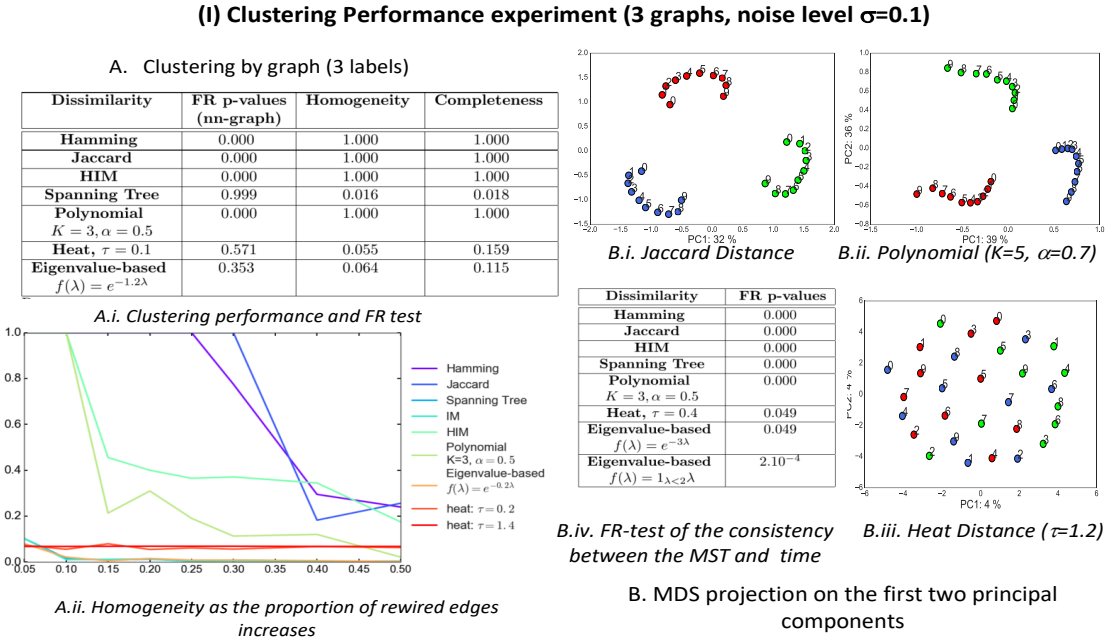
- Consequently, each cuisine's connected component only accounts for a small fraction of the $1,530$ nodes (Figure C1c). As such, the distance between graphs can be considered very small and the graphs very similar, since more than 80% of the nodes are disconnected from the maximum connected component in both graphs. To account for this imbalance, the distance between cuisine A and B is computed only with respect to the ingredients appearing in either A or B's connected component.

A discussion of the different distances' application to this dataset can be found in the main body of the article. We propose here to discuss an additional benefit of the heat distance over the others–that is, its granularity, which allow to capture in which parts of the graphs the most important changes occurs. Indeed, as detailed in section 2.9 and in equation 2.4.3, the heat distance compares in fact the graphs' "topological" signatures, and tracks the variation of the nodes' topological roles from one graph to the other. As such, it is easy to go back to the node level to understand where the variation from one graph to the other is the strongest. We also note that, as described in [49], these structural signatures can be enriched to contain information at multiple scales, yielding a richer "multi-scale" representation of the topological role of each node. Table A.1 shows the list of 10 ingredients present in the connected components of two cuisines whose representations have changed the most (from one cuisine to the other). In this case, we have used a multi-scale representation of the signatures (with scale $\tau \in \{1, \cdots, 29\}$).

Ingredient comparisons (heat-wavelet based distances)		
Cuisine	Neighbor	top changes (char. distance)
Middle Eastern	Indian	mustard, dill, bread, thyme, oregano, feta cheese, walnut sesame seed, coconut, olive
	Moroccan	chive, nut, red wine, feta cheese, cane molasses, yogurt, rose, oregano, fennel, walnut
	Spanish	apricot, lentil, mint, zucchini walnut, pork sausage, feta cheese, sesame seed, lamb, yogurt
Chinese	Asian	black bean, oyster, turmeric, cumin, lime juice, nira, coconut, basil, beef broth, lime
	Japanese	lemon, oyster, salmon, buckwheat enokidake, tuna,radish, barley, kelp, katsuobushi
	Thai	peanut butter, mint, roasted peanut, fenugreek, turmeric, lime juice, cumin, coconut, basil, lime

Table A.1: Identification of the ingredients that change the most from one graph to another

A.4 Results for the synthetic experiments



Dissimilarity	FR p-values (nn-graph)	Homogeneity	Completeness
Hamming	0.000	1.000	1.000
Jaccard	0.000	1.000	1.000
HIM	0.000	1.000	1.000
Spanning Tree	0.999	0.016	0.018
Polynomial $K=3, \alpha=0.5$	0.000	1.000	1.000
Heat, $\tau=0.1$	0.571	0.055	0.159
Eigenvalue-based $f(\lambda)=e^{-1.2\lambda}$	0.353	0.064	0.115

Dissimilarity	FR p-values
Hamming	0.000
Jaccard	0.000
HIM	0.000
Spanning Tree	0.000
Polynomial $K=3, \alpha=0.5$	0.000
Heat, $\tau=0.4$	0.049
Eigenvalue-based $f(\lambda)=e^{-3\lambda}$	0.049
Eigenvalue-based $f(\lambda)=1_{\lambda<2}\lambda$	2.10^{-4}

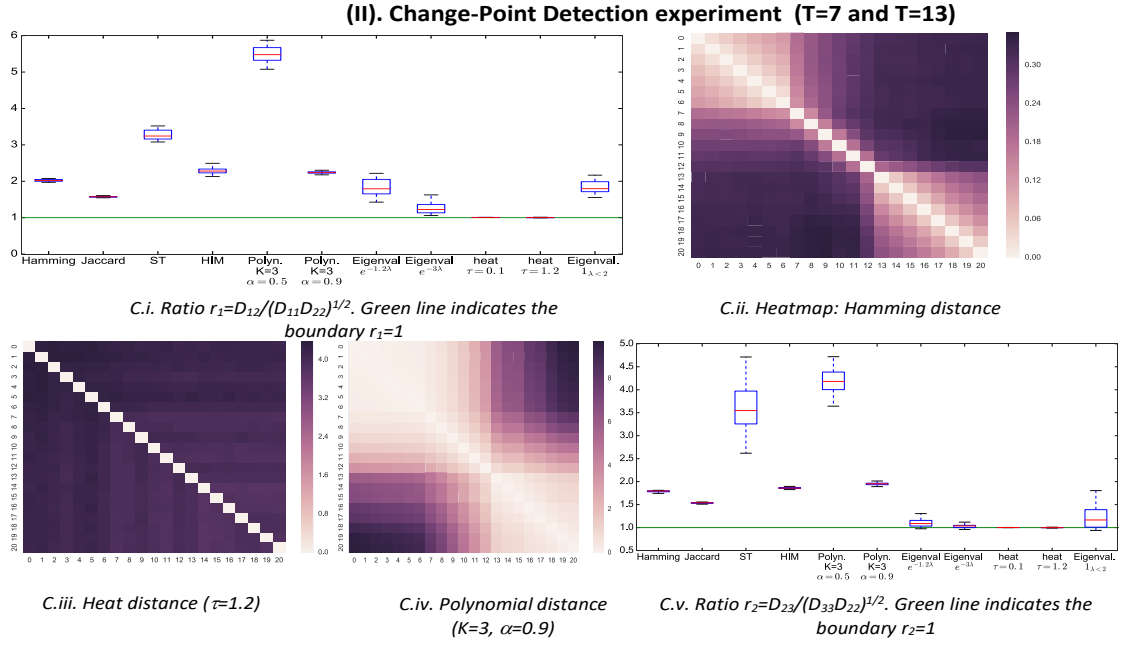
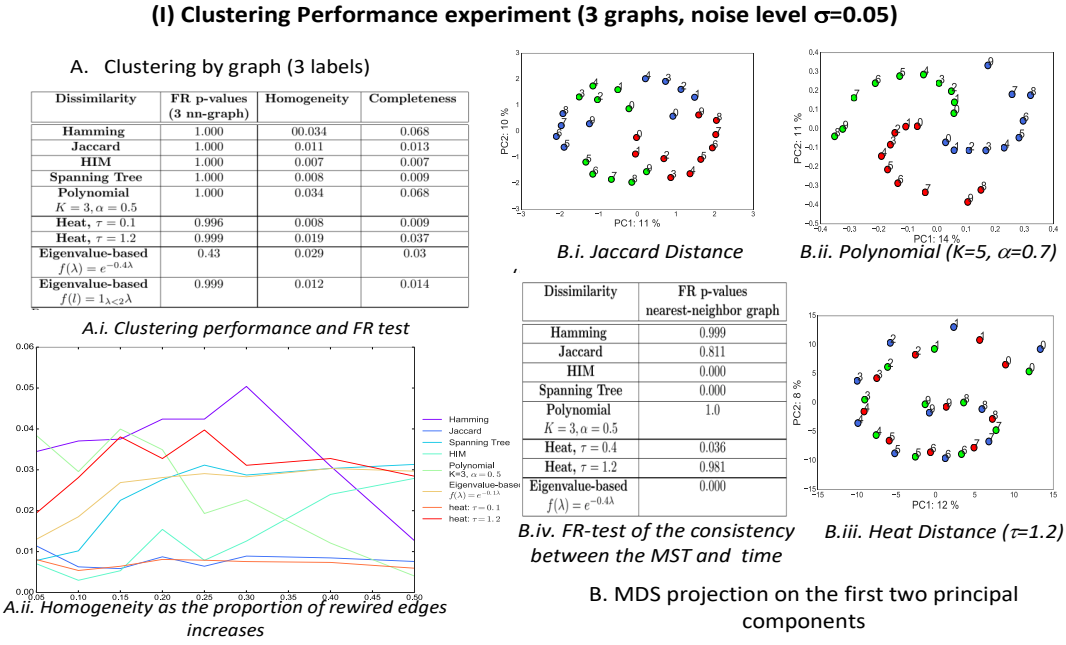


Figure D1: Results for the Erdös-Rényi topology. **Top Row:** Comparison of the smooth dynamics (no change point), with 0.1% edges rewired at each time step. **Bottom Row:** Change point detection experiment.



Dissimilarity	FR p-values (3 nn-graph)	Homogeneity	Completeness
Hamming	1.000	00.034	0.068
Jaccard	1.000	0.011	0.013
HIM	1.000	0.007	0.007
Spanning Tree	1.000	0.008	0.009
Polynomial $K=3, \alpha=0.5$	1.000	0.034	0.068
Heat, $\tau=0.1$	0.996	0.008	0.009
Heat, $\tau=1.2$	0.999	0.019	0.037
Eigenvalue-based $f(\lambda)=e^{-0.4\lambda}$	0.43	0.029	0.03
Eigenvalue-based $f(l)=1_{\lambda<2}\lambda$	0.999	0.012	0.014

Dissimilarity	FR p-values nearest-neighbor graph
Hamming	0.999
Jaccard	0.811
HIM	0.000
Spanning Tree	0.000
Polynomial $K=3, \alpha=0.5$	1.0
Heat, $\tau=0.4$	0.036
Heat, $\tau=1.2$	0.981
Eigenvalue-based $f(\lambda)=e^{-0.4\lambda}$	0.000

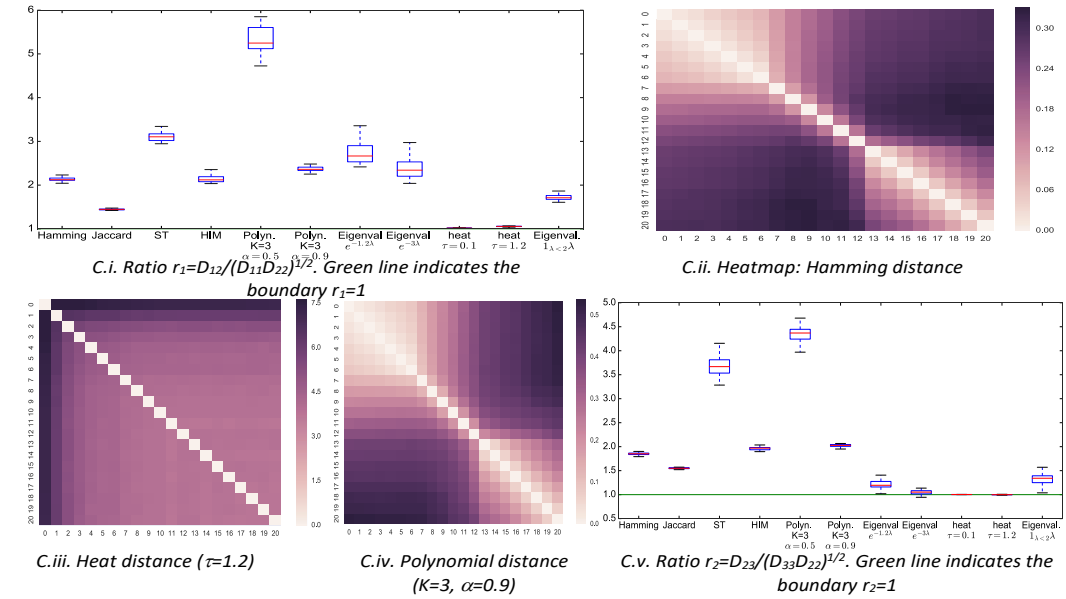


Figure D2: Results for the Preferential Attachment topology. **Top Row:** Comparison of the smooth dynamics (no change point), with 5% edges rewired at each time step. **Bottom Row:** Change point detection experiment.

A.5 Understanding the HIM parameters

In this appendix, we give details about the effect of the parameter choices for several of the distances described in this paper.

Ipsen-Mikhailov distance

Returning to the definition of the Ipsen-Mikhailov distance provided in Eq. 2.3.3. The main drawback of this characterization is that it relies on two parameters, K and γ, whose values can be approximated but for which the literature (to the best of our knowledge) does not provide any motivation – giving this interesting metric a black-box flavor. To be more explicit and understand how this distance behaves for different types of graphs, let us explicitly determine approximations (in the limit of large N) of its parameters.

Normalization constant K. To begin with, the normalization constant can be written as:

$$\begin{aligned} K\int_0^\infty \rho(\omega,\gamma)d\omega = 1 &\iff K\sum_{i=1}^{N-1}\int_0^\infty \frac{1}{\gamma}\frac{1}{1+\left(\frac{\omega-\omega_i}{\gamma}\right)^2}d\omega = 1 \\ &\iff K\sum_{i=1}^{N-1}\underbrace{\left[\arctan\left(\frac{\omega-\omega_i}{\gamma}\right)\right]_0^\infty}_{=\pi/2+\arctan(\omega_i/\gamma)} = 1 \\ &\iff K = \frac{1}{(N-1)\frac{\pi}{2}+\sum_{i=1}^{N-1}\arctan(\omega_i/\gamma)} \end{aligned} \tag{E2}$$

This yields:

- **for the empty graph $\mathcal{E}_n$**, which has eigenvalue 0 with multiplicity N:

$$K_{\mathcal{E}_n} = \frac{1}{\frac{(N-1)\pi}{2}}$$

- **for the complete graph** $\mathcal{F}_n$, where $\omega_0 = 0$, and $\omega_1^2 = \cdots = \omega_{N-2}^2 = \omega_{N-1}^2 = N$:

$$\begin{aligned} K_{\mathcal{F}_n} &= \frac{1}{(N-1)(\frac{\pi}{2} + \arctan(\sqrt{N}/\gamma))} = \frac{1}{(N-1)(\pi - \arctan(\gamma/\sqrt{N}))} \\ &= \frac{1}{(N-1)(\pi - \gamma/\sqrt{N}[1 - \frac{1}{3}(\gamma^2/N) + o(1/N)])} \\ &= \frac{1}{(N-1)\pi}[1 + \frac{\gamma}{\pi\sqrt{N}}[1 - \frac{1}{3}\frac{\gamma^2}{N} + o(\frac{1}{N})]) + \frac{\gamma^2}{\pi^2 N} + \frac{\gamma^3}{\pi^3 N^{3/2}} + o(\frac{1}{N^{3/2}})] \\ &= \frac{1}{(N-1)\pi}[1 + \frac{\gamma}{\pi\sqrt{N}} + o(\frac{1}{\sqrt{N}})] \end{aligned} \tag{E3}$$

Hence:

$$\boxed{K_{\mathcal{F}_n} \approx \frac{1}{2} K_{\mathcal{E}_n}} \tag{E4}$$

The Intuition behind the Scale Parameter. Through properties of the Lorenz distribution, the scale parameter γ is equal to half the interquartile range. It is thus a measure of the "probable measurement error" with respect to the mode ω_i. Here, as proposed by [87], we have chosen γ such that the spectral distance between the empty graph and the complete graph is 1: $\epsilon_{\bar{\gamma}}(\mathcal{E}_N, \mathcal{F}_N) = 1$. By definition of the Ipsen-Mikhailov distance, and using Eq.E4 to estimate the different normalizing constants:

$$\begin{aligned} \rho_{\mathcal{F}_N}(\omega,\gamma) - \rho_{\mathcal{E}_N}(\omega,\gamma) &= \frac{N-1}{\gamma} K_{F_n}[\frac{1}{(\frac{\omega-\sqrt{N}}{\gamma})^2 + 1} - \frac{2}{(\frac{\omega}{\gamma})^2 + 1}] \\ &= \frac{N-1}{\gamma} K_{F_n}[\frac{\gamma^2}{N}\frac{1}{1 - 2\frac{\omega}{\sqrt{N}} + \frac{\omega^2}{N} + \frac{\gamma^2}{N}} - \frac{2}{(\frac{\omega}{\gamma})^2 + 1}] \\ &= \frac{1}{\pi\gamma}\Big(\frac{\gamma^2}{N}[1 - \frac{2\omega}{\sqrt{N}} + o(\frac{1}{N^{1/2}})] - \frac{2}{(\frac{\omega}{\gamma})^2 + 1}\Big) \quad \text{since } (N-1)K_{F_n} \approx \frac{1}{\pi} \\ &= \frac{1}{\pi\gamma}[-\frac{2}{(\frac{\omega}{\gamma})^2 + 1}] + o(\frac{1}{N}) \end{aligned} \tag{E5}$$

Hence:

$$\begin{aligned} \int_0^\infty [\rho_{\mathcal{F}_N}(\omega,\gamma) - \rho_{\mathcal{E}_N}(\omega,\gamma)]]^2 d\omega &= \int_0^\infty \frac{1}{\pi^2\gamma^2}\frac{4}{((\frac{\omega}{\gamma})^2 + 1)^2} d\omega + o(\frac{1}{N}) \\ &= \frac{4}{\pi^2\gamma}\frac{1}{2}[\frac{\omega\gamma}{\omega^2 + \gamma^2} + \arctan(\frac{\omega}{\gamma})]_0^\infty \\ &= \frac{1}{\pi\gamma} \end{aligned} \tag{E6}$$

Hence, here: $\bar{\gamma} \approx \frac{1}{\pi}$

Comparison with other spectral distances

The advantages Ipsen-Mikhailov distance has two advantages over the other spectral distances: it has a "physical" interpretation as the joint behavior of a system on N strings. Secondly, empirically, it outperforms other spectral distances in capturing regime changes in the dynamics of the network (see the experiments detailed in section 5). However, the running time required by this distance increases dramatically with the number of nodes, making it a less practical tool to work with on large graphs.

The advantages of the Ipsen-Mikhailov distance over the spectral distances previously introduced is linked to its use of the ℓ_2 norm of the spectral densities rather than raw eigenvalues themselves: the integration of this density makes the distance more robust to small perturbations that have little structural effect, making this distance more appropriate to the analysis of complex systems.

We can assess the impact of the perturbation of a graph G, with adjacency matrix A. By supposing that the perturbation is small enough that each vibration frequency $\omega_i = \sqrt{\lambda_i}$ of the Laplacian is perturbed by an amount ϵ_i . Writing $\tilde{\omega}_i = \omega_i + \epsilon_i$, we have:

$$\begin{aligned}
\frac{1}{(\frac{\omega-\omega_i-\epsilon}{\gamma})^2+1} =& \frac{1}{(\frac{\omega-\omega_i}{\gamma})^2+1-\frac{2\epsilon(\omega-\omega_i)}{\gamma^2}+\frac{\epsilon^2}{\gamma^2}} = \frac{1}{(\frac{\omega-\omega_i}{\gamma})^2+1-\frac{2\epsilon}{\gamma^2}((\omega-\omega_i)-\frac{\epsilon}{2})} \\
=& \frac{1}{(\frac{\omega-\omega_i}{\gamma})^2+1}\Big[1+\frac{\frac{2\epsilon}{\gamma^2}((\omega-\omega_i)-\frac{\epsilon}{2})}{(\frac{\omega-\omega_i}{\gamma})^2+1}+4\frac{\epsilon^2(\omega-\omega_i)^2}{\gamma^4((\frac{\omega-\omega_i}{\gamma})^2+1)^2}+o(\epsilon^2)\Big] \\
=& \frac{1}{(\frac{\omega-\omega_i}{\gamma})^2+1}\Big[1+\frac{2\epsilon(\omega-\omega_i)}{\gamma^2((\frac{\omega-\omega_i}{\gamma})^2+1)} \\
& -\frac{\epsilon^2}{\gamma^2((\frac{\omega-\omega_i}{\gamma})^2+1)}[1-4\frac{(\omega-\omega_i)^2}{\gamma^2((\frac{\omega-\omega_i}{\gamma})^2+1)}]+o(\epsilon^2)\Big]
\end{aligned}$$

Hence, the IM distance between the spectra of graphs G and $\tilde{G}$ becomes:

$$
\begin{aligned}
&\int_0^\infty [\frac{1}{(\frac{\omega-\omega_i-\epsilon_i}{\gamma})^2+1} - \frac{1}{(\frac{\omega-\omega_i}{\gamma})^2+1}]^2 d\omega \\
&= \int_0^\infty \frac{4\epsilon^2}{\gamma^2} \frac{1}{[(\frac{\omega-\omega_i}{\gamma})^2+1]^4} \Big[\frac{(\omega-\omega_i)}{\gamma}\Big]^2 d\omega + o(\epsilon^2) \\
&= \int_{-\omega_i/\gamma}^\infty \frac{\gamma 4\epsilon^2}{\gamma^2} \frac{x^2}{[x^2+1]^4} dx + o(\epsilon^2) \\
&= \frac{4\epsilon^2}{\gamma} \times \frac{1}{2\times 48} \Big[\frac{(x(-3+8x^2+3x^4)}{(1+x^2)^3} + 3\text{arctan}[x])\Big]_{-\omega_1/\gamma}^\infty + o(\epsilon^2) \\
&= \frac{\epsilon^2}{24\gamma} \Big[3\pi/2 - (3\text{arctan}(-\omega_i/\gamma) - \frac{(\omega_i/\gamma)(-3+8(\omega_i/\gamma)^2+3(\omega_i/\gamma)^4)}{(1+(\omega_i/\gamma)^2)^3})\Big] + o(\epsilon^2) \\
&= \frac{\epsilon^2}{16\gamma} \Big[\pi + 2\text{arctan}(\omega_i/\gamma) + \frac{2}{3}\frac{(\omega_i/\gamma)(-3+8(\omega_i/\gamma)^2+3(\omega_i/\gamma)^4)}{(1+(\omega_i/\gamma)^2)^3}\Big] + o(\epsilon^2) \\
&= \frac{\epsilon^2}{16\gamma} \Big[\pi + 2\text{arctan}(\omega_i/\gamma) + \frac{2(\omega_i/\gamma)}{3}[\frac{-8}{(1+(\omega_i/\gamma)^2)^3} + \frac{2}{(1+(\omega_i/\gamma)^2)^2} \\
&+ \frac{3}{1+(\omega_i/\gamma)^2}]\Big] + o(\epsilon^2)
\end{aligned}
$$

Summing over all perturbed eigenvalues and taking the square root yields:

$$d_{IM}(A,\tilde{A}) \propto \sqrt{\sum_i \epsilon_i^2} + o(||\epsilon||_2)$$

and the variations are of the order $||\epsilon||_2$. By comparison, provided that the perturbation of each frequency ω_i is small enough and since $\tilde{\lambda}_i = \lambda_i + 2\epsilon_i\sqrt{\lambda_i} + \epsilon_i^2$, from Eq. 2.3.1, the standard ℓ_p-distances are such that:

$$d(A,\tilde{A})^p = \sum_{i=1}^{N-1} |2\sqrt{\lambda_i} f'(\lambda_i)|^p \epsilon_i^p + o(\sum_{i=1}^{N-1} \epsilon_i^p).$$

As such, the distance between G and $\tilde{G}$ puts more emphasis on changes in the eigenspectrum where the product $\sqrt{\lambda_i} f'(\lambda_i)$ is large – which might result in putting too much weight on "noisy" components of the signal, to continue with the signal frequency analogy of section 2.3.1. Moreover, if we now want to compare the IM distance with the spanning tree distance, we note that the spanning tree (ST) distance is such that:

$$
\begin{aligned}
d_{ST}(A,\tilde{A}) &= \sum_{i=1}^{N-1} |\log(\lambda_i + \epsilon_i^2 + 2\sqrt{\lambda_i}\epsilon_i) - \log(\lambda_i)| \\
&= \sum_{i=1}^{N-1} |\log(1 + \frac{\epsilon_i^2}{\lambda_i} + \frac{2\epsilon_i}{\sqrt{\lambda_i}})| \approx \sum_{i=1}^{N-1} \frac{\epsilon_i}{\sqrt{\lambda_i}} \qquad (**)
\end{aligned}
$$

where (**) holds provided that the perturbation remains small compared to the magnitude of the corresponding eigenvalues. However, in the case of sparse graphs, the second smallest eigenvalue (the algebraic connectivity of the graph) is typically very small (and bounded below by $\frac{4}{ND}$ where D is the diameter of the graph). This eigenvalue might thus actually be of the order of the perturbation, thus yielding high variability in the proposed log-ST distance.

The Ipsen-Mikhailov distance can be interpreted as providing an embedding of the distances between graphs in a probabilistic setting, where distributions over eigenvalue densities are compared. This explains its increased robustness to small changes or local perturbations.

Appendix B

Convex Clustering

B.1 Derivation of the ADMM updates

In this appendix, we provide the derivations of the ADMM algorithm used to benchmark our FISTA-based approach in section 3.4.

Description of the algorithm

The Alternating Direction Method of Multipliers [15] is a popular algorithm for solving convex optimization problems with a large number of constraints. Indeed, with a guaranteed speed of convergence in $O(\frac{1}{k})$ iterations, this algorithm has become the work-horse of convex problems with coupling constraints. However, contrary to the parameter-free implementation of convex clustering with FISTA, ADMM requires the selection of the parameter ρ, whose choice has been shown to considerably affect the speed of convergence [15, 149]. In what follows, in order to simplify the notations, denoting as e the vectors of the Cartesian basis, we introduce the pairwise-difference matrix $\delta \in \mathbb{R}^{N\times N^2} : \delta_{k,ij} = \mathbf{e}_{k,i} - \mathbf{e}_{k,j}$ Introducing the variables $Z_{ij} = \pi_i - \pi_j$ and dual variables u_{ij}, the ADMM-augmented Lagrangian can be written as:

$$\begin{aligned} \min_{\pi\in\Delta_N} \frac{1}{2}\mathrm{Tr}(\pi^T K\pi - 2K^T\pi) + \frac{\rho}{2}\sum_{ij}||\pi\delta + u_{ij} - Z_{ij}||^2 \\ + \lambda\sum_{ij} K_{ij}(\alpha||Z_{ij}||_1 + 1 - \alpha)||Z_{ij}||_2) \\ \text{s. t. } \pi \in \Delta_N, \quad \forall i,j, \quad \pi_i - \pi_j = \pi\delta_{ij} = Z_{ij} \end{aligned} \tag{E1}$$

The full algorithm and derivation of the updates are provided in the following subsection and the whole procedure is summarized in Alg. 3, and the corresponding updates are derived in the following paragraphs.

Input: Similarity matrix K, regularization parameter λ
Output: Optimal solution $\pi^{(\lambda)}$
Initialization; $Z, U = \mathbf{0} \in \mathbb{R}^{N \times N^2}, t = 0$
while not converged **do**
  $\pi^{t+1} = \text{Update}_{\pi}(Z^t, U^t)$ {explicited in Algorithm 4}
  $Z^{t+1} \leftarrow \text{SoftThreshold}_{\frac{\alpha\lambda||Z^t||}{\rho||Z^t||+(1-\alpha)\lambda}} \left[\frac{\pi^{t+1}\delta+U^t}{(1+\frac{(1-\alpha)\lambda}{\rho||Z^t||})}\right]$
  $U^{t+1} \leftarrow U^t + (\pi^{t+1}\delta - Z^{t+1})$
  $t \leftarrow t+1$
end while
Return $\pi^* = \Pi_{\Delta_N}(\pi)$

Algorithm 3: ADMM

Updates

Updating π. The objective in Eq. E1 reads as quadratic linear optimization problem in π. The updates in X are unfortunately not computable in closed form. However, provided that we have access to an efficient projection on the set of doubly stochastic matrices Δ_N, we can solve the corresponding update using an accelerated Proximal Descent algorithm. In particular, the gradients with respect to π are given by:

$$\nabla_{\pi} F(\pi, Z, u) = K\pi - K + \rho(\pi\delta + U - Z)\delta^T$$

We note that these gradients are in particular Lipschitz (with respect to π, all other variables being fixed):

$$\nabla_{\pi} F(\pi_1, Z, u) - \nabla_{\pi} F(\pi_2, Z, u) = K(\pi_1 - \pi_2) + \rho(\pi_1 - \pi_2)\delta\delta^T$$

We also have:

$$\delta\delta^T = \Big(\sum_{ij}(e_{ki} - e_{kj})(e_{li} - e_{lj})\Big)_{kl} = 2\Big(\sum_{j}(e_{lk} - e_{lj})\Big)_{kl} = 2\Big(ne_{lk} - e_{ll}\Big) = 2nI - 2\mathbf{1}\mathbf{1}^T$$

$$\begin{aligned}
&\implies ||\delta\delta^T||_F^2 = \text{Trace}[4n^2 I - 8n\mathbf{1}\mathbf{1}^T + 4n\mathbf{1}\mathbf{1}^T] \\
&\implies ||\delta\delta^T||_F^2 \leq 4n^3 \\
&\implies ||\nabla_{\pi} F(\pi_1, Z, u) - \nabla_{\pi} F(\pi_2, Z, u)||_F \\
&\quad \leq \sqrt{||K||_F^2 + \rho^2||\delta\delta^T||^2}||\pi_1 - \pi_2||_F \\
&\quad \leq \sqrt{||K||_F^2 + 4\rho^2 n^3}||\pi_1 - \pi_2||_F
\end{aligned} \tag{E2}$$

Hence, to solve for π, we can use an accelerated proximal method (such as FISTA), with constant step-size $L = \sqrt{||K||_F^2 + 4\rho^2 n^3}$. Since, the projection onto Δ_N does not have a closed form solution either, the literature typically resorts to fixed-point algorithms such as the one proposed in [108], yielding the procedure described in Algorithm 4.

Algorithm: Updates for $\pi^t(\lambda)$
Input:(fixed) variables K, Z and U initialization: $t_k = 1$
while not converged **do**
  $\pi^k = \Pi^{\Delta_N}_{\text{1-round}}(Y_k - \frac{1}{\sqrt{||K||_F^2 + 4\rho^2 n^3}} \nabla_\pi F(Y^k, Z^t, U^t))$ (Projection $\Pi^{\Delta_N}_{\text{1-round}}$ described in Alg. 5)
  $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
  $Y_{k+1} \leftarrow \pi_k + \frac{t_k - 1}{t_{k+1}}(\pi_k - \pi_{k-1})$
end while

Algorithm 4: Updates for π.

Objective: Projection on Δ_N: $\Pi^{\Delta_N}_{\text{1-round}}$
Input:square matrix Y
Initialization: $P = Y$
while not converged **do**
  $P \leftarrow P + (\frac{1}{n} I + \frac{\mathbf{1}^T P \mathbf{1}}{n^2} I - \frac{1}{n} P)\mathbf{1}\mathbf{1}^T - \frac{1}{n} 1 1^T P$
  $P \leftarrow \frac{P + |P|}{2}$
end while
Return P

Algorithm 5: One-round of the fixed point iterative algorithm ($\Pi^{\Delta_N}_{\text{1-round}}$) for projecting unto Δ_N as proposed in [108].

Updating Z. The updates in terms of Z are more explicit, since Z is simply the solution to a denoising problem with an elastic net penalty.

Taking the gradient with respect to Z yields:

$$\rho(Z - X\delta - U) + \lambda\alpha \text{sign} Z + \lambda(1-\alpha)\frac{Z}{||Z||} = 0$$

We solve the later through a set of sequential updates:

$$\begin{aligned} (1 + \frac{(1-\alpha)\lambda}{\rho||Z^{t-1}||})Z^t &= X\delta + U + \frac{\lambda\alpha}{\rho}\text{sign} Z^t \\ \implies Z^t &= \text{SoftThreshold}_{\frac{\alpha\lambda||Z^{t-1}||}{\rho||Z^{t-1}|| + (1-\alpha)\lambda}} \Big[\frac{1}{(1 + \frac{(1-\alpha)\lambda}{\rho||Z^{t-1}||})}(X\delta + U)\Big]. \end{aligned}$$

B.2 Large Scale Computations: derivation of the linearization algorithm

We provide here an alternative way of solving the problem using gradient descent, which might be better suited to larger scale problems. We begin by reminding that the problem that we are solving has the following form:

$$\begin{aligned}
&\text{argmin}_\pi \text{Tr}[\pi^T K\pi - 2K\pi] + \lambda \sum_{i,j} K_{ij} ||\Phi\pi_{.i} - \Phi\pi_{.j}||^2 \\
\iff &\text{argmin}_\pi \text{Tr}[\pi^T K\pi - 2K\pi] \\
&+ \lambda \sum_{i,j} K_{ij} (||\Phi\pi_{.i}||^2 + ||\Phi\pi_{.j}||^2 - 2\pi_{.i}^T \phi^T \Phi\pi_{.j}) \qquad \text{(E1)} \\
\iff &\text{argmin}_\pi \text{Tr}[\pi^T K\pi - 2K\pi] \\
&+ 2\lambda \text{Tr}[\pi^T K\pi \text{Diag}(\tilde{K}\mathbf{1})] - 2\mathbf{1}^T(\pi^T K\pi \odot \tilde{K})\mathbf{1}
\end{aligned}$$

where $\mathbf{1}^T\pi = \mathbf{1}$, and $\tilde{K} = K - \text{diag}(K)$. Let us write $\Delta = \{\pi \in \mathbb{R}^{n\times n} : \mathbf{1}^T\pi = \mathbf{1}\}$ the space of row-wise stochastic matrices. To compute a solution, we propose using an iterative algorithm based on a linearization of the previous objective function. Introducing x such that $||x|| \leq \delta$, at each iteration t, we update π^t as $\pi^t = x + \pi^{t-1}$. All we need is thus to solve for the updates x at each iteration. Linearizing the previous equation with respect to x yields:

$$\begin{aligned}
&\text{argmin}_{||x|| \leq \delta} \text{Tr}[2\pi_{t-1}^T Kx - 2Kx] \\
&+ 2\lambda \text{Tr}[\text{Diag}(\tilde{K}\mathbf{1})\pi_{t-1}^T Kx] - 2\lambda \mathbf{1}^T((\pi_{t-1}^T Kx + x^T K\pi_{t-1}) \odot \tilde{K})\mathbf{1}
\end{aligned} \qquad \text{(E2)}$$

such that $||x|| \leq \delta$ and $\pi^t = \pi^{t-1} + x \in \Delta$. The gradient with respect to x of the previous objective function is:

$$\nabla_x \ell(x, \pi^{t-1}) = 2K\pi^{t-1} - 2K + 2\lambda K\pi^{t-1}\text{Diag}(\tilde{K}\mathbf{1}) + 4\lambda K\pi^{t-1}\tilde{K}.$$

Denoting as $\Pi_{\mathcal{B}_\delta}$ and Π_Δ respectively the projections on the ball of radius δ and on Δ, we can thus use projected gradient descent to solve the previous problem, as described in Alg. 6.

Objective: Solve Eq. E2
Input: $K, \tilde{K}$ and initialized π^0
Initialization: $x = 0$
while not converged **do**
  $x \leftarrow \Pi_{\mathcal{B}_\delta}(x - \eta\nabla_x \ell(x, \pi^{t-1}))$
  $\pi^t \leftarrow \Pi_\Delta(\pi^{t-1} + x)$
end while
Return π^t

Algorithm 6: Linearization algorithm

Appendix C

Bayesian ICA

C.1 fMRI-Studies

HNU - Hangzhou Normal University- dataset.[1]

This sample includes 30 healthy adults, aged 20 to 30. Each participant received ten scans across one month, one scan every three days. Five modalities (EPI/ASL/T1/DTI/T2) of brain images were acquired for all subjects. During functional scanning, subjects were presented with a fixation cross and were instructed to keep their eyes open, relax and move as little as possible while observing the fixation cross. Subjects were also instructed not to engage in breath counting or meditation. All imaging data was collected on 3T GE Discovery MR 75 using an 8-channel head coil [160].

A Functional fMRI Pre-processing

The fMRI data used in this chpater is the time series corresponding to the Craddock-200 atlas. Functional parcellation was accomplished using a two-stage spatially-constrained functional procedure applied to preprocessed and unfiltered resting state data corresponding to 41 individuals from an independent dataset (age: 18–55; mean 31.2; std. dev. 7.8; 19 females). A grey matter mask was constructed by averaging individual-level grey matter masks derived by automated segmentation. Individual-level connectivity graphs were constructed by treating each within-gm-mask voxel as a node and edges corresponding to super-threshold temporal correlations to the voxels 3D (27 voxel) neighborhood. Each graph was partitioned into 200 regions using normalized cut spectral clustering. Association matrices were constructed from the clustering results by setting the connectivity between voxels to 1 if they are in the same ROI and 0 otherwise. A group-level correspondence matrix was constructed by averaging the individual level association matrices and subsequently partitioned into 200 regions using normalized cut clustering. The resulting group-level analysis was fractionated into

[1] All the information in this subsection was gathered on the study's website.

functional resolution using nearest-neighbor interpolation.

The pre-processing was done according to the Configurable Pipeline for the Analysis of Connectomes (C-PAC). This python-based pipeline tool makes use of AFNI, ANTs, FSL, and custom python code.

As per the ABIDE's detailed explanations, this pipeline includes:

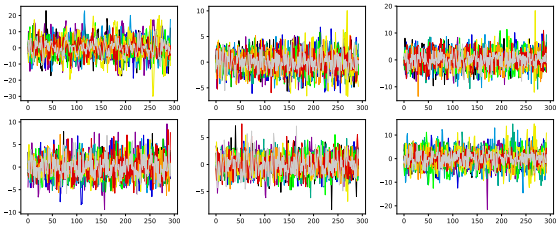
- **Structual Preprocessing**
  1. Skull-stripping using AFNI' 3dSkullStrip.
  2. Segment the brain into three tissue types using FSL's FAST,
  3. Constrain the individual subject tissue segmentations by tissue priors from standard space provided with FSL.
  4. Individual skull stripped brains were normalized to Montreal Neurological Institute (MNI)152 stereotactic space (1 mm^3 isotropic) with linear and non-linear registrations using ANTs.
- **Functional Preprocessing**
  1. Slice time correction using AFNI 3dTshift
  2. Motion correct to the average image using AFNI's 3dvolreg (two iterations)
  3. Skull-strip using AFNI 3dAutomask
  4. Global mean intensity normalization to 10,000
  5. Nuisance signal regression was applied including motion parameters:
     - 6 head motion parameters, 6 head motion parameters, and the 12 corresponding squared items
     - top 5 principal components from the signal in the white-matter and cerebro-spinal fluid derived from the prior tissue segmentations transformed from anatomical to functional space
     - linear and quadratic trends
  6. Band-pass filtering (0.01-0.1Hz) was applied for only for one set of strategies
  7. Functional images were registered to anatomical space with a linear transformation and then a white-matter boundary based transformation using FSL's FLIRT and the prior white-matter tissue segmentation from FAST
  8. The previous anatomical to standard space registration was applied to the functional data in order to transform them to standard space.

HNU 1

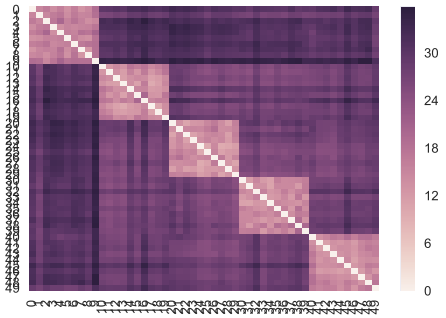
	Anatomical	Rest	DTI
Manufacturer	GE	GE	GE
Model	Discovery MR750	Discovery MR750	Discovery MR750
Headcoil	8 Chan	8 Chan	8 Chan
Field Strength	3T	3T	3T
Sequence	3D SPGR	EPI	-
Flip Angle [Deg]	8.0	90	-
Inversion Time (TI) [ms]	450	-	-
Echo Time (TE) [ms]	Min Full	30	Min
Repetition Time (TR) [ms]	8.06	2000	8600.0
Bandwidth per Voxel (Readout) [Hz]	125	3437.5	-
Parallel Acquisition	ASSET x 2	-	-
Partial Fourier	Off	-	-
Number of Slices	180	43	68.0
Slice Orientation	Saggital	Axial	-
Slice Phase Encoding Direction	Anterior to Posterior	Anterior to Posterior	Right to Left
Slice Acquisition Order	Interleaved Ascending	Interleaved Ascending	Interleaved Ascending
Slice Thickness [mm]	1.0	3.4	1.5
Slice Gap [mm]	0	0	0.0
Field of View [mm]	250	220	192.0
Acquisition Matrix	250x250	64x64	128x128
Slice In-Place Resolution [mm^2]	1.0x1.0	3.4x3.4	1.5x1.5
Number of Measurements	-	300	33
Acquisition Time [min:sec]	5:01	10:00	-
Fat Suppression	None	Yes	Yes
Number of Directions	-	-	-
Number of B Zeros	-	-	-
B Value(s) [s/mm^2]	-	-	-
Averages	-	-	-

Figure E1: Scan Parameters for the HNU1 dataset.

Fig. E2 shows the raw fMRI time series of the data, as well as the distance matrix between a few adjacency matrices. As indicated by the block diagonal structure in Fig. E3b, adjacency matrices corresponding to the anatomical white matter connections between brain regions exhibit a strong subject effect: structural connectomes are much similar within each subject.



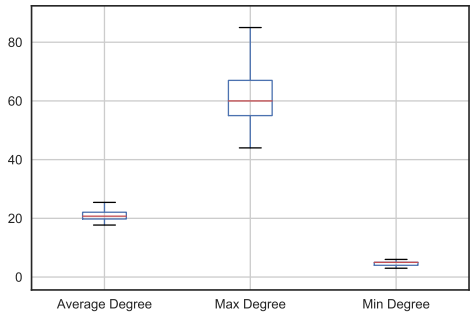
(a) Time Series for a randomly selected subset of nodes across different subjects and scans



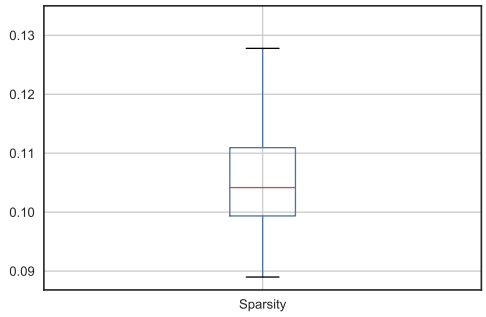
(b) Distances between adjacency matrices for 5 scans across 3 different subjects.

Figure E2: A few visualization of the data

Fig. E3 also highlights some of the properties of the graphs that we have at hand. In particular, the structural sparsity in this dataset (that is, the number of edges in the graph over the total number of possible edges) is 0.113. The structures that we work with are thus relatively sparse.



(a) Properties of the node degree distributions



(b) Average sparsity in the structural connectomes

Figure E3: Structural Graph properties

A Structural Processing Pipeline: NDMG

The structural connectomes that we use in this paper were obtained from the neurodata's website and have been processed through **NDMG**. As per the package's website, the Python pipeline **NDMG**

[95] provides end-to-end "robust and reliable estimates of MRI connectivity at 1mm resolution [..] from 48 to 72,000 nodes, all in MNI152 standard space." The details can be found in the associated paper [95], but, for the sake of clarity and completeness, we summarize here the four key components of this pipeline.

(1) Registration. NDMG begins by using FSL to perform a series of linear "standard" registrations on minimally preprocessed DWI and T1W images and aligns them to the MNI152 atlas.

(2) Tensor Estimation. As per [95], the MNI152-aligned "diffusion volumes and b-values/b-vectors files are transformed into a 6-dimensional tensor volume. A fractional anisotropy map of the tensors is provided for QA, again using multiple depths in each of the three canonical image planes."

(3) Tractography. As per [95], a deterministic tractography algorithm (Dipy's EuDX [65]) is used to generate streamlines. Each voxel at the boundary of the brain mask "is used as a seed-point in EuDX and fibers are produced and then pruned based on their length".

(4) Graph Generation. Fibers are traced through predefined parcellations and "an undirected edge is added to the graph for every pair of regions along the path, where the weight of the edge is the cumulative number of fibers between two regions"[95].

Here, consistently with the choice of the parcellation for the fMRI pre-processing, we have chosen the structural connectomes for the Craddock 200 atlas.

C.2 Numerical approximation

We begin by providing a numerical approximation to the constrained ICA approach. While this approach does not allow us to get confidence intervals for the parameters that we are estimating, it will nonetheless produce ICA components that are consistent with our 3 hypotheses of section 4.2: non-negative, sparse and localized. This method can be used either for warm-starting the Bayesian HMC algorithm, or by itself if point estimates are sufficient to the analysis.

To do so, we take a step back from the Bayesian model, which requires to estimate both S and A, while only A is of true importance in the subnetwork discovery problem. Since by assumption, $\frac{1}{T}\mathbb{E}[S^T S] = I$, we propose integrating S out of the model altogether, and focusing on finding a solution for A that reconstructs the correlation of the observed time series:

$$\frac{1}{T}Y^T Y \approx A^T \frac{S^T S}{T} A + \text{Diag}(\gamma^2) \approx A^T A + \text{Diag}(\gamma^2)$$

This leads us to consider the Stein loss associated to the matrix $A^T A$. Following [112], the Stein loss between two Positive Definite Symmetric (PDS) matrices Σ and $\hat{\Sigma} \in \mathbb{R}^{N \times N}$ is defined as:

$$L(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1}) - \log(|\hat{\Sigma}\Sigma^{-1}|) - N$$

Using $\Sigma = S = \frac{Y^T Y}{T}$ the sample correlation matrix, and $\hat{\Sigma} = A^T A + \text{diag}(\gamma_i^2)$, we want to minimize:

$$\begin{aligned} &\text{Minimize}_{\hat{\Sigma}} tr(\hat{\Sigma} S^{-1}) - \log(|\hat{\Sigma} S^{-1}|) - N \\ &\text{such that } \hat{\Sigma} = A^T A + \text{diag}(\gamma_i) \\ &\qquad \text{diag}(\hat{\Sigma}) = 1 \\ &\qquad \forall i \leq N, \quad \gamma_i \in (0, 1) \\ &\qquad ||A||_1 \leq s \qquad \text{(sparsity)} \\ &\qquad \forall k, A_k L A_k^T \leq \rho \quad \text{(connectedness)} \end{aligned} \tag{E1}$$

Formulated as such, this problem amounts to finding an optimal low-rank decomposition under Stein loss that is also aligned with our original sparsity and localization assumptions. However, this loss is convex as a function of $\hat{\Sigma}$ but not a a function of A. To solve this new problem, following [**?**], we use a composite algorithm [**?**, **?**], which linearizes the previous loss and provides a series of convex objectives to sequentially optimize:

$$\begin{aligned} &\text{Minimize}_{\hat{\Sigma}} tr(\hat{\Sigma} \Sigma^{-1}) - \log(|\hat{\Sigma} \Sigma^{-1}|) - p \\ &\text{such that } \hat{\Sigma} = A^{(k)T} A^{(k)} + A^{(k)T} X + X^T A^{(k)} + \text{diag}(\gamma_i) \\ &\qquad \text{diag}(\hat{\Sigma}) = 1 \\ &\qquad \forall i \leq N, \quad \gamma_i \in (0, 1) \\ &\qquad \forall j \leq K, \quad ||X_{j\cdot}||_2 \leq \delta \\ &\qquad A^{(k+1)} = A^{(k)} + X \geq 0 \qquad \text{(non negative entries)} \\ &\qquad ||A^{(k+1)}||_1 \leq s \qquad \text{(sparsity)} \\ &\qquad \forall k, A^{(k+1)} L A^{(k+1)T} \leq \rho \qquad \text{(connectedness)} \end{aligned} \tag{E2}$$

Each iteration in Eq. E3 is now convex and we thus know that there exists a global optimal solution. Note that while each iteration is convex, the overall problem remains non-convex, and the solution might thus heavily depend on global initialization. We suggest use as warm start the absolute value of the Vanilla ICA loadings, as these can typically be very efficiently computed and provide good starting points[2].

The algorithm operates as follows:

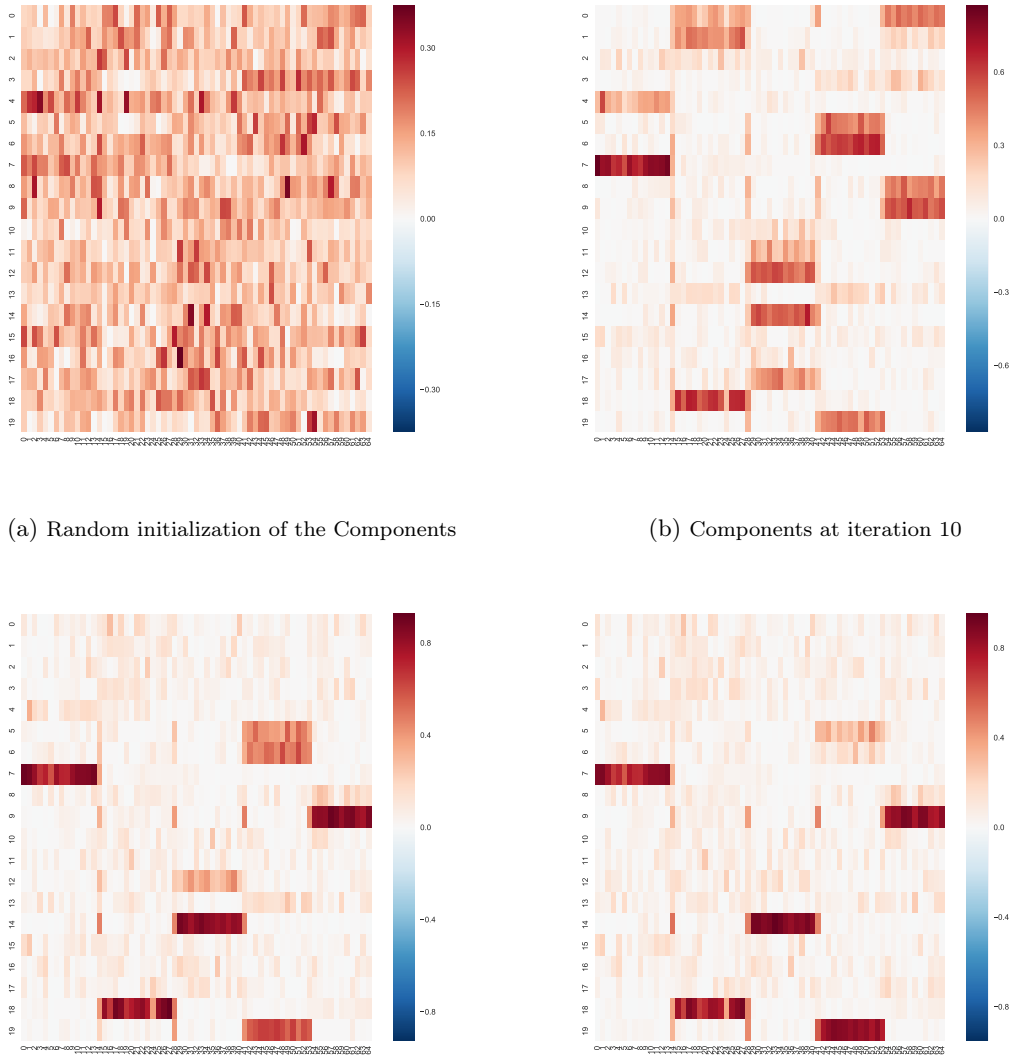
1. Optimize Eq. E3 with respect to X and γ^2
2. Solve for $A^{(k+1)}$
3. Iterate until convergence

(A) Accurate resolution of the problem using CVXPY. As a proof of concept, we solve the previous problem using CVXPY [43]. CVXPY allows to find accurate solutions to the previous

[2]The code is publicly available at `https://github.com/donnate/ConstrainedICA`

problem, and we use it here to check if the algorithm works efficiently on our synthetic example of section 3.4.

Fig. E4 allows us to validate the method: we see that, as the number of iteration increases, the recovered components become sparser and sparser. In particular, it is interesting to note that after only 10 iterations, the Ground-Truth components are recovered (the colored blocks in E4b match the localized ground truth components that we expect to see). However, the components are redundant. As the number of iteration increases, redundant components fade away and the algorithm recovers both the number of ground truth components (5, in this case) and the ground truth components themselves accurately (Fig. E4c and E4d).



(a) Random initialization of the Components

(b) Components at iteration 10

(c) Components at iteration 75

(d) Components at iteration 200

Figure E4: Visualization of the recovered components at various steps of the optimization process.

(B) Scalable resolution of the problem. While, for smaller problems, this can accurately be solved using numerical solvers such as CVXPY, for large problems, we need to find more efficient solvers, and potentially compromise between accuracy and scalability. The previous paragraph has highlighted the potential of the algorithm to extract accurate components, and we now turn to a more scalable derivation of the updates of our two-step algorithm.

Step 1: Gradient descent to solve for X and γ^2. The first step focuses on solving the following

objective:

$$\begin{aligned}
&\text{Minimize}_{\ \hat{\Sigma}} tr(\hat{\Sigma}S^{-1}) - \log(|\hat{\Sigma}|) \\
&\text{such that } \hat{\Sigma} = A^{(k)T}A^{(k)} + A^{(k)T}X + X^T A^{(k)T} + \text{diag}(\gamma_i) \\
&\qquad \text{diag}(\hat{\Sigma}) = 1 \\
&\qquad \forall i \leq N, \quad \gamma_i \in (0,1) \\
&\qquad \forall i \leq N, \quad ||X_i|| \leq \delta \\
\iff &\text{Minimize}_{\ X,\gamma} 2tr(XS^{-1}A^{(k)T}) + tr(\text{Diag}(\gamma)S^{-1}) \\
&\qquad - \log\Big(|A^{(k)^T}A^{(k)} + ((A^{(k)})^T X + X^T A^{(k)} + \text{Diag}(\gamma)|\Big) \\
&\text{such that } \text{diag}\Big((A^{(k)})^T A^{(k)} + (A^{(k)})^T X + X^T A^{(k)} + \text{diag}(\gamma_i)\Big) = 1 \\
&\qquad \forall i \leq N, \quad ||X_i||_2 \leq \delta \\
&\qquad \forall i \leq N, \quad \gamma_i \in (0,1)
\end{aligned} \tag{E3, E4}$$

We solve the previous problem using a gradient descent method. In particular, here, we propose using FISTA [6] updates, which provide an optimal step size (thus foregoing the need for the practioner to select it) and are typically fast to converge. Derivations of the sequential steps are provided in the subsequent paragraphs.

Computing the gradient with respect to X. The objective in Eq. E4 are convex with respect to X, and each column of X is in the ball of radius δ. All we need to show now if that the gradients are Lipschitz in order to apply the FISTA updates.

We begin by deriving the gradients and updates for X and γ in Eq. E4

$$\begin{aligned}
&\log(|(A^TA + A^T(X+tH) + (X+tH)^TA + \text{Diag}(\gamma))|) \\
=&\log(|M + t(H^TS + A^TH)S^{-1}|) \quad \text{where } M = (A^TA + A^TX + X^TA + \text{Diag}(\gamma)) \\
=&\log(|M|) + \log(|I + tM^{-1}(H^TS + A^TH)|) + o(t) \\
=&\log(|M|) + t\text{Trace}[M^{-1}(H^TA + A^TH)] + o(t) \\
=&\log(|M|) + t\text{Trace}[H^TAM^{-1} + M^{-1}A^TH] + o(t) \\
=&\log(|M|) + 2t\text{Trace}[H^TAM^{-1}] + o(t)
\end{aligned} \tag{E5}$$

Thus, noting that M is symmetric:

$$\nabla_X(-\log(|A^TA + 2A^T(X+tH)|)) = -2(M^{-1}A^T)^T = -2AM^{-1},$$

and the full gradient with respect to X is thus given by:

$$\nabla_X L = 2A(S^{-1} - M^{-1})$$

Now, defining $K = A^T A + \text{Diag}(\gamma)$, we also have:

$$\begin{aligned} M &= A^T A + X^T A + A^T X + \text{Diag}(\gamma^2) \\ &= K^{1/2}[I_N + K^{-1/2}(X^T A + A^T X)K^{-1/2}]K^{1/2} \\ \implies M^{-1} &\approx K^{-1/2}[I_N - K^{-1/2}(X^T A + A^T X)K^{-1/2}]K^{-1/2} \\ \implies AM^{-1} &\approx AK^{-1/2}[I_N - K^{-1/2}(X^T A + A^T X)K^{-1/2}]K^{-1/2} \end{aligned}$$

Hence:

$$\begin{aligned} ||\nabla_X L(X_1, \gamma) - \nabla_X L(X_2, \gamma)|| &\leq 2||AK^{-1}(X_1 - X_2)^T AK^{-1} + AK^{-1}A^T(X_1 - X_2)K^{-1}|| \\ ||\nabla_X L(X_1) - \nabla_X L(X_2)|| &\leq \sqrt{8}(||AK^{-1}|| \times ||K^{-1}| \times ||A||) \times ||X_2 - X_1|| \end{aligned}$$

Thus the gradient of the loss in Eq. E4 is Lipschitz with a constant upper-bounded by $L_X = \sqrt{8}(||AK^{-1}|| \times ||K^{-1}| \times ||A||)$.

Computing the gradient with respect to γ. Similarly:

$$\begin{aligned} &\log(|A^T A + X^T A + A^T X + \text{Diag}((\gamma + th))|) \\ &= \log(|M + t\text{Diag}(h)|) \\ &= \log(|M|) + \log(|I + tM^{-1}\text{Diag}(h)|) \\ &= \log(|M|) + tTr(M^{-1}\text{Diag}(h)) + o(t) \\ \nabla_\gamma L &= \text{diag}(S^{-1}) - \text{diag}(M^{-1}) = \text{diag}(S^{-1} - M^{-1}) \end{aligned} \tag{E6}$$

The gradient of the loss with respect to γ is thus also Lipschitz, with constant upper-bounded by $L_\gamma = \max\{\text{Diag}(K_2^{-1})\}^2$:

$$\Big(\nabla_\gamma L(X, \gamma_1) - \nabla_\gamma L(X, \gamma_2)\Big) \leq \max\{\text{Diag}(K_2^{-1})\}^2 ||\gamma_1 - \gamma_2||$$

where $K_2 = A^T A + A^T X + X^T A$. The FISTA update for X are summarized in Alg. 7.

Updated $X^{(k)}$
while not converged **do**
$X_k = p_{L_X}(Y_X)$
$X_k = \Pi_{\mathcal{B}_\delta}(Y_X - \frac{1}{L_X}\nabla_X L)$
$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
$Y_X = X_k + \frac{t_k - 1}{t_{k+1}}(X_k - X_{k-1})$
end while

Algorithm 7: FISTA Updates for X

$\gamma^{(k)}$
while not converged **do**
$\gamma_k = p_{L_\gamma}(Y_\gamma)$
$\iff \gamma_k = \Pi_{[0,1]}(Y_\gamma - \frac{1}{L_\gamma}\nabla_\gamma L)$
$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
$Y_\gamma = \gamma_k + \frac{t_k - 1}{t_{k+1}}(\gamma_k - \gamma_{k-1})$
end while

Algorithm 8: FISTA Updates for γ

Step 2: Updating $A^{(k+1)}$. Step 2 can be easily solved by introducing the augmented Lagrangian of the problem and solving:

$$\begin{aligned} &\min_A \frac{1}{2}||(A^{(k)} + X) - A||^2 + s||A||_1 + \frac{\rho}{2}Trace(ALA^T) \\ &\text{such that } \text{Diag}(A) + \text{Diag}(\gamma) = 1 \end{aligned} \tag{E7}$$

We propose to solve this problem using FISTA updates, and projecting onto the subspace $\Delta = \{M \in$

$\mathbb{R}^{N\times N} : \forall i, M_{ii} = 1\}$ of matrices with diagonal equal to 1. The gradient here is given by:

$$\nabla L_2 = (A - (A^{(k)} + X)) + \rho AL$$

which is thus Lipschtiz with constant $L_A = ||I + \rho * L||$ The FISTA updates for solving Eq. E3, combined with those for X and γ are summarized in Algorithm 9.

---

subnetworks A

$A = |A_{\text{Vanilla ICA}}|$

while not converged **do**

$X_k = \Pi_{\mathcal{B}_\delta}(Y_X - \frac{1}{L_X}\nabla_X L)$

$\gamma_k = \Pi_{[0,1]}(Y_\gamma - \frac{1}{L_\gamma}\nabla_\gamma L)$

$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$

$Y_X = X_k + \frac{t_k - 1}{t_{k+1}}(X_k - X_{k-1})$

$Y_\gamma = \gamma_k + \frac{t_k - 1}{t_{k+1}}(\gamma_k - \gamma_{k-1})$

$A_k = \Pi_\Delta(Y_A - \frac{1}{L_A}\nabla_A L)$

$Y_A = A_k + \frac{t_k - 1}{t_{k+1}}(A_k - A_{k-1})$

end while

Algorithm 9: FISTA-based Numerical ICA for localized and sparse components

Appendix D

Inference on Graphs: An Epidemics Example

D.1 Validation: Synthetic Experiments

Given that the problem of estimating R_0 is completely unsupervised and we have no accurate way of finding proxies for the ground truth, it is necessary to run synthetic experiments to assess (a) the accuracy of the recovery of the parameters by our method given the model and (b) the sensitivity of the method to the modeling assumptions — before deploying it on real-world data.

We generate a fake epidemic time series with 10 independent outbreak groups as described in Algorithm 10. Note that Algorithm 10 generates data that follows exactly the model assumed in Eq. 5.2.3 and is thus the more amenable setting to the evaluation of R_0. Each cluster is populated with $N_k = 5 \times 10^7$ individuals (region size), and an initial number of infected cases following a Poisson distribution with parameter $\lambda = 10$.

Epidemic Time series
Fix $\tau = 0.1, \quad \bar{c}_1 = 2$
Generate $\forall g \geq 2, \quad \bar{c}_g \sim \Gamma(5, 1)$
Fix K = 15
for g = 1: G **do**
  for t = 2 : T **do**
    $N_{g,t-1} \sim \text{Poisson}(R_0^{(g)} w_s^T N_{g,(t-K):)t-1)})$
  end for
end for

Algorithm 10: Generative Mechanism using the model by Fraser[61] and Cori et al [32].

We run the MCMC algorithm using Rstan, with 8 chains, a warmup of 10,000 iterations and a

sampling phase of 1,000. All further details are provided in the Github repository along with the code.

Comparison with current estimates. To compare the strength of our approach, we compare it against the estimates provided by `earlyR`, which does not assume a hierarchical structure and is thus unable to leverage strength across the different clusters.

Figures E1 shows the credible densities for the R_0s obtained by our model (the ground truth values are shown by the vertical black bar), compared to ones recovered by `earlyR` (where the ground truth interval is indicated by the triangles) in Figure E2. We note that while in this "easy" case (as the data is generated exactly by the same mechanism as per assumed by the recovery process in both methods), our method achieves 100% coverage of the R_0, and better confidence intervals that those projected by `earlyR` . The Bayesian model is thus efficient at retrieving the values for R_0 for each cluster.

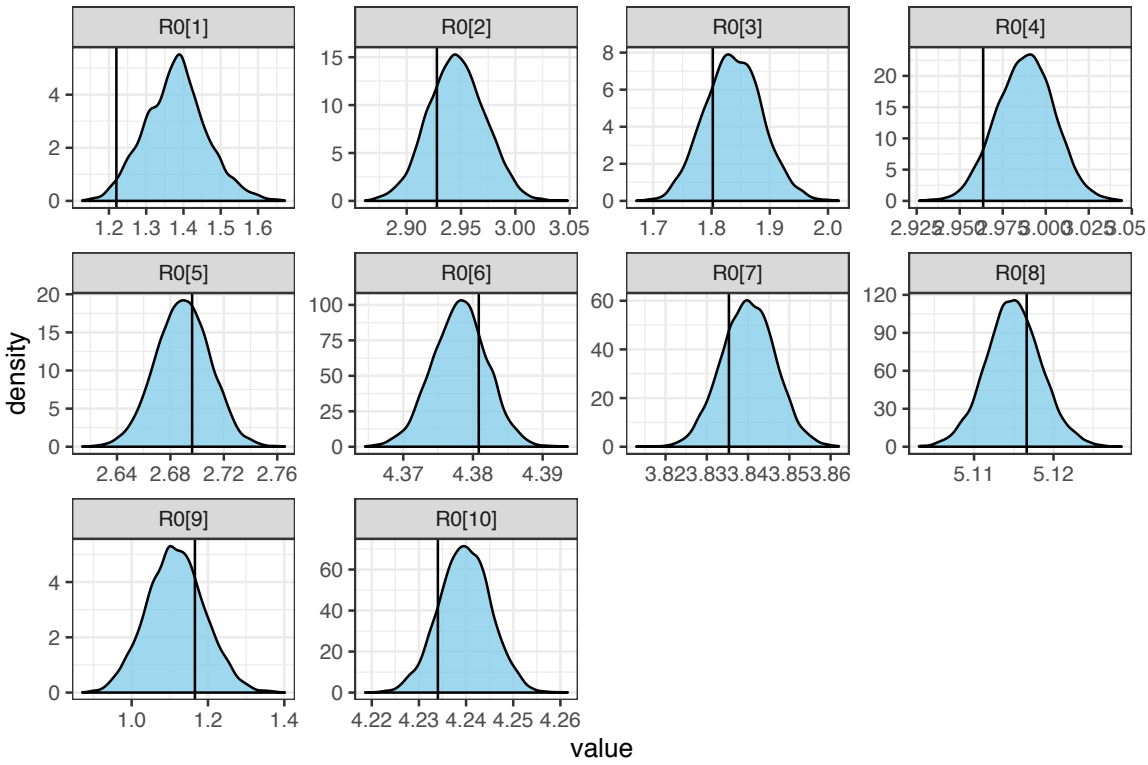


Figure E1: Recovered Credible intervals for the Fraser Model

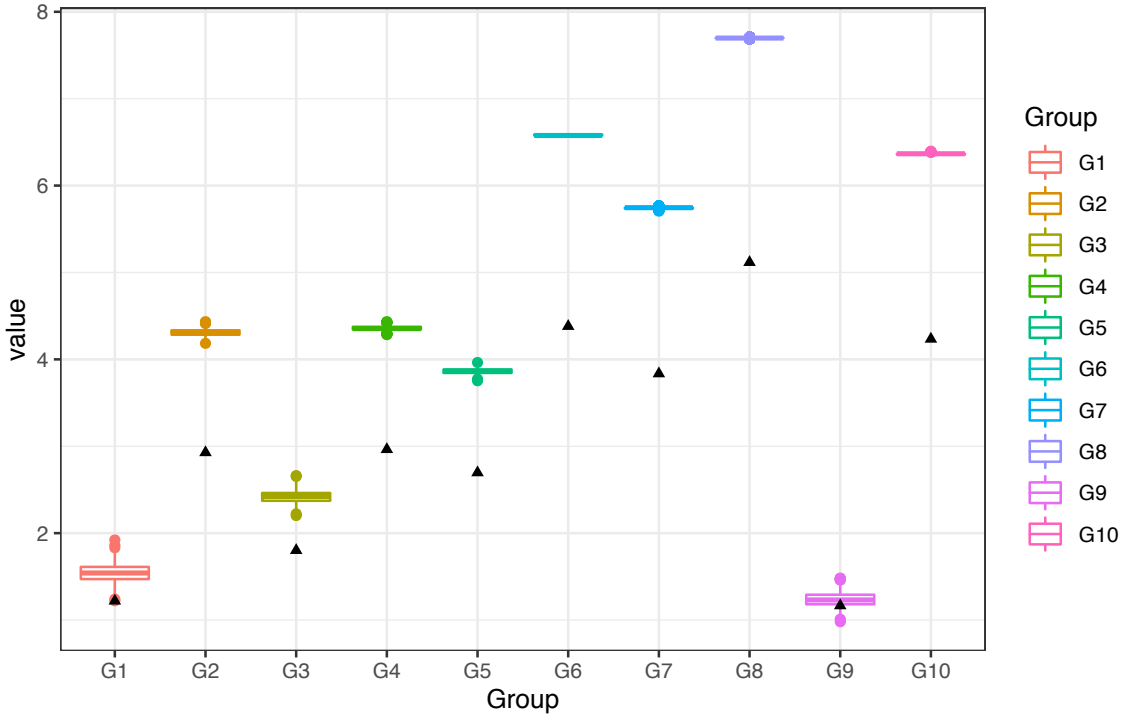


Figure E2: Boxplot of the confidence intervals obtained using the `earlyR`package using the data generated in Algorithm 10.

D.2 Comparison of the Bayesian model with its deterministic counterpart

Formal Comparison of the traditional and hierarchical approaches. We assess how much a modeling of the heterogeneity in the reproductive numbers R could make in terms of uncertainty quantification. We begin by formally writing down the log-likelihood of the model. Letting C be a constant depending only on the data (and not on the parameters of the model), the log-likelihood associated to our Poisson Model can be written as:

$$\ell(\theta) = \sum_{g=1}^{G}\sum_{t=1}^{T}\left(X_{g,t}\log(R^{(g)}) + X_{g,t}\log(\Lambda_{g,t}) - R^{(g)}\Lambda_{g,t}\right) + \text{prior on } R^{(g)} + C \tag{E1}$$

with $\Lambda_{g,t} = \sum_{s=1}^{K} w_s X_{g,t-s}$. The paper by Cori et al [32] puts a Gamma prior on their estimate of $R^{(g)}$ — to make the model more amenable to comparison, we write their reproductive number $R = D_I\tau_0\bar{c}$ where $\tau_0 = 1/D_I$, so that putting a gamma prior on $R^{(g)}$ is equivalent to considering D_I and $\tau_0 = 1/D_I$ fixed, and putting a gamma prior on $\bar{c}$. The updates at the b^{th} iteration of $\bar{c}$ are

performed independently and can be rewritten as:

$$\bar{c}^b \sim \Gamma(a + \sum_{g=1}^{G}\sum_{t=1}^{T} X_{g,t}, b + \sum_{g=1}^{G}\sum_{t=1}^{T}\sum_{s=1}^{K} w_s X_{t-s}) \tag{E2}$$

In our case, the prior has a little more structure, since $R_0^{(g)} = \bar{c}_g \tau$.

$$\begin{aligned}\ell(\theta) = &\sum_{g=1}^{G}\sum_{t=1}^{T}\Big(X_{g,t}(\log(\tau) + \log(c_g)) + X_t \log(\Lambda_{g,t}) - \tau c_g D_I \Lambda_{g,t}\Big) \\ &+ \sum_{g>1}\Big((\alpha - 1)\log(c_g) - \beta c_g\Big) + (\alpha_0 - 1)\log(\tau) + (\beta_0 - 1)\log(1 - \tau)\end{aligned} \tag{E3}$$

In particular, the conjugate updates of our model are given by:

$$\begin{aligned}c_g^b &\sim \Gamma(\alpha + \sum_{t=1}^{T} X_{g,t}, \beta + D_I \tau \sum_{t=1}^{T} \Lambda_{g,t}) \\ \tau &\sim \beta(\alpha_0 + \sum_{g=1}^{G}\sum_{t=1}^{T} X_{g,t}, \beta_0 + D_I \sum_{t=1}^{T}\sum_{g=1}^{G} c_g \Lambda_{g,t})\end{aligned} \tag{E4}$$

where the last update in τ follows that by assuming that the transmissibility is small, the following approximation holds: $\log(1 - \tau) = -\tau + o(\tau)$. Thus, by the law of total variance and linearization around $\tau_0 = \frac{1}{D_I}$, the variance of the R_0 is:

$$\begin{aligned}\mathrm{Var}[R_0^{(g)}] &= \mathbb{E}[\tau^2 D_I^2 \mathrm{Var}(c_g|\tau)] + D_I^2 \mathrm{Var}[\tau^2 \mathbb{E}[c_g|\tau]^2] \\ &= D_I^2 \mathbb{E}[\tau^2 \frac{\alpha + \sum_{t=1}^{T} X_{g,t}}{(\beta + D_I \tau \sum_{t=1}^{T} \Lambda_{g,t})^2}] + D_I^2 \mathrm{Var}[\tau^2 \frac{(\alpha + \sum_{t=1}^{T} X_{g,t})^2}{(\beta + D_I \tau \sum_{t=1}^{T} \Lambda_{g,t})^2}] \\ &= \frac{\alpha + \sum_{t=1}^{T} X_{g,t}}{(\beta + \sum_{t=1}^{T} \Lambda_{g,t})^2} + \frac{(\alpha + \sum_{t=1}^{T} X_{g,t})^2}{(\beta + D_I \sum_{t=1}^{T} \Lambda_{g,t})^2} \\ &+ \mathbb{E}\Big[(\tau - \frac{1}{D_I})(2\frac{1}{D_I}\frac{\beta(\alpha + \sum_{t=1}^{T} X_{g,t})}{(\beta + \sum_{t=1}^{T} \Lambda_{g,t})^3}\Big] \\ &+ \mathrm{Var}\Big[(\tau - \frac{1}{D_I})(2\frac{1}{D_I}\frac{\beta(\alpha + \sum_{t=1}^{T} X_{g,t})^2}{(\beta + \sum_{t=1}^{T} \Lambda_{g,t})^3}\Big]\end{aligned} \tag{E5}$$

This allows us to quantify the variability in the spatial reproductive number R. The bias in the variance is thus proportional to the deviation of τ from $\frac{1}{D_I}$. Since we are considering exponential models (where roughly, R_0 governs the slope of the exponential curve), this could lead to some substantial deviations in the predictive scenarios that are drawn.

D.3 Appendix: Added variability in the infectious profile

Here, we show the performance of the model when modeling w_s as a random variable. We add on top of the baseline model uncertainty in the infecticious profile. The infectious profile is assumed to be sampled from an ordered dirichlet distribution, initialized with the serial interval with men 3,96 and standard deviation 4.26, as detailed in some of the current reports. The mean length of the confidence intervals is 747.5, which is a little broader than in the case where w_s is considered to be fixed, but not substantially so.

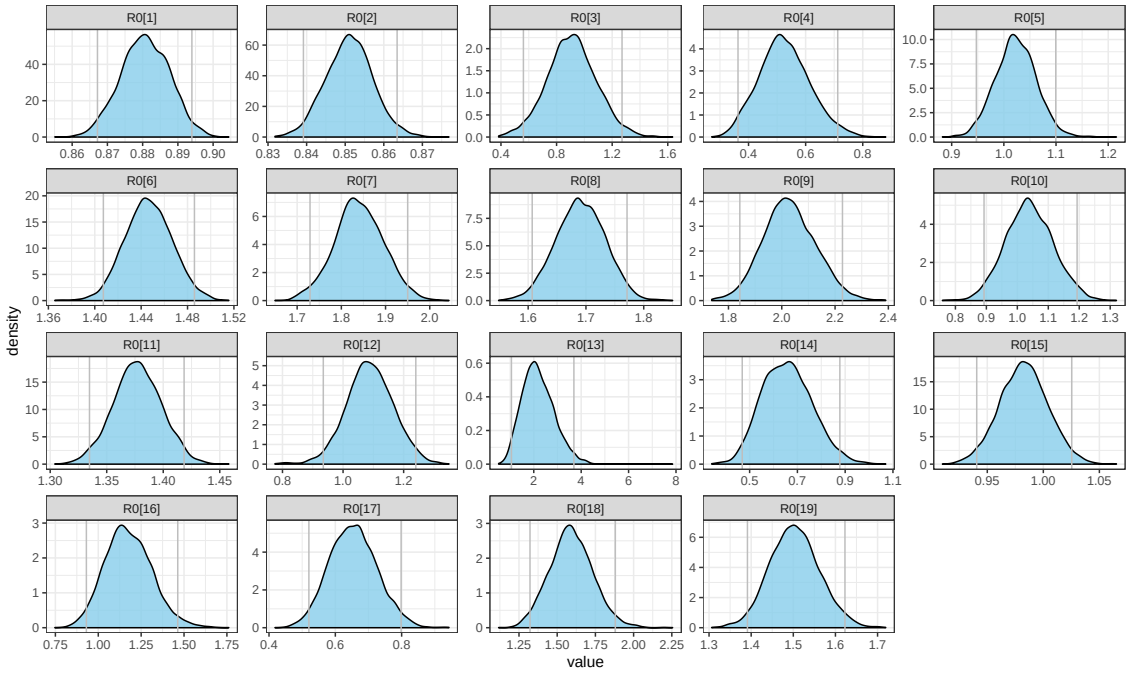


Figure E3: R_0 for the spatial Random-Effects with Dirichlet estimated infectivity profile.

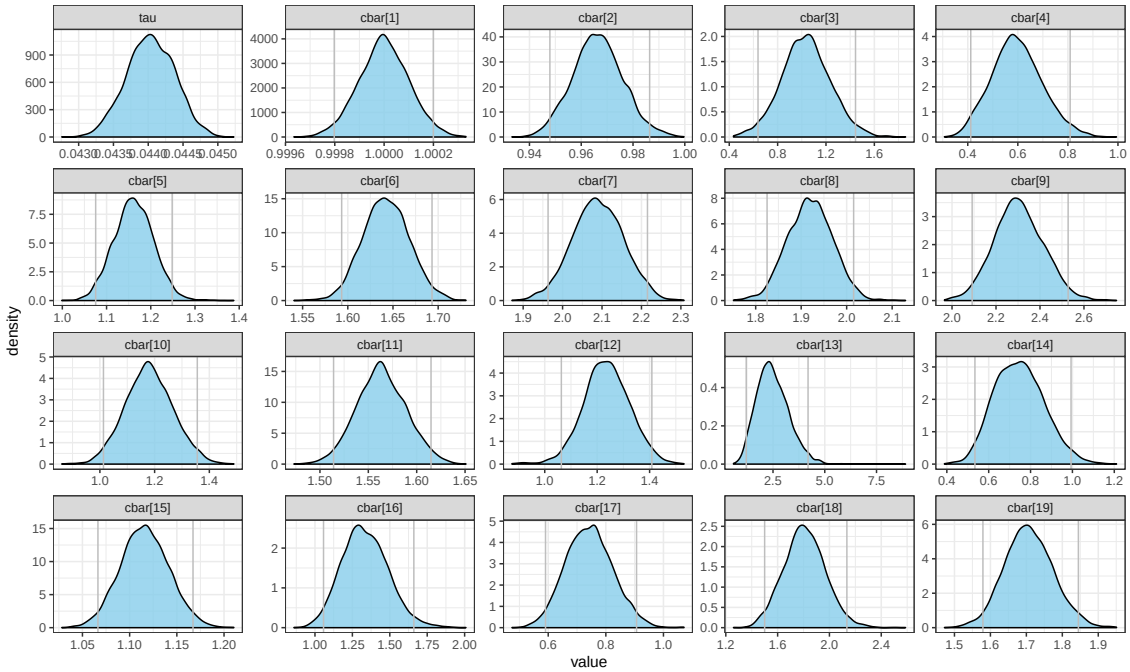
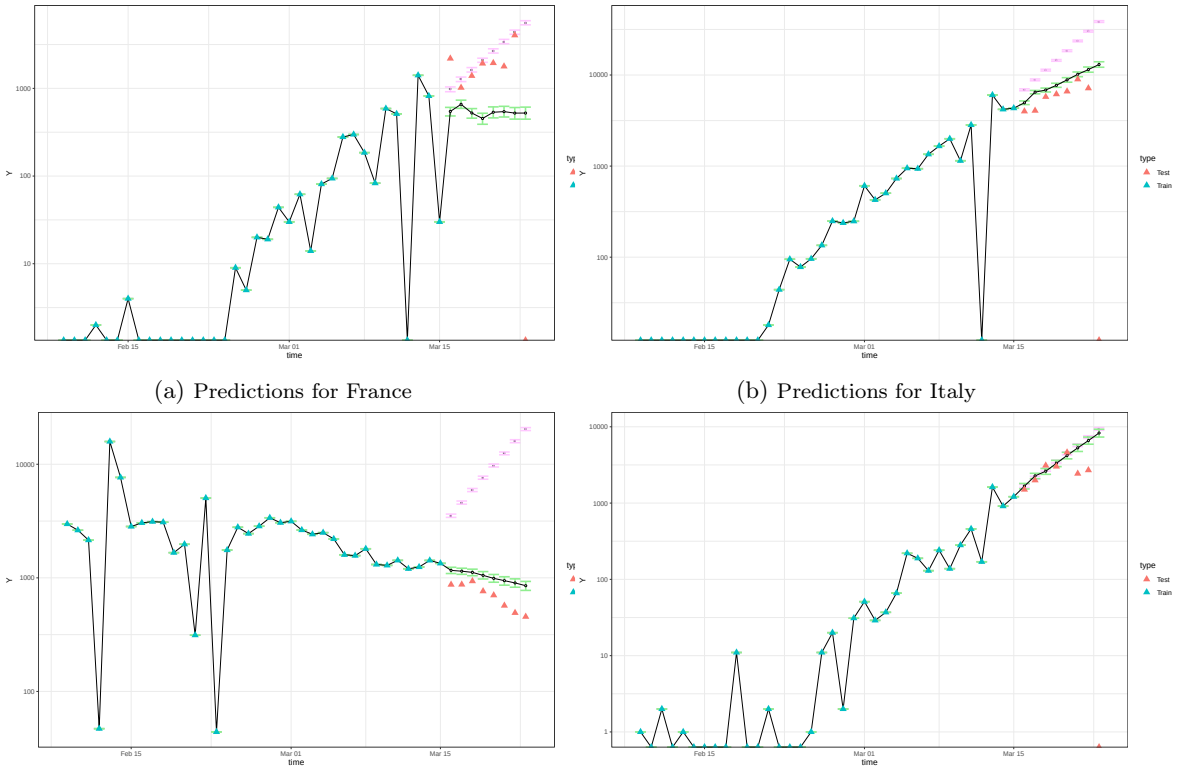


Figure E4: $\bar{c}$s for the spatial Random-Effects with Dirichlet estimated infectivity profile.



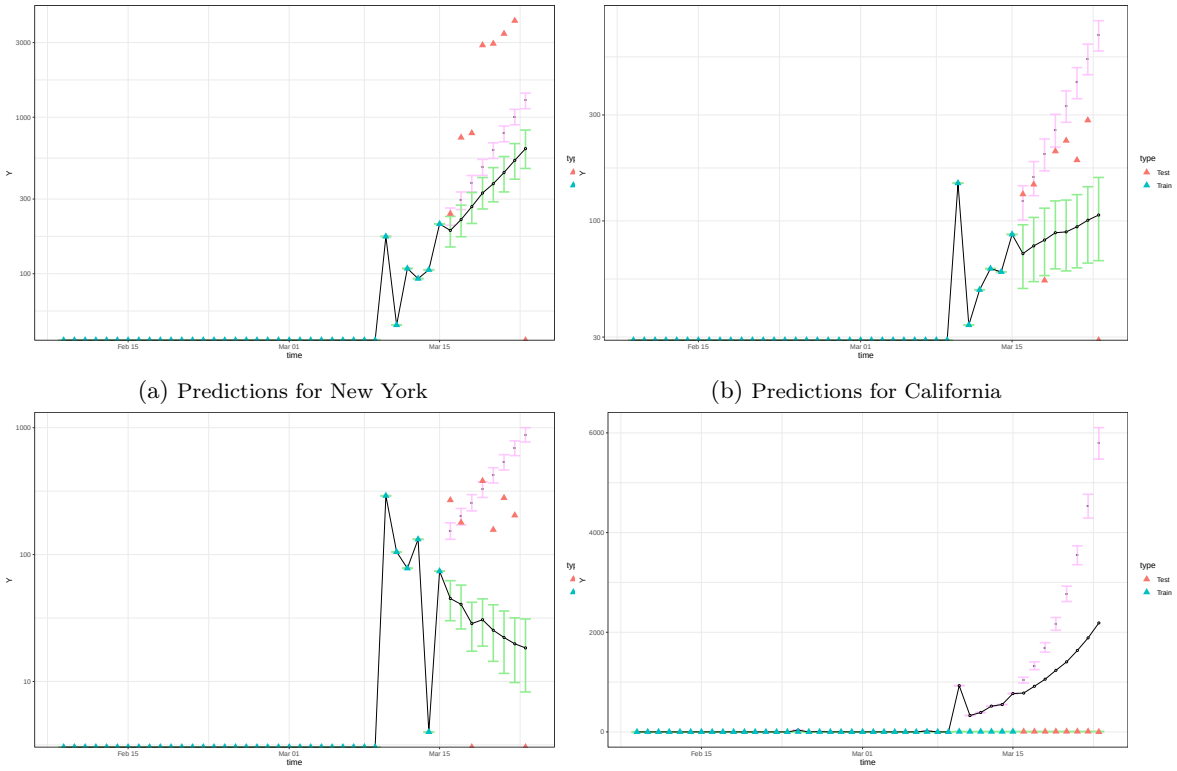
(a) Predictions for France

(b) Predictions for Italy

(c) Predictions for Hubei 1

(d) Predictions for Germany

Figure E5: Projection accuracy for a few of the European groups for the spatial Random-Effects with Dirichlet estimated infectivity profile.



(a) Predictions for New York

(b) Predictions for California

(c) Predictions for Washington

(d) Predictions for the US

Figure E6: Projection accuracy for groups in the United States for the spatial Random-Effects with Dirichlet estimated infectivity profile.