Spatial Community Structure and Epidemics

Marta Sarzynska
St John's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2015

Acknowledgements

I would like to thank my supervisor, Mason Porter, and long-time collaborator Elizabeth Leicht, for their ideas, guidance and help throughout my DPhil. Thanks to Dr Gerardo Chowell from Arizona State University for collaboration, providing the epidemic data and biological comments on my results.

Thanks to my office mates Lucas Jeub and Marya Bazzi, for providing a source of inspiration, daily help and support. I thank the all the other group members, in particular SangHoon Lee, Valentin Danchev, Mikko Kivela, SeWook Oh, Vladimir Murevics, Puck Rombach, and Martin Gould for their companionship, discussions and day-to-day help. Warm thanks to Vittoria Colizza, Chiara Poletto, and other members of their research group, for biological comments and feedback. I also thank the people involved in the MTBI summer school: Carlos Castillo-Chavez, Steve Wirkus, Erica Camacho and all the other faculty and graduate students. Thanks go to all my college friends and rowing friends, who made my life brighter when things were not going so well. Last but not least, thanks to Andrew Elliott for always being there for me and providing support, suggestions, discussions, collaboration, and reading over many incarnations of this manuscript.

I would also like to thank the Peruvian Health Ministry's General Office of Epidemiology, the Peruvian National Institute of Statistics and Informatics, the Mexican Institute for Social Security, the Chilean Ministry of Health, the WHO, and Office for the Coordination of Humanitarian Affairs in West & Central Africa for providing the data that I use in this thesis. I also thank the Bodleian Library staff for help with obtaining additional geographical and population data, and Caitlin Rivers and all the anonymous volunteers who collated the Ebola Datamarket data set. I thank the authors of MATLAB code used in this research:, Sumeet Agarwal, Danielle Bassett, Dan Fenn, Lucas Jeub, Peter J. Mucha, Yulian Ng and Tom Prescott.

I gratefully acknowledge funding from the EPSRC via a studentship at the Systems Biology DTC at the University of Oxford, the James S. McDonnell Foundation, and the European Commission FET-Proactive project PLEXMATH. I also thank St John's College, Oxford for hosting me and for its generous support.

Abstract

Networks are a useful quantitative representation for complex systems of interacting entities arising in fields such as biological, physical and social sciences. A network representation provides a degree of simplification while capturing key connectivity patterns. This thesis focuses on two main themes: the study of community structure, an important mesoscopic feature of many networks, and its application to study spatiotemporal spread of infectious diseases.

Community detection seeks to partition a network into dense sets of nodes that are connected sparsely to other dense sets. The notion of denseness is often relative to some "null model" that describes baseline connectivity that can be construed to occur randomly. In the first part of the thesis, we discuss the incorporation of spatial information into null models for community detection. We develop a spatial null model based on the radiation model of mobility. We test different spatial null models using static and temporal (multilayer) spatial benchmarks with planted partitions that represent interactions between human populations. Our results indicate that it is important to incorporate spatial information into null models for community detection, but it is best to incorporate only relevant information into null models, as extraneous information can lower performance.

In the second part of the thesis, we present the results of community detection with different null models on disease-correlation networks generated form real and synthetic time series of disease occurrence. We use data sets for endemic diseases (established in a region, with occasional epidemic outbreaks) and emerging diseases (newly-discovered or introduced into a region for the first time). We study the spatial and temporal organization of partitions. Finally, we apply community detection with different null models to synthetic time series generated from an agent-based model (ABM) simulating the spread of endemic and emerging diseases between spatially-embedded cities with a planted, transport-based community structure. We compare the findings on real and synthetic data sets, and we searched for model parameter regimes in which we are able to detect planted partitions or other interesting communities.

For emerging diseases, we find spatial communities that are associated with the first times the infection reached a node in both ABM and disease data. For endemic diseases, we are unable to find planted or spatial communities in the ABM data, but we detect spatial communities for two of the three disease data sets. For these diseases, we also

detect temporal communities corresponding to some of the important time points in disease history.

We hope that these results show that community structure of disease-correlation networks appears to be more complicated than simple spatial patterns and is a fascinating topic to study.

Publications

Some of the work in this thesis has been submitted for publication and is awaiting a final review. Details of the publication are given below.

[P1] M. Sarzynska, E. Leicht, G. Chowell, and M. A. Porter, *Null models for community detection in spatially-embedded, temporal networks*, arXiv:1407.6297, 2014

I also performed other research (largely during teaching on a summer school at Arizona State University) that I do not include in this thesis due to its lack of successful results ([P2]) and collaborative nature and unrelated topic ([P3]). The results are published in technical reports on the arXiv server, the details of which are given below.

[P2] M. Sarzynska, O. Udiani, and N. Zhang, *A study of gravity-linked metapopulation models for the spatial spread of dengue fever*, arXiv:1308.4589, 2013

[P3] V. Cheathon, A. Flores, V. Suriel, O. Talbot, D. Padilla, M. Sarzynska, A. Smith, and L. Melara, *Dynamics and control of an invasive species: The case of the Rasberry crazy ant colonies*, arXiv:1308.3675, 2013

Statement of Originality

The research in this thesis is a result of collaboration between myself and my coauthors on the listed publications. My collaborators have helped develop the ideas described in this thesis, but I have performed all of the analysis leading to the results that I present.

Contents

List of Figures

List of Tables

Glossary of Terms

Term	Description	Page
b	bin size	40
$c_i, \overline{c_{is}}$	community of node i — static and multilayer in layer s	31
d_{ij}	distance between nodes i and j	18
$f(d)$	deterrence function for the gravity or radiation null model	18
$g_i, \overline{g_{is}}$	degree of node i — static and multilayer in layer s	24
$k_i, \overline{k_{is}}$	strength of node i — static and multilayer in layer s	24
l	size of a spatial benchmark — size of the $\{1, 2, \ldots, l\} \times \{1, 2, \ldots, l\}$ lattice on which nodes are placed in space	62
m	number of layers in a multilayer network	25
n_i	population of node i	18
p	fraction of nodes that change community assignment between a pair of layers for temporally-evolving spatial benchmarks	64
p_{ij}^{dist}, p_{ij}^{flux}, p_{ij}^{distpop}	probability of an edge between i and j in "distance", "flux" and "distance and population" benchmarks	62
t_c, t_{c1}, t_{c2}	"critical time" defining a temporal partition with the highest z-Rand score in a multilayer network (t_c for a partition into 2 communities, and t_{c1} and t_{c2} for 3 communities)	49
v_i, v_i^s	i-th eigenvalue — static and multilayer in layer s	44
$2w, 2\overline{w}$	total edge weight in a static and multilayer weighted networks	31
$2z$	total edge weight in a static unweighted network	33
$A_{ij}, \overline{A_{ijs}}$	edge in an unweighted network between nodes i and j — static and multilayer for layer s	23
B, $\overline{B}$	modularity matrix — static and multilayer; $B = A - P$ for unweighted and $B = W - P$ for weighted networks	36
C	correlation matrix	43
$C^{(r)}, C_s^{(r)}$	"random component" of the correlation matrix C — static and multilayer for layer s	44
$C^{(g)}, C_s^{(g)}$	"group mode" component of the correlation matrix C — the meaningful correlations	44
$C^{(m)}, C_s^{(m)}$	"market mode" component of the correlation matrix C — static and multilayer for layer s	44
$D_i(t)$	number of disease cases for province i, at time t	25
$E_i^{(s)}$	subset of a disease time series for province i for a time window defined by (τ_s, Δ)	26
G_{ij}	gravity model between locations i and j	18
$H(X)$	entropy of partition X	53
$I(X, Y)$	Mutual Information between partitions X and Y	52
N	number of nodes in a network	23

Contd.

Term	Description	Page
$\hat{N}$	number of nodes that "exist" in a network, i.e., they experienced disease in the corresponding time window	27
$\mathrm{NMI}(X, Y)$	normalized mutual information between partitions X and Y	53
$\mathrm{NVI}(X, Y)$	normalized variation of information between partitions X and Y	53
$P_{ij}, \overline{P_{ijs}}$	null model for community detection — — static and multilayer networks for layer s	31
$Q, \overline{Q}$	modularity — static and multilayer	31
R	Rand coefficient	47
z-Rand, z_R	z-score of the Rand coefficient	47
T_{ij}	flux between nodes i and j predicted by the radiation model	20
$\hat{T}_{ij}$	mean flux predicted by the radiation model between nodes i and j	42
T	length of a time series	25
$VI(X, Y)$	variation of information between partitions X and Y	53
$W_{ij}, \overline{W_{ijs}}$	edge in a weighted network between nodes i and j — static and multilayer for layer s	23
Z_1, Z_2	normalization constants in spatial benchmarks	62
α	infection rate in a compartmental model of disease	13
β	recovery rate in a compartmental model of disease	13
γ	resolution parameter for community detection	32
$\delta_{ij}, \delta_{c_i c_j}$	Kronecker delta — for edges and communities	27
Δ	time-window width	26
η	intercommunity connectivity in the agent-based disease model	163
$\eta(c_i, c_j)$	function controlling intercommunity connectivity in the agent-based disease model	163
2ζ	total edge weight in a multislice unweighted network	34
$\overline{\kappa_{js}}$	multislice strength for node j in layer s	34
λ_d	intercommunity connectivity in a spatial benchmark	62
$\lambda(c_i, c_j)$	function controlling intercommunity connectivity in a spatial benchmark depending on community assignment	62
μ	edge density in a spatial benchmark	62
ρ_{ij}	Pearson correlation coefficient between disease time series E_i and E_j [in a time window defined by (τ_t, Δ)] for provinces i and j	27
$\sigma_i^{(s)}$	standard deviation of the disease time series E_i for province i in layer/network s	27
τ	beginning of all of the time windows used disease correlation network creation: $\tau = (\tau_1, \tau_2, \ldots, \tau_m)$	26
υ	difference between starts of time windows ($\tau_2 = \tau_1 + \upsilon$)	26
ϕ	probability of movement in the agent-based disease model	162
$\Psi_i^{(s)}$	indicator of whether a node exists in a given layer/network s	27
ω	parameter governing the interlayer edge weight	25

Chapter 1

Introduction

1.1 Networks

The study of complex systems is an emerging discipline that has provided a new approach to studying systems that have previously been difficult to understand and predict the behavior of, such as the human brain and the world economy [35, 43, 89, 184]. Although there is no precise definition, a complex system, roughly speaking, is a system consisting of multiple parts whose highly individual behaviors are shaped by their interactions [243,248]. Further, many authors agree that a true complex system should possess emergent properties: ones that are not describable by a single rule or due to central control, and instead they arise from the interactions of individual components [7, 128].

Traditionally, researchers tend to try to summarize the components of the system and their dynamics in terms of their lowest common denominators. However, a complex system cannot be fully understood by studying the parts in isolation and it is essential to consider the interactions between parts. Thus, a holistic approach to problems lies at the heart of complexity science [35].

One important aspect is the pattern of connections between the individual components of a system. It influences their interactions and is thus crucial to the overall behavior or the system. Often, relationships among a set of objects can be modeled as a *network* (or *graph*) in which nodes represent the objects, and edges represent the relationships between them [3, 200, 204]. A network is a simplified representation that captures the key connectivity patterns; this makes the analysis of systems with multiple components and interactions simpler and more tractable, but it also means that information can be lost in the simplification process.

The mathematical study of networks originates within *graph theory*, which dates back to the 17th century, when Euler published a solution to the Königsberg bridge problem [37]. Initially, graph theory focused on regular graphs, but since the 1950s graph theorists have also investigated the properties of random graphs [38]. In parallel to the mathematical developments, social scientists have been using network concepts since the 1920s to represent

the relationships between people [94,278]. They were often interested in answering questions relating to the meaning of edges in the networks [105]. Network science has experienced explosive growth since the late 1990s [204]. This has been sparked by the computerization of many aspects of life, which led to the increased availability of large data sets related of various fields, and the increased computational resources available to analyze these data sets. These factors caused a surge of interest in network science from researchers from new fields interested in both theoretical research into properties of networks (e.g., statistical physicists and statisticians) and empirical study of network data sets (e.g., biologists and economists).

The most traditional network representation is a static graph, in which nodes represent entities and edges represent pairwise connections between nodes. However, many networks are time-dependent [129, 130] or include multiple types of connections between nodes [73, 146]. Moreover, the structure of networks that are embedded in space is influenced profoundly by spatial effects [21]. In order to avoid discarding such potentially important information, which can lead to misleading results, it is crucial to develop methods that incorporate features such as time-dependence, multiplexity, and spatial embeddedness in a context-dependent manner [21, 129, 146]. Because of the wealth of new, rich data, it is now also possible to validate more complicated network structures and methods using empirical data.

Mesoscale structure of networks The structure and composition of networks is often so complicated that it is impossible to study their (or the individual node and edge) properties individually. It is thus common to study aggregate features of networks, or their substructures. One such approach is to investigate a mesoscale network structure known as *community structure*, which is the methodology we will use in this thesis due to its demonstrated usefulness and flexibility [91, 219].

A *community* is a set of nodes with dense connections among themselves and with only sparse connections to other communities in a network [91, 219]. Communities (also called "clusters" or "modules") are believed to share common properties and/or play similar roles within a network [91]. Communities arise in numerous applications: for example, social networks typically include dense sets of nodes with common interests or other characteristics [269], networks of legislators have been shown to contain dense sets of individuals who vote in similar ways [279], and protein-protein interaction networks include dense sets of nodes that constitute functional units [166].

Several studies have successfully applied community detection to uncover new properties in data sets. Traud et al. [269] studied online friendship networks of American students, finding that students at Caltech and Rice tend to establish friendships within their "House affiliation" (an arrangement similar to the Oxford Colleges), while students at universities

lacking this structure tend to make friends mainly within their year group. They also found that high school plays a greater role in the social organization of larger universities (where there are typically more people from the same high school). Such intuitively expected results highlight the usefulness and validity of the method.

Thiemann et al. [262] and Ratti et al. [222] used community detection to find novel features in spatially-embedded data: they used community detection on proxy networks of mobility (based on banknote circulation and mobile phone calls) to search for effective regional boundaries in countries. Both studies found structures that were similar to the actual administrative regions.

Expert et al. [82] went a step further: they used a novel spatial null model to remove the expected increased connectivity of neighboring regions in a mobile phone network from Belgium. They showed that while standard community detection produces spatial partitions that are similar to regional boundaries, the spatial null model recovered the well-known linguistic division of Belgium into two groups (Flemish and French).

Community detection has also been applied to time series data with success in identifying groups of entities that have similar time series patterns. This approach is particularly common for financial time series [84, 171], where it has been observed that groups of correlated stocks evolve in time, and they change following large-scale events in the market [84]. It also produced some novel insights into asset correlations: communities often partially overlap with industrial sectors and related types of goods or services, which may be useful for portfolio optimization [171]. Outside the financial domain, community detection has been used on correlation networks for such applications as detecting changes in areas of brain activity from fMRI time series [24], spatial groupings in climate time series [80, 173] and functional groupings in gene expression data [31, 150].

Methodological considerations for spatial and temporal networks Myriad community detection methods have been developed [91, 219]. The most popular family of methods entails the maximization of a quality function known as *modularity* [199, 202]. To optimize modularity, one compares actual network structure versus some *null model*, which quantifies what it means for a pair of nodes to be connected "at random".

Traditionally, most community detection applications using modularity have only randomized network structure in some way (while preserving some structural properties). The standard null model for modularity optimization is the "Newman-Girvan" (NG) null model, in which one randomizes edge weights, such that the strength of the nodes is preserved (where node strength is the sum of the edge weights of all its adjacent edges) [199, 202]. The NG null model has become very popular due to its simplicity and effectiveness, and it has been derived systematically through the consideration of Laplacian dynamics on networks [155, 156].

However, the NG null model is also a naïve choice as it does not incorporate domain-specific information. The choice of a null model is an important consideration because (1) it can have a significant effect on the community structure obtained via optimization of a quality function and (2) it changes the interpretation of communities [26, 108, 171]. The best choice for a null model depends on the particular data set and scientific question.

Most existing research on community detection does not incorporate metadata about the nodes or information about the timing and location of interactions between nodes [91, 219]. However, with the increasing wealth of space-resolved and time-resolved data sets, it is important to develop community detection techniques that take advantage of the additional spatial and temporal information (and of domain-specific information, such as generative models for human interactions [82]). Indeed, community detection in temporal networks has become increasingly popular [23, 29, 48, 50, 83, 140, 194], but the majority of methods use networks constructed from either static snapshots of data or aggregations of data over time windows. Few investigations of community structure in temporal networks have used methods that take advantage of the temporal structure (see, e.g., [23, 194]). More work is being done on the influence of space on community structure [22, 47, 82, 119, 242], most of it motivated by application to particular data sets. To our knowledge, there has not yet been a review of the differences in community structures obtained using spatial and non-spatial null models, a problem that we explored in a working paper [235], both in a controlled setting of spatial benchmarks and on disease-correlation networks.

In this thesis, we combine the application of community detection to a new topic (disease data) with the detailed study of the results of community detection by modularity maximization using the standard Newman-Girvan null model as well as spatial and correlation-specific null models. In Chapter 4, we develop spatially-embedded benchmarks with (static and temporally-evolving) planted community structure that depends on distance between nodes and node properties such as population. We study the performance of different null models for modularity maximization on these benchmarks. In Chapters 6–8, we study community detection on disease-correlation networks that we generate from the time series of disease occurrence across several neighboring regions. In Chapter 9, we apply the same approach to disease-correlation networks generated from a spatially-embedded disease model with known disease parameters and connectivity between populations that is related to the distance between them and a planted community structure.

1.2 Infectious diseases

Infectious diseases are quintessential complex systems, with many interacting components including hosts (infected individuals, both human and non-human), disease vectors (individuals that transmit disease but do not get infected themselves, e.g., mosquitoes), host behavior, immunological reactions and disease evolution. Infections can also be influenced

by external factors such as weather. Their large-scale spread can be notoriously difficult to predict, as one infected individual traveling a long distance can seed an entire disease outbreak [281].

Infectious diseases are also an important topic to study. They are the second largest cause of death worldwide (15% of all deaths in 2008 [283]). Epidemics of infectious diseases have been documented throughout history, and some, like the Black Death, have had a huge impact on the history of mankind. Further, the economic impact of even relatively mild epidemics can be huge – for example, the annual economic burden of influenza on the USA has been estimated to range between $26.8 to $166.5 billion [183, 186]. Despite medical developments, infectious disease rates are rising due to the emergence of new and resurgent pathogens, and changes in factors such as climate, human behavior, population density and transportation [283]. Understandably, studying disease mechanisms and preventing disease epidemics is an important subject of research in both academic and policy-making settings.

Moreover, studying the spatial aspects of disease occurrence is crucial: in a globalized world, many diseases rapidly spread from their origins in global epidemics. Avian H5N1 influenza (1997–present) spread from Hong Kong to Europe and Africa through poultry and migrating birds [144, 280]. The global spread of SARS in 2002–2003 has been traced to a series of airline travels. In 2009, H1N1 influenza ("swine flu") spread to 41 countries within about 3 months from the first outbreak in La Gloria, Mexico [93]; its spread was also linked to airline transportation [143, 174].

The world is also experiencing an expansion of vector-borne diseases into new territories. For example, dengue fever has spread across South and Central America since 1981 [32, 117, 134, 154], and malaria incidence is increasing in the high altitude highlands in Africa and the Amazon [64, 168, 225]. The initial transport of infected mosquitoes has been linked to shipping [259] and plane travel [154, 274], but there is still much controversy as to what is causing their increased persistence [118, 124, 134, 211].

Factors influencing disease spread An infectious disease is transmitted through time and space from one individual to another through a contact network. Understanding disease-transmission mechanisms is important for the control and prevention of epidemics. Many factors are known to influence the patterns of disease spread.

On the micro-scale, within one disease outbreak, the spread of disease is most influenced by the characteristics of the population and the location where the disease is spreading. Such characteristics include age, health status, susceptibility, contact patterns, mobility, social networks, population size, population density and others (and the respective properties for the vector population for vector-borne diseases) [40, 195].

On the macro-scale, when the disease is spreading between populations, spatial factors such as distance, transport, and differences in climate and socioeconomic factors play a

larger role [237]. Short-distance travel often results in traveling wave infection patterns, e.g., for the Black Death in medieval Europe [195] and dengue fever sub-strains in Vietnam [221]. Modern long-distance transportation speeded up disease dissemination, e.g., through air travel of infected individuals [174, 281] or shipping of infected mosquito eggs [110, 123]. Climate is another spatial factor thought to affect the spread of disease. Vector-borne diseases are especially influenced [97, 191], but other diseases such as influenza can be affected as well [256].

Socioeconomic factors can also affect epidemic spread: the unprecedented 2014–2015 Ebola epidemic in West Africa linked to local burial practices and the quality of healthcare provision [5], human activities and land usage can influence the amount of vectors for vector-borne diseases [147, 226], and the use of face masks and hand hygiene (the uptake of which varies between populations, age groups, or even workplaces) strongly affects the transmission of influenza [2].

In Chapter 5, we will review the known evidence for the influence of spatial, climatic and other external factors on the spread of each of the diseases that we study.

Methodology There are many ways to approach the study of infectious diseases. Experimentalists usually focus on studying particular aspects of the mechanisms of disease infection and spread in detail; they examine the infectious agents (e.g., viruses and bacteria), immune response, disease progression etc. However, this kind of a bottom-up empirical approach is often unable to describe or discover all aspects of the enormous complexity of infectious diseases. Moreover, experiments are relatively expensive and time-consuming.

Mathematical studies are commonly used alongside experimental knowledge to better understand disease spread and to predict and counteract epidemics [40,195]. With the rapid increase in available computing power in the last 20 years, modelling disease spread is a fast-growing field, and model results are increasingly applied in policy-making settings, for example during the 2001 foot-and-mouth disease epidemic in the UK (several approaches reviewed in [237]), the 2009 "swine flu" (Influenza A/H1N1) pandemic [36] and the 2014–2015 Ebola epidemic [182]. Mechanistic models often take the form of differential equation models (which can give analytical results and mechanistic insights, but are difficult to fit to noisy, large-scale data sets), and agent-based models (which represent agent heterogeneity and their interactions explicitly and are thus well-positioned to take into account the multiple aspects of complex systems, but they are computationally expensive and difficult to validate).

A different class of approaches comes from statistical and data-analysis techniques. These models usually cannot offer much insight into the mechanisms of a disease, as they only identify associations in the data — and these can arise from a combination of many different mechanisms. However, they do provide potentially useful clues about the nature

of the disease and epidemiological processes, which can be further examined by explicit modelling or experimental approaches. Many data-analysis studies seek to identify the influences of external factors, such as climate variables, on the disease spread in a population (e.g., [75, 133, 136, 141]). Such approaches have the advantage of the lack of mechanistic assumptions about the disease in question, and statistical methods are broadly applicable and reusable. However, statistical approaches are heavily dependent on the availability of good quality disease data, which is not always a given, especially when dealing with emerging diseases in developing countries.

Our community-detection approach attempts to find spatiotemporal patterns in the spread of disease. It thus addresses external influences on the spread of disease from factors such as distance, climate and socioeconomic factors, and aims to identify times when the patterns of infection change. It does not require knowledge of the disease mechanisms (although such knowledge can illuminate the analysis of results). It is also widely reusable — in fact, it could be applied to any time series whether originating from disease surveillance or a generative model. However, our methodology suffers from the same limitations as statistical methods: it is only able to identify associations in the data rather than give mechanistic insights, and it is heavily dependent on the availability of good quality data.

We will further review and compare the main kinds of approaches to studying infectious diseases using mathematical and statistical methods, and the known applications of network science in epidemiology, in Chapter 2. In Chapters 6–9, we will focus on the results of community detection on disease-correlation networks that we generate from the time series of disease occurrence across several neighboring regions (using real disease time series for endemic and emerging diseases, and synthetic data from an agent-based model). We will study the algorithmic network partitions in the context of known factors that influence disease spread. The two types of diseases present different data sets to work with. The endemic disease time series contain many data points compared to the number of provinces. For the majority of time the provinces experience a low level of disease, and there are several larger epidemic outbursts. The data sets for emerging diseases refer only to one epidemic wave, and if data are collected early enough, they describe the spread of the disease into susceptible populations. In Chapter 9 we will study the potential usefulness of our methodology to a wider range of diseases. We will do this by applying our community-detection methodology to the output of an agent-based model of the spread of disease between spatially-embedded cities, and testing a variety of disease and connectivity parameter values.

1.3 Outline

The main aim of this thesis is to investigate the applicability of using correlation networks and community detection to study the geographical spread of disease using benchmarks, synthetic time series, and real data sets. Additionally, we thoroughly investigate the effects

of using different null models for community detection using modularity maximization on spatially-embedded and temporally-evolving networks. The rest of this thesis is organized as follows.

Chapters 2 and 3 present the background and methodology of the disciplines of epidemiology and networks. In Chapter 2, we review the mathematical and statistical approaches that are commonly used to study infectious diseases. We focus on the methods that take the spatial and temporal aspects of infections into account, and we present the current use of network science methodology for studying disease spread. In Chapter 3, we give an overview of network science and community detection, and we present the pipeline that we use to generate disease-correlation networks from time series. We give a detailed discussion of null-model choices for community detection, and we introduce a novel radiation null model. We then present the methods that we use to examine the properties of algorithmic network partitions. We introduce two statistics for quantifying the degree of spatial organization in a network partition and a statistic to measure whether community assignments in disease-correlation networks are related to the first infection times for provinces (the times when the disease first reached a province). Finally, we summarize our approach to studying the spread of disease, and we discuss the reasoning behind it, its place in the wider fields, and the possible alternatives.

In Chapter 4, we study the effects of incorporating spatial and temporal interactions into null models for community detection on synthetic spatially-embedded networks with planted community structure. Synthetic benchmark networks are a common test-bed for new community detection methods. We develop novel spatial benchmarks where edge weights between spatially-embedded nodes are based on distance and population (through relationships based on gravity and flux-like interactions) and on membership in planted communities. We develop both static and multilayer benchmarks, and we incorporate temporal evolution into the multilayer planted partitions. We compare the performance of NG, gravity and radiation null models on these benchmarks.

In Chapters 5–8, we work with disease data sets. In Chapter 5, we compare the diseases and data sets that we use in this thesis, and we present the procedures and the parameter choices for creating networks from all data sets, and the effects of these choices on the structure of the networks. We then embark on analyzing the disease data sets. In Chapter 6, we present detailed results of using community detection on a data set about the occurrence of dengue fever in Peru. This vector-borne disease has previously been shown to exhibit strong spatiotemporal patterns related to climate, and its patterns have changed profoundly over the duration of data collection. This led us to expect community structures to have a degree of spatial organization and detectable temporal changes. We thus use this data set to study the strengths and weaknesses of our methodology and to motivate some of the choices for experiments presented in the following two chapters. In Chapter 7, we present

our results of using the same methodology on the data sets about another endemic disease (one that is established in its environments and leads to recurring epidemics): rubella in Peru; we also present results for another endemic disease — seasonal influenza in Chile — in Appendix B.3. In Chapter 8 we use the same methodology to study the spread of emerging diseases (newly-discovered or introduced into a region for the first time) across space, using two data sets of different length and quality related to the 2014 Ebola epidemic in West Africa. We also present results from applying the same methodology to the 2009 epidemic of a new strain of influenza (H1N1, also known as "swine flu") in Mexico in Appendix C.3.

In Chapter 9, we attempt to evaluate the applicability of our methodology to disease-correlation networks in a more general setting. We develop an agent-based model of disease spread between a number of cities within a known spatial environment, in which the probability of an individual traveling between cities is related to the distance between the cities and to a planted community structure. We apply our community detection methodology (modularity maximization using various null models) to synthetic time series produced by this model. By using a "burn-in period" in the model, we study two cases representing the endemic and emerging diseases. We also investigate various parameter values: infection rates, recovery rates and transport rates, as well as different degrees of inter-community mixing. We identify the parameter regimes where community detection using different null models detects spatial partitions and we attempt to explore the reasons for their spatial appearance. We also study the ability of our method to detect the planted partitions.

Finally, in Chapter 10, we summarize our results and suggest some directions for future research.

Chapter 2

Motivation: Studying disease spread

This chapter is a literature review and contains no original research.

2.1 Introduction

The use of statistical approaches to study infectious diseases dates to the 18th century when Bernoulli analyzed life expectancies and death rates while studying the effectiveness of variolation as a public health tool [42, 78]. In the 19th century, Philip-Charles Alexandre Louis in France introduced the design of clinical trials and William Farr in Britain pioneered medical statistics [42]. This merger between epidemiology and social justice gave rise to the statistical hygienic movement, which dominated the field until the end of the 19th century [42].

Following the proof that germs are responsible for infections, researchers began to develop mechanistic models of infection — such as the now very familiar SEIR (Susceptible – Exposed – Infected – Removed) models [42, 241], formulated using deterministic differential equations. Many of these simple compartmental models are designed to model a single population and/or focus on a single aspect of infection, and thus they assume homogeneity for various parameters.

In the past few decades, epidemiology appears to be re-embracing the statistical identification of risk factors, with an increasing number of studies using statistical approaches to elucidate factors that influence disease infections and spread [42]. With the increase in computing power and interest in the field from mathematicians and physicists, new powerful methods have become available for modelling and data analysis [42, 188].

Most epidemiological studies to date sought to understand the effects of introduction of disease into a population, and to focus on the most urgent questions such as the numbers and timings of new infections, rather than their exact locations and ways of transmission. While useful in estimating these numerical data, such single-population models do not explicitly address the causal factors in epidemic development, such as the path the disease

took and the factors that influenced it. The development of computer technology and the increased availability of disease-related spatial data have made possible the development of increasingly complex models and invention of new approaches that include individual activity, travel, contact patterns and other spatial aspects of infection.

In this chapter, we review established approaches to studying infectious diseases, from both modelling and data-analysis perspectives. We focus on methods that can be used to explore the factors that influence the spatiotemporal spread of infections.

2.2 Mechanistic approaches to disease modelling

2.2.1 Compartmental models

Probably the most well-known type of epidemic models are *compartmental models*, in which a population is divided into compartments based on their disease status, and the nature and rate of transfer from one compartment to another are described using parameters based on biological characteristics of the disease [195]. Most commonly, the rates of transfer between compartments are expressed mathematically as derivatives of compartment sizes with respect to time and as a result, models are formulated as differential equations [40,195]. Other possibilities include implementing discrete-time models using difference equations, and stochastic models using branching processes [40].

The canonical compartmental models deal with the spread of disease when introduced into a large susceptible population. The two most widely studied types are the SIS (Susceptible – Infected – Susceptible) and SIR (Susceptible – Infected – Removed) models.

The SIR model represents diseases in which hosts can be removed from the possibility of being infected, either through lasting immunity after recovery (e.g., infantile diseases such as measles, mumps and rubella) or through death (e.g., plague, rabies, and other animal diseases). In the SIR model, the population is divided into the following classes:

- $S(t)$, the number of susceptible individuals at time t (who are not infected);
- $I(t)$, the number of infected individuals at time t (who can transmit the disease by contact with susceptibles);
- $R(t)$, the number of removed individuals at time t (who have been removed from the possibility of being infected through immunization, isolation, or death).

In the canonical SIR model, these disease states are modeled by a set of coupled ordinary differential equations (ODEs) for the spread of disease in time:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI, \\ \frac{dI}{dt} &= \beta SI - \alpha I, \\ \frac{dR}{dt} &= \alpha I\,, \end{aligned} \tag{2.1}$$

where β is the infection rate of the disease, α is the recovery rate, and there is no entry or departure from the population (except possibly through death from the disease).

The SIS model represents diseases that do not confer lasting immunity, such as the common cold, most bacterial diseases, and most sexually transmitted diseases. Its canonical, ODE version is defined in an analogous manner to the SIR model in Eq. (2.1). However, after an infected individual recovers, he/she returns to a susceptible state. The disease states for the SIS model can be described by the following equations:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI + \alpha I \\ \frac{dI}{dt} &= \beta SI - \alpha I \,, \end{aligned}$$

where the parameters α and β are the same as in the SIR model above.

Compartmental models can be used to study the progress of a disease within a population. After validation on past data, they can also be used as generative models to predict likely epidemic outcomes and to estimate disease properties such as the *basic reproduction number* R_0 (the expected number of secondary cases that one primary case generates when the disease is introduced into a susceptible population) [77]. This quantity determines the potential for epidemic spread, in the sense that an epidemic will result from the introduction of an infectious agent if $R_0 > 1$ [40, 77, 237]. Models validated with disease data can also be used to give mechanistic insights into the factors that influence infection and disease spread.

In formulating the models in terms of derivatives, it is assumed the number of members in each compartment is a differentiable function of time, and that the epidemic process is deterministic. For smaller compartment sizes, this may not be a good assumption, as stochastic effects are then more important, and this is where the stochastic compartmental models are most useful [40].

The SIR and SIS models form a simple starting point to epidemic modelling, and they can be modified for different diseases and research questions. Some of the generalizations include:

- adding more compartments (e.g., $E(t)$ — exposed but not yet infectious individuals, or $Q(t)$ — quarantined individuals),
- introducing vaccination and/or temporary immunity,
- vector transmission,
- nonhomogenous mixing within populations,
- age-structured populations,
- variable infectivity,
- introducing births and deaths,

and many others. Describing all of them is beyond the scope of this thesis, but there are detailed discussions in many books, e.g., [40, 77, 195].

Incorporating space into compartmental models The majority of spatial compartmental models belong to either "metapopulation" (e.g., patch) or "spatially continuous" models [237].

In a *metapopulation model*, the population is distributed into a collection of spatially discrete groups. Usually, one assumes that the individuals within a group are well-mixed and that the groups are coupled to one another through a "contact matrix" or explicit mobility rates. Disease transmission occurs within groups that contain both resident and visitor individuals. Many models simplify the assumptions as much as possible in order to focus on particular questions, e.g., by assuming symmetry and equal rates of contact between populations. However, with better computing infrastructure, numerical studies of more realistic models are also often being pursued [16, 273].

Spatially continuous models assume that a population is continuously distributed through space. They are partial differential equation models in which the spatial location is represented by a variable x, and densities of susceptible and infected individuals at time t throughout the region are represented by $S(x,t)$ and $I(x,t)$. The basic models assume disease transmission is a local event, so rate of transmission at any point in space depends only on the densities of S and I individuals at this point. Population mixing can be introduced by allowing individuals to move at random through, for example, spatial random walks or diffusion [196, 237].

Summary The mechanistic insights of compartmental models can be extremely useful in aiding our understanding of the mechanisms of infection and spread of diseases. However, the reductionist compartmentalization approach is not able to effectively capture the heterogeneity of the various components of the chain of infection, such as hosts and vectors, the individuality of behavioral patterns based on past experiences, disease evolution, and other aspects arising from the enormous complexity of the system. Further, fitting data to compartmental models is a daunting task, especially for large and noisy data sets.

2.2.2 Agent-based models

Agent-based models (ABMs, also referred to as individual-based models) are computer representations of complex systems that consist of a collection of agents that interact with each other and their environment over discrete time steps [13]. The "agents" are anything that alters its behavior in response to input from other agents and the environment. For disease models agents usually represent individual people or disease vectors, but collectives

such as farms, offices, or government departments can also be agents. The scale of agent-based models varies from very small — representing a hospital, a school or a couple of villages — to extremely large, representing cities, provinces, or whole countries [49, 244]. Often, larger models necessitate simplifications in the agent rules because of computational constraints.

The interaction of many different agents with individual traits and initial behavior rules that organize their actions and interactions on a local scale in an ABM can give rise to complex and emergent group-level phenomena that are not specifically encoded in the rules of the model and might be counter intuitive. Furthermore, each aspect of the model rules can be calibrated and validated separately, making ABMs a relatively natural and flexible form of modeling. ABMs are considered to be particularly suitable for situations with local, complex interactions, heterogenous agents, where the phenomenon has inherent temporal aspects and where agents are adaptive [244].

In a typical epidemic ABM, each individual is modeled explicitly, together with their disease status (e.g., S, I, R as defined in Section 2.2.1) and behavioral rules. Stochasticity can be included in the assignment of agent characteristics and in the interactions. An ABM is run over time, and each simulation is commonly repeated numerous times in order to obtain a distribution of possible outcomes for the system. The important differences between ABMs and other models are the discrete representation of each individual, the explicit representation of their unique characteristics and behavior, and the explicit modelling of interactions that lead to disease transmission.

The individual-based approach of an ABM often gives different results than population-based modelling, due to the dependence on explicit interactions and the strong influence of stochasticity on the model, especially at low individual numbers. Most ABMs are not used as predictive models; instead, they are tools to study particular aspects and mechanisms of disease spread.

Summary Thanks to increases in computing power, ABMs have become increasingly popular in the past 10 years, and they have been used to study several diseases [85, 169]. Both researchers and policy makers are turning to ABMs, as they are more detailed than traditional models, yet more cost-effective (and arguably more ethical) than empirical observations, for example in studying the effect of various interventions. Due to their complexity and stochastic nature, they have the potential to better reflect the intertwined and complex issues involved in epidemic spread. They can take into account details such as the transportation infrastructure, population mobility and demographics, and epidemiological aspects such as the evolution of disease within a host and transmission between hosts. However, due to the explicit modelling of each agent, they are computation-intensive compared

to traditional models. Additionally, due to their complexity and computational nature, they are typically more difficult to interpret and to validate [237].

In this thesis, we use an agent-based model of disease spread in Chapter 9 to generate synthetic time series for the spread of disease through a group of interconnected cities. We use these time series for testing the applicability of our community-detection methodology to a wide range of disease types, and in particular we examine whether it generates networks in which we are able to detect spatial partitions. We also plant transport-based partitions and test our ability to recover them through community detection. We want to use a stochastic model, as we are especially interested in modelling diseases which periodically die out and reinfect locations. We choose to use an ABM rather than a compartmental model due to its greater suitability to our modelling skills and the limited time-frame for this study. However, other stochastic models reviewed in this section would be applicable to the task as well.

2.3 Statistical and data analysis approaches

The second main approach to studying infectious diseases is through statistical analysis. Statistical modelling is usually used either for exploratory data analysis in order to obtain the main characteristics of disease data, or for confirmatory data analysis in order to test a hypothesis about the association of disease occurrence with chosen variables. However, statistical approaches are not able to provide definitive information about underlying epidemiological mechanisms, because a given pattern of association between disease time series can be generated by a wide variety of different mechanisms.

Multiple statistical approaches to studying infectious disease have been proposed. An in-depth review of these methods is beyond the scope of this work; a good starting point is given by the reviews in Refs. [42, 198]. Many studies use a combination of the different "standard tools" such as correlation, linear regression, Poisson regression, and other non-linear models to disentangle the various (e.g., climatic and socioeconomic) factors that influence disease occurrences [62, 75, 198]. Below, we highlight some of the more specialized tools for studying the spatial and temporal aspects of disease time series.

Temporal approaches Two types of statistical models of particular interest to us are time-series models and spectral analysis; both of these approaches are used to study patterns in time-series data.

Time-series models are based on the assumption that the past behavior of a system allows one to predict future behavior; they use moving average and/or autoregressive components to predict future time series, and the addition of space-time covariances adds dependency on neighboring areas [57]. These kinds of models have been used for infectious disease modelling since the 1980s [127], including to identify climatic influences on dengue

fever [76, 198] and in disease forecasting [6]. However, classical time-series analysis techniques can only be used for time series in which the statistical properties do not vary with time — i.e., they are stationary — and epidemiological time-series are typically noisy, complex and strongly non-stationary, requiring data filtering. Other approaches, such as dynamic linear models [44], support vector machines and artificial neural networks [288] are also used to forecast disease time series.

Spectral analysis can be used to search for patterns in data, taking into account the time dependences, trends, and cycles in epidemiological time series. Fourier transforms have been used to decompose filtered data sets with relation to climate and weather variability [45,228]. Wavelet analysis can be used to detect periodicity in non-stationary disease data [45, 198].

Geographical approaches Most geographical approaches to studying infectious disease spread are descriptive and retrospective. They use statistical methods to answer three key questions: identifying patterns of disease distribution, detecting disease clusters and projecting the future spread of epidemics [237].

Patterns of disease distribution and factors that influence disease occurrence can be studied by constructing a set of maps that describe disease occurrence, associated risks and the direction and magnitude of spread. Most mapping techniques focus on smoothing data in order to simplify them and to bring important features to attention. Further inspection using methods such as regression analysis can identify factors that influence the timing and extent of disease spread [237]. The map of the spatial distribution of disease risk can then be used to predict future spread of disease through approaches such as Monte Carlo simulation and Bayesian estimations [237].

"Disease clustering methods" aim to uncover unusual concentrations of disease occurrence in space and time. Global measures of spatial autocorrelation (e.g., Moran's I [189] and Geary's c [98]) can assess whether any spatial correlation is present in a whole data set. Local measures of spatial autocorrelation (e.g., Local Indicator of Spatial Association [10] and Kuldorff's spatial scan statistic [152, 258, 285]) additionally detect cluster location [111, 277].

Summary In general, statistical models tend to be flexible with respect to the format of input data and parameter selection, which makes them very useful, especially for preliminary data analysis. They require few assumptions about the nature of the disease process, which makes them easily reusable as long as the data format fits the methodology. However, the requirement for high-quality data is a potential drawback of statistical approaches, and it can be particularly problematic for emerging infectious diseases or for epidemics in developing countries, where data collection is limited, such as the Ebola epidemic in west Africa that we study in Sections 8.2 and 8.3. Additionally, for more advanced methods,

computation complexity can be an issue. Finally, these models can only identify correlations rather than causal relationships.

2.4 Mobility models

Understanding the patterns of mobility of individuals is a common research topic, and its applications range from city planning to public health. Quantitative studies of human mobility have suggested that human movements follow statistically predictable patterns and different models have been developed to study it. In recent years, mobility models have increasingly been used as part of large-scale metapopulation models of disease spread [265,270]. The two main types of models used for this purpose are gravity models and radiation models. The intervening-opportunities model [254] is another notable mobility model that is focused on predicting migration in terms of the opportunities available at potential destinations, but it has not to our knowledge been used in disease modelling.

Gravity models Gravity models have been used in various fields to describe interactions in a way similar to Newton's law of gravity. A gravity model assumes that the interaction between two locations is proportional to their importance (e.g., mass, population, GDP), but it decays with distance. For modelling population-related quantities such as disease spread and transport rates, the population of a location (e.g., a city or a province) is usually used as the measure of importance. Thus, the interaction between locations i and j with respective populations n_i and n_j that are a distance d_{ij} apart is

$$G_{ij} = n_i^{\alpha} n_j^{\beta} f(d_{ij}) \,, \tag{2.2}$$

where the "deterrence function" $f(d)$ describes the effect of space on node interactions. Common choices for the deterrence function use inverse proportionality to distance (i.e., $f(d_{ij}) = 1/d_{ij}$), inverse proportionality to squared distance (i.e., $f(d_{ij}) = 1/d_{ij}^2$) exponential decay (i.e., $f(d_{ij}) = e^{-d_{ij}}$), and other interactions of the form $f(d_{ij}) = d_{ij}^{\kappa}$ [21], where α, β, and κ are parameters that one can determine using regression [21]. Gravity models have been successfully employed during the past half century to model spatial interactions such as human mobility (across multiple scales) [251,252,282,289], population migration [15,21,164], and trade [79].

Gravity models have also been used in infectious disease modelling. Scholars have used them to predict the spread of H1N1 influenza from Mexico (taking into account extra variables such as population size, per capita GDP and distance [167]), and to represent transport for models of diseases including influenza in the US [273] and measles in the UK [284].

Gravity models include multiple parameters that need to either be chosen arbitrarily or estimated from data. Moreover, by their design, gravity models are unable to predict

different fluxes between locations that are the same distance apart but which have regions with different population densities between them. For example, due to the higher availability of susceptible hosts, one would expect a higher flux of infectious disease between two locations separated by a region with high population density than between locations separated by a region with low population density [138]. By contrast, one would expect a smaller commuting flux between such locations due to higher availability of nearby jobs, which reduces people's willingness to commute for longer distances [245]. Finally, gravity models lack foundation in theory related to human behavior and disease spread [49, 245]. However, calibration data is increasingly available, and the simplicity of gravity models is a major advantage for implementation and analysis.

Radiation model The radiation model [245] attempts to address some of the issues faced by the gravity models, and the model has shown promise in matching commuting and travel data [103,178,245]. In the radiation model, the commuting probability depends only on the origin and destination populations (n_i and n_j respectively) and on the population r_{ij}: the population q_{ij} residing in the circle centered in i with radius d_{ij}, minus the populations at the origin and destination. This population is denoted by $r_{ij} = q_{ij} - (n_i + n_j)$.

The derivation of the radiation model aims to describe the number of people moving between the two locations (e.g., counties) in terms of their long-term decisions to choose their jobs that then generate the daily commuting patterns. The model assumes that job selection consists of two steps:

1. An individual searches for jobs across all counties, including their home county. Assuming that there is one job opening for every n_{jobs} individuals, the number of jobs in county i is proportional to the resident population n_i. Each job has its benefits quantified by z, randomly chosen from as distribution $p(z)$.

2. The individual chooses the closest job to their home with higher benefits z than the best offer available in their home county. Thus lack of commuting has priority over the benefits, that is, individuals are willing to accept lesser jobs closer to their home.

Applying this process in proportion to the size of the population of each county assigns work locations to each potential commuter, which in turn determines the movement fluxes across the whole country. The model has three unknown parameters: the benefit distribution $p(z)$, the job density n_{jobs}, and the total number of commuters, N_c. However, the commuting fluxes T_{ij} are independent of $p(z)$ and n_{jobs}, and the remaining free parameter, N_c, does not affect the flux distribution, making the model parameter-free. As the model can also be formulated in terms of radiation and absorption processes (see Supplementary Information of [245]), we refer to it as the radiation model.

The average commuting flux between i and j is then

$$T_{ij} = T_i \frac{n_i n_j}{(n_i + r_{ij})(n_i + n_j + r_{ij})}, \tag{2.3}$$

where $T_i \equiv \sum_{j \neq i} T_{ij}$ is the number of commuters who reside in location i. The authors of Ref. [245] assume that the quantity T_i is proportional to the population of i, so $T_i = n_i(N_c/N)$, where N_c is the total number of commuters and N is the total population of the country.

There have been several subsequent modifications to the radiation model, including normalization for finite systems [178] and a generalization into a framework that includes radiation, gravity, and intervening opportunities models [246].

The radiation model has been used to model the spread of infectious disease in several applications ranging from influenza in UK and US to Ebola in West Africa [67, 106, 270], and it has been shown to match census mobility data well for epidemic modelling applications [265].

Summary Mobility models have been used to model the spread of infectious diseases in recent years. They are also a useful starting point for incorporation the expected influence of distance and population sizes on connectivity patterns between nodes into null models for community detection. The gravity model has been incorporated into a gravity null model in Ref. [82], which we define in Section 3.3.3. In this thesis, we develop a null model based on the radiation model, which we define in Section 3.3.5. We then test the results of using these null models on spatial benchmark networks in Chapter 4, on disease-correlation networks constructed from real disease data in Chapters 6– 8, and on disease-correlation networks constructed from synthetic time series from an agent-based model of disease spread in Chapter 9.

2.5 Network models

Network models of the spread of infections are usually based on the assumption that infectious diseases follow identifiable paths such as interpersonal contact networks or transportation links. Thus, a contact network is placed between entities (individuals, cities, etc.) and a disease model is ran on this connectivity structure. The model can be a compartmental model, or an agent-based model (where nodes are individuals), a metapopulation model (where nodes are populations) or any other model type. With the increased popularity of network science, recent years witnessed a number of novel network approaches to epidemic modelling, and many of these studies are reviewed in [212].

Several models previously described in this chapter can be thought of as network models, ranging from agent-based models that incorporate a contact network between individuals [49, 244] through to global metapopulation models with populations joined by a transportation network [16, 273].

Individual-based network models of epidemic spread often take the form of a dynamical system (e.g., a compartmental model) placed on a network where nodes represent people (with each node having a defined disease status), edges represent their contacts, and the nodes have update rules that govern how the states change [213, 218]. On one end of the spectrum, multiple studies focus on the properties of epidemic spread on idealized networks, yielding general insights into the mechanics of epidemic spread between people (for a review see [142]). On the other end of the spectrum, data-driven simulation studies often use realistic contact networks — e.g., networks of sexual [227] or hospital contact data [163], or social networks on a citywide scale [232]. Many studies have used temporal networks, and they illustrated that the temporal structure of the network is important to disease spread. Furthermore, temporal structure can be used to devise vaccination strategies [129].

Population-based network models often represent large-scale metapopulation models with some transportation network [16, 59, 61, 131]. Advancements in theoretical methods and computational power — and the interest from physicists and mathematicians — are beginning to drive the development of increasingly sophisticated models (e.g., combining the individual and population levels of contact and transportation [16, 28, 41]). Such studies include investigation of the influence of features such as spatial location, climate, and strength of transport links on synchronization of disease spread and disease persistence. Studying the temporal evolution of disease patterns on a larger scale is also of interest, as it can provide useful epidemiological information.

One of the major challenges to modelling disease spread on realistic networks is the requirement for data about the properties of nodes and their connectivity patterns. Population-scale studies require some combination of: population, census, land use, transportation, and economic and other data, which is sometimes collected by governmental departments or may need to be inferred or modeled. Individual-level models require contact data which is hard to directly acquire (and may be unreliable), or may need to be modeled based on population-level information. This additional level of complexity adds to the difficulty in applying network models to predict disease spread, and to the potential for biases and errors.

2.6 Summary

In this chapter we have surveyed the mathematical, statistical and computational approaches currently used to model infectious diseases, with a focus on studying their spatial spread. The two most popular approaches to modelling infections are differential equation-based models (working on homogeneously mixed populations), and agent-based models

(ABMs — computer-based models where individuals are represented separately). The equation-based approach can allow the proof of analytical results, but an analytical approach is only applicable to relatively simplified scenarios. Further, compartmentalization forces one to remove much of the heterogeneity that is important to the complex phenomenon of infectious disease spread. In contrast, ABMs can be very flexible in terms of realistic simulation of disease spread scenarios. With increasing availability of computing power, ABMs can be used to study disease spread on any scale from immune response within-individuals to global pandemic spread. However, because of their complexity, these models are harder to analyze and validate, especially if one seeks mechanistic insights. Both mechanistic model types suffer from the need to simplify them for computational and/or analytical tractability, and the need to identify the inputs beforehand. They also need to be designed and parametrized for each new disease/location studied.

The effects of space and transportation can be incorporated into predictive disease models (both equation and ABM-based) in several ways: through the use of partial differential equations, metapopulation models, gravity models, and network models. These models have been used to study the effects of contact and travel patterns on the spread of diseases. However, these approaches use explicit information about contact/transport patterns, and thus are relying on the availability of reliable empirical or model data to inform parameter choices.

Statistical analysis and data analysis have pioneered a different approach to studying disease epidemics. These studies use regression, correlation, time series analysis, Fourier analysis and other methods to find properties of known disease data sets. One advantage of such approaches is that they do not require detailed knowledge or mechanistic assumptions about the disease in question, and they are broadly applicable and reusable. However, they are heavily dependent on the availability of good quality disease data, which is not always a given, especially when dealing with emerging diseases in developing countries. Further, they are only able to detect associations, rather than mechanistic insights.

The community detection methodology that we present in the next chapter is similar to the data analysis methods in its aims — it is able to detect patterns in the data and suggest avenues for further investigation rather than delivering mechanistic insights. Its advantages include broad applicability and no reliance on a particular data distribution, and the fact that it only requires limited knowledge of the disease properties to effectively use. However, is is very dependent on good-quality data.

Chapter 3

Methodology: Networks and community structure

The majority of this chapter consists of a literature review. Sections 3.1.1, 3.4.1, 3.4.2, and 3.5 present particular methodologies of constructing, visualizing, summarizing and examining partitions of disease-correlation networks that we choose to use, and Section 3.6 puts these choices into context of other approaches to studying disease time series that we presented in Chapter 2.

The novel null model for modularity maximization in Section 3.3.5 appears in a working paper by MS, E. Leicht, G. Chowell, and M. A. Porter [235]. The statistics described in Sections 3.4.4.1 and 3.4.6 are part of a working paper with A. Elliott and M. A. Porter that is not yet published. The remainder of the chapter forms an introduction to the field of networks and the particular techniques we use.

3.1 Networks

Some terminology Relationships among a set of objects can be modeled as a "network" in which *nodes* (or vertices) represent the objects, and *edges* (or links) represent the relationships between them. In its simplest form, a network is simply a graph consisting of a set V of N nodes and a set E of M edges. Such a graph can also be represented as an $N \times N$ adjacency matrix A, where $A_{ij} = 1$ if there exists an edge between node i and node j and $A_{ij} = 0$ otherwise . In this thesis, we refer to graphs as "static" networks, in contrast to the temporally-evolving multilayer networks that we will describe later.

The networks that we study have at most one edge between any pair of nodes (i.e., there are no *multiedges*). Additionally, in our network representation there are no edges between a node and itself (i.e., no *self-loops*).

Networks can be unweighted ($A_{ij} = 1$ if there is an edge that connects nodes i and j, and 0 otherwise), or weighted, in which the weight W_{ij} of an edge represents the strength of the relationship between the entities i and j. Some (usually weighted) networks are *fully connected*, which means that all the possible edges are present.

Networks can be *undirected*, meaning that $A_{ij} = A_{ji}$ for all i and j. Alternatively, their edges may be *directed* from one node to another (where we denote an edge from i to j as W_{ij}). In this case, the resulting adjacency matrix is typically asymmetric. In this thesis, we use undirected networks.

The *degree* g_i of a node i is defined as the number of neighbors (i.e., the number of edges that are incident to the node), and the *strength* k_i is the sum of the weights of all edges incident to the node: $k_i = \sum_j W_{ij}$. The *degree distribution* $P(g)$ of a network is then defined to be the fraction of nodes in a network with degree g, and the *strength distribution* $P(k)$ is the fraction of nodes in a network with strength k.

A graph is *connected* when there is a path (i.e., a sequence of vertices $\{v_1, ...v_n\}$ s.t $A_{v_i v_{i+1}} = 1$ for $1 \geq i \geq n-1$) between every pair of nodes. A *tree* is an undirected, connected graph with no cycles or self-loops, and a *spanning tree* of a graph with N nodes is a subset of $N-1$ edges that form a tree. The *minimum spanning tree* (MST) is a spanning tree with minimum total cost with respect to a cost function on the edges (commonly the edge weight W_{ij} in networks related to distance or a function of edge weight (commonly $1/W_{ij}$ or $\sqrt{2(1-W_{ij})}$) when the edge weight is a similarity measure [175].

Network representations in this thesis We represent the disease time series data as *similarity networks*, which are constructed by defining edges based on some form of similarity (e.g., a correlation measure) between each pair of nodes. Similarity networks tend to be fully connected (or almost fully connected) and weighted, except when they have been deliberately thresholded. In this thesis, we use Pearson correlation coefficients between time series to construct undirected networks from disease time series. Raw Pearson correlation values lie between -1 and 1. However, following examples from previous studies [26, 83, 84] for most experiments (except the correlation null model [171], see Section 3.3.6), we linearly reweigh the networks to lie between 0 and 1. All edges then have the same sign, which allows us to apply the standard null models for community detection via modularity maximization. This means we lose the ability to treat the negative edge weights differently than the positive edge weights (a feature of the null model for modularity maximization on signed networks [107]). However, as we only have a very small number of negative edge weights, this does not have a large influence on the results.

We wish to study the temporal evolution of network structure through the detection of "communities" — a type of mesoscale network structure formally defined in Section 3.2. Two main approaches have been used for community detection in temporal networks. The first is to construct a static network by aggregating temporal snapshots of the network into a single network (e.g., by taking the mean or total edge weight for each edge across time; this can be problematic if the set of nodes varies in time). The second approach is to use static community-detection techniques on each element of a time-ordered sequence of

networks at different times over different time intervals, and then tracking the communities across the sequence [83, 84]

Both of these approaches allow one to use static community detection methods and thus provide a good starting point for the development and investigation of new methods — which, in our case, entails how to incorporate spatial information into null models for community detection via modularity maximization. However, static networks do not take full advantage of the temporal information contained in data that changes in time. For example, it can be hard to track the identity of communities in temporal sequences of networks [194].

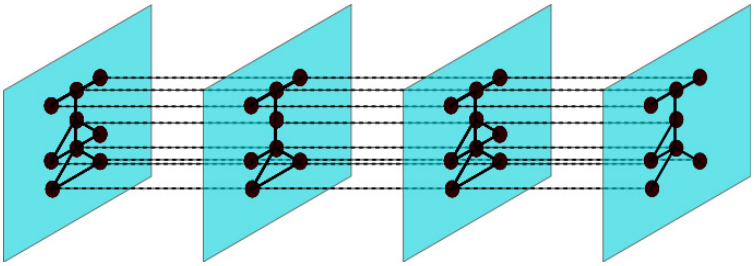


Figure 3.1: A multilayer network with 9 physical nodes (representing provinces). Intralayer edge weights (solid lines) indicate connectivity between the provinces (e.g., Pearson correlation between each pair of disease time series in a time window for the disease-correlation networks). A set of such correlations gives one temporal layer. We connect copies of each node in neighboring layers with interlayer edges of uniform weight $\omega \in [0, \infty)$ (dashed lines).

To mitigate this, we also use a type of *multilayer network* [146] that is known as a *multislice network* [194]. In this case, we have an $N \times N \times m$ adjacency tensor $\overline{W}$ that has m "layers" (or slices) and N nodes in each layer. Such a multislice network represents the connections between N *physical nodes* (entities), and it has $(N \times m)$ *multilayer nodes* (i.e., node-layer tuples), each of which corresponds to a specific (node, time) pair. The intralayer edges in the network are defined in exactly the same manner as they would be for a sequence of static networks. The tensor element $\overline{W}_{ijs}$ gives the weight of an intralayer edge between nodes i and j in layer s. Additionally, each layer has a copy of node i, which is connected to its counterparts in consecutive layers s and r using interlayer edges of weight U_{isr}. In this thesis, we suppose for simplicity that $U_{isr} \in \{0, \omega\}$ where $\omega \in [0, \infty)$, but one can also consider more general situations [73, 146]. This structure makes it possible to detect temporally evolving communities in a natural way.

3.1.1 Network construction methodology

In this section, we present the methodology that we use to construct correlation networks from disease time series.

Each data set D consists of N time series of weekly disease counts $(D_1, D_2, \ldots, D_N)$ over T weeks: $D_i(t)$ is the number of disease cases in province i at time $t \in \{1, \ldots, T\}$. (See

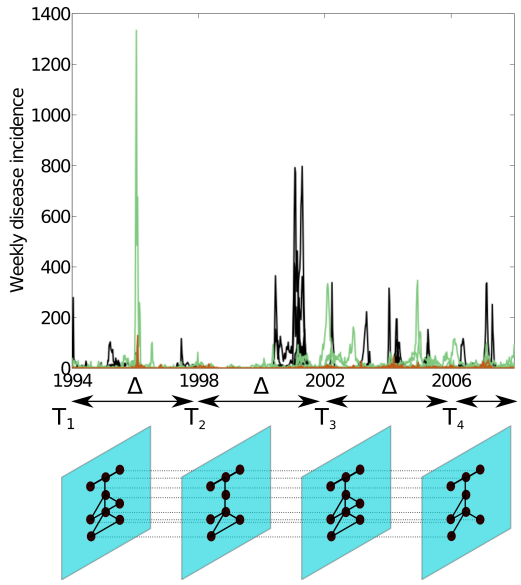


Figure 3.2: Construction of multislice correlation networks from disease time-series data. The top panel shows the dengue fever time series for the 79 provinces of Peru. We color the provinces by climate: coastal provinces are in black, mountainous provinces are in brown, and jungle provinces are in green. The bottom panel shows an example of the multislice network construction for 9 nodes with starting point $\tau_1 = 1$, $\upsilon = 208$ and time window width $\Delta = 208$. (The time points correspond to 1/1/1994, 27/12/1997, 22/12/2001, and 17/12/2005). The nodes represent provinces and each intralayer edge weight is given by a Pearson correlation between a pair of single-province time series in a given time window. One set of correlations gives one temporal layer, and we connect copies of each node in neighboring layers using interlayer edges of uniform weight $\omega \in [0, \infty)$ (dashed lines). The case $\omega = 0$ yields a set of static networks. (All other aspects of our network construction are the same.)

Fig. 3.2 for an example, in which we plot total number of disease cases versus time for the dengue fever data set together with a schematic illustration of the network construction.)

We use the term *time window* for a set of discrete contiguous time points. We seek to study the temporal evolution of the correlations between disease occurrence in provinces by constructing separate networks for different time windows — we either construct a set of static networks or a multislice network. We define a list of m starting points for the time windows $\tau = \{\tau_1, \tau_2, \ldots, \tau_m\}$. They are placed at fixed interval υ with respect to the first start point τ_1, such that $\tau_i = \tau_1 + (i-1)\upsilon$. To create networks from these starting points, we define Δ to be the width of the time window; thus, $(\Delta - \upsilon)$ describes the amount of overlap between the two time windows. We use uniform values of parameters Δ and υ for each set of static networks or a multislice network. We use $\tau_1 = 1$ unless we state otherwise.

The starting point τ_s, and time window width Δ define a subset of the disease time series that corresponds to network/layer number s. For example, for the time series of disease cases in province i, the time-series subset $E_i^{(s)} = \{D_i(\tau_s), D_i(\tau_s + 1), \ldots, D_i(\tau_s + \Delta)\}$ represents the numbers of disease cases in province i at times $\tau_s, \tau_s + 1, \ldots, \tau_s + \Delta$ where s is the layer number. By considering $E_i^{(s)}$ for all provinces, one can construct a static network or one layer of a multislice network.

We construct the networks using Pearson's correlation coefficient for all pairs of the

time series subsets $E_i^{(s)}$. However, some provinces do not experience disease in every time window. If a province does not experience disease cases in a time window, correlation is not defined as the corresponding time series has a standard deviation of 0. For this reason, we only construct static networks and multislice network layers using nodes that experience disease in the relevant time windows, and the remaining nodes are "dummy nodes" with 0 strength (and 0 interlayer connectivity in multislice networks). We define $\hat{N}^{(s)}$ as the number of nodes that "exist" in the network/layer s, i.e., they experienced disease in the corresponding time window and thus they have non-zero strength. We also define $\Psi_i^{(s)}$ as an indicator of whether a node exists in a given layer/network s. That is,

$$\Psi_i^{(s)} = \begin{cases} 1, & \text{if node } i \text{ exists in layer } s, \\ 0, & \text{otherwise}\,. \end{cases} \tag{3.1}$$

When we refer to static networks, we sometimes omit the network number s for simplicity.

For a static network, we define a set of N nodes $\{1, 2, \ldots, N\}$, where node i corresponds to province i. We generate networks using the Pearson correlation coefficient between the subsets of the time series ($E_i^{(s)}$) corresponding to the desired time window s. For the NG, gravity and radiation null model we reweigh networks to lie in the $[0, 1]$ interval, which simplifies community detection as we do not have to consider negative edge weights. Thus, the edge weight

$$W_{ij}^{(s)} = \begin{cases} \frac{1}{2}(\rho_{ij}^{(s)} + 1) - \delta_{ij}, & \text{if } \exists\ t, v \text{ st. } E_i^{(s)}(t) > 0 \text{ and } E_j^{(s)}(v) > 0, \\ 0, & \text{otherwise}\,, \end{cases} \tag{3.2}$$

represents the similarity between the time series $E_i^{(s)}$ and $E_j^{(s)}$, corresponding to network/layer number s, where the Kronecker delta δ_{ij} removes self-edges in layer s. The quantity ρ_{ij} is the Pearson correlation coefficient between the disease time series for provinces i and j. That is,

$$\rho_{ij}^{(s)} = \frac{\left(\langle E_i^{(s)} E_j^{(s)} \rangle - \langle E_i^{(s)} \rangle \langle E_j^{(s)} \rangle\right)}{\sigma_i^{(s)} \sigma_j^{(s)}}\,,$$

where $\langle \cdot \rangle$ indicates averaging over the time window under consideration and $\sigma_i^{(s)}$ is the standard deviation of $E_i^{(s)}$. Our construction yields a network W with elements $W_{ij} \in [0, 1]$. For the correlation null model, we construct networks using the raw correlation values ($\rho_{ij}^{(s)}$) in an analogous manner to Eq. (3.4).

In a multislice network for NG, gravity and radiation null models, the intralayer edge weights are

$$W_{ijs} = \begin{cases} \frac{1}{2}(\rho_{ij}^{(s)} + 1) - \delta_{ij}, & \text{if } \exists\ t, v \text{ st. } E_i^{(s)}(t) > 0 \text{ and } E_j^{(s)}(v) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

for each layer s, and for the correlation null models the intralayer edge weights are constructed in the same manner using raw correlations: $\rho_{ij}^{(s)}$. We connect each node i in the

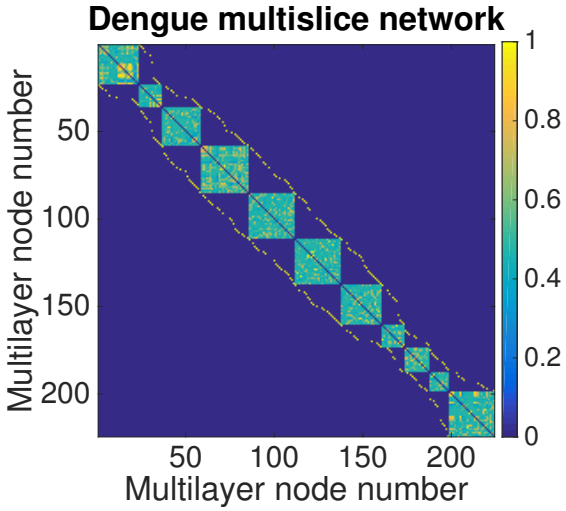


Figure 3.3: The multislice adjacency matrix for the dengue network, reweighted to lie in the interval $[0, 1]$, and shown as an adjacency-matrix representation where layers of the multislice network are placed on the main diagonal, and the inter-layer connections ω are on the off-diagonal, connecting nodes with copies of themselves across layers. We only show the non-zero nodes here — which demonstrates that many layers have only a small number of non-zero nodes present.

rth time window to copies of itself in an adjacent time windows $s \in \{r-1, r+1\}$ if the node "exists" in both layers, i.e., if there are disease cases in the corresponding time windows. We use interlayer edges of uniform weight

$$U_{isr} = \begin{cases} \omega \in [0, \infty), & \text{if } \exists\ t, v \text{ st. } E_i^{(s)}(t) > 0 \text{ and } E_i^{(r)}(v) > 0 \text{ and } |r - s| = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

The case $\omega = 0$ in the multislice network corresponds to a sequence of static networks.

This yields a weighted multislice correlation network, where only the nodes that experienced disease in adjacent time windows are considered. See Fig. 3.2 for a schematic that shows the construction of a multislice network, and Fig. 3.3 for an example of the multislice network for dengue fever, shown as a supra-adjacency matrix representation where all layers are represented as matrices on a diagonal.

Similar constructions of (both static and multislice) networks from time series have been employed for systems such as functional brain networks [23, 24], currency exchange-rate networks [83], gene expression networks [31, 150], climate networks [80, 173], and political voting networks [172, 193, 194]. When calculating a similarity network from a set of time series, the choice of similarity measure and the subsequent choices that one makes (e.g., uniform or nonuniform window length, and overlap or the lack of overlap if one uses a rolling time window) affect the values of the similarity measure and in consequence, the structure of the network.

When studying static networks, we typically use a set of overlapping static networks. We compute correlations using a rolling time window with a uniform window length and uniform amount of overlap for each data set. We report the specific values of the parameters Δ and v for each application in Section A.2.

To construct a multislice network, for dengue fever and both influenza data sets, we typically use $\upsilon = \Delta$ to create nonoverlapping time windows. We also briefly investigate the use of overlapping time windows for dengue fever in Section 6.6, and we use overlapping time windows for rubella and Ebola in Chapters 7 and 8. We report the specific values of the parameters Δ and υ for each application and the effects of these parameter choices on network structure in Section A.2.

There are a multitude of similarity measures in the literature, and the choice of the most suitable one depends on the particular data set and scientific question, and it is an active area of research [238, 250, 287]. The standard method of choice to create networks from disease time series is by calculating the Pearson correlation coefficient between each pair of time series. Other possible methods for network creation that we considered include partial correlation and lagged correlation, coherence (a spectral measure that is insensitive to lags, and has been used for fMRI data about brain activity [23]), and event correlation [173] (a measure based on pre-defined events and thus potentially suitable for "spiky" data). These and other methods are discussed in Ref. [250]. The use of lagged correlation, or a measure that is insensitive to lags such as coherence might be beneficial for this data set, as one can expect there to be a lag in the spread of disease across space. However, Pearson correlation is commonly used, especially in financial and biological contexts such as stock market prices and gene expression data [20, 26, 31, 83, 150]. Further, some papers have used community detection on correlation networks [26, 83, 84, 272]. Finally, lagged correlation would generate directed networks, and community detection using modularity maximization on directed networks is more complicated than on undirected networks and one has to carefully consider the meaning of edge directionality [145,165]. We choose to use Pearson correlation for simplicity.

Many features, such as the number of layers and the mean and variance of the Pearson correlation values, depend on the parameters that we use in constructing our networks. For example, it is important to consider the choice of the time-window width Δ. There is a trade-off between having many layers to obtain a good temporal resolution of events and ensuring that we construct each layer using enough time points to be confident of the statistical significance of the similarity values in each layer [24]. Larger values of Δ yield smaller variations in mean correlation across the years and lessen the effects of small, regional epidemics on the number of cases and on the correlation between disease profiles in different provinces. Therefore, we want to use a sufficiently large value of Δ so that we can examine long-term, repetitive disease patterns.

Additionally, studies based on random matrix theory (RMT) show that correlation matrices generated from time series that are shorter than the number of entities being analyzed (i.e., shorter than the number of nodes) are indistinguishable from the correlations that one calculates from short, uncorrelated sequences of noise [240]. This suggests that we

may want to use an additional constraint of $\Delta > \hat{N}$, where possible. However, choosing a value of Δ that is too large risks over-smoothing data and losing important information. See an analogous discussion of time-window choice in Ref. [84] in the context of financial networks. We will study the choice of time window width Δ for each disease data set in detail in Section A.2.

3.2 Community structure and modularity

Networks can be described using a mix of local, global, and intermediate-scale (mesoscale) perspectives. One of the key applications of network science is the development of summary statistics for analyzing and comparing complex structures and large networks [204]. The identification of mesoscale structures allows the observation of features that might not be apparent either at the local level of nodes and edges or at the global level of summary statistics.

One well-explored aspect of network structure is *community structure*, which searches for groups of densely-connected nodes. It is the methodology we will use in this thesis due to its demonstrated usefulness and flexibility [91, 219]. Other approaches to detecting mesoscale structures often take the form of various block-models [91], where nodes that have similar structural properties are placed in groups, however unlike in community detection these groups do not have to be more densely connected within groups than between groups. Another type of of mesoscale structure is core-periphery structure [65, 229], which entails identifying densely connected core nodes and sparsely connected peripheral nodes.

Loosely speaking, a *community* is a set of nodes that are relatively densely connected to each other but sparsely connected to other dense sets in a network [91, 219]. (See the toy examples in Fig. 3.4.) Although the notion of communities makes intuitive sense, a precise mathematical definition is difficult to pin down [91, 219, 224].

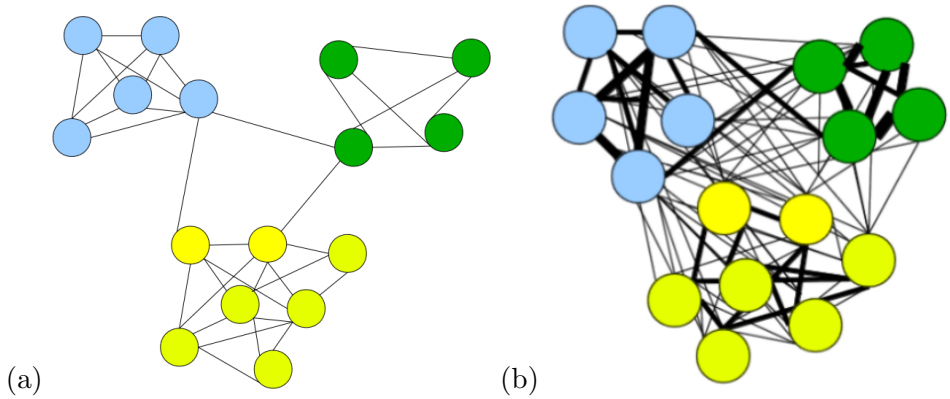


Figure 3.4: The community structure of (a) a toy unweighted network and (b) a toy weighted network. The three communities are in yellow, green, and blue. There are dense internal connections but sparse connections between communities.

We restrict ourselves to hard partitions, in which each node is assigned to exactly one community, and we use the term "partition" to mean "hard partition". It is also important, but beyond the scope of this thesis, to consider "soft partitions", in which communities can overlap (i.e., nodes can belong to more than one community) [91, 208, 209, 219].

A large number of community-detection methods exist, as reviewed in Ref. [68, 91, 205, 219]. Traditional methods adapted from other fields include the minimum-cut method and hierarchical clustering [219]. One of the first network-specific methods is the Girvan-Newman method [102] based on dividing the network by cutting edges with high betweenness centrality (i.e., edges that lie on a large number of paths between nodes). Some of the most popular approaches to date work through the optimization of some *quality function*, for example modularity [199] and stability [74]. A different class of methods are those based on information-theoretic ideas, such as Infomap [230] and related methods based on statistical inference [121, 162]. Another large class exploits links between community structure and dynamical processes taking place on networks, such as random walks, Potts models or oscillator synchronization [91, 205]. A very different set of approaches focuses on the detection of local communities [91, 135, 219]; these methods aim to find the community assignment of nodes without necessarily assigning communities to the whole network, which is useful for studying limited portions of larger networks.

In this thesis we will focus on modularity, as its approach of describing the "quality" of a particular network partition into communities in terms of its departure from a null model [199] is a useful (and underexploited) feature. In particular, modularity allows us to investigate the choices of different null models, from general ones [199] to application-specific spatial [82] and correlation-specific [171] null models, which we define in Section 3.3. Further, modularity is applicable to multilayer and time-dependent networks; something that is not present for the other methods at this point. Finally, algorithms developed to optimize modularity deliver some of the fastest and most widely used ways of community detection [33, 91, 219], which is important for our work on multislice networks, which can be computationally intensive.

For a weighted static network W, *modularity* [201] Q counts the total intracommunity edge weight minus the total edge weight that would be expected at random (given the choice of a null model):

$$Q = \frac{1}{2w} \sum_{ij} (W_{ij} - P_{ij}) \delta_{c_i c_j} \,, \tag{3.5}$$

where $2w = \sum_{ij} W_{ij}$ is the total edge weight, c_i denotes the community that contains node i, $\delta_{c_i c_j}$ is the Kronecker delta, and P_{ij} is the ij-th element of the null-model matrix (which we describe in Section 3.3).

Modularity can be derived by combinatorial arguments [204] or by examining the statistical properties of random walks on a network [155], as we discuss in Section 3.2.1.

Modularity maximization has many limitations: e.g., a resolution limit on the size of communities [92], the inability to detect communities in networks with heterogenous distributions of cluster sizes [159], a detectability limit [197], and a huge number of nearly-degenerate local maxima [108]. Nevertheless, it is a very popular and well-studied method, and it has been used successfully in numerous applications [91, 219]. It is thus a natural choice for our application.

The resolution limit of modularity can be partially mitigated by using a multi-resolution version of modularity, such as the approaches proposed by Reichardt and Bornholdt [224] Arenas, Fernández and Gómez [12], and Traag and Van Dooren [267]. In these variations, one can examine different scales of community structure by incorporating a resolution parameter, although the resolution limit applies to any fixed value of the resolution parameter. We use the multi-resolution method of Reichardt and Bornholdt, as this method allows us to change null models, and it is defined for multislice networks. For changing null models, this method has an added benefit: in some sense the value of the resolution parameter (which we refer to as γ) determines the importance that one assigns to the null model relative to the observed network. Smaller values of γ tend to yield larger communities, and vice versa.

This yields the formulation

$$Q = \frac{1}{2w} \sum_{ij} (W_{ij} - \gamma P_{ij}) \delta_{c_i c_j} \,. \tag{3.6}$$

3.2.1 Modularity for multilayer networks

For multilayer (and multislice) networks, the usual procedure for defining modularity as a count of the intracommunity edge weight minus intracommunity edge weight expected at random ignores contributions from the interlayer edges. To tackle this problem, Mucha et al. [194] formulated a null model in terms of the stability of communities under Laplacian dynamics (as defined in [155, 156]), and they derived a generalization of modularity maximization to multilayer networks with separate intralayer and interlayer edges. To derive the objective function for multilayer modularity we will first derive an expression for stability in a standard undirected and unweighted network, as detailed by Lambiotte et al. [155,156]. We then will derive an expression for multilayer modularity in a similar fashion, as derived by Mucha et al. [194].

Lambiotte Stability Lambiotte et al. [155] defined stability of a partition of the unipartite, undirected network A in relation to a given continuous-time Markov process on the network. If we assume that there are independent, identical homogenous Poisson processes defined on each node of the network, such that the walkers jump at a constant rate from

each node, the corresponding continuous-time normalized Laplacian dynamics is

$$\dot{p_i} = \sum_j \frac{1}{k_j} A_{ij} p_j - p_i. \tag{3.7}$$

The dynamics above have a steady state given by $p_j^* = k_j/2w$. Using the operator $L_{ij} = A_{ij}/k_j - \delta_{ij}$ (where δ_{ij} is a simplified way of writing out the Kronecker delta), they calculated stability using the probability that a random walker remains in the same community after time t discounting the probability of such an event occurring by chance, at stationarity

$$R(t) = \sum_{ij} \left[\left(e^{tL}\right)_{ij} p_j^* - p_i^* p_j^* \right] \delta_{c_i c_j}, \tag{3.8}$$

where the contribution from an independence assumption of the walkers appears in the second term in brackets. Expanding the matrix exponential in Eq. (3.8) to first order in t, so that $\left(e^{tL}\right)_{ij} \approx \delta_{ij} + tL_{ij}$, yields the quality function [155]

$$Q(t) = \frac{1}{2z} \sum_{ij} \left[tA_{ij} - \frac{k_i k_j}{2z} \right] \delta_{c_i c_j}, \tag{3.9}$$

(where $2z$ is the total edge weight in the network,) up to d_{ij} factors that always contribute to the sum and thus do not contribute to identifying partitions that optimize $Q(t)$. This reduces to modularity with the standard Newman-Girvan null model (see Section 3.3.1) for $t = 1$ [155].

Setting $\tau = 1/t$ provides a direct interpretation of the resolution parameter $\gamma = 1/t$ when the quality function is written in the usual form: $Q = \frac{1}{2z} \sum_{ij} (A_{ij} - \gamma \frac{k_i k_j}{2z}) \delta_{c_i c_j}$ without changing the optimization.

Multislice modularity Mucha et al. [194] used this kind of Laplacian dynamics approach to recover the previously proposed null models for bipartite [19], directed [11, 165], and signed networks [107], and to derive null models for multilayer networks. They considered three key generalizations:

1. They replaced the independent contribution $p_i^* p_j^*$ in Eq. (3.8) by a conditional independent contribution $\rho_{i|j} p_j^*$, where $\rho_{i|j}$ is the conditional probability at stationarity of a walker moving from node i to node j along a specific type of edge.

2. They generalized the Laplacian dynamics to consider multiple types of connections, e.g., directed or signed edges.

3. They re-interpreted the stability under Laplacian dynamics to permit different spreading weights on different types of edges, which allowed them to obtain separate resolution parameters for different edge types, e.g., positive and negative edges.

They then combined these insights to derive a modularity formulation for multilayer networks. Recall that each network layer s consists of adjacency terms $\overline{A}_{ijs}$ between nodes i and j, and interlayer couplings denoted by U_{jrs} connect node j in layer r to its counterpart in layer s. Noting the strengths of each node individually in each layer by $\bar{k}_{js} = \sum_i \overline{A}_{ijs}$ and across layers by $u_{js} = \sum_r U_{jsr}$, Mucha et al. defined the multilayer strength of node-layer (j,s) by $\bar{\kappa}_{js} = \bar{k}_{js} + u_{js}$. They used the continuous-time Laplacian dynamics given by

$$\dot{p}_{is} = \sum_{jr} \frac{\left(\overline{A}_{ijs}\delta_{sr} + \delta_{ij}U_{isr}\right) p_{jr}}{\bar{\kappa}_{jr}} - p_{is}, \tag{3.10}$$

which respects the intralayer nature of $\overline{A}_{ijs}$ and the interlayer couplings encoded by U_{isr}. This dynamics has the steady-state probability distribution $p^*_{jr} = \bar{\kappa}_{jr}/(2\zeta)$, where $2\zeta = \sum_{jr} \bar{\kappa}_{jr}$ is the total edge weight in the multislice network. Then, the probability of a walker moving from node i in layer s to node j in layer r at stationarity conditional on whether the multilayer structure allows one to step from (j,r) to (i,s) is

$$\rho_{is|jr}p^*_{jr} = \left(\frac{\bar{k}_{is}}{2\bar{z}_s}\frac{\bar{k}_{jr}}{\bar{\kappa}_{jr}}\delta_{sr} + \frac{U_{jsr}}{u_{jr}}\frac{u_{jr}}{\bar{\kappa}_{jr}}\delta_{ij}\right)\frac{\bar{\kappa}_{jr}}{2\zeta}\,, \tag{3.11}$$

where $\bar{z}_s = \sum_j \bar{k}_{js}$. The first term in brackets in Eq. (3.11) describes the conditional probability of motion along intralayer edges, including the probability $\bar{k}_{jr}/\bar{\kappa}_{jr}$ of using an intralayer edge when leaving (j,r) and a restriction of motion to the given layer (δ_{sr}). The second term in brackets describes the conditional probability of motion between two layers and the resulting restriction to the same node (δ_{ij}).

Subtracting the conditional joint probability in Eq. (3.11) from the linear (in time) approximation of the exponential describing the Laplacian dynamics as shown for the static NG case above, Mucha et al. obtained a multilayer generalization of modularity with the Newman-Girvan null model (see Section 3.3.1):

$$\overline{Q}_{NG} = \frac{1}{2\zeta}\sum_{ijsr}\left[\left(\overline{A}_{ijs} - \gamma_s\frac{\bar{k}_{is}\bar{k}_{js}}{2\bar{z}_s}\right)\delta_{sr} + \delta_{ij}U_{jsr}\right]\delta_{\bar{c}_{is}\bar{c}_{jr}}\,, \tag{3.12}$$

where $2\zeta = \sum_{ijs}\overline{A}_{ijs}$ (the total weight in the multislice network, as before), and $\bar{c}_{is}$ denotes the community containing node i in layer s. Note that there can be different resolutions γ_s in each layer. The resolution parameter for the interlayer couplings was absorbed into the elements of U_{jsr}. In this thesis, for simplicity, we presume U_{jsr} to take binary values $\{0,\omega\}$, where $\omega \in [0,\infty)$, indicating the absence (0) or presence ($\omega > 0$) of interlayer edges. When $\omega = 0$, this reduces to a separate modularity optimization in each layer. At the other extreme, when ω becomes sufficiently large, the quality-optimizing partitions force the community assignment of a node to remain the same across all layers in which that node appears.

As for static networks, the definition of multilayer modularity in Eq. (3.14) can be generalized to allow any null model $\overline{P}_{ijs}$:

$$\overline{Q} = \frac{1}{2\zeta} \sum_{ijsr} \left[\left(\overline{A}_{ijs} - \gamma_s \overline{P}_{ijs} \right) \delta_{sr} + \delta_{ij} U_{jsr} \right] \delta_{\bar{c}_{is} \bar{c}_{jr}}. \tag{3.13}$$

However, the spatial and correlation null models that we use in this thesis (defined in Section 3.3) lack the Laplacian dynamics motivation that the multilayer NG null model possesses.

In this thesis, we use a formula of multilayer modularity for a weighted network $\overline{W}$ and we use the same resolution parameter γ for all layers for simplicity. We thus note multilayer modularity as

$$\overline{Q} = \frac{1}{2\overline{w}} \sum_{ijsr} \left[\left(\overline{W}_{ijs} - \gamma \overline{P}_{ijs} \right) \delta_{sr} + \delta_{ij} U_{jsr} \right] \delta_{\bar{c}_{is} \bar{c}_{jr}} , \tag{3.14}$$

where $2\overline{w} = \sum_{ijs} \overline{W}_{ijs}$ and $\overline{P}_{ijs}$ is the ij-th element of the null-model tensor in layer s [194].

3.2.2 Modularity optimization algorithms

To detect communities via modularity maximization, one searches over possible network partitions to try to find the partition with the highest modularity score Q. However, exhaustive search over all possible partitions is computationally intractable, as the number of possible partitions grows super-exponentially with the number of nodes [39]. Thus practical algorithms invariably use approximate optimization methods, and different approaches offer different balances between speed and accuracy [91, 219].

One of the most common methods used to optimize modularity are greedy algorithms, which offer a balance between high speed and a reasonable accuracy [91, 219]. Greedy algorithms perform the operation that maximizes the objective quantity without considering other factors, and thus they get trapped in local optima. The "Louvain" algorithm [33] is a particularly popular, locally greedy modularity-increasing optimization process, which is considerably faster than other related approaches and therefore is a very popular choice in practice, and we use it in this thesis as a most balanced approach with fast speed (needed for multislice networks) and reasonable accuracy.

In contrast, simulated annealing [116] can achieve a very good approximation of the maximum of Q, but it is slow [91]. Simulated annealing is a probabilistic procedure for global optimization that explores the space of possible states looking for the global optimum of modularity Q. Transitions from one state to another occur with probability 1 if Q increases after the change, otherwise with a probability $\exp(\beta \Delta Q)$, where ΔQ is the decrease of modularity after the change, and β is an index of stochastic noise which increases after each iteration. The noise term reduces the risk of getting trapped in local optima. The algorithm considers both local and global moves. At some stage, the system converges

to a stable state, which can be an arbitrarily good approximation of the maximum of Q, depending on how many states were explored and how slowly β is varied [91].

Extremal optimization achieves an accuracy comparable with simulated annealing, but with a substantial gain in speed [91]. Extremal optimization is based on the optimization of local variables, expressing the contribution of each node of the system to the global value of modularity [81]. Each node is assigned a fitness measure by dividing its local modularity (the value of the corresponding term in in global modularity) by its degree. The method starts from a random partition of the network into two equal-sized communities; at each iteration, the node with the lowest fitness is shifted to the other community, and the local fitness of nodes is recalculated. The process continues until no more improvements in modularity Q are possible, after which each community is treated as its own network and the process is repeated, as long as Q increases.

Spectral optimization is faster than extremal optimization and it is also slightly more accurate, especially for large graphs [91]. Spectral optimization methods decompose the network based on the eigenvectors of the modularity matrix $B_{ij} = W_{ij} - \gamma P_{ij}$ [203] . In its simplest form, spectral optimization groups the nodes into two communities according to the signs of the components of v_1, the eigenvector of B with the largest (positive) eigenvalue. This procedure is repeated for each of the communities separately, and the number of communities increases as long as modularity increases. The method can also be extended to use more of the positive eigenvectors of the modularity matrix, and thus group nodes into a larger number of clusters at once [91].

The methods presented above are the most common methods used to optimize modularity in practical applications. As modularity optimization is an active research field, a multitude of other techniques for modularity optimization have also been proposed, and many of them are reviewed in Ref. [91].

3.2.2.1 Community detection techniques used in this thesis

In the present thesis, we optimize modularity using the GenLouvain algorithm, a two-phase iterative procedure similar to the "Louvain method" [33]. We choose it due to its relatively good results, fast speed and the ability to substitute null models. The procedure differs from the Louvain method in that rather than using the adjacency matrix W, it works with the modularity matrix B with elements $B_{ij} = W_{ij} - \gamma P_{ij}$ for static networks, and with the modularity tensor with elements $\overline{B}_{ijs} = \overline{W}_{ijs} - \gamma \overline{P}_{ijs}$ for multilayer networks [139]. This GenLouvain algorithm is very appropriate for this work, as the method is general enough to allow us to use any null model that has the form P_{ij} (or $\overline{P}_{ijs}$ for multilayer networks).

The GenLouvain algorithm consists of two phases, which are repeated iteratively. It begins by placing each node in it own community. During phase 1, it considers the nodes one by one (in some order), and places each node in the community (including its own)

that results in the largest increase of modularity. It reconsiders moving single nodes one by one (in some order), until there are no more changes that increase modularity. At this point it enters a second phase: it creates a "meta-network" in which communities i and j are represented by nodes i and j, and the edge weight B_{ij} in the new modularity matrix is the sum of the edge weights between all pairs of nodes in communities i and j in the node-level modularity matrix. In particular, this results in the addition of self-loops onto the nodes with the weights equal to the total weight of edges inside the community represented in each node. The algorithm then iteratively repeats the process of single-node community reassignments and metanetwork creation until no more merging or reassignment can increase modularity further.

A key point of GenLouvain modularity optimization is that increasing modularity in the network of meta-communities also increases it on the original network. The GenLouvain method is very fast [91, 158] because the change in modularity in phase 1 can be computed very quickly without computing the modularity for the whole network, as each node only contributes to a small number of terms in the sum in modularity (Eq. 3.6 and 3.14). This is an important consideration in multislice networks, for which the total number of nodes is the number of nodes in each layer multiplied by the number of layers, and it grows very quickly.

We use a nondeterministic implementation of the GenLouvain algorithm, which randomizes the node order at the start of each iteration of phase 1 in order to avoid making the final partition a function of the node order. Further, the original version of the Louvain algorithm and of our code performs the community reassignment that maximally increases modularity every time. However, it exhibits a sudden change in behavior when the strength of the interlayer coupling ω approaches the maximum value of the intraslice modularity matrices [26]. We mitigate this by implementing random node moves, in which every time a node is considered to be moved, the target community is chosen uniformly at random from all moves that increase modularity. In addition to mitigating the undesirable behavior, this allows one to explore a larger sample of the modularity landscape.

In order to achieve robust results across multiple stochastic runs of the GenLouvain algorithm, to detect persistent features of the communities across experimental repeats, or to detect communities on the level of physical nodes in the multislice network, we use several slight modifications to the simple community detection pipeline presented above.

3.2.3 Details of community detection used in this thesis

Because of the stochasticity of the version of the GenLouvain algorithm we use [139] and the fact that modularity optimization tends to produce many nearly-optimal partitions [108], for each of our numerical experiments, we apply the computational heuristic 50 times to obtain a *consensus community structure* [160]. We do this by constructing an

association matrix A^{rep} (where the entries A_{ij}^{rep} represent the fraction of times that the pair of nodes i and j are classified together in the 50 partitions; for multislice networks, we work on the level of multilayer nodes). We then perform community detection on A^{rep} using the uniform null model $P_{ij}^{U} = 2w/[N(N-1)]$ [23], which we describe in Section 3.3.2. This null model emphasizes the block-diagonal structure present in the association matrix [26], resulting in detection of a community structure that includes community pairings that are repeatedly detected by the original algorithm and is thus considered representative.

For multislice networks, the procedure above yields an assignment of each multilayer node (i.e., node-layer tuple) to a community. Sometimes we are also interested in a partition of the original entities (i.e., physical nodes). For example, to compare the result of algorithmic community detection to known partitions, we group physical nodes (i.e., regions) by climate, population, administrative region, etc.

The simplest method to detect province-level community structure is to perform community detection on a fully-aggregated network, i.e., one static network generated from the whole time series. We will investigate the performance of this method against our *province-level community detection* from a multislice network. The province-level community detection proceeds in two rounds: (1) we detect communities in the multislice network using any null model of choice; (2) we use this partition to construct an $N \times N$ province-level association matrix (a matrix A^{province} where entries A_{ij}^{province} represent the fraction of times that the pair of nodes i and j are classified together in all layers), and we detect province-level communities by maximizing modularity on the association matrix using the uniform null model. As stated above, this null model emphasizes the block-diagonal structure present in the association matrix, allowing us to detect the most temporally persistent community structure in the association matrix — one that is detected repeatedly in multiple layers.

3.3 Null models for modularity maximization

The choice of null model is vital for the detection of communities using modularity maximization [23, 108]. In this section, we review the standard (i.e.,"Newman-Girvan") null model and other possible choices that are relevant to spatial networks and correlation networks. We then introduce a novel spatial null model based on the radiation model of human mobility [245].

3.3.1 The Newman-Girvan null model

The standard null model for modularity optimization is the Newman-Girvan (NG) null model, in which one randomizes edge weights, such that the expected strength distribution is preserved [199, 202]. It is thus related to the classical configuration model [204]. The NG null model has become very popular due to its simplicity and effectiveness, and it has been

derived systematically through the consideration of Laplacian dynamics on networks [155], as discussed in Section 3.2.1.

For static networks, the NG null model is given by

$$P_{ij}^{\mathrm{NG}} = \frac{k_i k_j}{2w}\,, \tag{3.15}$$

where $k_i = \sum_j W_{ij}$ is the strength of node i and $2w = \sum_{ij} W_{ij}$ is the total edge weight in the network.

For multislice networks, the NG null model is [194]

$$\overline{P}_{ijs}^{\mathrm{NG}} = \frac{\bar{k}_{is}\bar{k}_{js}}{2\overline{w}_s}\,, \tag{3.16}$$

where $\bar{k}_{is} = \sum_j \overline{W}_{ijs}$ is the intralayer strength of node i in layer s and $2\overline{w}_s = \sum_{ij} \overline{W}_{ijs}$.

Despite its popularity and demonstrated effectiveness in many situations, the NG null model is naïve in the sense that it does not incorporate problem-specific information (such as spatial embeddedness). It only takes node strengths into account, which may not be suitable for all applications — what one considers to be connected "at random" depends fundamentally on the research question of interest. It is often important to incorporate additional (domain-specific or even problem-specific) information when they are known.

3.3.2 Uniform null model

The uniform null model is another simple null model that has been used, mainly in the context of weighted networks [23, 26]. It is defined as

$$P_{ij}^{U} = a = 2w/[\hat{N}(\hat{N}-1)]\,, \tag{3.17}$$

where $\hat{N}$ is the number of nodes that "exist" in the network, i.e., have non-zero entries. This null model represents the expected edge weight in the weighted network, discounting self loops (otherwise the denominator of Eq. (3.17) would read $2\hat{N}$ rather than $\hat{N}(\hat{N}-1)$). Modularity maximization with a uniform null model is equivalent to a block-diagonalization of the adjacency matrix with a penalty on the size of communities. As one increases the resolution parameter γ, one favors smaller sets of nodes with stronger internal connectivity. This null model simply emphasizes the block-diagonal structure present in the adjacency matrix [26]. It is thus the most suitable for detecting consensus community structure from an association matrix describing the co-classification of nodes in several realizations of community detection on the same network, or the co-classification of physical nodes in layers of a multislice network, as described in Section 3.2.3.

3.3.3 Spatial null models: Gravity model

In many spatially embedded networks, proximity has a strong effect on connections between nodes, as (all else held equal) neighboring nodes are more likely to be connected to each other (and their connections are likely to have to have higher weights) than nodes that are far away [21,82]. Moreover, proximity can mask other underlying influences. Consequently, incorporating the expected influence of proximity on edge weights into null models for community detection (via modularity maximization) should make it possible to discover new and important types of structures. Furthermore, for multilayer networks with elements of spatial and temporal organization, it is conceivable that using a spatial null model might allow one to uncover otherwise undetectable elements of the temporal structure.

Expert et al. [82] proposed a spatial null model that was inspired by the "gravity model" that we described in the context of modeling epidemics in Section 2.4 and Eq. (2.2). Recall that a gravity model assumes that the interaction between two locations grows with their importance (e.g., population), and that it decays with distance. Its most general form is defined as

$$G_{ij} = I_i^{\alpha} I_j^{\beta} f(d_{ij}) \,, \tag{3.18}$$

where I_i is the importance of location i (e.g., its population) and $f(d_{ij})$ describes the effect of distance on node interactions.

The simplest form of a gravity-like interaction in Eq. (3.18), with $\alpha = \beta = 1$, was incorporated by Expert et al. into a *gravity null model* [82] to give

$$P_{ij}^{\text{grav}} = I_i I_j f(d_{ij}) \,, \tag{3.19}$$

where I_i is the importance of node i, and the "deterrence function" $f(d)$ is estimated from data:

$$f(d) = \frac{\sum_{\{k,l|d_{kl}=d\}} W_{kl}}{\sum_{\{k,l|d_{kl}=d\}} (I_k I_l)} \,. \tag{3.20}$$

Expert et al. [82] used the population $I_i = n_i$ as the node importance. After briefly experimenting with variations such as using population density or the logarithm of the population (i.e., $\log(n_i)$) and observing no significant differences in the empirical results of community detection on the disease-correlation network generated from the dengue fever data set, we will follow their lead.

In order to achieve a reliable weighted average in the deterrence function $f(d)$ in Eq. (3.20), we require a minimum number of observations for each distance for which we calculate the expected weight of the null model. In our data sets, distances are often unique, so we choose to bin the data by their distance.

We bin the distances into equal-distance bins (e.g., every b km). For the disease-correlation networks, we study the effects of bin size on algorithmic community structure following the methods used by Expert et al. [82] by observing the effect of bin size on the

deterrence function as shown in Section A.4. We also observe the effect of bin size on the similarity of partitions to each other, as measured by normalized mutual information, but due to the similarity of partitions we do not show these results here. We select one bin size for each application, based on the smoothness of the deterrence function and the number of nodes in each bin, which we (arbitrarily) choose to be at least 5. Alternative binning methods could include binning into equal-sized bins (e.g., each bin containing at least b elements).

We only use the distances and the correlation values for the nodes that are present in each layer/network for calculating the bins and the null model. Combining Eq. (3.19) and (3.20) with Ψ_i (defined in Eq. 3.1) — an indication of whether a node exists in a given static network — we write the gravity null model as

$$P_{ij}^{\text{grav}} = \Psi_i \Psi_j I_i I_j \frac{\sum_{\{k,l|d_{kl}=d_{ij}\}} W_{kl} \Psi_k \Psi_l}{\sum_{\{k,l|d_{kl}=d_{ij}\}} (I_k I_l \Psi_k \Psi_l)} \,. \tag{3.21}$$

In the present thesis, we generalize the gravity null model to a multislice setting by calculating a separate gravity null model for each layer s, where Ψ_i^s is an indication of whether a node exists in layer s. The resulting multislice gravity null model is

$$\overline{P}_{ijs}^{\text{grav}} = \Psi_i^s \Psi_j^s I_i I_j \frac{\sum_{\{k,l|d_{kl}=d_{ij}\}} \overline{W}_{kls} \Psi_k^s \Psi_l^s}{\sum_{\{k,l|d_{kl}=d_{ij}\}} (I_k I_l \Psi_k^s \Psi_l^s)} \,, \tag{3.22}$$

where we have assumed that the population stays constant over time. If one has reliable information about the population changes in time, then one can incorporate such information into the null model (3.22) by substituting I_i with an analogous quantity I_{is} that depends both on the node i and on the layer s.

3.3.4 Spatial null models: Other models

The incorporation of spatial information into null models for community detection is an important problem, and several other ideas have been proposed recently. For example, Cerina et al. [47] focused on disentangling the correlation between node attributes and space. They used an exponential decay: $f(d_{ij}) = e^{-d_{ij}/\bar{d}}$, where $\bar{d}$ is the mean distance between nodes in a network. Shakarian et al. [242] focused on finding geographically-disperse communities. They introduced a decay constant θ such that $f(d_{ij}) = e^{-d_{ij}/\theta^2}$. Another recently-proposed null model was used to attempt to find geographically-proximate communities [119].

We do not know the exact nature of the influence of spatial distance on interactions in the disease data, so we chose to only use null models that can be directly measured from data rather than fitted using an arbitrarily chosen functional dependence.

3.3.5 A novel spatial null model: Radiation model

In this section we develop a novel null model for community detection via modularity maximization. Our null model is based on the "radiation model" [245], which was recently proposed as an alternative to gravity models for studying population mobility (see Section 2.4). Since a gravity model has successfully been used as part of spatial null models for community detection, and both gravity and radiation models have been used for modelling disease spread (as reviewed in Section 2.4), we want to investigate the effect of using a null model for community detection based on the radiation model for mobility and constructed in a similar manner to the gravity null model for community detection (defined in Section 3.3.3).

Recall that the radiation model flux from i to j with populations n_i and n_j located at distance d_{ij} has the form

$$T_{ij} = T_i \frac{n_i n_j}{(n_i + r_{ij})(n_i + n_j + r_{ij})}, \qquad \text{(2.3 revisited)}$$

where $T_i \equiv \sum_{j \neq i} T_{ij}$ is the number of commuters residing in location i and $r_{ij} = q_{ij} - (n_i + n_j)$ is the population residing in the circle centered in i with radius d_{ij}, minus the populations at the origin i and destination j.

We propose a novel null model for community detection based on the original formulation of the radiation model in Eq. (2.3) [245]. We use a similar formulation to the gravity null model in Eq. (3.21) in order to incorporate both the expected distance-dependent flux and the actual network structure. We keep the null model as simple as possible in order to focus on the effect of incorporating space into it. To avoid creating a directed null model, we use the mean flux $\hat{T}_{ij} = (T_{ij} + T_{ji})/2$ between nodes i and j to construct the null model. Further, a directed null model would not be appropriate for the undirected networks that we generate using Pearson correlations, but if we decided to use a lagged correlation and therefore create a directed network, a directed null model might become more appropriate.

We thereby construct the *radiation null model*:

$$P_{ij}^{\text{rad}} = \Psi_i \Psi_j \hat{T}_{ij} \frac{\sum_{\{k,l|d_{kl}=d_{ij}\}} W_{kl} \Psi_k \Psi_l}{\sum_{\{k,l|d_{kl}=d_{ij}\}} \hat{T}_{kl} \Psi_k \Psi_l}\,. \qquad (3.23)$$

Following the approach in [245], we assume that commuters are distributed uniformly across the nodes. We can then simplify Eq.(2.3) by substituting $T_i = T_f n_i$, where T_f is the fraction of commuters in the whole population. Because the quantity T_f is present in both the numerator and denominator of Eq. (3.23), we can now cancel it out. However, when commuting data are available, it is desirable to incorporate thems to improve the radiation null model.

We also extend the radiation null model to a multislice setting in an analogous manner to the gravity null model. The multislice radiation null model is

$$\overline{P}^{\text{rad}}_{ijs} = \Psi^s_i \Psi^s_j \hat{T}_{ij} \frac{\sum_{\{k,l|d_{kl}=d_{ij}\}} \overline{W}_{kls} \Psi^s_k \Psi^s_l}{\sum_{\{k,l|d_{kl}=d_{ij}\}} \hat{T}_{kl} \Psi^s_k \Psi^s_l} \,. \tag{3.24}$$

Again, one can incorporate temporal data about population sizes and thereby replace T_{ij} with T_{ijs} to improve the null model.

3.3.6 Correlation null model

Recently, MacMahon et al. [171] proposed a new null model specifically for community detection on correlation networks created from time series. They used ideas from random matrix theory (RMT) [180] to generate a null model that represents the "random" component of a correlation matrix and can take into account the single most strongly influential factor on the correlation structure. In the context of the financial systems that they studied, this is called a "market mode".

Because of the special structure of correlation matrices, modularity maximization using the standard NG null model gives more importance to pairs of nodes i and j whose direct correlation is larger than the product of the correlations of the time series for each node ($D_i = D_i(1), D_i(2), \ldots, D_t(T)$) with a common signal $D_{\text{tot}} = \{D_{\text{tot}}(1), D_{tot}(2), \ldots, D_{\text{tot}}(T)\}$, which is sum of all the time series: $D_{\text{tot}}(t) \equiv \sum_{i=1}^{N} D_i(t)$. In contrast, as discussed in Ref. [171], the correlation null model uses ideas from RMT to detect communities of nodes that are more connected than expected under the null hypothesis that all time series are independent of each other.

Based on RMT, for a given correlation matrix C that is constructed from N time series of length T each (with $T/N > 1$), any eigenvalues that are smaller than the maximum eigenvalue predicted for a correlation matrix created from the same number of entirely random time series [$\lambda_+ = (1 + \sqrt{N/T})^2$] are deemed to be due to noise [180]. Additionally, for many empirical correlation matrices, the largest eigenvalue λ_m is much larger than the others, and its corresponding eigenvector has all positive entries [171]. This has been interpreted as a common factor — called the "market mode" — that influences all of the time series [249].

Recall from Section 3.1.1 that the correlation values for the nodes which do not experience disease in a given time window are not defined. For this reason, we construct the correlation null model only using the submatrix corresponding to the nodes that exist in a given layer/network. For the nodes which do not experience disease in a given time window, we define the null model to be 0 (in a manner similar to the network construction). This also allows us to use the number $\hat{N}$ of nodes that exist in a network rather than the overall number of time series N, so for a network corresponding to a time window of length Δ, $\lambda_+ = (1 + \sqrt{\hat{N}/\Delta})^2$.

One can thus decompose a correlation matrix C into parts that correspond to the "market mode" and the noise as follows:

$$C = C^{(r)} + C^{(g)} + C^{(m)}\,, \tag{3.25}$$

where $C^{(r)}$ is the "random" component of the matrix, $C^{(m)}$ is the "market mode", and $C^{(g)}$ is the "group mode" — a component that embodies the meaningful correlations between time series. Using bra-ket notation following Ref. [171], we thus write

$$C^{(m)} \equiv \lambda_m |v_m\rangle\langle v_m| \tag{3.26}$$

and

$$C^{(r)} \equiv \sum_{i:\lambda_i \leq \lambda_+} \lambda_i |v_i\rangle\langle v_i|\,, \tag{3.27}$$

where λ_i and v_i are an eigenvalue and its corresponding eigenvector and λ_m is the maximum eigenvalue of the correlation matrix C.

For each of the layers s in the multilayer setting, we write

$$C_s^{(m)} = \lambda_m^s |v_m^s\rangle\langle v_m^s|\,, \tag{3.28}$$

and

$$C_s^{(r)} = \sum_{\{i:\lambda_i^s \leq \lambda_+^s\}} \lambda_i^s |v_i^s\rangle\langle v_i^s|\,, \tag{3.29}$$

where λ_i^s and v_i^s are an eigenvalue and its corresponding eigenvector for layer s.

We can construct a correlation null model either by removing both the random component of the matrix and the influence of the market mode or by only removing the random component. For each layer/network s we define a mapping function $\Phi_s(x)$ such that $\Phi_s(x) = y$ maps the physical node ID (e.g., province ID $x \in \{1, 2, \ldots, N\}$) to the layer ordering $y \in \{1, 2, \ldots, \hat{N}\}$; for static networks we sometimes omit the s subscript for simplicity.

The correlation null model for static networks with only random component removed is

$$P_{ij}^{\mathrm{cR}} = \begin{cases} \gamma [C^{(r)}]_{\Phi(i)\Phi(j)} & \text{if } \Psi_j \Psi_j = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.30}$$

where γ is the usual resolution parameter. For each of the layers s in the multilayer setting, we write

$$P_{ijs}^{\mathrm{cR}} = \begin{cases} \gamma [C_s^{(r)}]_{\Phi_s(i)\Phi_s(j)} & \text{if } \Psi_j^{(s)} \Psi_j^{(s)} = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.31}$$

The correlation null model with both the market mode and the random component of the network removed is

$$P_{ij}^{\mathrm{cM}} = \begin{cases} \gamma \left([C^{(r)} + C^{(m)}]_{\Phi(i)\Phi(j)} \right) & \text{if } \Psi_j \Psi_j = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.32}$$

where γ is the resolution parameter. For each of the layers s in the multilayer setting, we write

$$P_{ijs}^{\mathrm{cM}} = \begin{cases} \gamma \left([C_s^{(r)} + C_s^{(m)}]_{\Phi_s(i)\Phi_s(j)} \right) & \text{if } \Psi_j^{(s)} \Psi_j^{(s)} = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.33}$$

In this thesis, we use the correlation null model with only the random component removed (P^{cR}): the "market mode", which might represent some of the seasonal patterns in disease incidence, could contain useful information here. Further, many of the correlation matrices that we deal with in this thesis have only a small number of non-random eigenvalues, so removing the market mode might prevent us form performing community detection using the correlation null model at all. If presented with a suitable data set, one could verify our intuition about the meaning of market mode in this setting, and further examine the effects of community detection after removing both the market mode and the random component of the matrix.

3.4 Measures to study network partitions

The results of community detection are often too large and complicated to effectively study by naked eye, and visualization can be misleading. Hence, in addition to visualization, we use various summary statistics to examine the results of community detection. We also compare the algorithmic community structure versus manual partitions — e.g., partitions by climate or administrative region for disease networks, or known planted partitions for benchmarks and partitions of time series resulting from disease models with known properties. The choice of summary statistics and comparison methods depends on the particular network type and question, and we discuss some methods in the next several sections.

3.4.1 Partition visualizations

The typical way to start examining the results of community detection is to examine visualizations. One can represent nodes with dots, draw edges that connect relevant nodes, and use node color to signify community membership (see Fig. 3.4). There are various methods to define the arrangement of nodes for best visualization; in particular, spatially embedded networks in which nodes are assigned locations can be represented on a map, as seen in Fig. 3.5(a)-(b); we do not show the edges in this case. For Peru, we use MATLAB to plot maps using province boundary data downloaded from the Gadm website [1]. For the other data sets, we use the `plot_google_map` function from MATLAB file exchange website to plot a scatter plot over a Google map [18].

We visualize the temporal evolution of community structure in multislice networks on a two-dimensional plot with nodes on the vertical axis and layers on the horizontal axis [see Fig. 3.5(c)]. We can also select layers to visualize on a map in the same manner as for static networks.

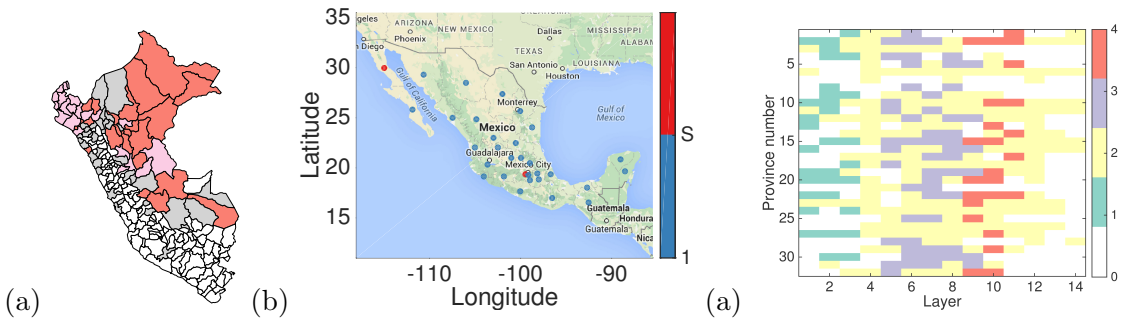


Figure 3.5: Visualizing the results of community detection. (a) Community structure of the dengue fever network in Peru on a map in which we color provinces according to their community assignment. White provinces are ones in which our data do not include any reported cases of dengue fever in the indicated time window. (b) Community structure of the H1N1 influenza network in Mexico on a map in which we color provinces according to their community assignment. We omit provinces with no disease cases. (c) Multislice community structure of the H1N1 influenza network in Mexico with time on the horizontal axis and node number on the vertical axis. We sort nodes by their location (north to south) and we color nodes by community membership, with white (community 0) representing the nodes with no disease cases at a given time. Community number is indicated on the colorbar.

3.4.2 Partition summaries

For both static and multislice disease networks, we study several properties of community structure:

- number of communities
- number of singleton communities (i.e., communities with 1 node)
- size of the largest community
- "intra-community distance" — the mean distance between a pair of nodes in each community
- "community spread" — the mean intra-community distance over all communities
- the mean population size for nodes in a community
- climate composition of communities
- position of large cities and other known important provinces
- patterns of the disease time series for nodes that are assigned to each community
- "first infection times" — the times when the disease first reached nodes that are assigned to each community

For multislice networks, one can study these properties per layer (which allows us to visualize the evolution of such properties over time) or for the whole network (in which case we take the mean across all layers), or one can seek a province-level partition (on the level of physical nodes, as described in Section 3.2.2.1). Further, in a multislice context, we need to

distinguish between singletons (where a node is placed entirely on its own), and "temporal singletons" (where a node is only connected to other multilayer nodes that correspond to the same physical node across time).

Additionally, for multislice networks it is desirable to search for "change points" – time points when the community structure has changed. We do this using z-Rand scores which we define in Section 3.4.3.

3.4.3 Quantifying partition similarity: z-Rand scores

In this section, we present the z-score of the Rand coefficient, so-called "z-Rand score" [268]. The Rand coefficient is a commonly used pair-counting similarity measure that is good at detecting similarities in coarse structure [268, 269] but is less sensitive to minor changes (such as one node changing community assignment). Thus, we use it to quantify the general agreement of an algorithmic community structure with a planted or manual partition.

The Rand coefficient is defined as

$$R = (w_{11} + w_{00})/M\,, \tag{3.34}$$

where w_{11} is the number of pairs classified together in the two partitions, w_{00} is the number of pairs classified differently in the two partitions, and M is the total number of pairs. Although it is easy to interpret, the Rand coefficient suffers from a skew towards higher values when there are more communities. Hence, the value of R that one obtains for one pair of partitions is not, in general, directly comparable with the value for another pair.

Traud et al. [268] also identified similar problems with other (more complicated) pair-counting methods. To overcome this issue, they compared the value calculated for the measure versus what would be expected at random. This yields a z-score of the Rand coefficient (which we call a *z-Rand score*). Large z-Rand scores point towards a statistically significant value for R and hence significant agreement between the two partitions. For classifying partitions as similar (or not) to a manual partition, we use the standard cutoff of +/-1.96 as statistically significant with a 95% confidence level to guide our assessment of partitions using the z-Rand score tests. We then focus on the highest-scoring partitions for further study.

The formula for the calculation of the z-Rand score is given as follows. Let $w := w_{11}$, and let M_1 and M_2 respectively denote the number of pairs classified the same way in the first and second partitions. The z-Rand score is

$$z_R = \frac{1}{\sigma_w}\left(w - \frac{M_1 M_2}{M}\right), \tag{3.35}$$

where the standard deviation σ_w is expressed in terms of M_1, M_2, M, and n [268].

We use z_R to assess the similarity of the community structures for the disease data versus spatial partitions such as climate or administrative divisions. For the disease data, we do

not possess ground-truth partitions, so we seek to evaluate broad organizational similarities in the algorithmic and manual partitions rather than attempting to conduct a fine-grained evaluation of community structure versus a known ground truth. We thereby aim to inform our understanding of the structural influences on spatiotemporal patterns of disease spread.

In the manual "climate partitions" for the Peruvian data sets of rubella and dengue fever, we group nodes according to the topography of their associated provinces — jungle, coastal, and mountainous provinces — and then subsequently divide the coastal and mountainous communities into northern, central, and southern provinces [see Fig. 3.6(a)-(b)]. In the 19-community "administrative partition" of Peru, we assign each node to its associated administrative region [see Fig. 3.6(c)]. For the multislice networks, we compare each algorithmic partition versus a manual partition by taking the same manual partition of nodes in each layer [see Fig. 3.7(a)]. We use the term "spatial partitions" to describe partitions that yield z-Rand scores greater than 1.96 in comparison to the climate or administrative manual partitions, and we use the term "highest-scoring partitions" in relation to climate and administrative partitions to describe the network partitions that have the highest z-Rand score of all the partitions we consider in a particular comparison (across a range of parameter values, and/or across a range of time points).

We do not possess definitive climatic or administrative data for countries other than Peru, so for these studies we focus on assessing the spatial aspect of community structure in other ways (described in Section 3.4.4).

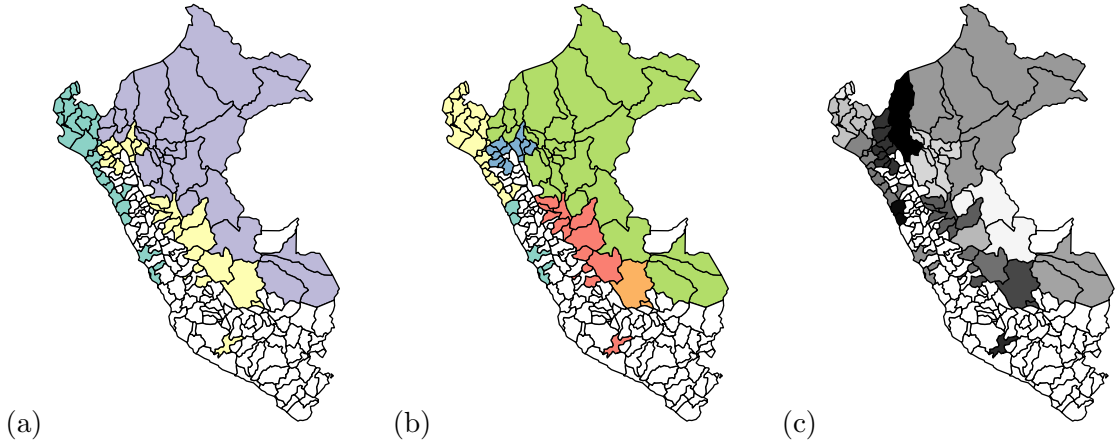


Figure 3.6: Visualization of three different topographical partitions of Peru's provinces on a map. (a) Broad climate partition into coast (green), mountains (yellow), and jungle (blue); (b) the further division of coast and mountains into northern coast, central coast, southern coast, northern mountains, central mountains, and southern mountains; and (c) the administrative partition of Peru.

For multislice networks, we can search for times when there was a large change in the structure of the network (e.g., for disease-correlation networks, a change in pair-wise synchronization of provinces in the data set). We do this by comparing algorithmic partitions to partitions that contain a planted temporal change in community structure. For these

comparisons, we group the multilayer nodes into ones that occur before or after a "critical" time point t_c [i.e., partitions into two communities: "pre-t_c" and "t_c and post-t_c" [see Fig. 3.7(b)]. We test all of the times $\tau = \left\{1, 1+\upsilon, 1+2\upsilon, \ldots, 1+\upsilon \times \left(\lfloor \frac{T}{\upsilon} \rfloor - 1\right)\right\}$ that mark the beginnings of layers in the multislice network (where υ is the distance between the starting points of adjacent layers of the network, defined in Section 3.1.1), and we report the time with the highest z-Rand score as the critical time point t_c. We call these partitions the "single critical time point" partitions. We also test for pairs of critical time points by examining all possible pairs of critical time points t_{c1} and t_{c2} in the same manner — we compare the algorithmically-detected community structure to all possible "two critical time points" manual partitions [see Fig. 3.7(c) for an example]. We use the term "temporal partitions" to describe algorithmic partitions of the disease-correlation networks that yield z-Rand scores greater than 1.96 in these comparisons, and we report the numbers of layers in which the new communities begin in our descriptions as t_c for single critical time point, and (t_{c1}, t_{c2}) for partitions with two critical time points. We also refer to the "highest-scoring" temporal partitions, meaning the partition with the highest z-Rand score against all tested temporal partitions, when searching over a range of parameters such as γ and ω.

Most of the current methods for detecting changes in temporal networks have been developed for networks in which edges represent functional connectivity rather than correlation, so they would not necessarily be applicable to our case [20,214]. For time series data, when the observed node characteristics are independent and normally distributed, methods exist to detect changes in the multivariate normal mean or covariance [122,286]. Barnett et al. [20] designed a method specific for correlation networks, that searches for a change point that maximizes the difference between the covariance matrices before and after it, which may be an alternative to our methodology.

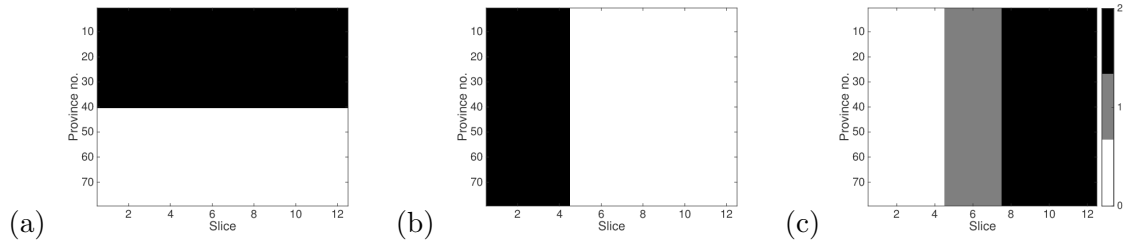


Figure 3.7: Visualization of the types of manual partitions of a multislice network: (a) a spatial partition, (b) a temporal partition with one critical time point $t_c = 5$, and (c) a temporal partition with a pair of critical time points (5 and 8). Community number is indicated on the colorbar.

3.4.4 Measures of spatial organization of communities

In this section, we present the measures that we use for assessing the extent of spatial organization in the community structures that we identify algorithmically for data sets for

which we do not possess definitive climatic or administrative data.

We first considered commonly-used measures of spatial autocorrelation that we discussed in Section 2.3 in the context of detecting disease clusters, e.g., Moran's I [189], Geary's c [98], Local Indicator of Spatial Association [10], and the Kuldorff's spatial scan statistic [152]. However, these statistics were not suitable for our purpose, as they are designed for locating spatial clusters within data in the first instance (and thus coming up with potential partitions) rather than assessing the extent of spatial clustering in predefined partitions.

3.4.4.1 Definitions of the distance and MST tests

We decided to test two simple measures to assess the extent of spatial clustering in the community structure on the dengue fever network [This is the data set with which we worked with to the greatest depth. Additionally, as we see in Sections 6.2 and 6.3, this network appears to show some spatial communities visually and when tested using the z-Rand scores, and we used this fact to verify the scores for the distance and MST tests — see Section 6.7.] The two measures are the distance test and the MST test defined below:

- The test statistic in the *distance test* is the "total intra-community distance" (the sum of the distances between the nodes that are in the same community): $T_D = \sum_{ij} d_{ij}\delta_{c_i c_j}$, where d_{ij} is the distance between nodes i and j, δ is the Kronecker delta and c_i denotes the community containing node i.

- The *MST test* requires us to create a set of weighted networks, one for each community, containing the nodes that were assigned to that community and with edges between nodes i and j weighted by d_{ij}. We then calculate the minimum spanning tree (MST) for each of these networks using Prim's algorithm [220], which is a common choice. We then calculate the "total MST cost" (the sum of total cost of the MSTs for all communities in the network partition) and use it as the test statistic.

For both of these test statistics, we quantify the extent of spatial clustering in the community structure using Monte Carlo tests with an appropriate null distribution that we construct from the original network, discussed below. The resulting p-values allow us to discern whether the observed values are unusually large or small. This calculation compares the observed value to the upper and lower tails of the null distribution:

$$p^{\text{upper}} = \frac{N_{GE} + 1}{N_{\text{runs}} + 1}, \tag{3.36}$$

$$p^{\text{lower}} = \frac{N_{LE} + 1}{N_{\text{runs}} + 1}, \tag{3.37}$$

where N_{runs} is the total number of Monte Carlo simulations (we use $10,000$ unless stated otherwise), N_{GE} is the number of simulations for which the diagnostic was greater than or equal to the observed diagnostic, and N_{LE} is the number of simulations for which the diagnostic was less than or equal to the observed diagnostic. We add 1 to the numerator and denominator because the observed value is included in the reference distribution. Unless stated otherwise, we use the p-values of 0.05 as the significance level, and we refer to partitions that score below it as "spatial".

For both of the Monte Carlo tests, the choice of a suitable null distribution for the randomization is a key question that needs to be addressed before applying the methodology. For static networks, this is relatively simple: we uniformly randomize the community assignment of all nodes while keeping the size of each of the communities constant.

If one wants to calculate partition-wide scores for whole multislice networks, the randomization issue becomes more complicated. We considered three options for obtaining a null distribution, in order of increasing complexity:

1. Randomize the location of nodes, i.e shuffle the multislice community structure using the same reordering for each layer.

2. Randomize the community assignment of nodes independently for each layer, only taking into account the nodes that experienced disease in the corresponding time period and thus are present in that layer.

3. Randomize the identities of the parts of the whole time series that correspond to each layer independently and reassemble them into a length-T time series. Then generate a multislice network and perform community detection on it using the same null model and parameter set as for the original network.

We chose to use option (2) for both tests. We discarded option (1) as it cannot distinguish the artificially high spatial scores due to localization of a disease in particular parts of the country from genuine spatial effects in communities among the nodes that do experience disease. In several of our disease data sets the disease localizes to a small collection of spatially-clustered nodes for prolonged periods of time. Using either the distance or MST test with this null distribution would consider the communities in the layers corresponding to these time periods as highly spatial, because the test would compare the communities from that spatially-constrained layer to randomly assigned same-sized communities composed of nodes sampled from the whole network. We also discarded option (3) due to the computational complexity of the generation of additional adjacency matrices and community detection for the very large number of Monte Carlo runs. The choice of option (2) is a compromise between complexity and accuracy, as we note that both options (1) and (2) do not take into account the temporal dependence of the community assignments of a physical

node in one layer on the community assignments of the same physical node in preceding and following layers that is introduced by the inter-layer connectivity ω between these two corresponding nodes. This might result in some community structures appearing artificially significant.

This randomization allows us to calculate combined "partition-wide" score that describes the level of spatial organization of the overall multislice network; this takes the sum of the test statistics for all layers as a test statistic (sum of total intra-community distances for all layers for the partition-wide distance test, and sum of the total MST costs for the partition-wide MST test).

When examining multislice networks in more detail, we also report the p-value scores for individual layers of the multislice network. We obtain them by using the same randomization in option (2) as for partition-wide scores, but then calculating a test statistic and p-value for each layer. We refer to these scores as "per-layer" scores.

Following testing on the dengue data set in Chapter 6, we chose the simple distance test as the most useful measure, which we use to study the extent of spatial clustering in multislice networks of disease data in Chapters 7 (endemic diseases) and 8 (emerging diseases) and on networks generated from our spatial model of disease spread in Chapter 9.

3.4.5 Quantifying partition similarity: Information-theoretic measures

In this section, we present several similarity measures based on information theory. These measures are relatively sensitive to small differences in partitions, such as the move of a single node from one community to another, compared to pair-counting measures such as the Rand coefficient and z-Rand scores [268]. This sensitivity makes them well-suited for judging the quality of performance of community detection on benchmark networks based on well-defined, planted partitions that represent a ground truth.

Normalized mutual information (NMI) is an information-theoretic similarity measure that is one of many normalized versions of mutual information (MI) [181]. Both MI and NMI are based on the concept of *information entropy*, which is a measure of uncertainty. MI measures the amount of information that one can predict about one random variable (which in the present thesis is a partition of a network into communities) based on another random variable. For a partition $X = \{X_1, X_2, \ldots, X_K\}$ with K communities and a partition $Y = \{Y_1, Y_2, \ldots, Y_L\}$ with L communities, MI is defined as

$$\mathrm{I}(X,Y) = \sum_{k=1}^{K}\sum_{l=1}^{L} P(k,l)\log_2\left[\frac{P(k,l)}{P(k)P(l)}\right], \tag{3.38}$$

where $P(k)$ and $P(l)$ are the marginal probabilities of observing communities k and l in partitions X and Y, respectively; and $P(k,l)$ is the joint probability of observing communities k and l simultaneously in partitions X and Y. MI takes values between 0 and

$\min\{\mathrm{H}(X), \mathrm{H}(Y)\}$, where $\mathrm{H}(X) = -\sum_{k=1}^{K} P(k) \log_2 P(k)$ is the entropy associated to the partition X.

Normalized mutual information (NMI) [255] is defined as

$$\mathrm{NMI}(X, Y) = \frac{\mathrm{I}(X, Y)}{\sqrt{(\mathrm{H}(X)\mathrm{H}(Y))}} \in [0, 1]. \tag{3.39}$$

The normalization to lie within the range $[0, 1]$ facilitates interpretation and comparison across different situations.

Normalized variation of information (NVI) is another popular information-theoretic similarity measure [151,181]. In contrast to NMI, variation of information (VI) and its normalized version are metrics in the mathematical sense. Both measures are related to mutual information. VI is defined as

$$\mathrm{VI}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y) - 2\mathrm{I}(X, Y). \tag{3.40}$$

VI takes values between 0 (if the two partitions are identical), and $\log N$. Normalized variation of information is defined [151] as

$$\mathrm{NVI}(X, Y) = 1 - \frac{VI(X, Y)}{H(X, Y)} \in [0, 1]\,, \tag{3.41}$$

where $\mathrm{H}(X, Y) = -\sum_{k=1}^{K} \sum_{l=1}^{L} P(k, l) \log_2 P(k, l)$ is the joint entropy of X and Y. See Refs. [151,181] for additional discussion of the different similarity measures based on information theory.

We use NMI for the spatial benchmarks in Chapter 4, and we obtain the same qualitative conclusions using VI. See Section 4.3.6 for selected examples. We use NMI exclusively for comparing algorithmically detected community structure with the planted structure in the synthetic time series generated from our spatial model of disease spread in Chapter 9.

3.4.6 The start-time test

For emerging diseases, we also want to test whether the algorithmic community assignment is a function of the first time that the disease is observed in each province. We assess this using a "start-time test". We only use this test for static networks covering the whole emergence of the disease — the Ebola WHO data set in Section 8.3 and the networks generated from the agent-based model synthetic time series in Chapter 9.

We calculate the standard deviation of the first infection time (which is defined as the first time point that a disease is observed at a node) for each community in a network partition. We then use the sum of these standard deviations for all communities in a network partition as a test statistic. If this sum is small in comparison to a null random distribution, then the network partition appears to be related to the start times of each community.

The choice of a suitable null distribution is important to this question. We use the same randomization as for the distance test for static networks (see Section 3.4.4.1): we uniformly randomize the community assignment of all nodes while keeping the size of each of the communities constant. As calculating all permutations of a community structure is computationally prohibitive, we use Monte Carlo sampling from the distribution to compute an estimated p-value in the same manner as for the distance and MST tests defined in Section 3.4.4.1.

3.5 A methodological "pipeline" to study disease-correlation networks

In this section, we present the "pipeline" that we use to analyze the community partitions arising from repeated community detection on the disease-correlation networks. Unless stated otherwise, we generate 50 repeats of community detection, leading to one consensus structure. We do this over parameter variation of $\gamma \in \{0.1, 0.2, \ldots, 3\}$ (and $\omega \in \{0.1, 0.2, \ldots, 3\}$ for multislice networks). This results in 30 result sets for each set of static networks (for each null model), and 900 result sets for each multislice network and null model combination. We use the descriptive statistics and measures of spatial organization that we presented in the previous sections to select a small number of partitions for detailed examination.

For static networks, we first plot the distance and MST test results for all the networks and γ parameter values. We use this to select the γ parameter value for which network partitions have p-values smaller than 0.05 in the distance and MST tests (after Bonferroni correction for multiple comparisons) the most consistently across time from all the γ parameter values that we test. We then plot the number of communities and the spatial spread of communities for that γ value, and we select one or more layers for further detailed study — usually, the layers that combine a low p-value in one or both spatial tests with an interesting partition visually, an interesting number of communities, or other features that we describe in the text. In order to examine the real-word properties of the nodes assigned into communities, we plot the time series for the nodes belonging to each community separately, and we plot a box plot of populations of all the nodes in the community, and a box plot of distances between nodes in each community (which we refer to as "intra-community distances", and we refer to the mean intra-community distance as the "community spread"). We then select one example partition for visualization in the thesis.

For multislice networks, we first select interesting (γ, ω) parameter pairs that tend to generate interesting partitions of the network for further study. We use the partition-wide distance and MST tests to select sets of parameters (γ, ω) for detailed visual examination of their respective network partitions. Unless stated otherwise, we select the parameter regimes that generate the most significantly spatial partition-wide scores (i.e., score the

lowest p-value in the distance test), and the partitions that score high in the temporal z-Rand scores. We then examine community structure of these partitions visually, and we plot the per-layer values for the number of communities, the spatial spread of communities, and the p-values for the distance and MST tests.

3.6 Alternative approaches

As reviewed in Chapter 2, there are several alternative methods that we could have used for studying the spatial spread of disease and analyzing the disease time series. In this section, we motivate the choice of using network science and community detection as our methodology.

Perhaps the most typical approach to the mathematical study of infectious diseases is to construct mechanistic disease models, based on one of the model types that we reviewed in Section 2.2 (the typical choices being compartmental models or agent-based models). One can use this kind of approach in two main ways:

1. One can generate a model with realistic parameters (with a level of detail depending on the question of interest), and study its behavior on parameter variation (analytically or numerically) without fitting its output to data; this is often done for detailed insights into the mechanisms of infection or for qualitative insights into the features of epidemic spread such as the timing of infection in different locations. For insights into our topics of interest (the influence of space and climate on disease spread), one could input realistic transport data and study the qualitative features of the resulting model in the spirit of the metapopulation models of influenza and other diseases [231]; for the dengue data set that we use in this thesis, the PhD thesis of Torre [266] showed a two-compartment model for the influence of climate on dengue fever in Peru. We chose not to pursue this route as we were interested in finding insights directly from the disease data without aggregating it, building reusable models, and in gaining general information based on comparison between data sets rather than focusing on detailed study of a small number of them.

2. One can attempt to fit a model with the multiple locations to the data sets. However, doing this on large, noisy data sets like the ones presented in this thesis without aggregating the number of locations is extremely difficult. We have attempted to match a metapopulation model to the dengue disease time series, with little success due to the internal variation between the time series and the difficulty of fitting the very large model. This work forms part of a report published online [236], but it is not presented in this thesis.

Another approach for analyzing disease time series could come from the field of time series analysis, e.g., using moving average (MA) and/or autoregressive (AR) models, or a combination thereof, e.g. autoregressive integrated moving average (ARIMA) models. These kinds of models require transformation of the disease time series to remove non-stationarity, and they are limited to the forecasting of either long-term non-seasonal, or repeating (e.g., seasonal) effects over a long period of time [6]. In contrast, the approach presented in this thesis does not require data modification and it is applicable both to the long-term study of endemic diseases, and to the study of newly-spreading epidemics, as shown in Chapters 7 and 8 respectively.

Another approach could come from the field of statistics or geographical approaches to studying infectious disease. Using geographical approaches allows one to identify patterns of disease distribution and detect disease clusters [237] — a similar aim to community detection. However, the data sets available to us do not have a high enough spatial resolution to permit a detailed analysis of disease risk in space. Further, this kind of approach would not allow us to distinguish any organization within the risk clusters, as community detection does.

Even once we decide on a similarity matrix approach, multiple questions remain. Other similarity measures may be more suitable than Pearson correlation (which we chose for simplicity) [238,250,287]. Further, there are multiple alternative approaches proposed so far to infer some form of modular or hierarchical organization from multiple time series. These are often based on financial time series (of asset prices and stock returns). These approaches tend to be based either on the introduction of thresholds (which is arbitrary, results in loss of information and does not provide a clear answer), or a geometric embedding in some metric space with predefined properties and calculating minimal spanning trees or planar maximally filtered graphs (which also result in information loss) [26,171]. Further, random matrix theory has been used as an exploratory tool to infer the proportion of financial correlation matrices that is non-random. This allows for the identification of groups of time series that rise and fall together. The correlation null model for community detection [171], which we described in Section 3.3.6, builds on these ideas for community detection.

The aim of using methodology described in this thesis is to explore the usefulness of community-detection methods that have been successful in other fields related to the analysis of time-series (such as financial markets), to the time series describing the numbers of disease cases in different locations. To our knowledge, this kind of approach has never been applied to epidemiology. Through the research presented here, we hope to shed some light on the advantages and disadvantages of a network science approach, and in particular of using the methodology of community detection using modularity maximization, to studying infectious diseases.

3.7 Summary

In this chapter, we have defined the network science terminology used in this thesis. We have reviewed the methods used for community detection in networks, and described modularity — the method that we use in this thesis for community detection on spatially embedded, temporal disease-correlation networks. We have also presented several null models for community detection using modularity maximization that are particularly relevant to this application, i.e., spatial null models and null models defined specifically for correlation networks. Each of the null models has its advantages and disadvantages. The NG null model is the most common choice in the literature dealing with applications of community detection; however, the interpretation of the NG null model for correlation networks may not be the one desired for the application — an issue that was addressed by the correlation null model. The spatial null models on the other hand aim to remove the impact of spatial proximity on connections, which may mask other relationships. The gravity null model has been proposed previously and is based on well known gravity models that are commonly used for modelling population mobility and disease spread. The radiation null model is based on the recent radiation model, which has shown promise both for mobility and disease modelling. See the summary of the advantages and disadvantages of the four null models in Table 3.1.

In the following chapters, we will study the effects of using different null models on the results of community detection on spatially-embedded networks. For this purpose, we will create novel spatially-embedded benchmark networks, which represent idealized relationships between spatial entities ("cities") that depend on a combination of distance, population, and flux. We test the NG and spatial null models on these benchmarks in Chapter 4. We will also test all of the null models on correlation networks created from disease incidence time series in Chapters 6 (dengue fever), 7 (endemic diseases), 8 (emerging diseases) and 9 (synthetic time series from an agent-based model).

We have also presented the methodologies that we use in this thesis to visualize, summarize, and study the spatial and temporal aspects of community structure. Our three main methods of assessing the spatial organization of community structure are z-Rand scores, NMI, and Monte Carlo tests using total intra-community distance and total MST score as the test statistics. The choice of methodology depends on the network (and data set) and the question asked. NMI is preferred over both z-Rand scores and distance/MST tests for the benchmarks, for which we know the ground-truth partition, and (at least for low inter-community mixing) the expected differences between planted and algorithmically detected partitions are small. For real data sets, the main advantage of the z-Rand scores over the Monte Carlo measures is using a known ground-truth spatial partition, and thus the relative

Table 3.1: Null models for community detection that we use in this thesis. Abbreviations: "Grav." — gravity, "Rad." — radiation, "Corr." — correlation.

Name	Purpose	Advantages	Disadvantages
NG	General	• Derivable from first principles • Well tested • Simple interpretation of results	• Spatial proximity masks other relationships • Interpretation issue for correlation networks
Grav.	Spatial networks	• Removes spatial influence • May uncover new information • Based on well-known gravity models • Relatively simple calculation c.w. radiation n.m.	• Assumes distance-based relationships • May not be optimally-suited for disease spread • Distance binning adds complications to the calculation
Rad.	Spatial networks	• Removes spatial influence • Radiation model should work better than gravity model for population fluxes • May be more suitable for disease spread than gravity n.m.	• Binning and flux calculation add complications • May be overcomplicated
Corr.	Correlation networks	• Null hypothesis of time series independence for correlation networks • Derivable from first principles	• Forces long time windows • Less clear network-science interpretation of the null model than NG

ease of interpretation of the results. However, z-Rand scores have low sensitivity, so partitions with low visual similarity to the manual partition might score as statistically similar. On the other hand, they might misclassify some partitions, in particular they those that exhibit spatial clusters that are shaped differently than the planted partition. Due to this, and to the fact that we do not possess ground-truth partitions for the disease data sets, the fact that the distance and MST tests do not require a manual partition to compare against is an advantage for them. These and other considerations are summarized in Table 3.2.

Table 3.2: Different ways to examine algorithmically-obtained community structure that we use in this thesis. Abbreviations: "MC tests" — Monte Carlo tests (i.e, distance and MST tests).

Name	Purpose	Advantages	Disadvantages
z_R	• General structural similarity to (spatial and temporal) manual partitions	• Simple • Uses relevant data (e.g., climate) • Relatively simple interpretation	• Data availability • Low sensitivity — worry about false positive similarity scores
NMI	• Comparison to ground-truth partitions • Detecting small changes in structure	• Easy interpretation	• Quickly drops off with errors • Limited suitability to data with no ground truth
MC tests	• Detecting spatial organization of community structure • Use when no manual spatial partitions	• No need for manual spatial partitions	• New and unproven

Chapter 4

Synthetic benchmark networks

This chapter consists of original work, most of which appears in a working paper by MS, E. Leicht, G. Chowell, and M. A. Porter [235].

4.1 Introduction

In this chapter, we develop novel synthetic benchmark networks that represent idealized, spatially-embedded, modular networks with a known (planted) community structure. We then test the performance of the gravity null model, the radiation null model, and the NG null model on these benchmarks across a range of parameter values.

Many researchers have studied the performance of community detection methods on synthetic benchmarks with known structure. The simplest benchmarks are variations of stochastic block models, often with 4 groups of 32 nodes [68]. The LFR benchmark is an extension of the 4-group benchmark that includes heterogeneous distributions of node degree and community size, making community detection harder [161], and the LF benchmark extends it to generate directed and weighted networks [157]. In contrast, benchmarks for spatially-embedded and temporally-evolving networks are only in their infancy. For spatial networks, Expert et al. [82] introduced the notion of simple spatial benchmarks in which the connectivity between nodes is based on distance and community membership. For multilayer networks, De Domenico et al. [72] extended the notion of benchmarks to multilayer benchmarks in which layers are obtained by rewiring an increasing percentage of the edges of the first layer, and Granell et al. [109] introduced temporal benchmarks based on grow/shrink and split/merge dynamics.

In this chapter, we design spatial benchmarks that combine both spatial and temporal featured and we use them to test the results of community detection using modularity maximization with different null models.

4.2 Benchmark construction

To construct the benchmarks, we assign N nodes uniformly at random to positions on the lattice $\{1, 2, \ldots, l\} \times \{1, 2, \ldots, l\}$. Each node i (i.e., an idealized "city" or "province") is assigned a population of n_i. We create two versions of each benchmark: "uniform population" and "random population". In the "uniform population" version, all nodes have the same population of 100; in the "random population" benchmarks, nodes are assigned integer populations $(n_1, n_2, \ldots, n_N)$ uniformly at random from $\{1, \ldots, 100\}$. We also assign the nodes uniformly at random to one of two equally-sized communities; this is the "planted" community structure. We then use the community assignment together with node locations and populations when allocating edges between nodes. We place edges between nodes at random up to a defined density, according to edge probability distributions that depend on benchmark type. We create both static (i.e., single-layer), and multilayer benchmarks that represent the temporal evolution of community structure. Most of the benchmarks that we study are "temporally stable" benchmarks, in that the underlying planted community structure that determines the edge-placement probabilities is the same for the independently drawn benchmark layers. However, we also try "temporally evolving" benchmarks, in which the planted community structure evolves over time.

In what we call the *distance benchmark*, the probability of an edge between two nodes depends only on the geographical distance between nodes and on their community assignments. In the uniform population distance benchmark, all nodes are assigned the same population $n_i = 100$ (this version corresponds to the benchmark in Expert et al. [82]). In the random population distance benchmark, nodes are assigned integer populations $(n_1, n_2, \ldots, n_N)$ uniformly at random from $\{1, \ldots, 100\}$. In the distance benchmarks, the probability p_{ij}^{dist} that an edge is placed between nodes i and j is inversely proportional to the distance d_{ij} between them:

$$p_{ij}^{\text{dist}} = \frac{\lambda(c_i, c_j)}{Z_1 d_{ij}}, \tag{4.1}$$

where c_i is the community that contains node i and the function $\lambda(c_i, c_j) = 1$ if nodes c_i and c_j are in the same community and $\lambda(c_i, c_j) = \lambda_d$ otherwise. The "inter-community connectivity" λ_d controls the amount of mixing between communities. When $\lambda_d = 0$, only nodes in the same community are connected; when $\lambda_d = 1$, there are no distinct communities. The normalization constant Z_1 ensures that $\sum_{i>j} p_{ij}^{\text{dist}} = 1$. We independently place $L = \mu N(N-1)/2$ edges, where we place an edge between nodes i and j with probability p_{ij}^{dist} each, and the parameter $\mu \geq 0$ determines the network's edge density. We interpret multiple edges as weights. We normalize the network by taking $W = W/\max(W)$, where $\max(W)$ is the largest edge weight in the network. This guarantees that all edge weights lie in $[0, 1]$. This makes the multilayer benchmarks comparable with the disease correlation networks in terms of the relative weights of interlayer edges and intralayer edges. For

Table 4.1: Parameters used in benchmark creation. Notation: rand($\{a, b\}$) signifies that we select a number uniformly at random from the set $\{a, a+1, \ldots, b\}$. Notes: * We tested values of 10 and 100 — see Section 4.3.1. ** We tested values of 10, 50, 90 and 100 — see Section 4.3.3. *** We tested values between 0.1 and 100 — see Section 4.3.2.

Parameter	Description	Typical Values
l	lattice size	10^*
N	number of nodes	50^{**}
n_i	node population	r: rand($\{1, 100\}$); u: 100
p_{ij}	edge probability	benchmark-dependent (see Table 4.2)
λ_d	inter-community connectivity	varied between 0 and 1
μ	edge density parameter	100^{***}

the distance benchmark, the gravity null model performs best with uniform populations, and all null models perform comparably with random populations (see Figs. 4.6 and 4.7 in Section 4.3.4).

The *flux benchmark* aims to mimic the spread of disease on a network. We allocate its edge weights depending on the mean flux predicted by the radiation model (defined in Section 3.3.5) between pairs of nodes. We place N nodes uniformly at random on the lattice $\{1, 2, \ldots, l\} \times \{1, 2, \ldots, l\}$, and we assign populations and communities in the same manner as for the distance benchmark. As with the distance benchmark, we consider both uniform-population and random-population versions of the flux benchmark. Now, however, the edge placement probability p_{ij}^{flux} is directly proportional to the mean predicted radiation model flux $\hat{T}_{ij}$ (Eq. 2.3), which is turn is inversely proportional to distance d_{ij} between nodes i and j. The edge placement probabilities are defined as:

$$p_{ij}^{\text{flux}} = \frac{\lambda(c_i, c_j)\hat{T}_{ij}}{Z_2}\,, \tag{4.2}$$

where Z_2 is a normalization constant to ensure that $\sum_{i>j} p_{ij}^{\text{flux}} = 1$. The radiation null model performs best on this benchmark (see Figs. 4.6 and 4.7 in Section 4.3.4).

The *distance and population* spatial benchmark is a distance-based benchmark that also incorporates population information into edge placement by taking $p_{ij}^{\text{distpop}} = \frac{n_i n_j \lambda(c_i, c_j)}{Z_3 d_{ij}}$, similar to a gravity model. This brings back the advantage of the gravity null model for both uniform and random populations (see Fig. 4.8 in Section 4.3.5). The radiation null model performs second-best on the distance and population benchmarks, and its performance is better for the random population case than the uniform population case. In Table 4.1, we summarize the parameters that we use in benchmark creation. In Table 4.2, we summarize the different types of benchmarks.

We create both static (i.e., single-layer) and multilayer benchmarks. The static benchmarks enable us to study the performance of modularity maximization using a given null model in a simple setting without the additional complications of a multilayer network. However, the multilayer benchmarks are ultimately more appropriate for disease data because they can incorporate temporal evolution.

Table 4.2: Population and edge probability, and the best performing null models, for the main types of benchmarks. Notation: rand($\{a, b\}$) signifies that we select a number uniformly at random from the set $\{a, a+1, \ldots, b\}$. Additionally, $\lambda(c_i, c_j) = 1$ if nodes c_i and c_j are in the same community and $\lambda(c_i, c_j) = \lambda_d$ otherwise; d_{ij} is the distance between nodes i and j in space; $\hat{T}_{ij}$ is the mean predicted radiation model flux between nodes i and j in space; n_i is the population of node i; and Z_1, Z_2, and Z_3 are normalization constants.

Benchmark	Population	Edge probability	Best null model	Main figure
Distance	100	$p_{ij}^{\text{dist}} = \frac{\lambda(c_i,c_j)}{Z_1 d_{ij}}$	Gravity	Fig. 4.6
Distance	rand($\{1, 100\}$)	$p_{ij}^{\text{dist}} = \frac{\lambda(c_i,c_j)}{Z_1 d_{ij}}$	All do badly	Fig. 4.7
Flux	100	$p_{ij}^{\text{flux}} = \frac{\lambda(c_i,c_j)\hat{T}_{ij}}{Z_2}$	Radiation	Fig. 4.6
Flux	rand($\{1, 100\}$)	$p_{ij}^{\text{flux}} = \frac{\lambda(c_i,c_j)\hat{T}_{ij}}{Z_2}$	Radiation	Fig. 4.7
Distpop	100	$p_{ij}^{\text{distpop}} = \frac{n_i n_j \lambda(c_i,c_j)}{Z_3 d_{ij}}$	Gravity	Fig. 4.8
Distpop	rand($\{1, 100\}$)	$p_{ij}^{\text{distpop}} = \frac{n_i n_j \lambda(c_i,c_j)}{Z_3 d_{ij}}$	Gravity	Fig. 4.8

We generate simple, "temporally stable" multilayer benchmarks (see Fig. 4.1) by connecting m layers that each have an intralayer structure that we determine independently from a static benchmark using the same starting conditions (N, l, μ, λ_d). Independent generation of each layer based on the same starting conditions represents the variation expected between observations due to noise and experimental variation.

We also introduce "temporally evolving" multilayer benchmarks, in which the planted community structure evolves throughout time. We change the community assignment of a defined fraction p of nodes. We select the nodes that change communities uniformly at random, and for each node we select the starting layer uniformly at random; For each, we select a new community assignment uniformly at random from the remaining communities, and we change its community assignment for all remaining layers.

For both types of multilayer benchmark, we set the interlayer edges between consecutive layers to be $\omega \in [0, \infty)$. For each benchmark, we obtain a consensus community structure over 50 runs (calculated as described in Section **??**. Each of the results of community detection that we report is a mean score (e.g., NMI) for the consensus community structures, each calculated from one of 50 instances of a benchmark with the same starting conditions (N, l, μ, λ_d and p if applicable) and parameter values (γ, ω).

We evaluate the performance of the NG, gravity, and radiation null models on our benchmarks by comparing algorithmic partitions with the planted community structure using normalized mutual information (NMI) [255], described in Section 3.4.5.

We use NMI in the following sections, and we obtain the same qualitative conclusions using variation of information [151], which is a different normalized measure of similarity. See Section 4.3.6 for our comparisons using VI and z-Rand scores [268] — the similarity measure that we use for detecting coarse structural similarities in disease data and synthetic time series from an agent-based model.

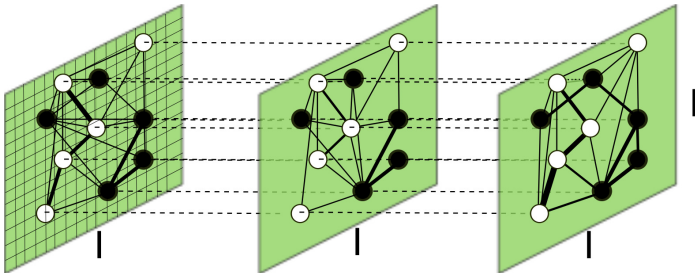


Figure 4.1: Construction of temporally stable multilayer spatial benchmarks. We assign N nodes uniformly at random to positions on an $l \times l$ lattice (which we show in layer 1) and divide them into two equal-sized communities (black and white) whose nodes we choose uniformly at random. Node i has a population of n_i, and each layer has the same set of nodes. For each layer, we allocate edges uniformly at random according to a probability distribution that depends on the type of benchmark; for details, see the text and Table 4.2. We interpret multiple edges as weights, and we visualize these weights using edge thickness. We connect copies of nodes in adjacent layers with interlayer edges of weight ω (dashed lines).

4.3 Results on static benchmarks

In this section, we present the results of community detection on the benchmarks defined in Section 4.2. We start by using different benchmark parameters and observing the effect (if any) on the ability to perform community detection on the resulting networks. This allows us to fix certain benchmark creation parameters before studying the effect of changing the parameters for community detection.

4.3.1 Benchmark size and bin size

To emphasize the difference between the gravity and radiation null models, we take $N = 50$ and $l = 10$ to obtain a relatively densely filled lattice. (See Section 4.3.3 for the results for a synthetic network with parameter values $N = 10$ and $N = 90$.) We first compare this benchmark versus a situation with parameter values $N = 100$ and $l = 100$ (which are the parameter values that were used in Expert et al. [82]). We test different bin sizes in uniformly-spaced bins using the parameter values $b \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.1\} \cup \{1, 2, \ldots, 10\}$, $l = 10$ and $b \in \{1, 2, \ldots, 100\}$, $l = 100$. We find that bin width makes a large difference on both benchmarks: $b = 1$ produces the highest NMI scores (i.e., it has the "best performance"), and increasing bin width leads to a decrease in performance of both spatial null models (see Fig. 4.2). This effect is especially pronounced for the gravity null model.

The performance of the spatial null models at optimal bin sizes (i.e., the bin sizes giving best performance of the ones tested) for $l = 10$ and $l = 100$ is similar, so we henceforth use the $l = 10$ benchmark with $b = 1$ to lower computational time and memory usage. However, one needs to keep the strong influence of bin size on algorithm results in mind for applications.

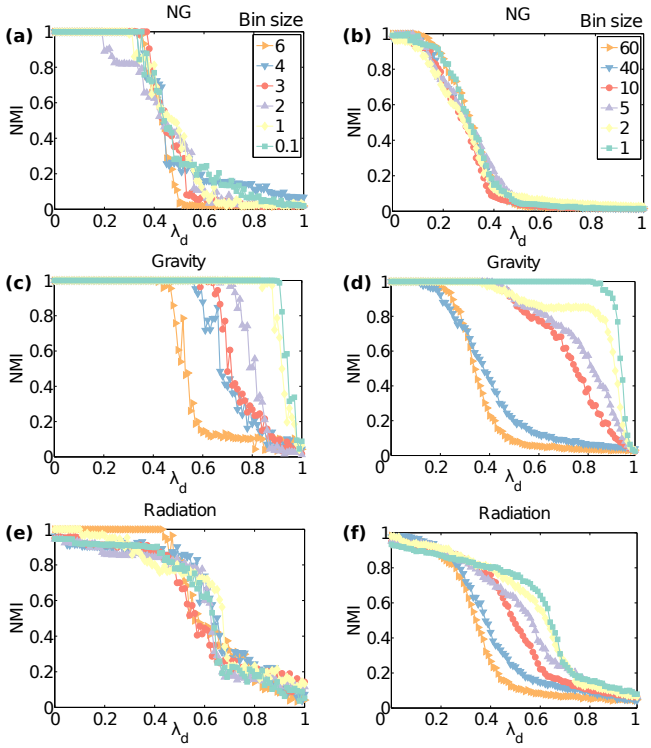


Figure 4.2: Uniform population static benchmarks: normalized mutual information (NMI) scores between algorithmically detected community structures and planted partitions in static uniform population distance benchmarks for (left) $l = 10$, $N = 50$ and (right) $l = 100$, $N = 100$. The edge density parameter is $\mu = 100$, and we use uniform populations of 100 for different bin sizes (colored curves). We detect communities by optimizing modularity using the (top) NG, (middle) gravity, and (bottom) radiation null models.

4.3.2 Variation of edge density parameter μ

We present the results of using different values of the edge density parameter μ. See Fig. 4.3 for uniform population static benchmarks, and Fig. 4.4 for random population static benchmarks. We observe that low edge density ($\mu < 1$) has a strong effect on the ability of the algorithm to detect communities, and for $\mu = 0.01$ none of the null models is able to detect any structure. For $\mu \geq 5$, the prevailing effect on the accuracy of community detection is the influence of the inter-community connectivity λ_d between communities. Low edge density appears to have the largest effect on the best-performing null models on each benchmark – on the gravity null model for the distance benchmark and on the radiation null model for the flux benchmark. At low edge density, their performance is lower than at higher edge density, and it is comparable to the other null models. Following these findings, we chose to focus on one high edge density ($\mu = 100$) for the rest of this thesis.

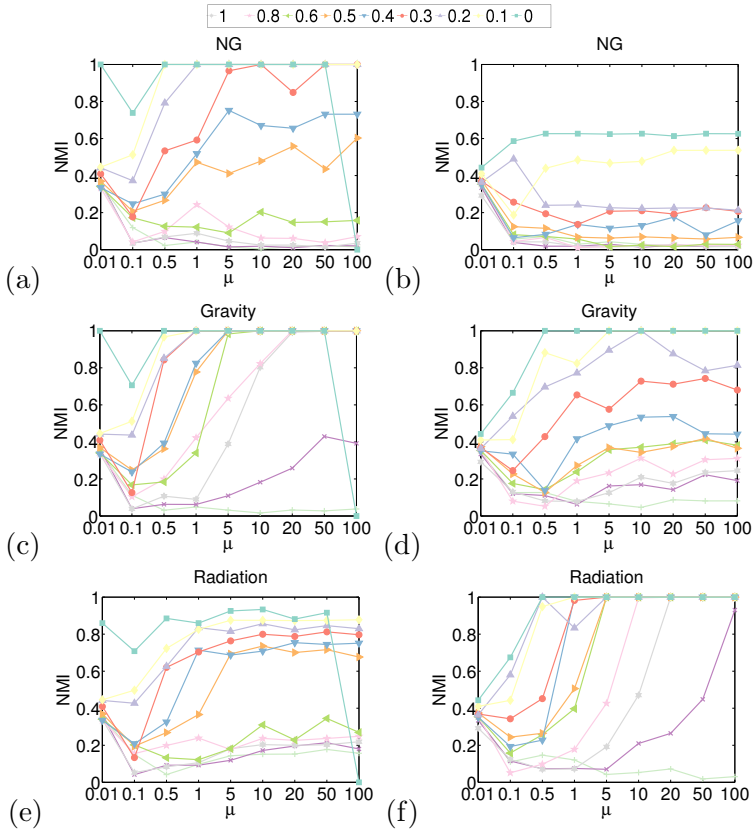


Figure 4.3: Uniform population static benchmarks: varying edge density. Normalized mutual information (NMI) between algorithmically detected community structures and planted partitions for uniform population static spatial benchmarks with $N = 50$, $l = 10$, a population of 100, and different inter-community connectivity λ_d (colored curves). We plot the NMI scores for community detection using $\gamma = 1$ as a function of the edge density parameter values μ for (left) the distance benchmark and (right) the flux benchmark.

4.3.3 Variation of the number of nodes

We present the results of placing increasing numbers of nodes, $N = \{10, 50, 90\}$, in the same-sized benchmark ($l = 10$), with edge density parameter $\mu = 100$ and uniform populations of

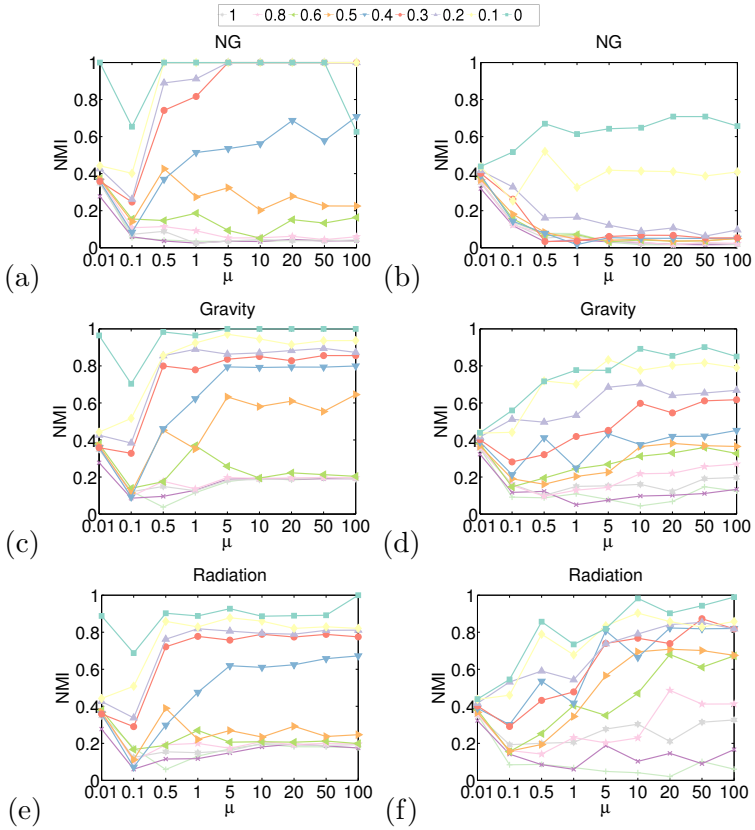


Figure 4.4: Random population static benchmarks: varying edge density. Normalized mutual information (NMI) between algorithmically detected community structures and planted partitions for random population static spatial benchmarks with $N = 50$, $l = 10$, a population n selected uniformly at random from $[1, 100]$ and different inter-community connectivity λ_d (colored curves). We plot the NMI scores for community detection using $\gamma = 1$ as a function of the edge density parameter values μ for (left) the distance benchmark and (right) the flux benchmark.

100 across a variety of γ parameters using the three null models (see Fig. 4.5). We observe no qualitative change other than an increase in variability at low N.

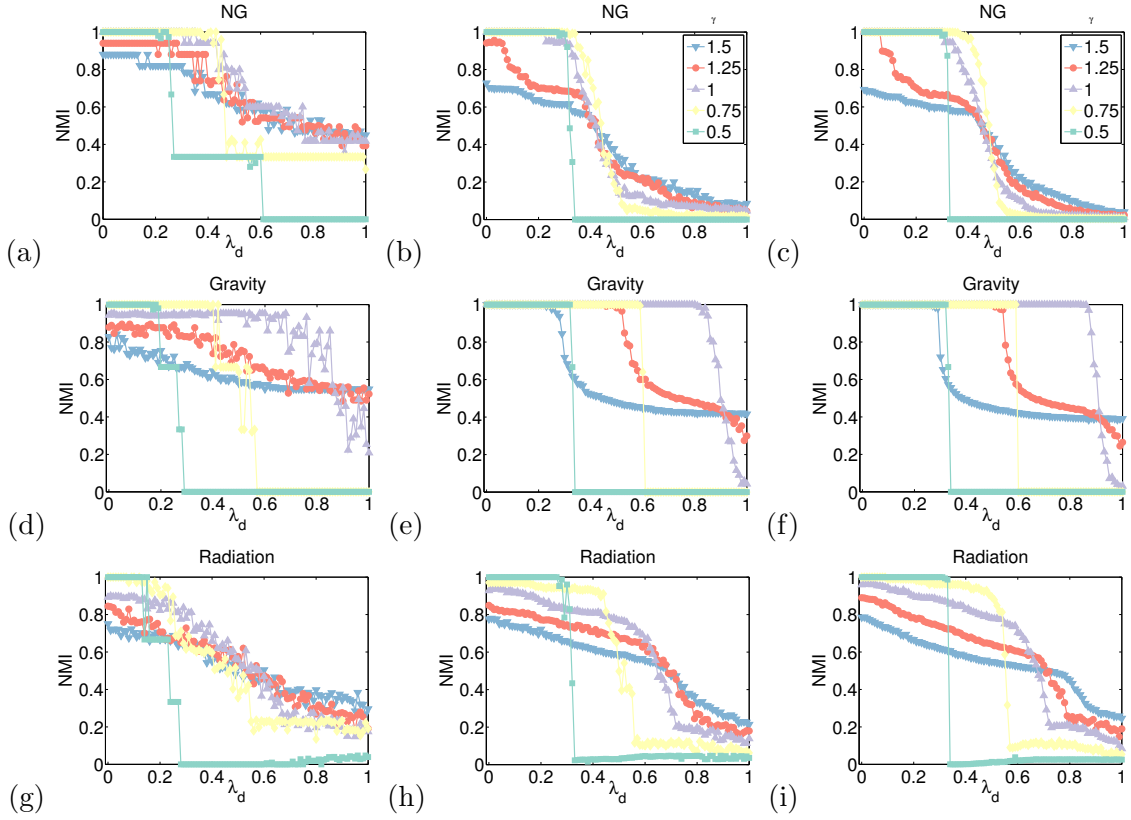


Figure 4.5: Uniform population static benchmarks: varying number of nodes. Uniform population static benchmarks: NMI scores between algorithmically detected community structures and planted partitions in static uniform population distance benchmarks with $l = 10$, $\mu = 100$, and uniform populations of 100 for different numbers of nodes in the same-sized space. We show plots for different values of γ (colored curves) for (left) $N = 10$, (middle) $N = 50$, and (right) $N = 90$ nodes. We performed community detection using (top) NG, (middle) gravity and (bottom) radiation null models.

4.3.4 Influence of the resolution parameter γ

We now fix the parameter values at $N = 50$ nodes, lattice size $l = 10$, bin size $b = 1$, and edge density parameter $\mu = 100$. We then study the performance of the three null models using several values of the resolution parameter $\gamma \in \{0.5, 0.75, 1, 1.25, 1.5\}$ and the inter-community connectivity $\lambda_d \in \{0, 0.01, \ldots, 0.99, 1\}$ on static benchmarks. Smaller values of γ tend to yield larger communities, and vice versa. Considering larger λ_d increases the level of mixing between the communities and makes it more difficult to detect planted communities successfully.

For the uniform population distance benchmark, the only factor that influences edge placement is the distance between nodes. On this benchmark, the gravity null model has the best performance, as it is able to find the correct partitions for $\lambda_d \lessapprox 0.82$ (see the left panel of Fig. 4.6). The radiation null model has the second best performance and is able to find partially meaningful partitions for $\lambda_d \lessapprox 0.74$, above which we observe a plateau of

"near-singleton" partitions in which most nodes are placed into singleton communities. (We use the term "singleton partition" to refer to a partition in which every node is assigned to its own community.) The NG null model, which does not incorporate spatial information, does much worse than either of the spatial null models; it suffers a sharp decline in performance at $\lambda_d \approx 0.4$.

The bad performance of the NG null model compared with the spatial null models demonstrates that incorporating spatial information into the null model for community detection by modularity maximization on networks where edges are related to space improves the performance of the algorithm. Further, the superior performance of the gravity null model over the radiation null model suggests that, although incorporating spatial influence is beneficial, it needs to be done intelligently. Here, using a null model that incorporates population information to study community structure in networks whose structure does not depend on population decreases the performance of community detection.

On the uniform population flux benchmark — in which we include the population density in the region between two nodes in the flux prediction (so the population density influences edge structure) — the radiation null model outperforms the other null models. The gravity null model comes in second place, and the NG null model is a distant third (see the right panel of Fig. 4.6).

For the random population distance benchmark, we observe a fast deterioration in quality of the detected communities for $\lambda_d \gtrapprox 0.4$ for all null models, and all null models reach the "near-singleton" plateau by $\lambda_d \approx 0.6$. The NG null model performs best for $\lambda_d \lessapprox 0.43$. Above that, the gravity null model performs best, although for $\lambda_d \gtrapprox 0.6$, the partitions are largely singletons (see Fig. 4.7).

For the random population flux benchmark, the radiation null model has the best performance of the three null models. The decrease in NMI values with the increase in λ_d is the slowest of the three null models. The gravity null model performs second best, and the NG null model fails even when there is no mixing between the two communities (see Fig. 4.7). However, note that the best performance is much worse on the random population benchmarks than on the uniform population benchmarks.

Among the parameter values that we consider ($\gamma \in \{0.5, 0.75, 1, 1.25, 1.5\}$), $\gamma = 1$ appears to give the best results (i.e., the largest NMI scores). In the near-singleton regime, $\gamma = 1.5$ outperforms it slightly in terms of NMI scores (see Figs. 4.6 and 4.7), although it yields near-singleton partitions that are very different from the planted partition. The optimal value of γ rarely changes for different null models. It does however change with the number of planted communities, as expected, with higher γ values optimal for benchmarks with a higher number of smaller communities (not shown).

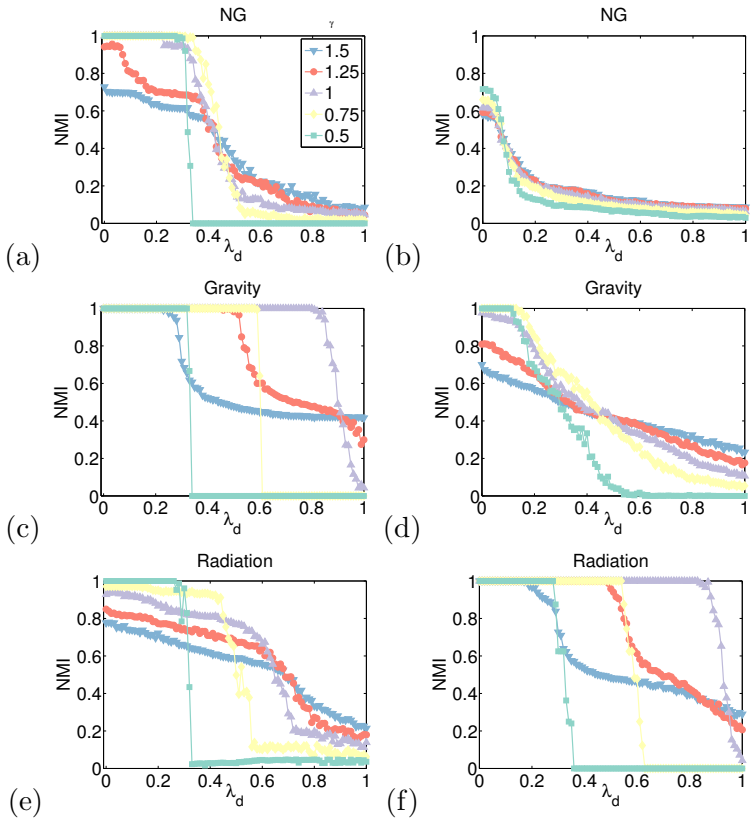


Figure 4.6: Uniform population static benchmarks: varying γ. Uniform population static benchmarks: normalized mutual information (NMI) scores between algorithmically detected community structures and planted partitions in static benchmarks with $l = 10$, $N = 50$, $\mu = 100$, and uniform populations $n_i = 100$. We show plots for different values of γ (colored curves) for distance benchmarks (left) and flux benchmarks (right). We performed community detection using (top) NG, (middle) gravity and (bottom) radiation null models.

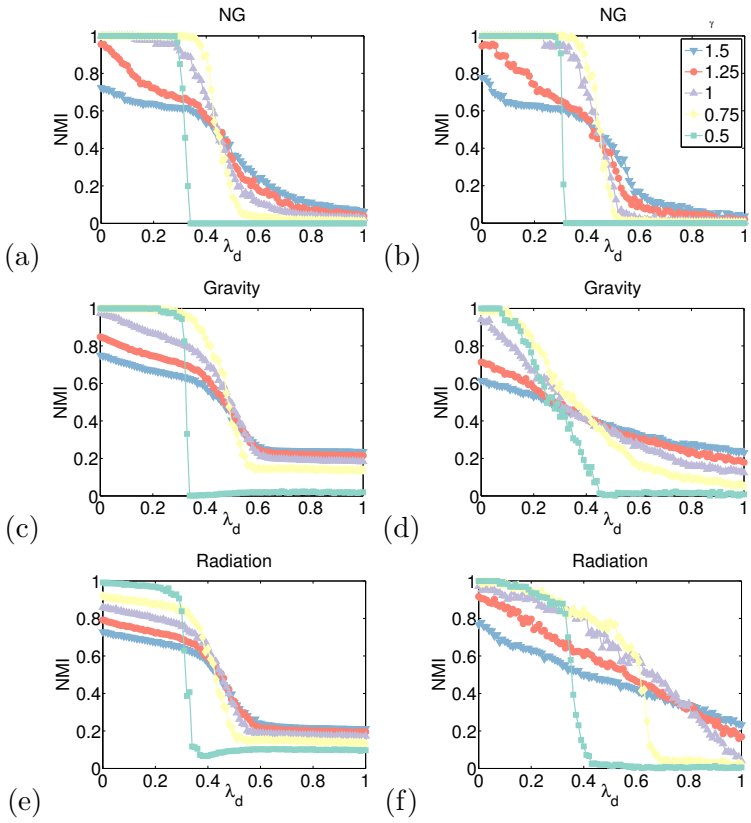


Figure 4.7: Random population static benchmarks: varying γ. Random population static benchmarks: normalized mutual information (NMI) scores between algorithmically detected community structures and planted partitions in static spatial benchmarks with $N = 50$, $l = 10$, $\mu = 100$ and populations n_i drawn uniformly at random from $\{1, \ldots, 100\}$, for different γ parameters (colored curves) for (left) the distance benchmark and (right) the flux benchmark. We performed community detection using (top) NG, (middle) gravity and (bottom) radiation null models.

4.3.5 "Distance and population" spatial benchmark

In this section, we present the "distance and population" spatial benchmark. Following the observation that none of the null models find the planted structure on a distance benchmark with random populations, we create this benchmark to test what happens if one incorporates the population sizes in the edge probabilities of the distance benchmark.

We build a benchmark in the same manner as the distance benchmark in Section 4.2, except that we include population into edge placement probability by taking $p_{ij}^{\text{distpop}} = \frac{n_i n_j \lambda(c_i,c_j)}{Z_3 d_{ij}}$. This brings back the advantage that the gravity null model had for both uniform and random populations (see Fig. 4.8). The radiation null model performs second best, with a better performance on the random-population version.

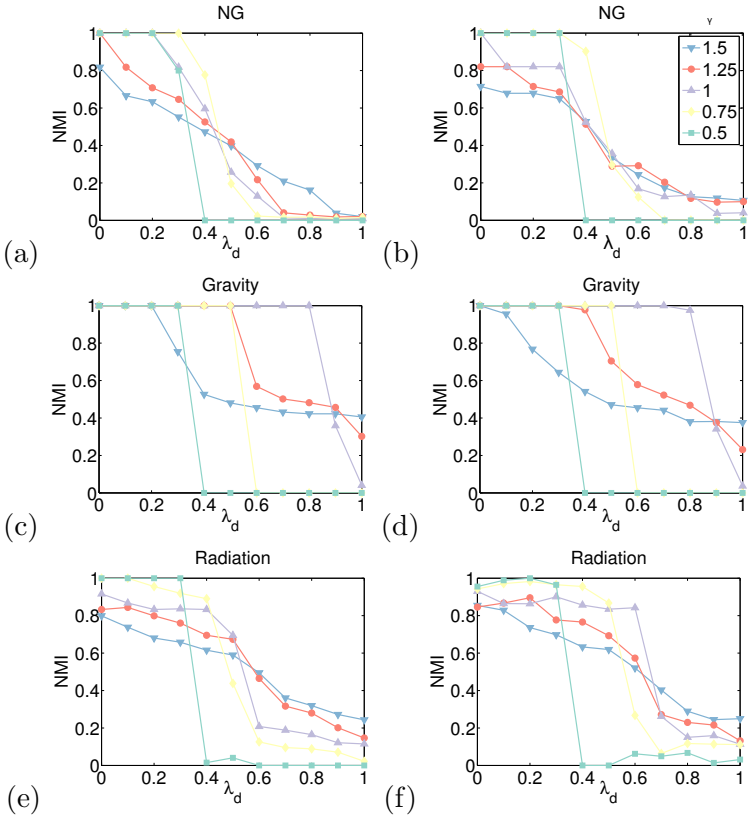


Figure 4.8: Distance and population static benchmarks: varying γ. Normalized mutual information (NMI) between algorithmically detected community structures and planted partitions in "distance and population" static spatial benchmarks with (left) uniform populations $n_i = 100$ and (right) random populations drawn uniformly at random from $\{1, \ldots, 100\}$, with parameters $N = 50$, $l = 10$, $m = 10$, $\mu = 100$, and $\gamma = 1$ for different values of γ (colored curves) and different values of λ_d. We performed community detection using (top) NG, (middle) gravity and (bottom) radiation null models.

4.3.6 Variation of information and z-Rand scores

For the spatial benchmarks, an alternative similarity measure between the planted partitions and the algorithmic partitions is normalized variation of information (NVI) [181], defined in Section 3.4.5. In contrast to NMI, variation of information (VI) is a metric in the

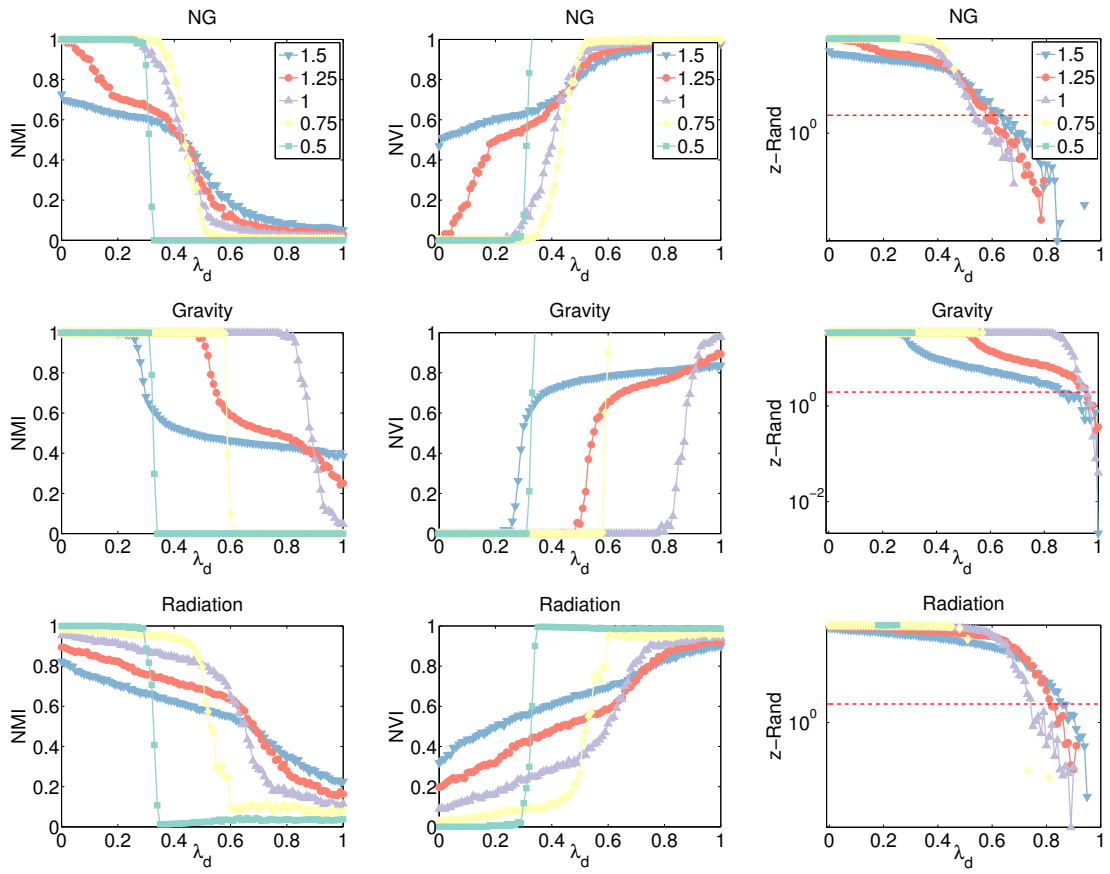


Figure 4.9: Uniform population static benchmarks: NMI, NVI and z-Rand scores. (Left) Normalized mutual information (NMI), (centre) normalized variation of information (NVI) and (right) $\log_{10}$ of z-Rand scores between algorithmically-detected partitions, which we obtain by maximizing modularity, and planted partitions in the uniform population distance static spatial benchmarks with $N = 50$ cities, a grid size of $l = 10$, and a density parameter of $\mu = 50$. We examine the partitions for different values of the resolution parameter γ as a function of inter-community connectivity λ_d using the (top) NG null model, (middle) gravity null model, and (bottom) radiation null model. For the z-Rand scores, we show the significance cutoff of 1.96 for guidance; scores of "Not a number" are not plotted.

mathematical sense, facilitating comparison of results across different conditions. We also compare the NMI and NVI results against z-Rand scores, a pair-counting similarity measure defined in Section 3.4.3. This measure is good at detecting similarities in coarse structure [268, 269] but is less sensitive to minor changes (such as one node changing community assignment) than NMI and NVI.

As one can see in Fig. 4.9, both NMI and NVI perform similarly and neither gives visibly better precision. Both NMI and NVI values change quite sharply above certain λ_d thresholds, when the algorithmic partitions start differing from the planted partitions by more than a few nodes. The z-Rand scores however continue scoring algorithmic partitions as similar to the planted partitions for larger values of inter-community connectivity λ_d, until the deterioration of algorithmic partitions into singletons (this is especially visible for the NG null model, see Fig. 4.9.)

4.4 Results on multilayer benchmarks

We now study the influence of the community-detection parameters γ and ω on the ability to detect planted partitions in multilayer benchmarks. We first study the performance of the NG, gravity, and radiation null models on temporally stable uniform and random population benchmarks with the same parameters as for the static benchmarks ($N = 50$ nodes, lattice size $l = 10$, bin size $b = 1$, and $m = 10$ layers) using $\gamma \in \{0.5, 0.75, 1, 1.25, 1.5\}$ and $\omega \in \{10^{-5}, 10^{-3}, 0.1, 0.25, 0.5, 0.75, 1\}$. We tested higher ω values ($\omega \in 2, 3$) but the results for these were not qualitatively different than $\omega = 1$ and were not included in the plots. We then perform the same experiments on temporally evolving benchmarks, and finally we experiment with province-level community detection from multilayer benchmarks.

For these experiments, we calculate the similarity measures (NMI, NVI and z-Rand scores) for the whole multilayer algorithmic partition against the planted multilayer partition, with both partitions defined in terms of multilayer nodes.

4.4.1 Temporally stable benchmarks: varying the resolution parameter γ

We first compare the results on multilayer benchmarks to findings from static benchmarks by varying γ and λ_d for fixed values of ω (see Fig. 4.10). We only show $\omega = 0.1$, as values of ω do not influence the results noticeably.

We obtain similar findings as with the static benchmarks. Once again, we find that the choice of γ has a large influence on the quality of algorithmic partitions, and (as with our findings for static benchmarks) $\gamma = 1$ seems to yield the best performance (i.e., the highest NMI scores) for low values of λ_d, whereas larger values of γ perform better at larger λ_d (see Fig. 4.10).

The results on random population benchmarks (Fig. 4.11) also resemble our findings from static benchmarks. Once again, we find that the choice of γ has a large influence on the quality of algorithmic partitions, and (as with our findings for static benchmarks) $\gamma = 1$ seems to have the best performance (i.e., largest NMI scores) for low values of λ_d, whereas larger values of γ perform better at larger λ_d (see Fig. 4.11).

4.4.2 Temporally stable benchmarks: varying the interlayer edge weight ω

We now examine the NMI of algorithmic versus planted partitions while varying ω and λ_d for fixed values of $\gamma = 1$. We expect that for larger values of ω it becomes more common that each node is assigned to the same community as its counterparts in other layers, which would hinder the detection of planted partitions. The point at which this begins to happen should depend on the strength of the intra-layer community structure, and on the value of ω. However, we did not observe this effect for the temporally stable benchmarks. As we show in Fig. 4.12, we find that the value of ω usually has very little effect on our ability to detect

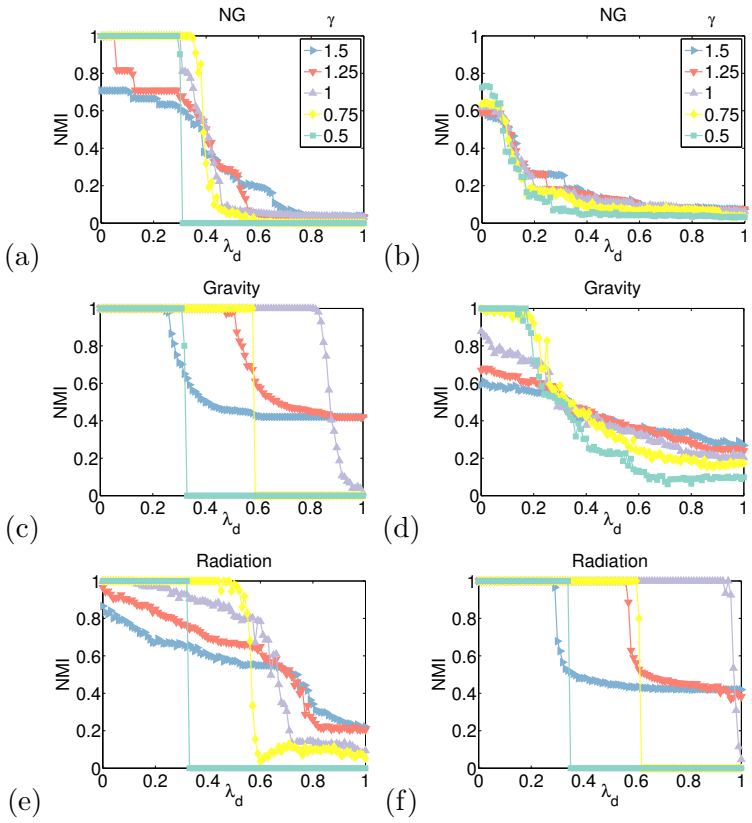


Figure 4.10: Uniform population multilayer benchmarks: varying γ. Normalized mutual information (NMI) between algorithmically detected community structures and planted partitions in uniform population ($n_i = 100$) multilayer spatial benchmarks with $N = 50$, $l = 10$, $m = 10$, and $\mu = 100$ for $\omega = 0.1$ and various values of the resolution parameter γ (colored curves), for different values of λ_d, for (left) the distance benchmark and (right) the flux benchmark. We performed community detection using (top) NG null model, (middle) gravity null model and (bottom) radiation null model.

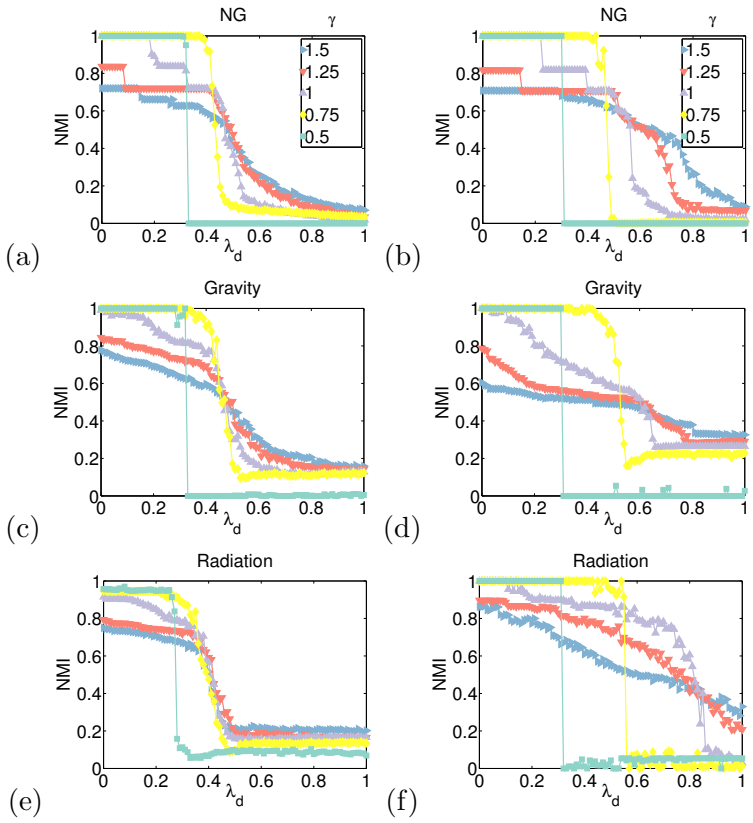


Figure 4.11: Random population multilayer benchmarks: varying γ. Normalized mutual information (NMI) between algorithmically detected community structures and planted partitions in random population multilayer spatial benchmarks (with the population selected uniformly at random from $\{1, \ldots, 100\}$) with $N = 50$, $l = 10$, $m = 10$, $\mu = 100$ for $\omega = 0.1$ and various γ parameters (colored curves) across different values of λ_d, for (left) the distance benchmark and (right) the flux benchmark using (top) NG null model, (middle) gravity null model and (bottom) radiation null model.

the planted communities via modularity maximization. This suggests that perhaps the small interlayer variation from the independent creation of layers is not enough to observe the nuanced influence of ω on community detection that we were hoping for, because the amount of signal form the planted partition is very large. This prompts us to experiment with temporally evolving benchmarks.

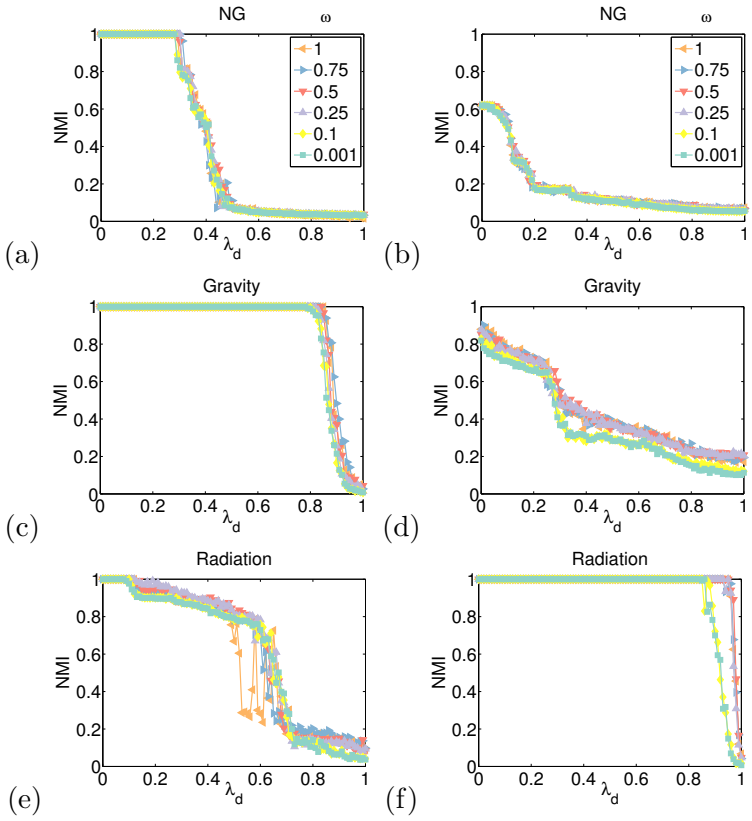


Figure 4.12: Uniform population multilayer benchmarks: varying ω. Normalized mutual information (NMI) between algorithmically detected community structures and planted partitions in uniform population ($n_i = 100$) multilayer spatial benchmarks with $N = 50$, $l = 10$, $m = 10$, and $\mu = 100$ for $\gamma = 1$ and different values of interlayer edge weights ω (colored curves) and different values of λ_d for (left) the distance benchmark and (right) the flux benchmark using (top) NG null model, (middle) gravity null model and (bottom) radiation null model.

Similarly, for random population multilayer benchmarks (see Fig. 4.13), we find that the value of ω usually has very little effect on our ability to detect the planted communities via modularity maximization.

4.4.3 Temporally stable benchmarks: "province-level" communities from multilayer benchmarks

We also perform a "province-level" community detection on the multilayer benchmarks in which we seek assignments of physical nodes (regardless of what layer they are in) to communities and compare the results to benchmark networks with planted community structure. This is analogous to trying to detect community structure in disease data that persists over time — where the community structure might be related to factors influencing

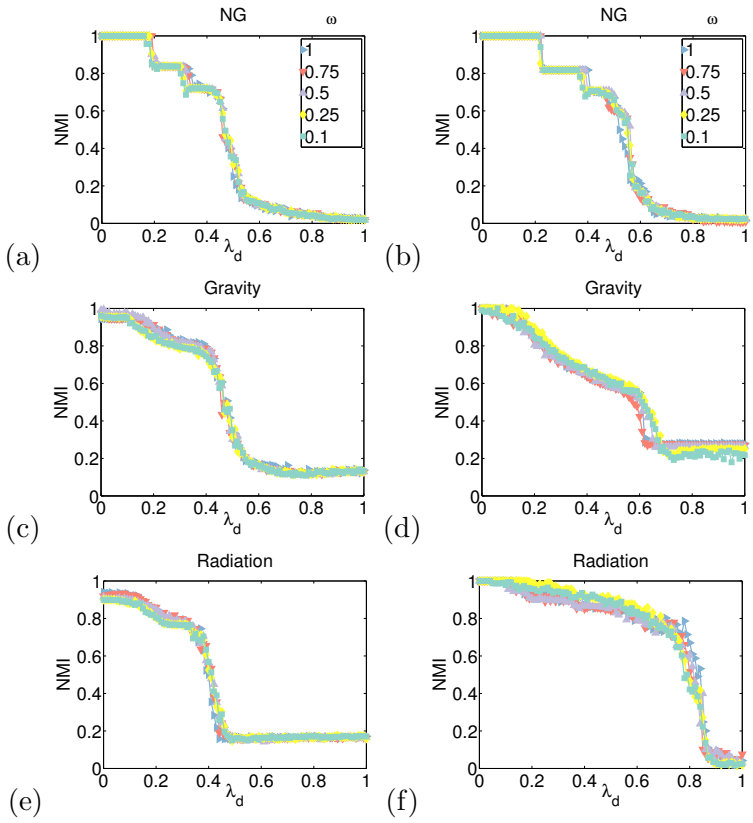


Figure 4.13: Random population multilayer benchmarks: varying ω. Normalized mutual information (NMI) between algorithmically detected community structures and planted partitions in random population (population selected uniformly at random from $\{1, \ldots, 100\}$) multilayer spatial benchmarks with $N = 50$, $l = 10$, $m = 10$, $\mu = 100$ and $\gamma = 1$ for different values of ω (colored curves) and different values of λ_d for (left) the distance benchmark and (right) the flux benchmark using NG (top), gravity (middle) and radiation null models (bottom).

disease patterns, such as climate, transport patterns, population, etc. As one can see by comparing Figs. 4.14 and 4.15 to Figs. 4.10 and 4.11, we obtain similar performance results as with the multilayer communities that we discussed in Section 4.4.2.

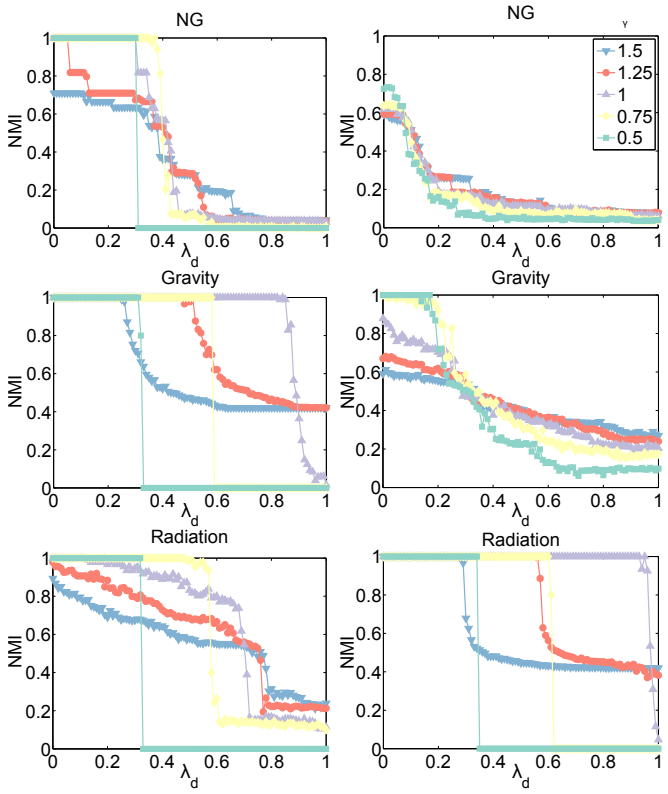


Figure 4.14: Uniform population multilayer benchmarks: province level community detection. Normalized mutual information (NMI) between algorithmically detected province-level community structures and planted partitions obtained from uniform population ($n_i = 100$ for all nodes i) multilayer spatial benchmarks with $m = 10$ layers and their respective single-layer planted partitions with $N = 50$ nodes, a grid size of $l = 10$, $\mu = 100$ for $\omega = 0.1$ and different values of the resolution parameter γ (colored curves) as a function of the inter-community connectivity λ_d, for (left) the distance benchmark and (right) the flux benchmark.

4.4.4 Temporally evolving benchmarks

To study the effect of the parameter ω in more detail, we examine the performance of the three null models on temporally evolving uniform population benchmarks. In this situation, it is perhaps more likely the interlayer edge weights influence the outcomes. We create benchmarks with parameter values of $N = 50$ nodes, a lattice parameter of $l = 10$, a fraction $p = 0.4$ of nodes that change communities over the whole timeline, and $m = 10$ layers. We show results for $\gamma \in \{0.5, 0.75, 1, 1.25, 1.5\}$ and $\omega = 0.1$ in the left panel of Fig. 4.16, and the results for $\omega \in \{10^{-3}, 0.1, 0.25, 0.5, 0.75, 1\}$ and $\gamma = 1$ in the right panel of Fig. 4.16.

Compare Fig. 4.16 to the left panels of Figs. 4.10 and 4.12. On temporally evolving benchmarks, varying ω makes a difference to the ability of modularity maximization to detect the planted partitions. The structures for $\omega \lessapprox 0.1$ for the gravity null model and

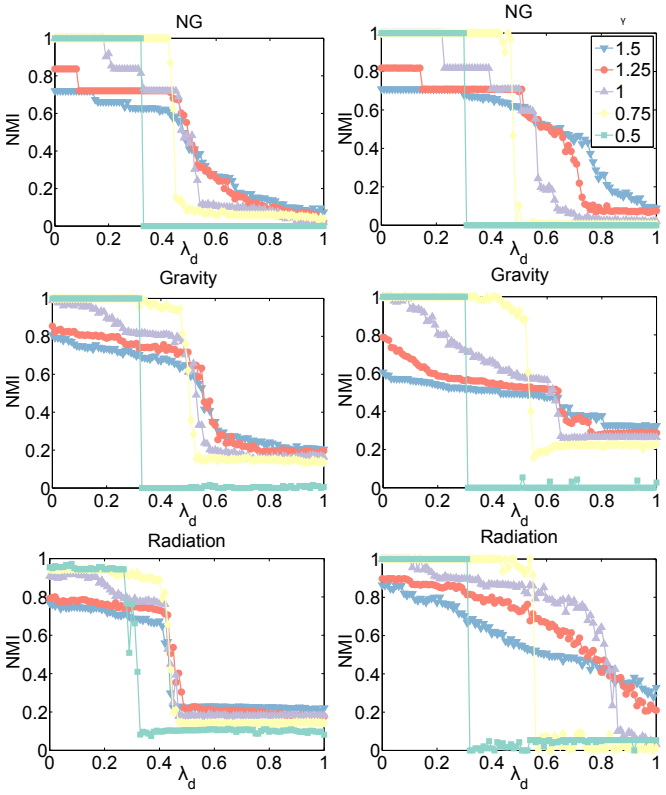


Figure 4.15: Random population multilayer benchmarks: province level community detection. Normalized mutual information (NMI) between algorithmically detected province-level community structures and planted partitions obtained from random population (where we select the population uniformly at random from $\{1, \ldots, 100\}$) multilayer spatial benchmarks with $m = 10$ layers and their respective single-layer planted partitions with $N = 50$ nodes, a grid size of $l = 10$, and $\mu = 100$ for $\omega = 0.1$ and different values of the resolution parameter γ (colored curves) as a function of the inter-community connectivity λ_d for (left) the distance benchmark and (right) the flux benchmark.

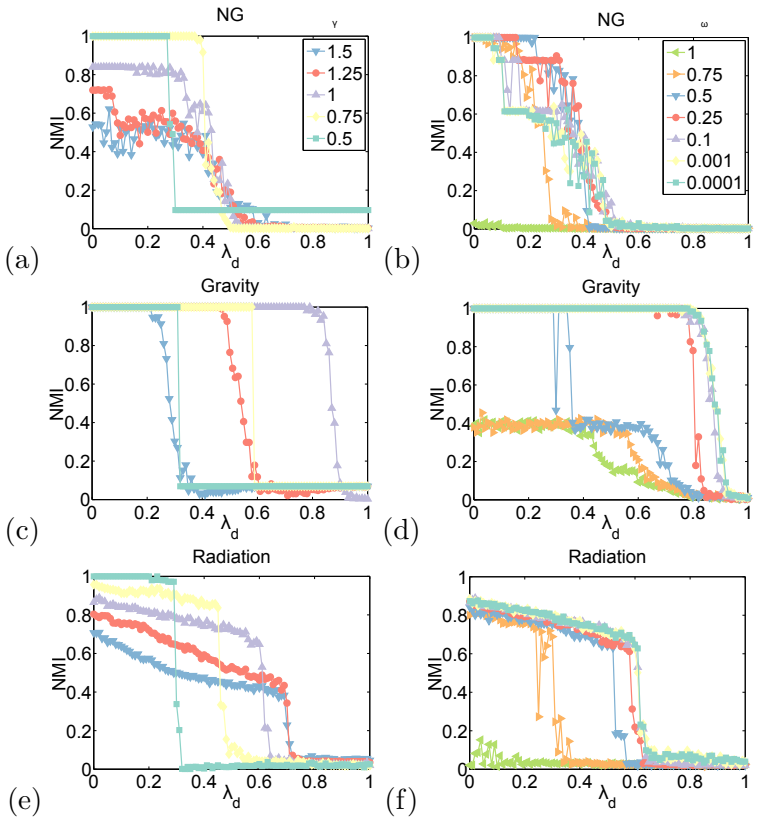


Figure 4.16: Uniform population multilayer benchmarks: temporally evolving benchmarks. Normalized mutual information (NMI) between algorithmically detected community structures and planted partitions in uniform population ($n_i = 100$) multilayer temporally evolving spatial distance benchmarks with $N = 50$, $l = 10$, $m = 10$ and $\mu = 100$ for (left) $\omega = 0.1$ and various γ resolution parameters (colored curves) and (right) $\gamma = 1$ and various ω interlayer weights (colored curves), plotted across different values of λ_d using NG (top), gravity (middle) and radiation null models (bottom).

$\omega \lessapprox 0.5$ for the radiation null model are the most similar to the planted partitions; the variation in ω appears to make less of a difference to the results using NG null model.

Finding critical ω values above which we struggle to detect planted partitions is in accordance with our expectation that algorithmically detected community structure becomes overly biased towards connecting counterparts of nodes across layers at higher ω values. From Fig. 4.16, it appears that the critical ω value depends on the null model used, suggesting that in practice one should test a wide parameter range.

4.5 Conclusions

Benchmark networks are commonly used in the community-detection literature to test new methods against established ones. However, the developments in community detection for spatial and temporal networks have been mostly driven by applications, and thus the development of benchmarks and the application of comprehensive and nuanced tests to compare the performance of null models has lagged behind the more general networks. In this chapter, we studied the outcomes of applying community detection using modularity

maximization with Newman-Girvan null model and two spatial null models on (static and temporally evolving) spatial benchmark networks.

In line with the simple investigation in the original paper that proposed a spatial null model for community detection [82], our results indicate that it is very important to incorporate problem-specific information such as spatial information into the null models for community detection. However, having tested the null models over a wider variety of benchmarks, our results also illustrate that there are many nuances to consider. That is, it is not simply a matter of incorporating spatial information in an arbitrary way but rather developing spatial null models that are motivated by application-appropriate generative models. For example, the NG null model performs better than the spatial null models (which both use population data) on the random population distance benchmark, where populations vary but edge weight does not depend on them. However, when we remove the variation in population or modify the benchmark to include population in edge placement probabilities, we find that the gravity null model performs best (as expected).

Parameter choices can also be extremely important, as demonstrated by the large influence of bin size (when binning distances for the spatial null models) on community detection results, the failure to find meaningful communities with any of the null models at low edge densities, and the strong influence of resolution parameter γ on the results.

To summarize, one needs to consider seriously what variables that influence the connections in a system of interest one wants to include in a null model, be careful about including spurious variables, and test how the results change for many parameter values. Finally, not incorporating space at all can be more appropriate than incorporating it in a manner that is overly naïve. In particular, incorporating population information into the null model for a benchmark where edge placement is independent of node population causes a community detection algorithm to fail to find the planted partitions, as observed for the random population benchmarks in Section 4.3.4.

The level of influence of different node properties or events (such as disease flux on edge placement) and the extent of mixing between communities is unknown for networks that are constructed from real data. For such networks, we recommend to try both spatial and non-spatial null models for a wide parameter range and to study the results carefully in light of any other known information about the network. In Chapters 6, 7 and 8, we will present examples of using such a procedure to study the community structure of correlation networks that are created from time series of disease cases, and in Chapter 9 we will study synthetic time series from a spatially-embedded agent-based model of disease spread that incorporates a distance-based transportation system with additional planted community partitions.

Chapter 5

The diseases, the data sets and the disease-correlation networks

This chapter consists of both a literature review and original work. Section 5.2 consists of a literature review of the diseases and the basic description and analysis of the data sets that we use. Section 5.3 and the accompanying Appendix A describe original work motivating the parameter choices for the construction of disease-correlation networks that we use in the following three chapters.

5.1 Introduction

In this chapter, we review the diseases that we study and the data sets that we use in this thesis. We present brief information about the history and importance of each disease. We review their main epidemiological properties, which can be relevant to the parameter choices that we make for generating disease-correlation networks — including the timescale of infection and the influence of external factors such as transport, climate, and urbanization level. We then present the data sets that we use in this thesis. We include a brief discussion of the distribution of disease cases in the context of the country from which the data originates. Finally, we present the disease-correlation networks that arise from each of the data sets, and we describe the reasoning for choosing parameters for network creation (time window width Δ, the step between time windows υ) and in the binning of distance data for the gravity and radiation null models. The full results of the experiments we performed to establish our parameter choices can be found in Appendix A.

5.2 The diseases and data sets

5.2.1 Endemic diseases

Endemic diseases are established in a region or population, i.e., the infectious agents (e.g., viruses or bacteria) are constantly present [217]. Some endemic diseases, such as rubella, affect many of the children present in the population, leading to an equilibrium such that

the adult population shows evidence of the disease much less commonly than children. Other diseases, such as dengue and influenza, which are composed of multiple strains with complex immunity patterns and variable prevalence in local subpopulations, can affect both children and adults.

The endemic disease data sets that we possess cover multiple seasonal epidemics each, and the patterns of infection may have changed during the time covered by them. We can use our methodology on such data sets to detect long-term patterns in disease prevalence (i.e., spatial partitions), and to detect time points corresponding to changes in the infection patterns (i.e., temporal partitions).

5.2.1.1 Dengue

The disease. Dengue is a human viral infection that is prevalent in most tropical countries. It is a vector-borne disease and is carried mainly by the *Aedes aegypti* mosquito [247]. There are four dengue strains (DENV-1–DENV-4); these viruses cause both dengue fever (DF) or the more dangerous dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS). Transmission occurs principally through a bite of an infected mosquito. Following infection, the *intrinsic incubation period* (the time between infection and symptom onset for the human host) is typically 4–7 days [117]. The *serial interval*, defined as the time between onset of symptoms in successive cases in a chain of transmission, is estimated at 15–17 days [4]. Infection with one dengue strain is often mild or asymptomatic; it produces lifelong immunity against that strain and short-term cross-protection against other strains. However, subsequent infection with another strain is usually (but not exclusively) associated with more severe disease, including DHF/DSS [8, 117].

Dengue has been known since the 10th century in Asia. It was historically associated with relatively rare but large epidemics, but its epidemiology changed in the 20th century. This change was due to the increased geographical distribution and movement of viruses leading to hyperendemicity (coexistence of all 4 dengue strains in one region), which in turn leads to recurrent epidemics [114]. Dengue is currently the second most important tropical infectious disease worldwide (after malaria) [113], affecting over 100 countries, and it has become a major public health problem, with significant economic, political, and social impact [114, 154]. The number of cases of dengue in the Americas has grown dramatically and is likely to continue growing, as more locations are becoming hyperendemic [154].

No effective antiviral agents to treat dengue infection are currently available. Due to the complex immune system response to the four strains of the disease, it has been difficult to develop a vaccine [46, 113, 247]. However, one vaccine (ChimeriVax from Sanofi Pasteur) passed its phase-3 trials in 2014, and it might be registered in dengue endemic countries in 2015 [234, 275]. At the moment however, mosquito control remains the only available method of disease prevention and control.

The spatial and temporal patterns of dengue infections Population growth, uncontrolled urbanization, and international transport (of both infected humans and mosquitoes) are thought to have contributed to the spread of dengue since World War II [113]. Transport and shipping from cities facilitate disease spread towards remote susceptible regions, causing epidemics that last until accumulation of host immunity begins to block disease flow [66, 274]. This, combined with cross-strain interactions, is thought to drive 3–5 yearly oscillations in epidemic size [63, 223].

Once the disease is imported to a new location, climatic variables affect the ability to establish repeated transmission [134]. In most endemic countries, dengue exhibits a seasonal pattern. Increased disease incidence is related to rainfall (mosquitoes often breed in rain-filled containers) and higher temperature (which shortens the time required for a newly infected female to become infectious, and the interval between consecutive bloodmeals) [86].

Human factors, such as urbanization, infrastructure, and human response to meteorological conditions, also play a role in disease persistence. Outside water containers used to deal with intermittent water supply during droughts and large rubbish items (e.g., tires) often serve as mosquito breeding sites [134]. Installing air conditioning, avoiding having stagnant water containers near buildings, and keeping doors and windows closed or screened reduces mosquito exposure. Note, however, that the adoption of these protective measures can be driven more by social or economic than by climatic factors [226].

The data set Our dengue data set consists of weekly measurements of the number of new disease cases across 79 out of 195 provinces of Peru collected by the Peruvian Ministry of Health [132] between 1 January 1994 and 31 December 2008. These data have previously been used by Chowell et al. to study the relationship between the disease attack rate and climate and population size of provinces [55].

Peru is located on the Pacific coast of South America. Its population of about 29 million people is distributed heterogeneously throughout the country. The majority live in the western coastal plain and much smaller population densities reside in the Andes mountains in the center and the Amazon jungle in the east. The climate varies from dry along the coast to tropical in the Amazon and cold in the Andes. The jungle forms a reservoir of endemic disease; it is likely that the disease spreads from there across the country in an epidemic every spring [55].

Peru is divided into 25 administrative regions, which are further subdivided into 195 provinces. Our data is gathered and analyzed mainly at the province level. Each province is classified as "Mountain", "Coast", or "Jungle", with a further classification of each of the first two into "northern", "central", and "southern" [See visualizations in Figs. 3.6(a)-

(b)]. We obtained census data, including population size per province, from the National Institute of Statistics and Informatics of Peru [132].

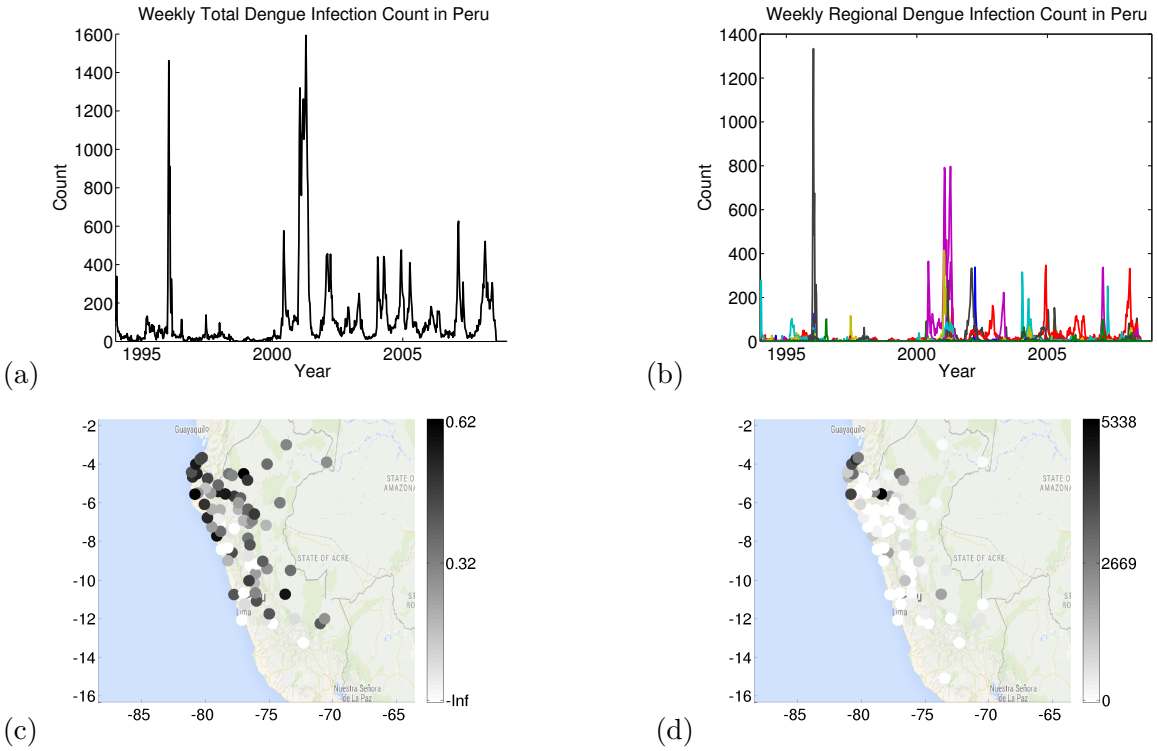


Figure 5.1: The number of dengue cases in Peru between 1994 and 2006: (a) country-wide, and (b) regional case numbers for the 79 affected regions versus time. (c) Base-10 logarithm of the total case numbers in the data set and (d) the number of cases per 100,000 people are plotted on a map of provinces.

In Peru, DENV-1 was first recorded in 1990 in Iquitos in the Amazon region [55]. Until 1995, the DENV-1 strain was the only present strain, causing mostly rare isolated outbreaks [51]. The DENV-2 strain was first observed in 1995–1996, when it caused an isolated large epidemic [149]. DENV-3 and DENV-4 entered Peru from Ecuador via the northern provinces in 1999, which led to a country-wide epidemic in 2000–2001 [187], and the hyperendemicity led to sustained yearly transmission afterwards (see Fig. 5.1). Notice that although the country-wide data give an appearance of yearly epidemics, these are largely due to a large number of disease cases affecting a small number of provinces, which are not the same year on year [see Fig. 5.1(a)-(b)].

The data set contains a total of 86,631 dengue cases that occurred between January 1994 and December 2008; most of them are in jungle and coastal provinces (47% and 49%, respectively), and only 4% of the cases occur in the mountains. The disease is present in 79 out of the 195 provinces. The number of disease cases per 100,000 people in the whole data set is the highest in the northern provinces that strongly experienced the 2000–2001 epidemic, and in the jungle [see Fig. 5.1(d)].

Despite the introduction of yearly epidemics on the country scale after 2000, we observe in our data that the subset of provinces that experience dengue epidemics changes every

year. This might be due to the complex issues surrounding immunity to the four strains of dengue and the time it takes to develop herd immunity. Local epidemics appear to be dominated by different strains in different years. For example, Iquitos experienced epidemics of dengue caused by DENV-1 in 1990—-1991 [192], an American strain of DENV-2 in 1995 [125], DENV-3 in 2001 [148], an Asian strain of DENV-2 in 2002 [192], and DENV-4 in 2008 [90].

5.2.1.2 Rubella

The disease Rubella is a childhood viral infection that usually presents with a rash and mild symptoms. Before the 1960s, it was endemic worldwide, with outbreaks every 4–7 years [30]. The discovery that rubella has an over 80% chance of causing stillbirths or birth defects during the first 12 weeks of pregnancy led to the quick development of a vaccine. Most developed countries introduced compulsory rubella vaccination in children, and rubella incidence has been almost absent there since then.

The primary epidemiologically-relevant properties of rubella are as follows: it is transmitted by aerosol and is less infectious than measles and influenza, requiring close contact for transmission. It has an incubation period of approximately 14 days, and patients can be infective for up to a week before symptoms begin and for 1–2 weeks after they disappear [17]. The serial interval is about 18 days [9, 17]. No antiviral drug is available for the treatment of rubella or to prevent congenital transmission [30]. The disease provides permanent immunity after recovery.

The spatial and temporal patterns of rubella infections To the extent of our knowledge, there does not appear to be any documented relationship between rubella spread and climate, weather, and mobility. However, this may be due to the relative shortage of studies related to this disease, which is rare in the developed world. As with any infectious disease, it is likely that factors such as human mobility, transportation, and population size and density do indeed influence its transmission [233]. Since rubella has a similar mode of infection to influenza, considerations about the seasonal changes in human behavior discussed in Section 5.2.1.3 can apply.

The data set The rubella data set consists of 673 time series recording the weekly number of new cases between 1 January 1997 and 31 December 2009, gathered by the Peruvian Health Ministry's Department of Epidemiology [132]. The data set contains 24,116 cases across 175 provinces; the remaining 20 provinces noted no cases in the period covered by our data. The data set contains multiple epidemics of the disease which occur roughly on a yearly basis. Two of the epidemics are much larger than the others, in 2000–2001 and 2005–2006.

Routine vaccination against rubella was only initiated in Peru in 2003–2005, but coverage during this period was estimated at less than 4% and is not thought to influence the disease spread in our data [55]. Enhanced routine vaccination was implemented in 2007, and the subsequent coverage ranges between 75% and 100% [55].

Country-wide outbreaks of rubella follow a predominantly annual pattern (see Fig. 5.2). At a finer scale, however, rubella incidence is variable in space and time, and the highly populated central regions of Peru have been affected by the disease more than other regions, both in terms of the total number of disease cases, and in the number of disease cases per 100,000 people [see Fig. 5.2 (d)].

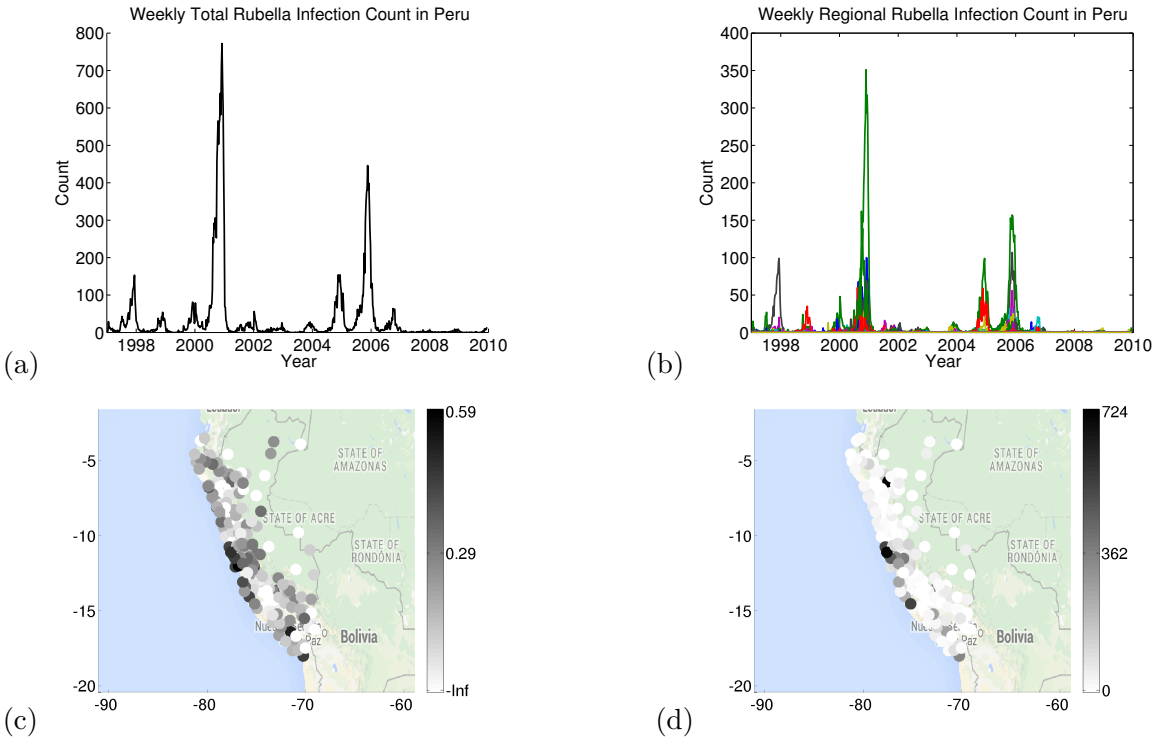


Figure 5.2: The number of rubella cases in Peru between 1997 and 2009: (a) country-wide, and (b) regional case numbers versus time. (c) Base-10 logarithm of the total case numbers in the data set and (d) the number of cases per 100,000 people are plotted on a map of provinces.

5.2.1.3 Influenza

The disease Influenza (commonly known as the flu) is a respiratory disease that has been known to humans for centuries. Influenza epidemics have often resulted in high rates of mortality: the 1918–19 pandemic killed about 50 million people [137], and even annual influenza kills about 35,000 people per year in the USA alone [263]. Although the impact of the disease is usually strongest on the young and the old, some strains of the disease impact adults as well.

The disease is caused by an RNA virus that is capable of fast evolution into new strains, crossing over between species and mixing between existing strains [87]. Consequently, a new

influenza vaccine that is specific to the popular strains for a given flu season is developed every year. This also fuels the recurring public fears of new, potent viral strains causing a large, fast-spreading global outbreak (a pandemic [190]).

The primary epidemiologically-relevant properties of influenza are as follows [87, 237]. The influenza virus can be transmitted through droplet, contact, and airborne transport of respiratory particles created by coughs and sneezes. The disease can be difficult to control, as transmission can begin within one day of being infected (a day before symptoms appear), and continues for 3–5 days afterwards. Symptoms usually include fever, runny nose, headaches, coughing, fatigue, muscle pain, and other illness. The serial interval is about 4 days [87]. Recovery takes about one week and leads to permanent immunity to the infecting strain and closely related viral strains. However, individuals remain susceptible to other strains of influenza.

In most cases, influenza gets better with or without treatment. Antiviral drugs can speed up recovery, but they are prescription-only and are only used in select situations, such as people with chronic medical conditions who may be at increased risk for complications [87].

The spatial and temporal patterns of influenza infections In temperate climates, influenza epidemics exhibit a distinct seasonality, with widespread infection during winter. In tropical regions with a rainy season, epidemics occur during the wet months. The disease shows a relative lack of seasonality in regions with little variation in temperature and precipitation [256].

The traditional explanation for the seasonality of influenza in temperate climates is the increased crowding and mixing during indoor activities prompted by the cold winter temperatures [95]. New results also suggest that the mode of disease transmission can be influenced by seasons, with the more infective airborne route dominant in cold temperatures [170, 179]. Further, host susceptibility might be increased during cold weather due to weakened immune system and exposure to other pathogens [95].

Traditionally, large-scale models of influenza spread assume that social factors such as transport, social networks, and contact patterns are the strongest influences on the spread of influenza [112]. However, some studies have suggested influenza epidemics might also be associated with the cold phase of El Niño-Southern Oscillation (ENSO) — the dominant mode of climate variability over the Pacific Ocean [133].

The influenza-in-Chile data set The Chilean influenza data set consists of weekly numbers of new hospitalized influenza-like illness (ILI) cases from the 15 regions of Chile over 365 weeks in the period between 1 January 2004 and 31 December 2010 obtained from a national surveillance system [56, 185].

A total of 61,393 cases were observed, with a maximum of 720 cases in any one week. The data set appears to show clear yearly epidemics in all years except 2006; it also shows very strong influenza activity due to the H1N1 ("swine flu") pandemic in 2009 (see Fig. 5.3). The majority of cases are located in the relatively highly populated central regions of the country. The northern regions exhibit large numbers of cases in some years (e.g., 2006) but not in others (e.g., 2004). The southern regions tend to get fewer cases than the centre and north of the country. The pattern is similar in terms of number of cases per person.

Chile covers a long, narrow strip in the southwestern part of South America between the Andes mountains to the east and the Pacific Ocean to the west. The climate ranges from desert in the north and Mediterranean in the centre, to rainy temperate in the south. The population of Chile is about 17 million and about 40% of the population is concentrated in the central region surrounding the capital, Santiago, which also experiences the highest number of disease cases per 100,000 people.

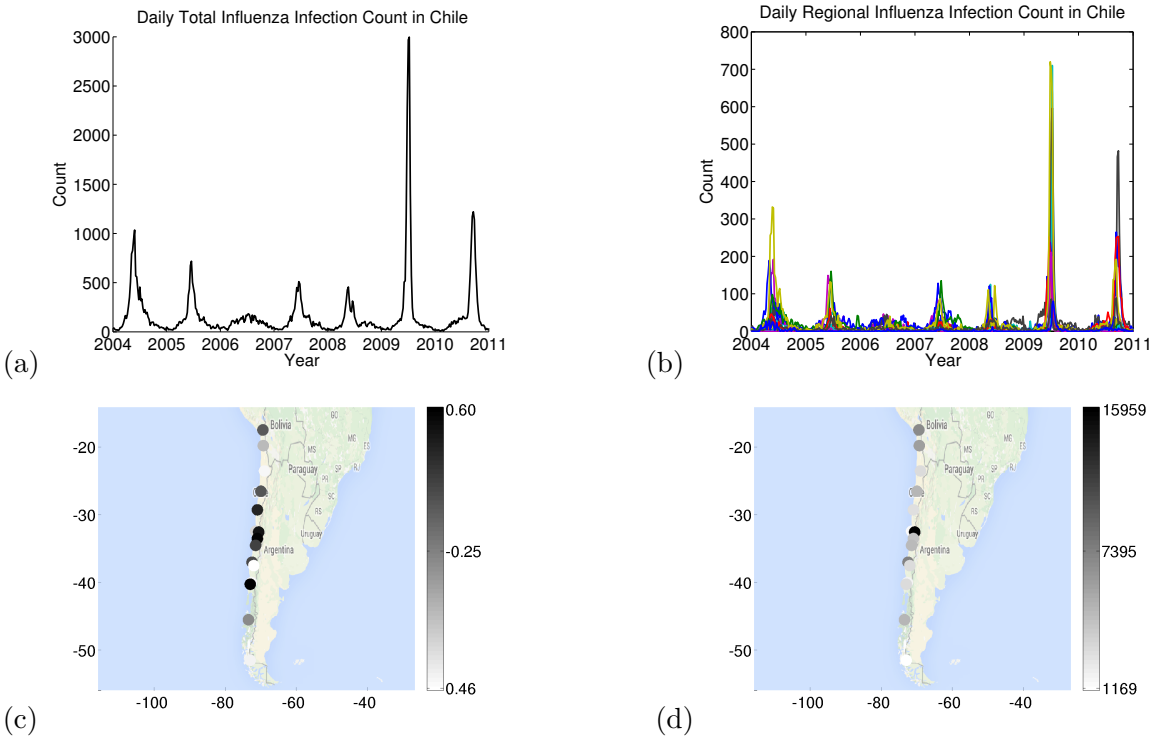


Figure 5.3: The number of influenza cases in Chile between 2004 and 2010: (a) country-wide, and (b) regional case numbers against time. (c) Base-10 logarithm of the total case numbers in the data set and (d) the number of cases per 100,000 people are plotted on a map of provinces.

5.2.2 Emerging diseases

Emerging diseases can refer to diseases that have been newly discovered, or ones that are novel in their outbreak ranges (geographic and host) or transmission mode [217]. The 2014 West African epidemic of Ebola virus disease (which we hereafter refer to as "Ebola" for conciseness) is the first time that the disease caused an outbreak in this region, and the

largest epidemic of this disease to date [88]. In contrast, the H1N1 influenza is a novel strain that emerged in Mexico in 2009, and it is thought to have originated in pigs [101].

The emerging disease data sets that we possess each consist of the time course of one epidemic that entered a susceptible population. We can use our community-detection methodology to study patterns in the initial spread of the epidemic, and examine the factors that influence them.

5.2.2.1 Ebola virus

The disease Ebola virus disease is a severe, often fatal, viral infection discovered in 1976 [27]. It is a zoonotic disease that is thought to originate from bats or other jungle mammals [176]. Most of the epidemics to date have occurred in small, remote communities and have been contained through a combination of early diagnosis, contact tracing, patient isolation and care, infection control, and safe burial [261]. The year 2014 saw the beginning of by far the largest epidemic of the disease to date. The epidemic covered most of Guinea, Liberia, and Sierra Leone, and for the first time Ebola reached densely populated cities, causing enormous challenges to the affected countries and eventually attracting an international response [261].

The main epidemiologically-relevant properties of Ebola are as follows [27]: human to human transmission of Ebola occurs through contact with bodily fluids from infected patients. The incubation period is usually between 5 and 9 days and patients are not considered infectious until they develop symptoms. The serial interval is about 15 days. There are typically three phases of illness: a few days of non-specific fever; a gastrointestinal phase, followed by either recovery or deterioration; with a subsequent phase (which is often fatal) that includes collapse, neurological manifestations, and bleeding. Case fatality rates range from 30% to 90% between epidemics [27].

There is no cure or vaccine for Ebola at the moment; treatment is currently focused on supportive care and infection control. Although experimental treatments are under development, they have not yet been fully tested for safety or effectiveness.

The spatial and temporal patterns of Ebola infections Transmission of the Ebola virus from animal reservoirs to human populations may be influenced by climate and spatial factors [5]. Some studies report an association with drier than normal conditions at the end of the rainy season, a situation that limits food availability and influences the population density and species distribution of wildlife [215]. However, other studies reported that outbreaks often begin during the transition from the rainy to dry seasons [25] or are associated with high humidity and low temperature [206].

Once Ebola has entered the human population, transmission is largely governed by personal contacts and behavioral factors (e.g., burial practices and adherence to safety

protocols) and the availability of safe care environments. As in case of the other infectious diseases that we discuss in this thesis, commuting patterns and migration routes are potentially linked to the spread patterns [5, 233].

The data sets We possess two data sets related to the Ebola epidemic in Western Africa.

The first data set was collected by the World Health Organization (WHO) and we obtained it from G. Chowell (Arizona State University). It contains weekly new case counts (a total of 15,878 cases) collected between 5 January 2014 and 11 January 2015 (54 weeks) in 63 provinces of Guinea, Liberia and Sierra Leone. The second data set comes the Office for the Coordination of Humanitarian Affairs in West & Central Africa. It was collated, cleared, and added to a database hosted on the Datamarket website by volunteers [70, 88]. The Datamarket data set consists of the total number of new cases per province, collected daily for 150 days starting on 24 March 2014.

The general outline of the epidemic agrees between the two data sets. The numbers of cases rose slowly until about July–August 2014, when the disease reached the most populated coastal provinces and the numbers of new infections shot up rapidly. The numbers of new cases being reported started falling by November 2014 (see Fig. 5.4 and Fig. 5.5). The broad spatial patterns of the number of cases are also similar for both data sets, with the majority of cases in the most populated provinces. Furthermore, Sierra Leone experienced a rise in the number of new disease cases later in the data set than Guinea and Liberia.

The WHO data set consists of 63 provinces and 54 weeks (see Fig. 5.4). It is shorter than recommended for correlation matrix creation from the random matrix theory point of view (which recommends that $T > \hat{N}$ where T is the time series length and $\hat{N}$ is the number of nodes with non-zero strength in the network or layer, as discussed in Section 3.1.1). This means that it is unclear how much information versus noise a correlation matrix generated from this data would hold. However, the WHO does not release more detailed data to everyone, and the temporal constraints on this project made it impossible to acquire clearance to use detailed data in time.

The Datamarket data set consists of 63 provinces and 150 data points describing the total case number up to each time point. The data points are recorded daily, however, in the first 45 days, many data points are missing and there are large increases in the total number of cases covered in the gaps. We decided to use the 105 data points after day 45 (16 August 2014) to increase the reliability of the data. Even within this shorter period, there are also a few clearly wrong data points. (1) The total number of disease cases in the original data set falls on some days, only to go back up again on the next day. We removed these data points from the data set. As a result, the remaining data set had a small number of missing data points, which we interpolated through cubic interpolation. We considered this simple method a good choice as we only want to connect two data points describing the

total number of cases before and after a missing input. (2) The original data set contained a very large spike in the daily new number of cases in two provinces in December 2014. After comparing this to the WHO data set and not finding such a spike, we assumed that it was a clerical error on the part of one of the volunteers and removed this data point for the two provinces affected and replaced it by a reduced value that corresponds to a daily increase by the mean of the daily case numbers over 14 days covering the preceding and following week. The modified data set contains a total of 13,714 cases (see Fig. 5.5). We calculate the number of daily new cases in the modified data set, and use this to create disease-correlation networks — in line with the methodology for all other data sets.

As we trust the new case numbers in the WHO data set more than in the volunteer-collated Datamarket data set, and we are interested in studying the early period when the disease was first spreading between provinces that is not available in the Datamarket data set, we will perform the community detection experiments on the WHO data set (in the format of a single static network) despite the fact that the length of the data is not optimal. We will also use the Datamarket data set (in the format of both static and multislice networks), as it enables us to identify the disease patterns in more detail thanks to its higher frequency of data collection.

The data quality issues that we faced with the two Ebola data sets showcase common issues with obtaining reliable disease data. Such issues are especially pronounced for emergent diseases, unexpected epidemics, and epidemics in developing countries with little health infrastructure. All of these factors influence our ability to obtain reliable data about the West African Ebola epidemic. Consequently, our results need to be treated with caution, as the WHO data set was shorter than suggested for correlation matrix creation, and the Datamarket data set has been manipulated arbitrarily. However, this investigation is useful for any future emerging diseases that we may want to investigate using this methodology.

5.2.2.2 "Swine flu" (H1N1 influenza)

A *pandemic* influenza usually refers to a novel, rapidly spreading, and highly infections strain that reaches a large spatial region quickly and poses a danger to the susceptible population (although there is no agreement on a formal definitions) [190]. The most recent influenza pandemic scare occurred in 2009, when a new virus, influenza A subtype H1N1 (colloquially known as "swine flu" due to its porcine origins), was isolated in Mexico in April. In light of the rapid spread of the virus all over the world and the apparent large number of children and young adults affected, the World Health Organization (WHO) raised a pandemic alert in June 2009 [101]. The H1N1 strain caused an unusual amount of influenza activity in 2009, especially in the northern hemisphere. It is estimated that about 11–21% of the population were infected [126], which is below the early modelling predictions of H1N1 infecting up to

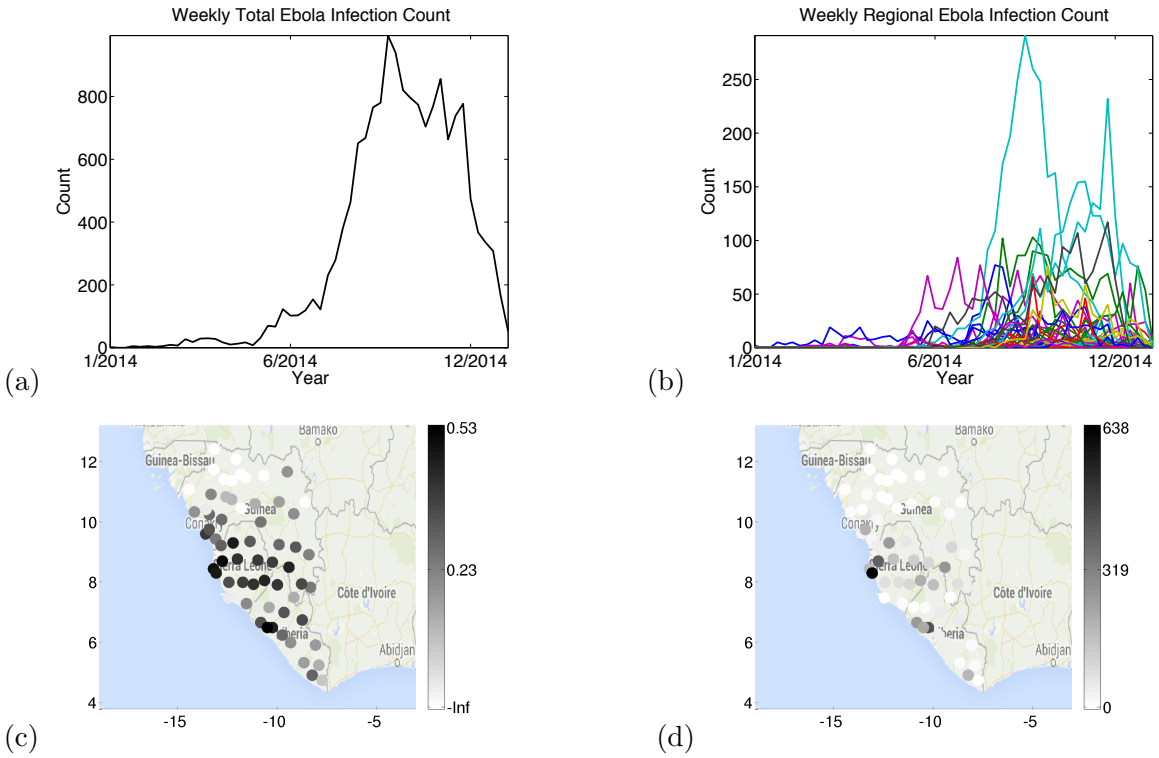


Figure 5.4: The number of Ebola cases in Western Africa in the 2014-2015 epidemic, from the WHO data set: (a) country-wide, and (b) regional case numbers against time. (c) Base-10 logarithm of the total case numbers in the data set and (d) the number of cases per 100,000 people are plotted on a map of provinces.

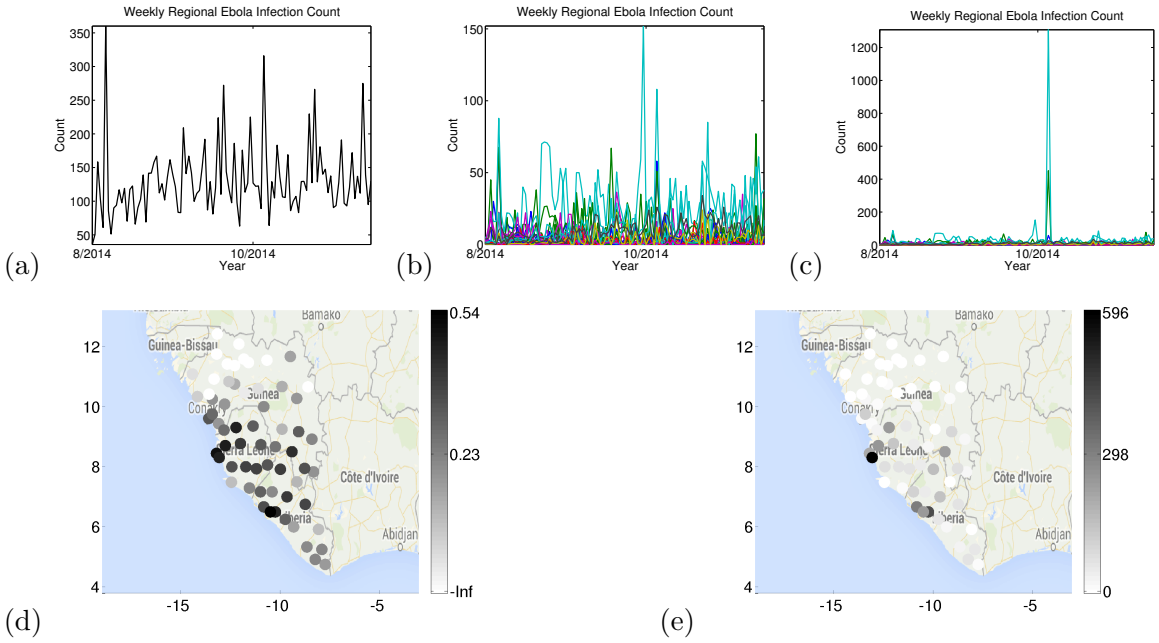


Figure 5.5: The number of Ebola cases in Western Africa in the 2014-2015 epidemic, from the Datamarket data set: (a) corrected country-wide, and (b) corrected regional case numbers against time, and (c) original case numbers before cleaning the data. In (d), the plot of base-10 logarithm of the total case numbers in the data set, and in (e) the case number per 100,000 people are plotted on a map of provinces.

45% of the population [101]. The H1N1 strain accounted for a large percentage of seasonal influenza infections in the following years, and it remains in circulation [257].

The transmission mode of H1N1 influenza is the same as for seasonal flu, although it has a slightly shorter serial interval of about 3 days [36]. It is mostly a mild, self-limiting upper respiratory tract illness [101]. Approximately 2–5% of the confirmed cases require hospitalization and antiviral treatment (mostly pregnant women and patients with underlying conditions).

Following the development of a vaccine, immunization against H1N1 influenza is now possible, though mass immunization projects were abandoned in late 2009 due to the lower than expected levels of infection [101].

H1N1 influenza shows similar spatial and temporal incidence patterns to seasonal influenza described in Section 5.2.1.3.

The H1N1-in-Mexico data set The H1N1 data consists of daily numbers of new hospitalized influenza-like illness (ILI) cases from the 32 states of Mexico over 1080 days in the period between 1 April 2009 and 28 February 2012. The data set was compiled by the Mexican Institute for Social Security (IMSS), and it covers 40% of Mexico's population. For this study, we select the first 430 days of data (April 2009–June 2010) due to the long break in disease occurrence in 2011–2012 that separates this period from the second epidemic in late 2012. This corresponds to 4,468 of the total 5,618 cases.

Mexico experienced a series of three H1N1 pandemic waves in the spring, summer, and autumn of 2009, followed by a large pandemic vaccination campaign towards the end of 2009 and a sharp fall in the number of new cases. The country also experienced a fourth wave in winter 2011–2012 [52,53] [Fig. 5.6(a)]. The three waves of the 2009 epidemic were spatially distinct: the first wave mostly affected central regions, the second wave was strongest in the southeast, and the third and fourth waves were geographically widespread [53].

Mexico is divided into 31 states and a Federal District. (We henceforth refer to them as 32 provinces for the sake of naming clarity.) Mexico has a surface of 2 million km^2 and a population of about 107 million people, who are heterogeneously distributed across the country. The climate ranges from tropical jungle in the southeast through temperate, rainy climate in the centre, to steppes and deserts in the north.

The majority of the population lives in the centre of the country, and the northern desert and southern rainforest have a much lower population density. The desert and jungle regions might be expected to act as natural barriers to the spread of disease because of their lower populations.

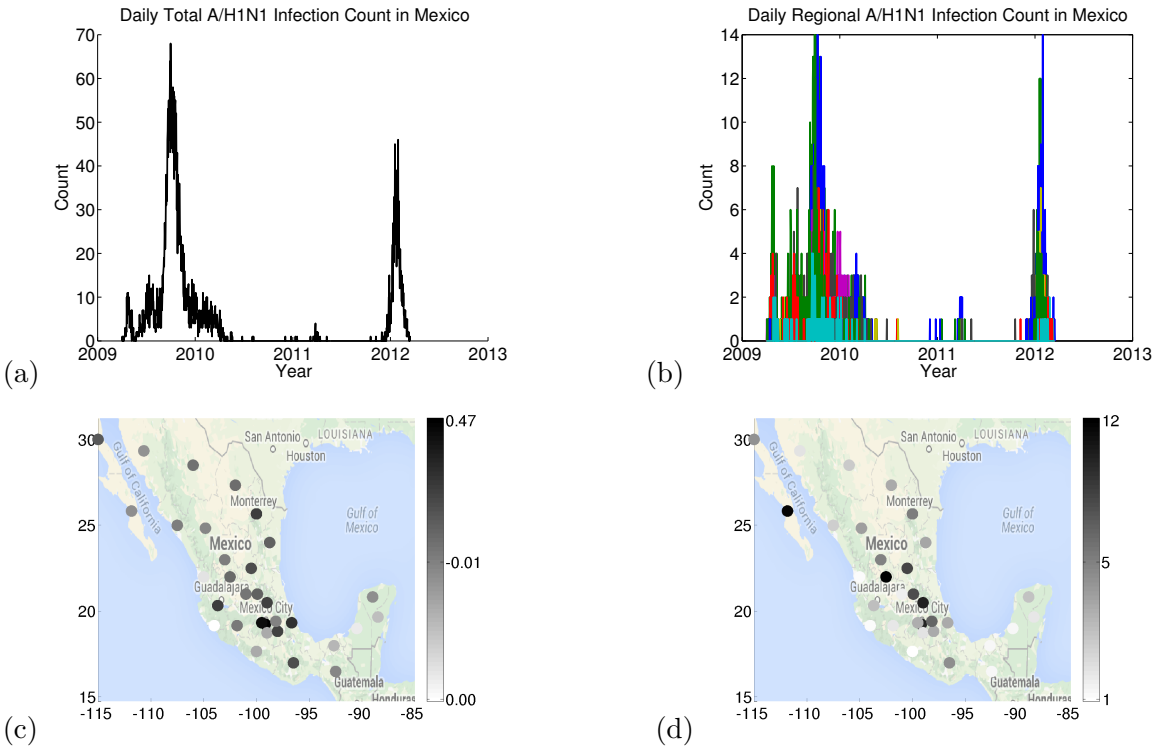


Figure 5.6: The number of influenza cases in Mexico: (a) country-wide, and (b) regional case numbers against time. (c) Base-10 logarithm of the total case numbers in the data set and (d) the number of cases per 100,000 people are plotted on a map of provinces.

5.3 Parameters for constructing disease-correlation networks

In this section, we discuss the reasoning behind choosing parameter values for network construction (time-window width Δ and the distance between starting points of layers υ) and community detection (bin width b for spatial null models).

To inform the choice of time-window width Δ, we investigate the effects of varying it on the properties of the windowed time series used for creating the network, and in turn on the disease-correlation networks. We want Δ to be high enough that the correlation matrices contain meaningful information, but low enough that the properties of disease time series and disease spread between provinces are smoothed as little as possible. The suggested minimum value for Δ can be based on the random matrix theory constraint described in Section 3.1.1 ($\Delta > \hat{N}$). For each of the disease data sets, we present the minimum width of the time window resulting form this constraint, and two longer time window widths. We study the the effect of the time window choices on the disease time series by examining the mean number of cases in the time window. We also study various properties of the disease-correlation networks constructed using the pipeline in Section 3.1.1 with each time window choice, and we use these results to inform our choice for the community detection experiments in Chapters 6, 7and 8.

The full results are presented in Appendix A. When we increase time window width Δ,

the features of the curve representing the mean number of cases within the time window become smoother, and as a result disease epidemics are increasingly difficult to distinguish for the endemic diseases, and the epidemic curve flattens for the emerging diseases. After experimenting with several values for time window widths, we chose the smallest Δ that complies with the requirements imposed by random matrix theory to generate meaningful correlation matrices. The resulting parameter choices are summarized in Table 5.1 below.

The choice of the difference between layer starting points, υ is mostly constrained by the fast increase of computational complexity with increased number of layers. Having a low υ increases computational complexity and makes data analysis difficult by generating a large number of structures to study. We thus often take υ of at least 4 weeks for the data sets that are collected weekly, and at least 7 days for the data sets that are collected daily. For the multislice networks the computational complexity becomes more of an issue, as the size of the adjacency matrix increases with each layer by the number of nodes present in the layer, and it becomes difficult to store and manipulate the matrix on a computer. For this reason, we choose larger values of υ for the multislice networks (see Table 5.1).

We investigate the influence of having overlapping versus non-overlapping time windows for the dengue data set in Section 6.6. For the numerical experiments on the remaining disease data sets in Chapters 7 and 8, we decide whether to use overlapping networks on a case-by-case basis depending on the comparison of the serial interval of the disease with the temporal resolution and length of the data set, as well as comparing the data set length with the number of provinces, and thus taking into account the size of the adjacency matrix and the computational complexity of the community detection. We use overlapping layers for rubella and Ebola multislice networks, and we use non-overlapping layers for influenza (both H1N1 and seasonal). Finally, for the Ebola Datamarket data set we choose a low υ of 5 days despite the serial interval of the disease being 21 days, because only low υ values allow us to generate a multislice network for this short data set. The parameter choices are summarized in Table 5.1.

We also investigate the influence of the spatial bin width b on the deterrent function $f(d)$ for the gravity null model (see discussion in Section 3.3.3). This bin size, which we measure in kilometers for the disease-correlation networks, determines the groupings of nodes for the deterrent function and for the null model. We investigate several options for each data set, and we use them to guide our final parameter choice or bin width b. After examination of results presented in Appendix A.4, we select the smallest bin sizes that ensure a minimum of 5 province pairs in each bin.

5.4 Conclusions

Although all of the diseases that we study are viral infections affecting humans, there are several differences between them that are important for choosing parameters when

Table 5.1: The parameter choices for time window width Δ and the step between consecutive time windows υ for all diseases, and the resulting networks: the number of static networks and the number of layers in multislice networks. N/A — not applicable.

	Static			Multislice		
Disease	Δ	υ	No. networks	Δ	υ	No. layers
Dengue	80	4	175	60	60	12
Dengue overlapping	N/A	N/A	N/A	80	24	30
Rubella	134	4	132	134	12	44
H1N1	30	7	58	30	30	14
Influenza	30	4	83	30	30	11
Ebola WHO	53	N/A	1	N/A	N/A	N/A
Ebola Datamarket	60	5	9	60	5	9

we generate disease-correlation networks and study their community structure. These are summarized in Table 5.2.

Property	Dengue	Rubella	H1N1 influenza	Seasonal influenza	Ebola virus
Transmission	mosquito	droplet	droplet	droplet	body fluids
Serial interval (days)	15–17	18	3	4	15
Origin of data	Peru	Peru	Mexico	Chile	West Africa
Number of provinces in data	79	195	32	15	63
Number of time points in data	780	673	1080	365	53/105
Measurement frequency	weekly	weekly	daily	weekly	weekly/daily
Epidemics	yearly	yearly	1	yearly	1

Table 5.2: Most important properties (for our purposes) of the diseases and data sets that we study in this thesis.

First, the different types of data force us to take slightly different approaches to generating and analyzing disease-correlation networks. The endemic disease data sets and the Chilean influenza data set consist of weekly measurements over many years, and they include at least two country-wide epidemics. In contrast, the emerging disease data sets contain measurements over a shorter period of time that covers the development of one large epidemic. This potentially enables a more detailed analysis, and this kind of data set may contain stronger spatial effects due to the expected large influence of transport on the infection patterns in the early stages of epidemics.

Spatial effects other than the influence of transport can also affect the spread of disease. Dengue, rubella, and influenza are more prevalent in urban settings than in rural ones, due to the increased interpersonal contact. Further, mosquitoes that spread dengue don't survive well in the mountains, leading to the smaller number of dengue cases in the mountainous regions of Peru and the mountains possibly acting as a barrier to the epidemic. In contrast, Ebola historically has mostly occurred in rural areas. During the West African epidemic

covered in our data set, it has reached the highly-populated coastal cities and it has since been more prevalent in the urban settings than the rural ones. However, the influence of urban vs. rural location on the transmission of Ebola virus remains unclear due to the close interpersonal contact needed for bodily fluid transmission.

The five diseases also have different time scales of infection: the serial interval (the time between two consecutive cases in a chain of infection) is about 2 weeks for dengue, rubella and Ebola, and only 3 days for influenza. This affects the minimal time resolution of data that is needed to effectively capture the spread of the diseases. In particular, we ideally want daily data for influenza in Chile, as it is a fast-spreading disease — which is a potential problem for this data set. Additionally, smaller time window widths are preferred for the creation of disease-correlation networks from the faster-spreading diseases. This is in order to capture short-term features of the epidemic patterns.

In Chapters 6, 7 and 8, and in Appendices B and C we will study these disease-correlation networks by performing community detection using Newman-Girvan, spatial, and correlation null models and we will then examine the structures of the resulting communities in context of the diseases and the data sets. In addition to using disease information to inform the construction of networks from disease data, we use knowledge about the disease types, epidemiological properties, and the external factors that are known to influence their spread when analysing the disease data sets through community detection.

Chapter 6

Application to dengue data

This chapter consists of original work, most of which appears in a working paper by MS, E. Leicht, G. Chowell, and M. A. Porter [235].

6.1 Introduction

In this chapter, we assess the performance of the Newman-Girvan (NG), gravity, radiation, and correlation null models (defined in Section 3.3) on static and multislice correlation networks that we construct from disease incidence data describing the spatiotemporal spread of dengue in Peru.

The dengue data set contains weekly new case count data from 79 of the 195 provinces of Peru over 15 years (780 weeks). The data show small isolated epidemics between 1994 and 2000, a large epidemic in 2000–2001 and subsequent yearly epidemics (see Fig. 5.1). The disease and the data set are described in detail in Section 5.2.1.1.

We generate sets of static networks and a multislice network from the dengue incidence time series using the procedure described in Section 3.1.1. Unless stated otherwise, for the (overlapping) static networks we use time window width $\Delta = 80$ with distance between time window starting points $\upsilon = 4$. We select the time window width to be long enough to obtain meaningful correlation matrices while preserving interesting disease patterns. We choose to start a new time window every 4 weeks rather than every week in order to lower the computational complexity of the calculations. Unless stated otherwise, we use $\Delta = 60$ with $\upsilon = 60$ for the non-overlapping multislice network (which then never has more than 59 nodes that experience disease and thus have non-zero strengths in any one layer). This is the minimum time-window width that we can have while complying with the $\Delta > \hat{N}$ constraint (where $\hat{N}$ is the number of nodes with non-zero strength in the network or layer) suggested by random matrix theory for obtaining meaningful correlation matrices — as discussed in Section 3.1.1. The reasons for parameter choices and the general properties of the time series and the static networks were described in Section 5.3 and Appendix A.

It is well-known that geographical distance has an important influence on disease spread [253, 260, 271]. The dynamics of vector-borne diseases such as dengue are strongly influenced by spatial factors, as the distance between regions is expected to affect the migrations of both humans and mosquito vectors [138]. Additionally, climate exerts a significant influence on the abundance of mosquito vectors and as a result on dengue infection patterns, and it is also necessary to consider Peru's particular topography (as its mountains form a barrier to disease spread) [51,55]. Therefore, we expect community structure in the disease-correlation networks to be strongly geographical. We also expect to observe large changes in community structure at certain times — such as when the introduction of new disease strains around 1999 led to a large countrywide epidemic in 2000–2001 and the onset of yearly countrywide epidemics thereafter [51].

In the following sections, we explore the similarity of algorithmically obtained community structures to spatial and temporal partitions of nodes for a range of null models, and for different parameter values.

6.2 Community structure using the NG null model

In this section, we use the most popular null model for community detection, the Newman-Girvan (NG) null model that we defined in Section 3.3.1, to study the disease correlation networks generated from the dengue data set using modularity maximization for a range of resolution parameters $\gamma \in \{0.1, 0.2, \ldots, 3\}$.

6.2.1 Static networks

We first study the community structures of the overlapping static networks formed by taking $\upsilon = 4$ and using $\Delta = 80$. We select the networks for which the algorithmic partitions score the highest versus manual spatial partitions of the network, as defined in Section 3.4.3, and ones that score as statistically significantly spatial in the distance and MST tests, as defined in Section 3.4.4.1, for detailed examination.

The community structures that we obtain from maximizing modularity have a strong spatial organization, as suggested by the high z-Rand scores when compared to topographical partitions. As one can see in Fig. 6.1(a), which shows a plot of the z-Rand scores versus climate partitions for resolution parameter values of $\gamma \in \{1, 1.1, 1.2, 1.3\}$ for each of the static networks, the spatial organization is especially evident for networks 70–100 (during the 2000–2001 epidemic).

This transition seems to occur around the time of the largest countrywide epidemic in the data, and the subsequent period includes recurring yearly epidemics that have been linked in prior studies to climatic patterns [51]. There are two periods of significantly spatial partitions ($z_R > 1.96$): one corresponds to the 2000–2001 epidemic, and it contains the spatial partition with the highest z-Rand score against climate [see Fig. 6.1(b)]; the

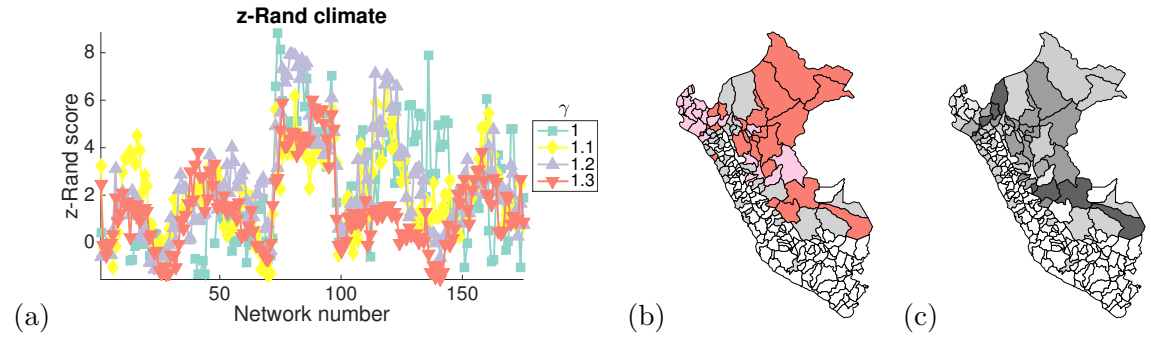


Figure 6.1: Dengue, static networks, NG null model — spatial partitions according to z-Rand scores. Properties of the algorithmic community structure: (a) Plot of the z-Rand scores versus the detailed climate partition for the static networks covering the whole time period (horizontal axis) for $\gamma \in \{1, 1.1, 1.2, 1.3\}$. (b) Community structure with the highest z-Rand score when compared to the climate partition. The resolution-parameter value is $\gamma = 1$, the network number is 73 (December 1999), and the z-Rand score is 8.85. (c) Community structure with the highest z-Rand score when compared to the administrative partition. The resolution-parameter value is $\gamma = 1.2$, the network number is 113 (October 2002), and the z-Rand score is 8.76. Visualizations in panels (b)-(c) use a map of Peru in which we color provinces according to their community assignment. White provinces are ones in which our data does not include any reported cases of dengue fever in the indicated time window.

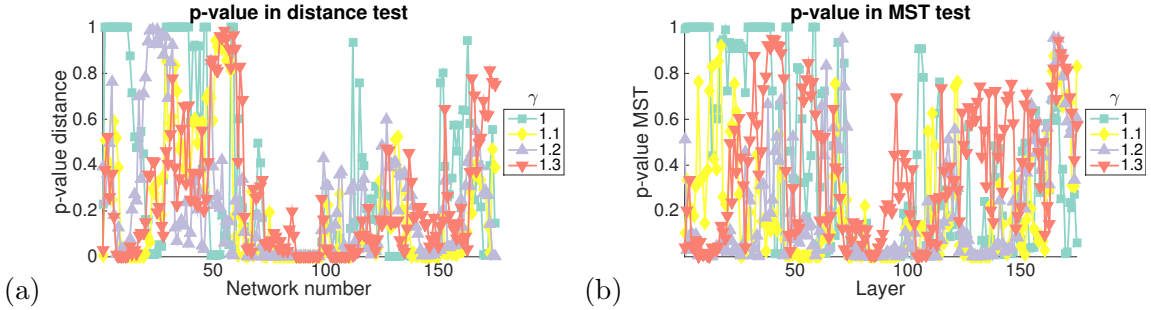


Figure 6.2: Dengue, static networks, NG null model — spatial partitions according to distance and MST tests. Properties of the algorithmic community structure for the static networks covering the whole time period (horizontal axis): (a) plot of the p-value in the distance test and (b) plot of the p-value in the MST test for $\gamma \in \{1, 1.1, 1.2, 1.3\}$.

second occurs in 2002–2004, and it contains the spatial partition with the highest z-Rand score against administrative partition [see Fig. 6.1(c)]. Note that the topographical z-Rand scores decrease after 2004 despite the continuing yearly dengue epidemics.

In Figs. 6.1(b)-(c), we plot the partitions that have the highest z-Rand scores with respect to the manual climate and administrative partitions. We observe that the high-scoring climate partition consists of one community that is dominated by the jungle (red community) and another community that is dominated by the coast (pink community), whereas the high-scoring administrative partition is composed of 7 smaller communities. The jungle nodes form the largest communities in both of these spatial partitions. There was a dengue epidemic in most of the provinces in these large communities during the time periods covered by the relevant networks. It is possible that their proximity and relatively stable climate drove the large amount of synchrony in the epidemic spread in these provinces.

We then study the spatial organization of the static networks using the alternative dis-

tance and MST Monte Carlo measures of spatial clustering that we defined in Section 3.4.4. These measures attempt to identify partitions with statistically significant spatial clustering without comparing them to manual partitions.

In Fig. 6.2 we observe that both of the measures detect spatial clustering in the same period as the highest z-Rand scores — roughly between networks 70 and 100, corresponding to the large countrywide epidemic. Network 91 (January 2002) scored the most consistently as significantly spatial ($p < 0.05$ after Bonferroni correction for multiple comparisons) on both tests, with $\gamma = 1$ achieving the lowest score across the comparisons. However, we see that there are whole sections of the network, notably between networks 120 and 140, where z-Rand scores are significant but the distance and MST measures are not.

The partition in network 91 for $\gamma = 1$ consists of 2 communities of similar sizes (see Fig. 6.3). Nodes in community 2 have slightly different disease patterns than community 1, with a period of low infection numbers preceding a large epidemic [see Fig. 6.3(b)].

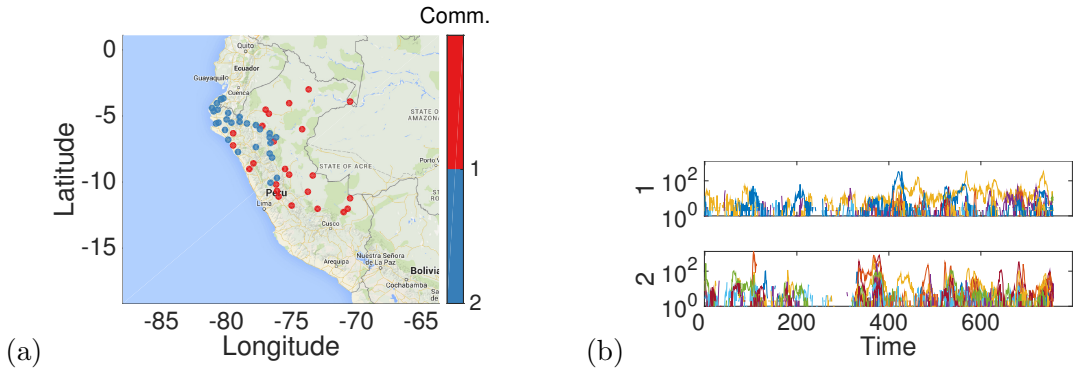


Figure 6.3: Dengue, static networks, NG null model — spatial partitions for $\gamma = 1$, network 91. In (a) we show a map of all the nodes colored by algorithmically detected communities for ($\gamma = 1$, network 91), and in (b) we show the time series of disease occurrence in these communities.

6.2.2 Multislice networks

We now consider community structure in the multislice disease network with non-overlapping layers that we construct using time window width $\Delta = 60$ and time between starting points $\upsilon = 60$. To find interesting parameter values, we compare the algorithmically computed community structure of the dengue multislice disease-correlation network to manual partitions across a range of $\gamma \in \{0.1, 0.2, \ldots, 3\}$ and $\omega \in \{0.1, 0.2, \ldots, 3\}$. For $\gamma \lessapprox 1$, all nodes are in one community.

For $\gamma \in [1, 1.8]$, the partitions score as significantly spatial in both z-Rand scores and in the distance test for spatial organization [see Fig. 6.4(a)-(b)]. The partitions for $\gamma \in [1, 1.2]$, and $\omega \lessapprox 1$ appear the most interesting, with relatively high z-Rand scores when compared to the temporal partition as well [see Fig. 6.5(a)]. These partitions exhibit a mixture of spatial and temporal features — see Fig. 6.4(c) for an example using $\gamma = 1$ and $\omega = 1$. This partition scores high in spatial (distance and MST) tests, especially around layers

6–7 (2001) and 10–11 (2006) [see Fig. 6.4(d)] but also shows several temporal changes in community structure, with communities dying and new communities being born.

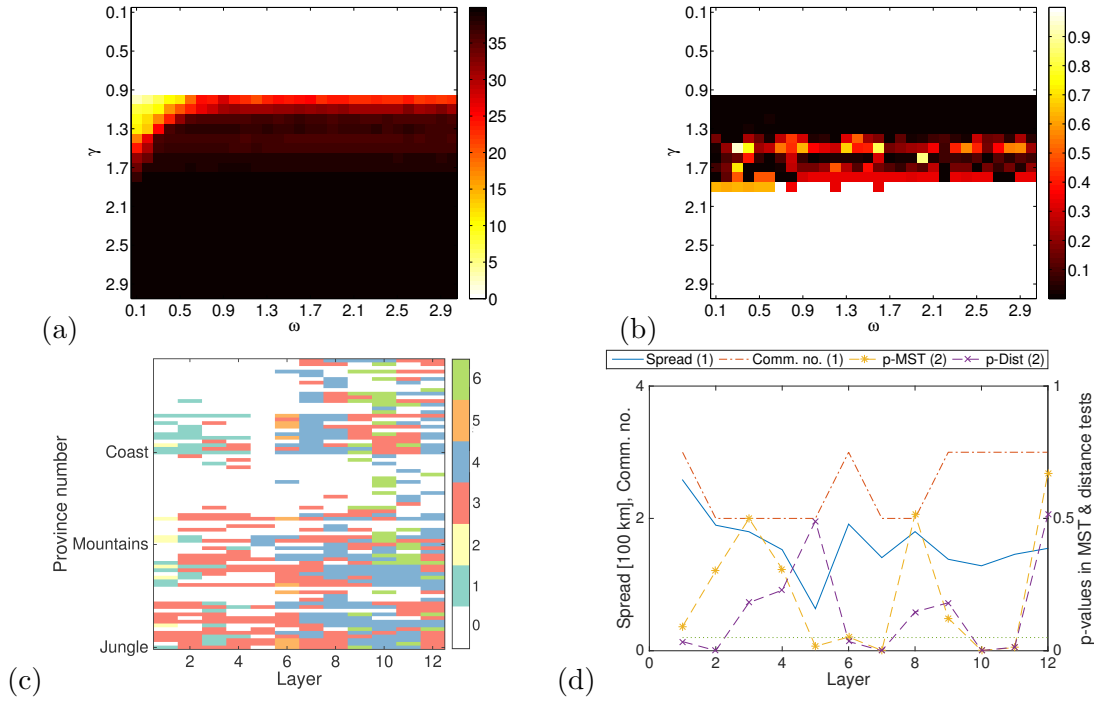


Figure 6.4: Dengue, multislice networks, NG null model — spatial organization of partitions. In (a-b) we show results of varying the parameters γ and ω: (a) z-Rand scores for similarity to "spatial" partitions by climate, in (b) the p-values for distances being smaller than expected at random in the distance test. In (c-d) we examine the multislice community structure for $\gamma = 1$, $\omega = 0.1$ against time on the horizontal axis. In (c) we plot nodes on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (d) on the left vertical axis we plot the mean community spatial spread and the number of communities, and on the right vertical axis we plot the p-values for the distance and MST tests for each layer.

When studying the qualitative features of the partitions for $\gamma \in [1, 1.2]$ (where the endpoints of this interval are approximate) and $\omega \lessapprox 1$, we observe that community detection repeatedly finds layer 6 (2001) as the highest-scoring single critical time point, and layers 6 and 10 (2001 and 2005) as the highest-scoring pair of critical time points [i.e., the strongest change points in temporal community structure]. See Fig. 6.5 (b)-(c). These times agree with our visual observations [e.g., they roughly correspond to the birth of the blue community and the growth of the green community in Fig. 6.4(c)].

This repeated finding suggests that a strong shift in the patterns of disease correlations occurred around these times. Indeed, Peru experienced a large countrywide dengue epidemic in 2000–2001, and this period also marked the onset of new yearly epidemic dynamics [51]. Thus, our method recovers the most important biological event in this data set in addition to providing additional information about spatial influences on disease spread. We also observe several other times when new communities are born, but we do not know the biological significance of these dates. Notably, in this parameter regime, community detection does not identify the large epidemic in the jungle Utcubamba province in 1996 (see the time

series in Fig. 3.2), which is the other known important event in the time period covered by this data set.

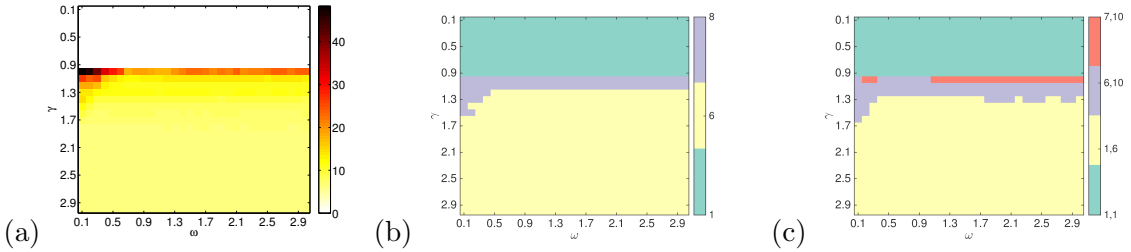


Figure 6.5: Dengue, multislice networks, NG null model — temporal organization of partitions. In (a-c) we show results of varying the parameters γ and ω on: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}, (b) the highest-scoring t_c in terms of the layer number (for a comparison against a single critical time point partition), and (c) pairs of highest-scoring t_{c1} and t_{c2} (for a comparison against a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

The NG null model has recently been shown to potentially have problems for correlation networks [171], as discussed in Section 3.3. In Section 6.3, we use a null model for community detection that has been specifically designed for correlation networks, and we compare the results of community detection using this null model to those for NG. Both of these approaches should inform our knowledge of the patterns of dengue spread in Peru.

6.3 Community detection using the correlation null model

In this section, we use the "correlation null model" that is specifically designed for community detection on correlation networks to study the disease correlation networks generated from the dengue data set. We defined this null model in Section 3.3.6.

6.3.1 Static networks

We test the performance of modularity maximization with the correlation null model with a range of resolution parameters $\gamma \in \{0.1, 0.2, \ldots, 3\}$ on the static correlation networks that we construct from dengue time series with $\upsilon = 4$ and $\Delta = 80$. In most of the static networks, community structures appear to be affected by spatial proximity — especially for post-2000 networks, as illustrated by the high z-Rand scores versus the climate partition (particularly in networks corresponding to 1995–1996, 2000–2001, 2003–2004, and 2005–2006) — see Fig. 6.6(a). These high z-Rand scores often result from (1) the classification of the majority of jungle provinces into one community and (2) the existence of a community that contains many of the northern coastal provinces [see Figs. 6.6(b)-(c)].

For the correlation null model, the distance test broadly agrees with the z-Rand score for spatial partitions [see Fig. 6.7(a)]. The MST test detects very few significantly spatial networks [see Fig. 6.7(b)]. We choose network 107 (April 2003), $\gamma = 0.9$ as an example, as its partition is significant in both z-Rand scores and distance and MST tests. This

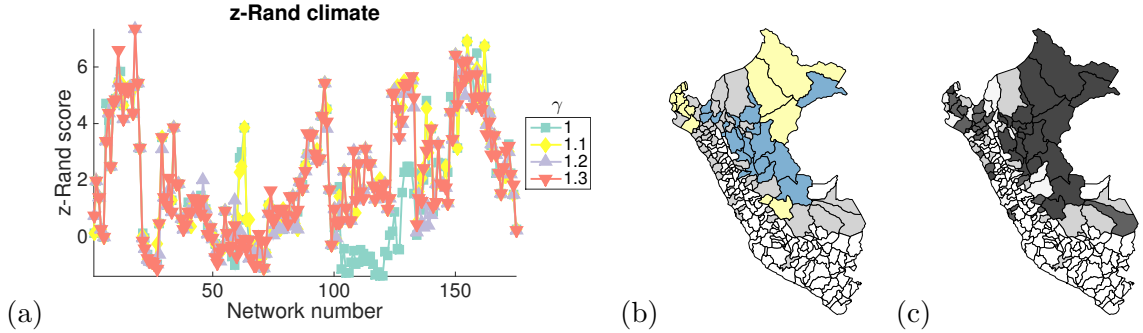


Figure 6.6: Dengue, static networks, correlation null model — spatial partitions according to z-Rand scores. Properties of the algorithmic community structure: (a) Plot of the z-Rand scores versus the detailed climate partition for $\gamma \in \{1, 1.1, 1.2, 1.3\}$. (b) Community structure with the highest z-Rand score when compared to the climate partition. The resolution-parameter value is $\gamma = 2.4$, the network is 9 (October 1995), and the z-Rand score is 9.04. (c) Community structure with the highest z-Rand score when compared to the administrative partition. The resolution-parameter value is $\gamma = 2.3$, the network is 43 (April 1996), and the z-Rand score is 8.9. Visualizations in panels (b)-(c) use a map of Peru in which we color provinces according to their community assignment. White provinces are ones in which our data does not include any reported cases of dengue fever in the indicated time window.

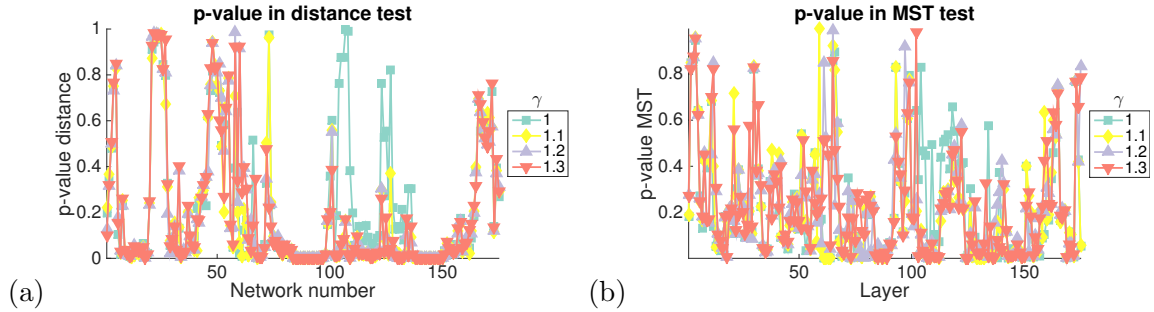


Figure 6.7: Dengue, static networks, correlation null model — spatial partitions according to distance and MST tests. Properties of the algorithmic community structure for the 700 static networks covering the whole time period (horizontal axis): (a) the p-value in the distance test and (b) the p-value in the MST test for $\gamma \in \{1, 1.1, 1.2, 1.3\}$.

partition contains a strong north-south division, and it consists of three communities, one of which [red in Fig. 6.8(d)] contains nodes that appear mostly to experience the highest epidemic periods, such as 2000–2001 (networks 80-100) and the yearly epidemics thereafter, another (colored blue) contains primarily jungle nodes that experience high disease burden throughout the study time, and the third [green in Fig. 6.8(d)] contains the northern regions that were first infected by the were DENV-3 and DENV-4 strains in 1999 which led to the 2000–2001 epidemic [187]; these regions also experienced a high number of disease cases relative to their population, both during the 2000–2001 epidemic and in the whole data set. This strongly spatial network partition appears to give us a clue as to the different disease patterns experienced by these regional groupings, highlighting that the difference between the northern nodes and the rest of the country is observable in periods other than 2000–2001 as well.

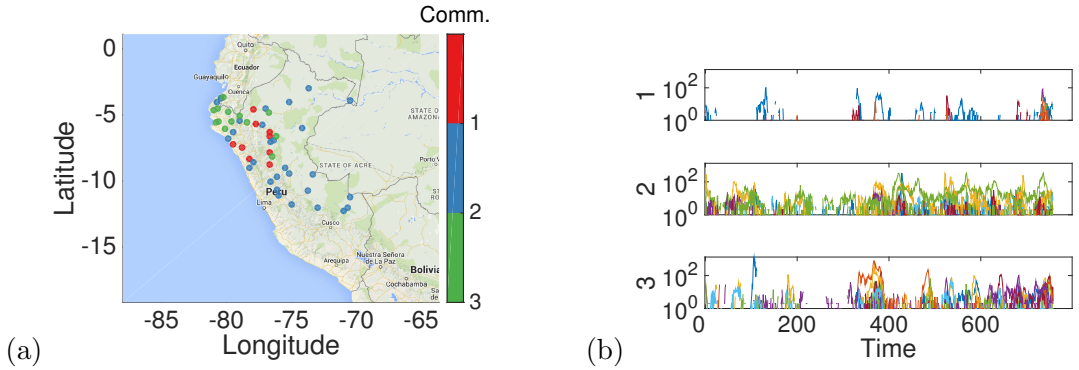


Figure 6.8: Dengue, static networks, correlation null model — spatial partitions for $\gamma = 1$, network 107. In (a) we show a map of all the nodes colored by algorithmically detected communities for ($\gamma = 1$, network 107), and in (b) we show the time series of disease occurrence in these communities.

6.3.2 Multislice networks

We also perform community detection on multislice networks constructed using $\Delta = 60$ and $\upsilon = 60$ using the correlation null model for $\gamma \in \{0.1, 0.2, \ldots, 3\}$ and $\omega \in \{0.1, 0.2, \ldots, 3\}$.

Nearly all parameter regimes show partitions that are detected as spatial by both the z-Rand score and the distance test [see Fig. 6.9(a)-(b)], but the climate z-Rand scores are lower than for the NG null model [compare Fig. 6.9(a) with Fig. 6.4(a)]. Additionally, all partitions appear significant in the z-Rand score against a temporal partition [see Fig. 6.10(a)].

We obtain a temporal partition with one critical time point at layer 8 for $\gamma \in \{0.6, 0.8, 0.9, 1\}$ and $\omega \lessapprox 0.2$. The other parameter regimes have different critical time points. For $\omega \lessapprox 1.5$ and a pair of critical time points, we obtain the highest temporal z-Rand score when the critical time points occur at layer 7 (i.e., January 2001) and at layer 10 (i.e., March 2004) — see Fig. 6.10(c). For $\omega \gtrapprox 1.5$, we obtain the highest temporal z-Rand scores with a pair of critical time points of layer 6 (i.e., June 1999) and 10, or with layers 1 and 8. (Note that

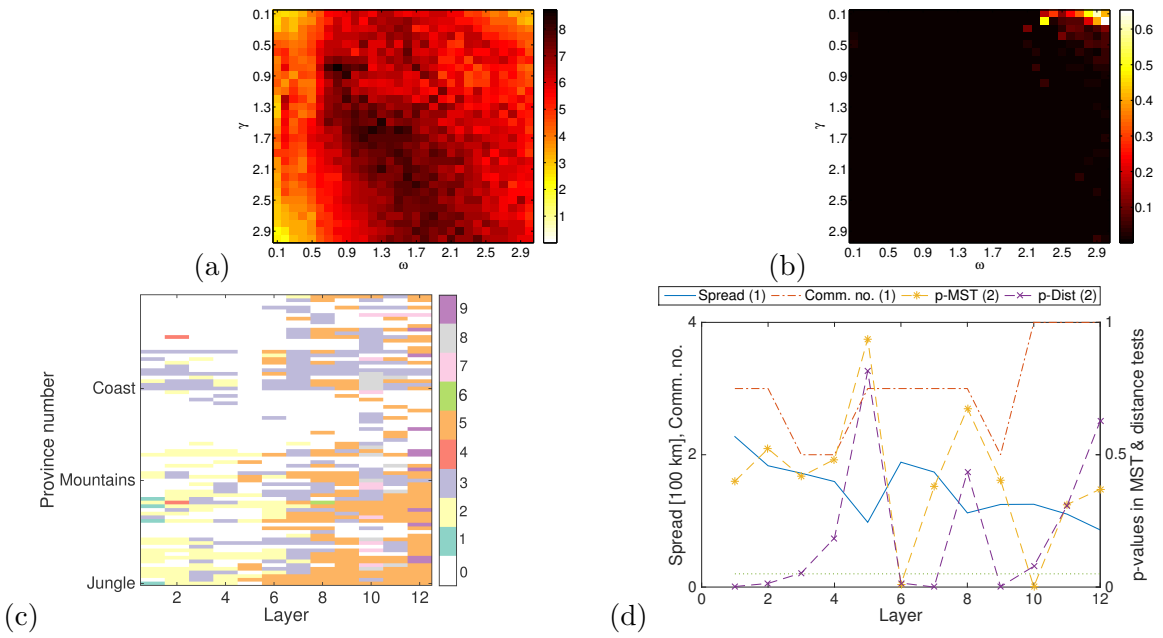


Figure 6.9: Dengue, multislice networks, correlation null model — spatial organization of partitions. In (a)-(b) we show results of varying the parameters γ and ω: (a) z-Rand scores for similarity to "spatial" partitions by climate, in (b) we plot the p-values for distances being smaller than expected at random in the distance test. In (c-d) we examine the multislice community structure for $\gamma = 1$, $\omega = 0.1$ against time on the horizontal axis. In (c) we plot nodes on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (d) on the left vertical axis we plot the mean community spatial spread and the number of communities, and on the right vertical axis we plot the p-values for the distance and MST tests for each layer.

these z-Rand scores tend to be lower than those for $\omega \lessapprox 1.5$.) The year 1999 corresponds to the arrival of new dengue strains into Peru, and 2000–2001 is when there was the first large, countrywide epidemic of dengue occurred — our method detects the two largest events in this data set.

We focus on the partitions for $\omega \leq 1$, in which we observe a mixture of temporal and spatial features. In Fig. 6.9(c), we show an example partition for $\gamma = 1$ and $\omega = 0.1$. This partition includes 9 communities. Although several communities coexist in each layer, the primary divisions appear to be largely temporal. For example, community 2 shrinks after layer 6 (January 2001). However, the highest z-Rand score versus a temporal partition (with either one or two critical time points) for this partition is one that has a single critical time point t_c at at layer 8 (i.e., in July 2002), corresponding to growth of community 5.

For both the NG and correlation null models, the community structure appears to be predominantly spatial. The strong influence of spatial proximity on the community structure is unsurprising, as spatial distance is an important influence on disease spread [253,271]. Previous studies have also noted that the community structure of spatial networks obtained by maximizing modularity using the NG null model tends to be strongly influenced by geographical factors [82,222,262]. If there are other interactions that shape the dengue disease-correlation network, they might be obscured by the strong influence of spatial proximity.

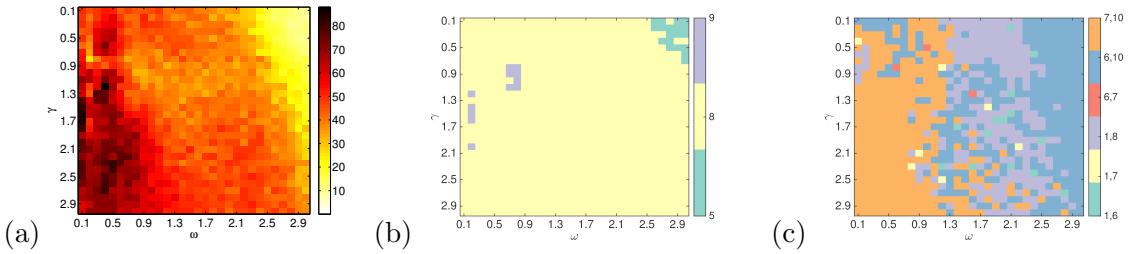


Figure 6.10: Dengue, multislice networks, correlation null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω on: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}, (b) the highest-scoring t_c in terms of their layer number (for a comparison against a single critical time point partition), and (c) pairs of highest-scoring t_{c1} and t_{c2} (for a comparison against a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

However, such interactions might be revealed by using a spatial null model that incorporates the expected effect of space on interactions. We pursue this idea in Section 6.4.

Further, on multislice networks for which the partitions contain both spatial and temporal organization, the use of a spatial null model and the removal of spatial organization might emphasize the temporal organization. We would hope that using a multislice version of a spatial null model would allow us to detect a larger number of possible critical time points, or to obtain higher z-Rand scores for temporal partitions and thus a stronger confirmation of the temporal divisions.

6.4 Community structure using spatial null models

In this section, we use the spatial null models: the gravity null model that we defined in Section 3.3.3 and the radiation null model that we defined in Section 3.3.5 to study the disease correlation networks generated from the dengue data set using modularity maximization for a range of resolution parameters $\gamma \in \{0.1, 0.2, \ldots, 3\}$.

We obtained province locations from the `Geonames.org` website [100] and obtained their populations from the National Institute of Statistics and Informatics of Peru [132]. We were only able to obtain the 1994 and 2007 populations. Due to the limited range of data and the several changes in Peruvian administrative structures between the two times, we only use the 2007 populations.

The maximum inter-province distance is about 1300 km. We report numerical experiments using a bin size of 400 km after testing the shape of the spatial deterrence function for several other sizes (ranging between 50 and 500 km) in the same manner as in Ref. [82], as described in Appendix A.4. We find that all of the bin sizes that we tested produce very similar partitions for both the gravity and radiation null models. Recall from Section 5.2.1.1 that only 79 of the 195 provinces reported cases of dengue in our data, so we use the location and population data only for those regions.

6.4.1 Static networks

We first study the community structure on static disease-correlation networks using the gravity and radiation null models. Both null models seem to remove most of the spatial organization of the community structures (including temporal variation in the spatial correlations), as indicated by low values of spatial z-Rand scores and lack of significance in distance and MST tests [see Fig. 6.11(a)-(c) for gravity and Fig. 6.12(a)-(c) for radiation null model]. The community structures typically contain one large community, that contains the majority of nodes at any given time, and several singleton communities [see Fig. 6.11(e) for gravity and Fig. 6.12(f) for radiation null model]. By examining the partitions directly, we see that the singleton communities tend to consist of the highest-populated nodes.

When studying algorithmic partitions of the static networks found using the gravity model in more detail using the z-Rand scores and the distance and MST tests, we see that only the MST test detects a small run of spatial partitions between networks 40 and 60 [see Fig. 6.11(a)-(c)] — a time for which no spatial partition was detected using the NG and correlation null models. Network 40 (February 1998) scores the lowest p-value in the distance test; this partition contains 24 singleton communities [all colored red in Fig. 6.11(e)] and one other community (colored blue). The singletons contain higher populated provinces than the one community [see Fig. 6.11(d)] and they experience higher numbers of disease cases over time [see Fig. 6.11(f)]. All other partitions using this null model look relatively similar, with one large community (a community with more than one node) and multiple singletons, often corresponding to the highest-populated provinces.

The algorithmic partitions for the radiation null model also score as not significantly spatial on z-Rand scores and distance and MST tests [see Fig. 6.12(a)-(c)]. The partition with the lowest p-value in the distance test, network 121 for $\gamma = 0.6$, contains one large community, one two-node community and three singletons. We observe no obvious pattern in the disease time series that would explain this partition; once again, the singletons are higher populated than the nodes in the larger communities. Partitions for higher γ contain more communities, with a degree of visual spatial organization [Fig. 6.12(e),(g),(i)]

6.4.2 Multislice networks

We also examine the community structures that we obtain by modularity maximization using spatial null models for multislice correlation networks constructed using $\Delta = 60$ and $\upsilon = 60$. We again obtain partitions with one large community that contains the majority of nodes [see Fig. 6.13(a)-(b)], and several nodes that correspond to provinces with the largest populations form temporal singleton communities, i.e., are only assigned together with their counterparts across time. This situation occurs for all of the tested parameter values. Additionally, we do not observe any clear pattern in the z-Rand scores and p-values

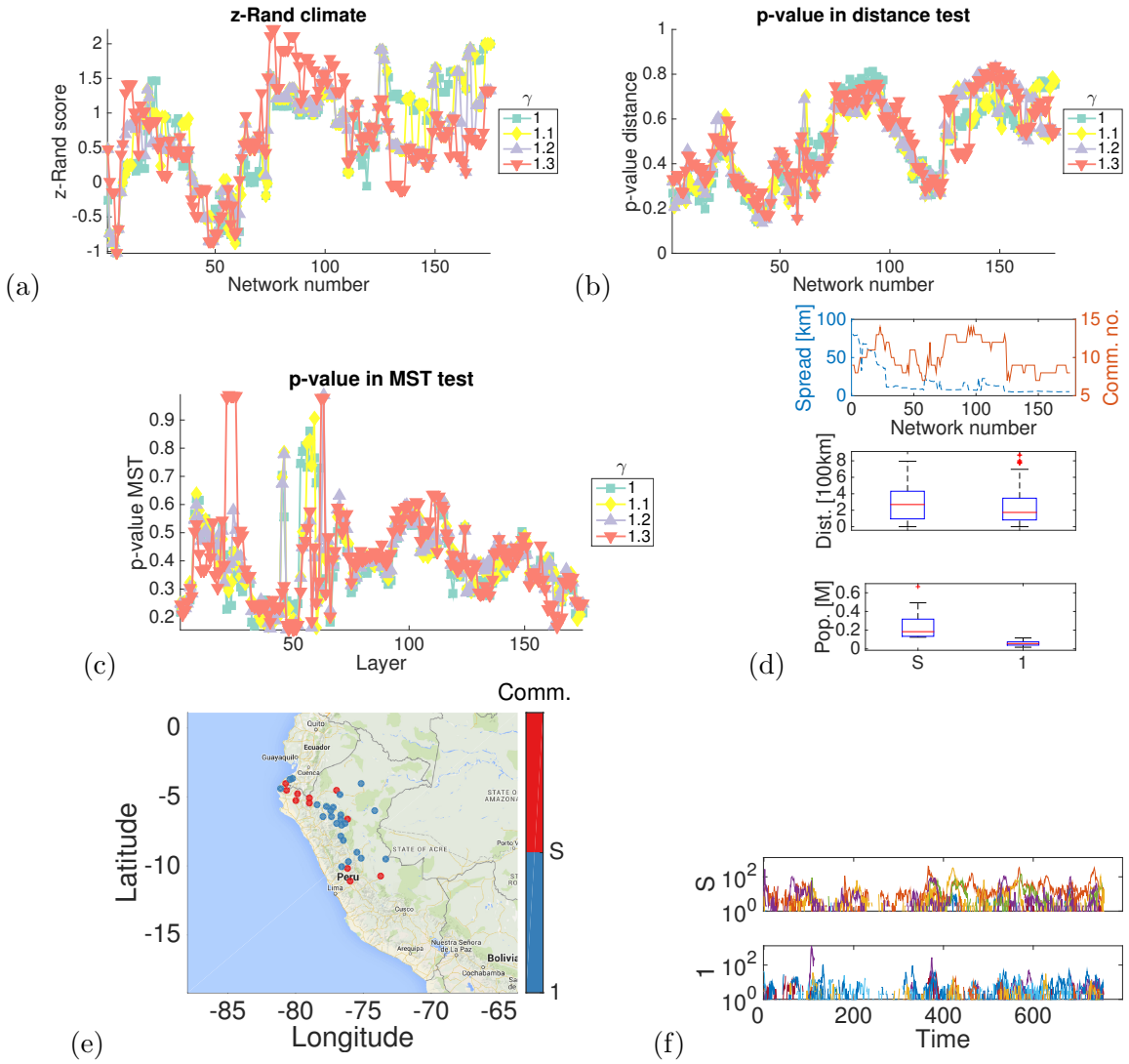


Figure 6.11: Dengue, static networks, gravity null model — spatial partitions according to z-Rand scores, distance test and MST test. Properties of the algorithmic community structure for the 700 static networks covering the whole time period (horizontal axis): (a) Plot of the z-Rand scores versus the detailed climate partition, (b) the p-value in the distance test and (c) the p-value in the MST test for $\gamma \in \{1, 1.1, 1.2, 1.3\}$. In (d), top: within-community spread and number of communities for all static networks at $\gamma = 1.9$, middle: box plot of within-community distance for the static network partition for $\gamma = 1.9$, network 40 (corresponding to February 1998), bottom: box plot of populations of communities for network 40 at $\gamma = 1.9$. In (e) we show a map of all the nodes colored by algorithmically detected communities for ($\gamma = 1$, network 40), and in (f) we show the time series of disease occurrence in these communities.

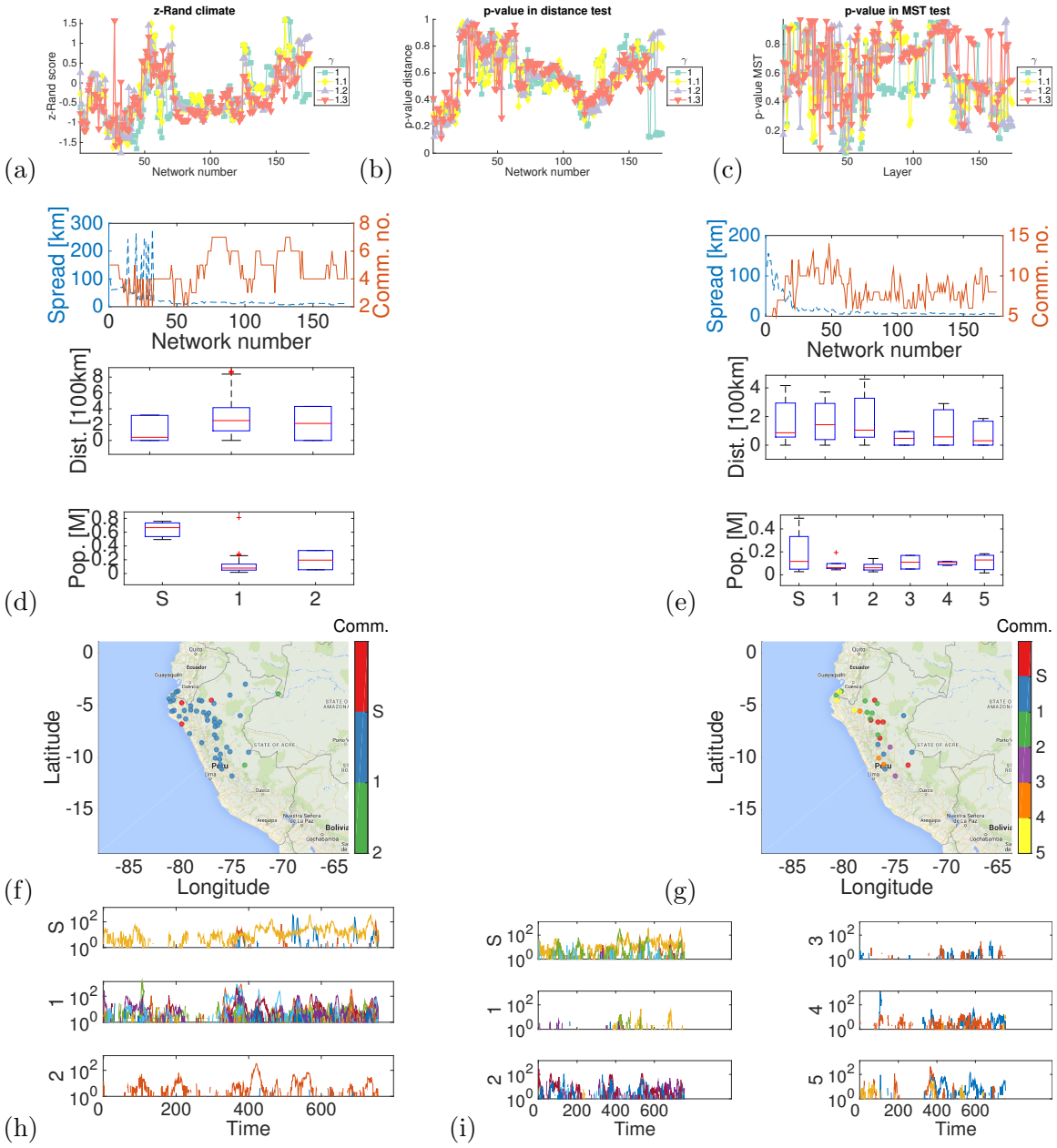


Figure 6.12: Dengue, static networks, radiation null model — spatial partitions according to distance and MST measures. In (a) we plot the z-Rand scores versus climate, in (b) we plot the p-value in the distance test, and in (c) we plot the p-value in the MST test for all static networks at $\gamma \in \{1, 1.1, 1.2, 1.3\}$. In (d), top: within-community spread and number of communities for all static networks at $\gamma = 0.6$, middle: box plot of within-community distance for the static network partition for $\gamma = 0.6$, network 121 (corresponding to May 2004), bottom: box plot of populations of communities for this partition. In (e) we show the corresponding plots for ($\gamma = 2.3$, network 47). In (f)-(g) we show a map of all the nodes colored by algorithmically detected communities for (f) ($\gamma = 0.6$, network 121), and (g) ($\gamma = 2.3$, network 47), and in (h-i) we show the time series of disease occurrence in these communities.

in the distance and MST tests as we change γ and ω, and thus we do not show the plots for these data.

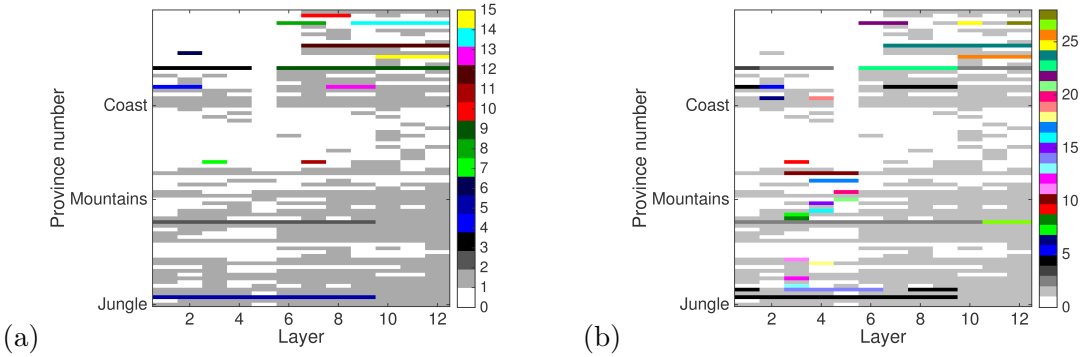


Figure 6.13: Dengue, multislice networks, gravity and radiation null models — community structure. Consensus community structure, which we obtain by maximizing modularity using (a) the gravity null model and (b) the radiation null model, of the dengue multislice disease-correlation networks. We use a resolution-parameter value of $\gamma = 1$ and consider $\omega = 0.1$. We plot layer numbers on the horizontal axis, and we plot the nodes on the vertical axis. We use color to indicate node community membership. Community number is also indicated on the colorbar.

Our findings suggest that the use of a spatial null model for modularity optimization might remove the majority of variation in the correlation structure of the dengue correlation networks, such that the influence of population size could be the only major factor that remains. There are only 5 provinces with populations over 500,000 people in Peru, and these provinces are often assigned to singleton communities when we use a spatial null model. This suggests that they have different disease patterns from the other provinces — as we have seen, they experience higher numbers of disease cases. This could potentially relate to them being above minimum population size required for sustained disease transmission, which for dengue has been estimated to lie between 10,000 and 500,000 [55, 153].

6.5 Province-level communities from the whole time series

We now examine what province-level information we can glean from the data. The simplest approach is to construct a single static network from the entire length-T time series, but our multislice approach allows us to aggregate data less severely. This, in turn, allows us to lose less information and to potentially obtain more informative community partitions.

6.5.1 Complete data aggregation

When we aggregate all time series to construct a single similarity network (i.e., when we choose $\tau = 1$ and $\Delta = 779$), the community structures that we obtain via modularity maximization with the spatial and correlation null models all consist of a single large community and a few singleton communities [see Figs. 6.14(a)-(c)]. Only the NG null model is able to detect meaningful-looking communities, especially for $\gamma = 1$ and $\gamma = 1.1$ [see

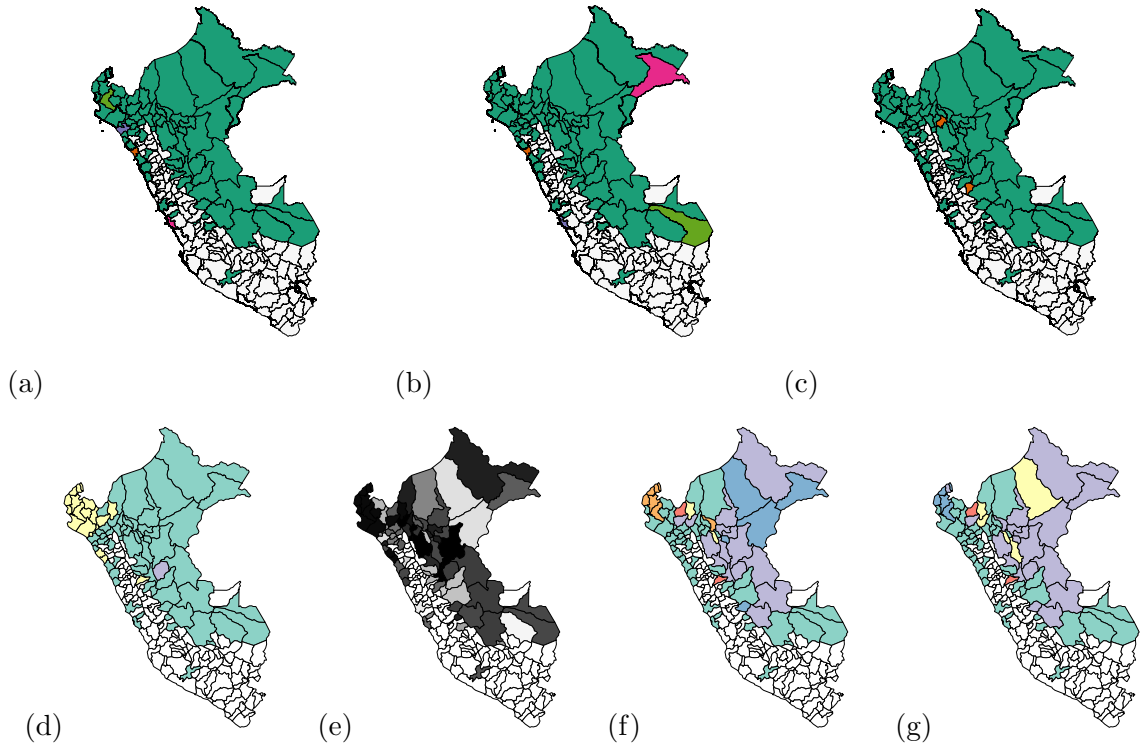


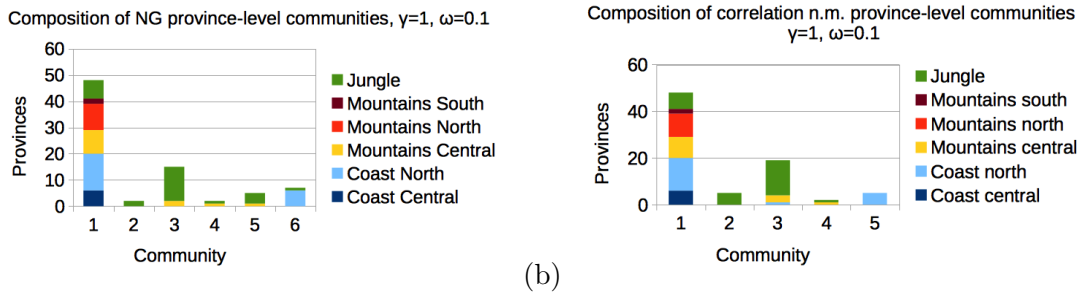
Figure 6.14: Dengue, province-level communities from fully aggregated and multislice networks. Province-level algorithmic community structure, which we obtain by maximizing modularity, for the static and multislice dengue correlation networks. We color the provinces according to their community assignments. White provinces are ones in which our data does not include any reported cases of dengue in the indicated time window. (a)-(e) Properties of the algorithmic community structure of the dengue static correlation networks that are fully aggregated (i.e., $\tau = 1, \Delta = 779$) for (a) the gravity null model, (b) the radiation null model and (c) the correlation null model for $\gamma = 1$, (d) the NG null model, (e) NG null model for $\gamma = 1.1$. (f)-(g) Properties of the algorithmic province-level community structure generated from multislice networks with a time window of width $\Delta = 60$ for (f) NG null model and (g) correlation null model. We use $\gamma = 1$ unless stated otherwise.

Figs. 6.14(d)-(e)]. For $\gamma = 1$, we find three communities; the smallest one is a singleton, and the middle-sized one consists almost exclusively (15 of 17 nodes) of northern coastal provinces. This partition has a z-Rand score of 7.3 versus climate. For $\gamma = 1.1$, using the NG null model yields 28 communities, many of them small. Nodes that are assigned to the community of the northern coastal provinces are the provinces of Peru that were most strongly involved in the 2000–2001 dengue epidemic; 15 nodes in this community experienced this epidemic, whereas only two other nodes experienced it. It appears that complete data aggregation over the whole time series results in the 2000–2001 epidemic dominating all other events in the time series. If we use the community structure of the temporally evolving multislice network to create the province-level structure, we might be able to obtain a more interesting partition.

6.5.2 Province-level communities from the multislice network

We now study the structure of province-level communities that we obtain from community detection using the uniform null model on an association matrix A^{province}. As we discussed in Section 3.2.2.1, we create this matrix by counting the number of physical nodes that are classified together in a consensus community detection in different layers of a multislice network. For simplicity, we present only results for the parameter values $\gamma = 1$ and $\omega = 0.1$.

Comparing the province-level communities that we obtain using the NG and correlation null models versus the broad topographical categories of coast, mountain, and jungle reveals large-scale climatic influences on disease patterns. The two null models yield similar results: more than 40 nodes are assigned to one large community that includes central coast, northwestern and southern jungle, and eastern jungle; and coastal north nodes form smaller, strongly spatial communities [see Figs. 6.14(f)-(g) and Fig. 6.15]. When we study the disease time series of the provinces grouped into the province-level communities (see Fig. 6.16), we observe distinct types of disease incidence patterns with jungle nodes split into several small communities. The community of northern coastal nodes (community 6 in NG, 5 in correlation null model) and the largest jungle community (community 3 for both null models) correspond to the majority of nodes that have been infected early in the disease time series. In contrast, the largest community combining the various climates (community 1 for both null models) corresponds to nodes that have only begun to experience the recurring epidemics post-2001. The NG null model finds one more type of temporal patterns in the jungle nodes than the correlation null model (community 5 which consists of nodes with late onset of disease).



(a) (b)

Figure 6.15: Membership of the consensus province-level communities, which we compute by maximizing modularity, in multislice dengue networks for $\gamma = 1$. We compare the climate composition of the communities using (a) the NG null model and (b) a correlation null model. Horizontal axis gives community number.

6.6 The effect of layer overlap

In this section, we compare our results from multislice networks with non-overlapping layers to the results of community detection on multislice networks with overlap between layers. Adding overlap between layers can allow us to study the temporal evolution of the network

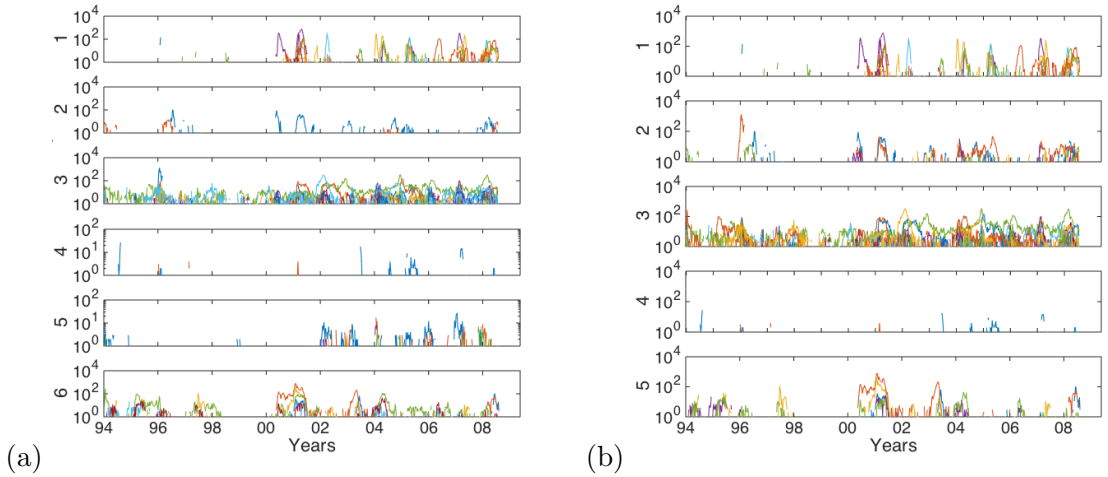


Figure 6.16: Time series for disease occurrences in the provinces that belong to consensus province-level communities, which we compute by maximizing modularity, in multislice dengue networks for $\gamma = 1$ using (a) the NG null model and (b) a correlation null model.

in more detail. For example, one can try to pinpoint the times of changes in the network structure to a greater level of detail — especially so if one is interested in specific, small-scale changes in community structure. This is especially useful where one possesses additional information about events that occurred at different times that are covered by the data. Furthermore, using overlapping layers allows one to construct multislice networks for short data sets. However, increasing the number of layers leads to a quick increase in computational complexity, so there is a trade off between the additional precision and resources and time required for the calculation.

We investigate the application of overlapping layers for the NG and correlation null models, as these null models showed the most interesting community structures. In Figs. 6.17 – 6.20 below, we show the same plots as in their respective sections for the networks with non-overlapping layers — compare the community assignment plots using the NG null model in Fig. 6.17 with Fig. 6.4 and for the correlation null model in Fig. 6.19 with Fig. 6.9.

From comparing the multislice community structures for $\gamma = 1$, $\omega = 0.1$, we see that the spatial organization of communities look very similar in both cases. This is further supported by plots of community number, spread and p-values in distance and MST tests showing qualitatively similar trends for both settings. When studying the broad structural features using z-Rand scores, we see that the results over the whole γ and ω parameter regime are remarkably similar. The one visible change is in the patterns of the critical time points for the correlation null model, which appear to be simpler in this setting [compare Fig. 6.20(c) with Fig. 6.10(c)], with layers 14 and 23 (December 2000 and January 2005) scoring the highest for $\omega \lessapprox 1.3$ (with some exceptions) and layer 19 (March 2003) scoring the highest for the other parameter values.

This comparison shows that the broad spatial and temporal structures that we are most interested in appear to be generally preserved between multislice networks with non-overlapping and overlapping layers. Using overlapping layers can provide additional detail as to the temporal evolution of network structure, but this comes at a cost of increased computational complexity. For the remaining disease data sets, we decide whether to use overlapping or non-overlapping layers based on the combination of the temporal resolution of the data set, the disease serial interval, and the computational complexity of community detection on the resulting disease-correlation networks.

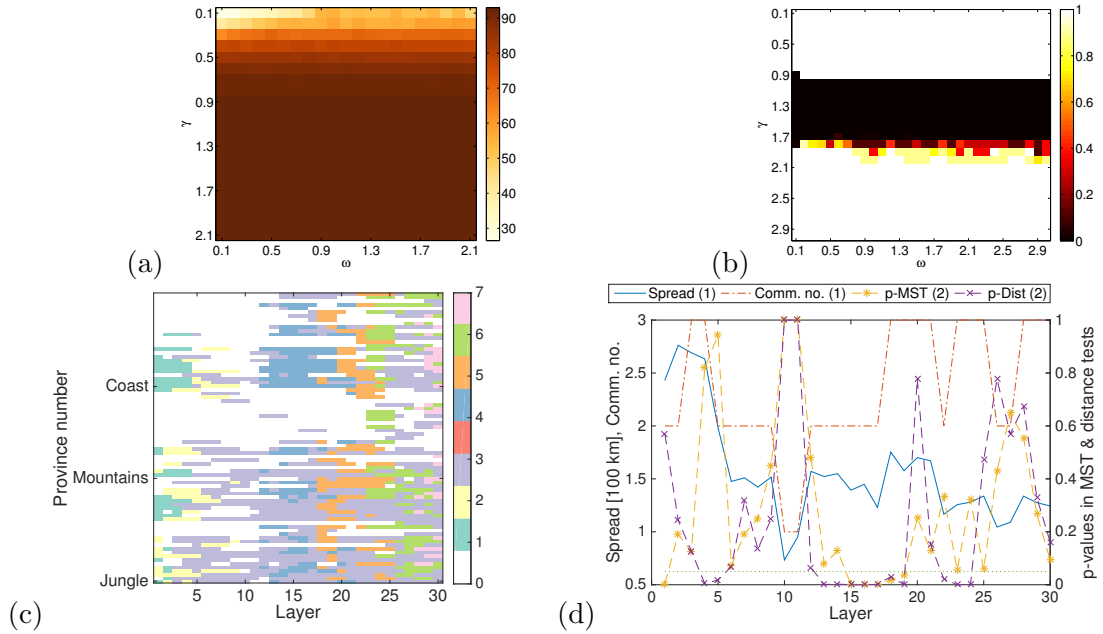


Figure 6.17: Dengue, multislice networks with $\Delta = 80$ and overlapping layers ($\upsilon = 24$), NG null model — spatial organization of partitions. In (a-c) we show results of varying the parameters γ and ω: (a) z-Rand scores for similarity to "spatial" partitions by climate, (b) the p-values for distances being smaller than expected at random in the distance test. In (c)-(d) we examine the multislice community structure for $\gamma = 1$, $\omega = 0.1$ against time on the horizontal axis. In (c) we plot nodes on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (d) on the left vertical axis we plot the mean community spatial spread and the number of communities, and on the right vertical axis we plot the p-values for the distance and MST tests for each layer.

6.7 Comparing measures of spatial clustering

In this section, we present the results of comparisons between measures of spatial organization of community structure based on using z-Rand scores to compare the algorithmic partitions against a manual spatial partition defined in Section 3.4.3 with the distance and MST tests defined in Section 3.4.4.1.

We first plot scatter plots for the two measures based on z-Rand scores (comparing against climate and administrative partitions) of network partitions detected with three select values of γ using all four null models. We observe the climate and administrative z-

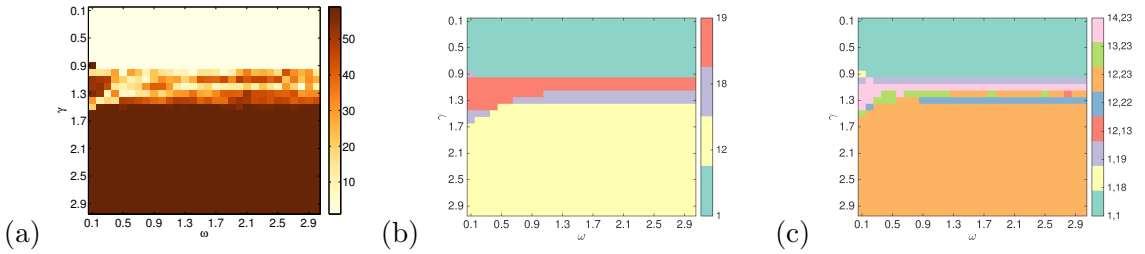


Figure 6.18: Dengue, multislice networks with $\Delta = 80$ and overlapping layers ($v = 24$), NG null model — temporal organization of partitions. In (a-c) we show results of varying the parameters γ and ω on: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}, (b) the highest-scoring t_c in terms of their layer number (for a comparison against a single critical time point partition), and (c) pairs of highest-scoring t_{c1} and t_{c2} (for a comparison against a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

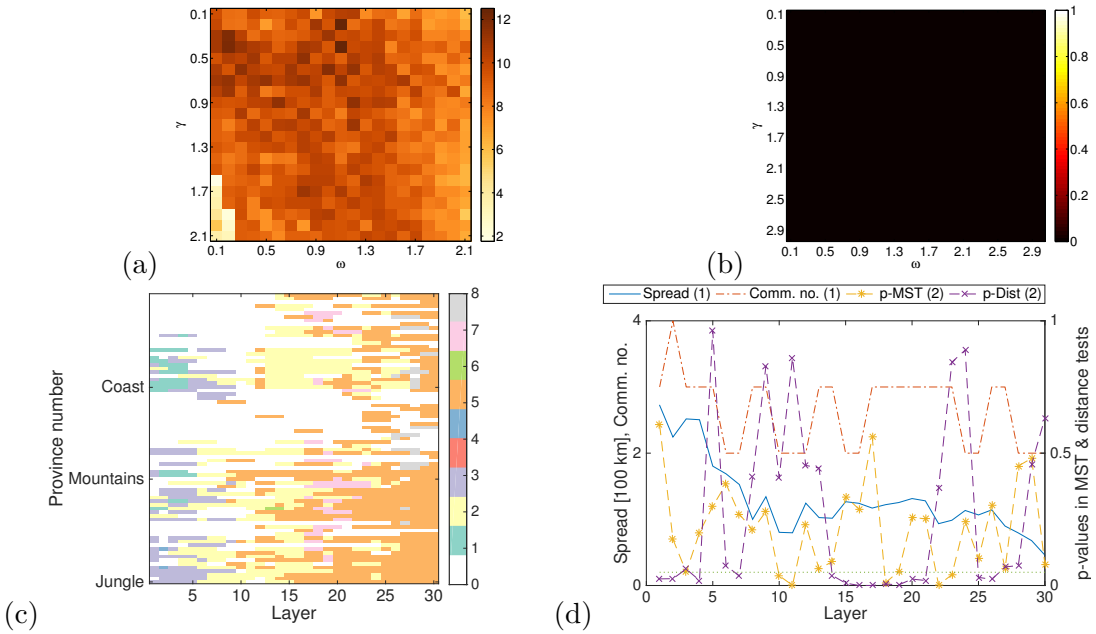


Figure 6.19: Dengue, multislice networks with $\Delta = 80$ and overlapping layers ($v = 24$), correlation null model — spatial organization of partitions. In (a-b) we show results of varying the parameters γ and ω: (a) z-Rand scores for similarity to "spatial" partitions by climate, (b) the p-values for distances being smaller than expected at random in the distance test. In (c)-(d) we examine the multislice community structure for $\gamma = 1$, $\omega = 0.1$ against time on the horizontal axis. In (c) we plot nodes on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (d) on the left vertical axis we plot the mean community spatial spread and the number of communities, and on the right vertical axis we plot the p-values for the distance and MST tests for each layer.

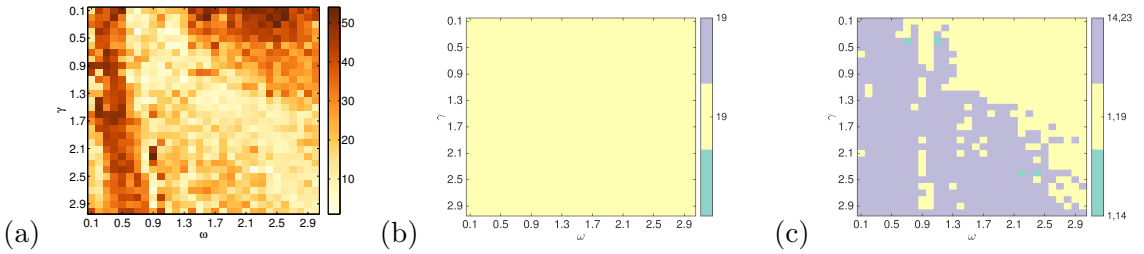


Figure 6.20: Dengue, multislice networks with $\Delta = 80$ and overlapping layers ($v = 24$), correlation null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω on: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}, (b) the highest-scoring t_c in terms of their layer number (for a comparison against a single critical time point partition), and (c) pairs of highest-scoring t_{c1} and t_{c2} (for a comparison against a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

Rand score measures perform similarly [see Fig. 6.21(a)]. Thus, we use the climate partition both for dengue in Chapter 6 and for rubella in Chapter 7; climate can influence the spread of both vector-borne diseases and diseases spreading by droplets, so it is a better-motivated choice of the two.

We then plot scatter plots of p-values in the distance and MST tests in the same manner. The majority of network partitions are uniformly classified as not significantly spatial ($p > 0.05$) by both tests, however there are some partitions that are differently classified by the two Monte Carlo tests.

We then plot scatter plots of p-values in the distance and MST tests against the z-Rand score versus a climate partition for the same null models and parameter values as above (see Fig. 6.21). When compared against the climate partition z-Rand score test, the majority of partitions found using gravity and radiation null models are classified by all statistics as non-spatial ($p > 0.05$ and $z_R < 1.96$), but there are more differences in classification for the partitions found using NG and correlation null models. In particular, some partitions with relatively high z-Rand scores are not classified as spatial by the distance and MST tests, which is a potential drawback of these tests. However, one should also note that both of the Monte Carlo tests can detect partitions that have a degree of spatial organization but are not organized in a manner similar to the climate or administrative partition. This freedom from assumptions as to the shape of spatial partitions is an advantage of these measures.

We now formalize the discussion of similarity between the measures of spatial organization by comparing the agreement in classification of partitions as significantly spatial by distance and MST tests ($p < 0.05$), with the classification by the climate z-Rand score ($z_R > 1.96$). We quantify the agreement in classification by the different methods using the Rand coefficient, defined in Section 3.4.3 (where a Rand coefficient of 1 indicates perfect agreement in classification). We show the histograms of the Rand coefficient values for the dengue and rubella data sets for the four null models in Fig. 6.22. The scores vary depending on the data set and null model, with more agreement between the two classifications

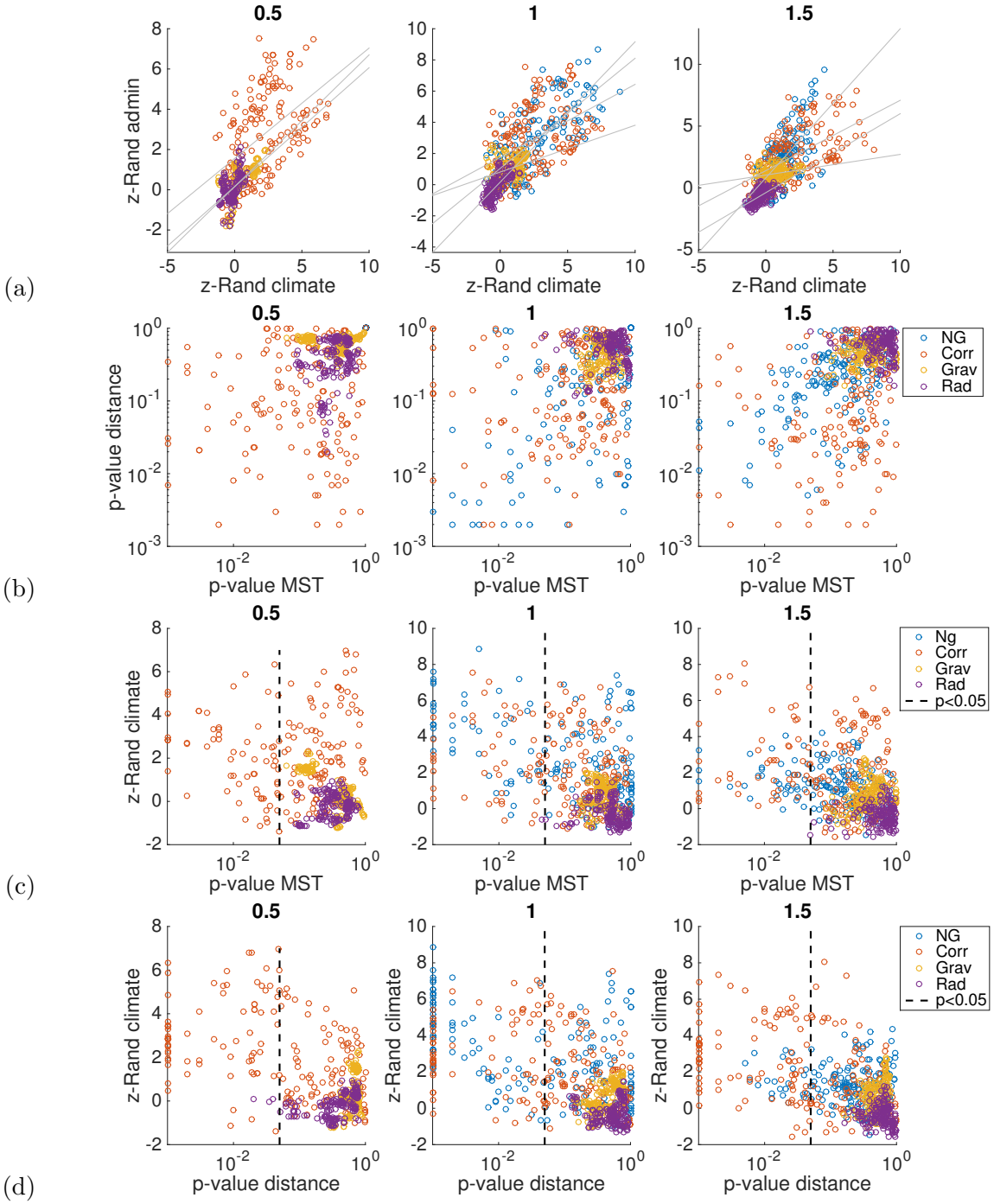


Figure 6.21: Scatter plot of the scores of spatial organization for the dengue network with different null models (colors), for (left) $\gamma = 0.5$, (centre) $\gamma = 1$, (right) $\gamma = 1.5$. (a) z-Rand scores against administrative and climate partitions, with the least-squares fit lines between each pair, (b) p-value in the distance and MST tests, (c) p-value in the distance test and z-Rand score against climate partition, and (d) p-value in the MST test and z-Rand score against climate partition. In (c)-(d) the p-value of 0.05 is plotted for visual guidance.

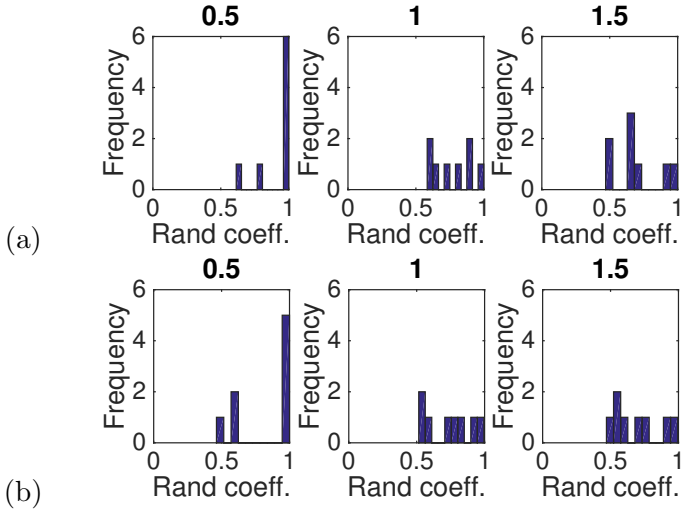


Figure 6.22: Histogram of the Rand coefficient scores between the classification as spatial and non-spatial partitions for pairs of measures of spatial organization, calculated for the dengue and rubella networks with all null models, for (left) $\gamma = 0.5$, (centre) $\gamma = 1$, (right) $\gamma = 1.5$. Rand score of 0 means no agreement in co-classification by the two methods, and a Rand score of 1 means perfect agreement. Plots show the frequency of Rand scores between (a) significant p-value in the distance test ($p < 0.05$) and significant z-Rand score against climate partition ($z_R > 1.96$), and (b) significant p-value in the MST test ($p < 0.05$) and significant z-Rand score against climate partition ($z_R > 1.96$).

for the gravity and radiation null models — see full results in Table 6.1.

Table 6.1: The Rand coefficient values between classification as spatial/non-spatial partitions by the distance and MST tests ($p < 0.05$) and classification by the z-Rand score on climate ($z_R > 1.96$).

Disease	Null model	Distance test			MST test		
		γ =0.5	γ =1	γ =1.5	γ =0.5	γ =1	γ =1.5
Dengue	NG	1.00	0.71	0.50	1.00	0.58	0.53
	Correlation	0.78	0.61	0.71	0.57	0.55	0.59
	Gravity	1.00	0.91	0.94	1.00	0.91	0.94
	Radiation	0.98	1.00	1.00	0.99	0.99	0.99
Rubella	NG	1.00	0.58	0.47	1.00	0.83	0.47
	Correlation	0.61	0.63	0.64	0.46	0.52	0.56
	Gravity	0.98	0.91	0.66	0.99	0.80	0.78
	Radiation	1.00	0.83	0.64	0.62	0.73	0.70

We select the distance test for assessing the spatial organization of disease-correlation networks from both disease data set and synthetic time series in the following three chapters. The level of agreement with the z-Rand scores versus a manual climate partition is greater for the distance test than the MST test. Further, based on visual observations for the dengue data set (e.g., Fig. 6.7), the dependence on parameter values and temporal variation is smaller for the distance test than the MST test.

6.8 Conclusions

In this chapter, we examined the results of community detection by modularity maximization on correlation networks constructed from the time series of dengue incidence in

provinces of Peru. We studied the effect of different null models on the network partitions — including ones that incorporate spatial information. This perspective is important, because we expect epidemic dynamics to be influenced strongly by spatial effects. We compared our results for the standard Newman-Girvan null model versus two null models that incorporate spatial information: a gravity null model [82] and a novel radiation null model, and versus a recently-developed correlation null model that is designed specifically for studying correlation networks that are derived from time series [171].

We observed for static networks that the NG and correlation null models find structures that are strongly spatial — especially for networks that correspond to the large epidemic in 2000–2001, and to a lesser degree after the onset of yearly dengue epidemics. In our study, we observed that spatial partitions are often dominated by large communities of neighboring jungle nodes that experience local epidemics during a time window.

On a multislice network, maximizing NG modularity can result in spatial or temporal partitions, or partitions with a mixture of both properties (depending on the parameter values). The spatial organization of partitions is often related to the climate of provinces, which has a known strong influence on dengue infection patterns. Temporal partitions successfully find the most important time point in the history of the disease — namely, the introduction of a new disease strain that caused a large epidemic in 2000–2001 and a subsequent shift in disease patterns — and several other potentially interesting time points and periods of high spatial correlation.

When studying province-level connectivity, we illustrated that obtaining consensus province-level communities from an association matrix constructed from the multislice network across time is a preferable approach to complete data aggregation. When aggregating into a static network, maximizing modularity using all null models except the NG null model failed to detect any meaningful communities; and even the NG community structure corresponds to only one large event: the 2000–2001 epidemic. Aggregating networks results in loss of information that is desirable when attempting to discern meaningful patterns [129, 146].

When we constructed multislice networks and computed consensus communities, we found "spatial" multislice partitions and province-level partitions that highlight the importance of climate to the disease patterns of dengue, as the jungle provinces are assigned to distinct communities from most mountainous and coastal provinces. This is sensible, as the yearly epidemic patterns tend (on average) to exhibit an earlier epidemic onset in the jungle [51, 55], and the jungle climate is rather distinct from the climate in coastal and mountainous provinces. The main climatic difference between jungle provinces and other provinces is temperature, and the influence of temperature on dengue transmission has been well documented [54, 55, 75, 136, 141].

The province-level communities that we detect using both the NG and the correlation null models yield distinct temporal disease incidence patterns. They separate the northern

coastal provinces that were repeatedly infected before the onset of yearly epidemics, and the non-jungle provinces that only began to experience recurrent disease post-2001. Community detection using both null models divides jungle nodes into separate communities, and the NG null model finds one more jungle community (with a late disease onset) than the correlation null model. A partition into communities with different yearly disease patterns can be useful to epidemiologists as it highlights the co-occurrence patterns of epidemics in different provinces — knowledge that could potentially provide epidemic warnings.

The assignment of different jungle nodes into separate communities hints that the factors that influence jungle epidemics may be different than those in other climates. Moreover, the variability in disease patterns between jungle provinces is high, as many provinces experience highly localized epidemics during the year-round disease season (in contrast to the existence of a summer disease season on the coast).

When we attempt to remove the influence of space by using the gravity and radiation null models, we obtain one large community that contains all but the highest-population provinces (which are assigned to singleton communities). This is different to the Belgian example in Ref. [82], where the spatial null model found a partition of the country into the French and Flemish speaking regions (which was undetectable using the NG null model). Our results suggest that for our disease-correlation networks the incorporation of space into the null model might account for the majority of the structure present in the network. The spatial structure that we removed likely includes the structure that corresponds to the variation in climate that causes different epidemic patterns in the jungle, coastal, and mountainous provinces. The structure removed by spatial null models might also include any influence of transport on the disease patterns, or this influence is lower than the internal disease dynamics. The only variable that we were able to identify as influencing community structure detected using spatial null models is province population: the highly populated (and typically coastal) provinces form singleton communities. These highly populated provinces are local economic centers, with many people traveling there from the other provinces and thereby transmitting the disease [120, 177, 207, 253]. They also possess populations above the minimum population size required for sustained disease transmission [55, 153].

One possibility is that these highly populated provinces could be the seeds of epidemics for the other coastal and mountainous provinces, leading to high correlations across atypically long distances compared with the majority of the data, which could in turn cause them to be assigned to singleton communities when using spatial null models. In fact, two studies have reported (so-called) "hierarchical" transmission of dengue from populous regions to those with low populations in both Peru and Thailand [55, 66]. The difference in disease patterns could also be due to internal dynamics of the highly populated provinces.

Population size influences dengue transmission: the basic reproductive number R_0 and disease persistence (i.e., the fraction of weeks with disease cases) are positively correlated with population size, and the attack rates are negatively correlated with it [51, 55].

We have seen that community detection is able to detect climatological patterns in a disease that is known to be influenced by climatic and weather patterns. In Chapters 7 and 8 we will use the same methodological pipeline to analyze data from other endemic and emerging infectious diseases.

In this chapter, we also tested two measures for detecting spatial organization in networks for which we do not possess ground-truth partitions (i.e., disease-correlation networks for countries other than Peru), or where such partitions do not exist (i.e, the agent-based model). By visual comparisons throughout this chapter and a formal comparison in Section 6.7, we found that the distance test appears to be preferable to the MST test. We will use the distance test in Chapters 7 and 8, as we do not possess "ground-truth" spatial partitions for the data sets that originate in countries other than Peru, and for the agent-based model in Chapter 9.

Chapter 7

Applications to endemic diseases

This chapter consists of original work by MS and M. A. Porter which is not yet published.

7.1 Introduction

In this chapter, we apply our community-detection methodology to a data set for the incidence another endemic disease: rubella in Peru. We present the most interesting results in this chapter, and the remainder is attached in Appendix B. This appendix also contains the majority of the results for a data set on seasonal influenza in Chile, in which we found a degree of temporal organization but we failed to find significant and informative spatial organization using our methods.

We use the same approach [community detection on static and multislice networks using modularity maximization with Newman-Girvan (NG), correlation, gravity and radiation null models] as we described in Section 3.5 and that we used for analyzing community structure of correlation networks created from dengue fever data sets in Chapter 6. We found partitions with spatial and temporal organization in the dengue data.

We examine the spatial organization of community structures detected in the static networks, first measuring the degree of spatial organization across values of the resolution parameter $\gamma \in \{0.1, 0.2, \ldots, 3\}$ using z-Rand scores against climate partitions (defined in Section 3.4.3) and the distance test (defined in Section 3.4.4). We then select particular parameter values and networks for which we study the partitions in more detail.

For the multislice networks, we study the spatial and temporal organization of the community structure that we detect algorithmically. We use the multislice versions of the z-Rand scores versus a climate partition and the distance test for detecting spatial organization. We search for critical time points when community structure changes using the z-Rand score methodology that we described in Section 3.4.3. We use these methods to select interesting partitions for further study from algorithmic partitions generated for parameter values $\gamma \in \{0.1, 0.2, \ldots, 3\}$ and $\omega \in \{0.1, 0.2, \ldots, 3\}$.

By maximizing modularity using the NG and correlation null models, we find communities in the rubella networks that have a comparable degree of spatial organization to the communities that we found in the dengue networks, but they do not appear to be related to climate. When we take into account the expected influence of distance on community structure through the use of a spatial null model, we once again observe one or two large communities and a small number of singleton communities containing the most populated nodes. When we perform modularity maximization on multislice networks, all four null models are able to detect temporal partitions of the networks at known times when disease patterns changed. The NG and correlation null models appear to be relatively reliable at finding major changes, and the gravity and radiation null models detect a wider range of critical time points with higher z-Rand scores. We present the results for NG, correlation and gravity null model here, and for radiation null model in Appendix B for completeness.

We find only a small degree of spatial organization in the Chilean influenza data set (and the structures show a strong dependence on parameter values). We find clear temporal partitions in the multislice networks for this data set that correspond to a year with the lowest number of influenza infections in our data set. For conciseness, we only present the select results of temporal partitions of multislice networks in this chapter, and we present the remaining results in Appendix B for completeness.

As discussed in Section A.3, both rubella and influenza data sets are affected by issues related to our ability to discern causal influence of disease spread from correlations. Issues can arise from the interactions between the sampling rate of the data, our choices of the parameter υ that describes the distance between adjacent time windows used for network creation, and the serial interval of the disease — leading to the inclusion of multiple infections during one step between time windows, which can obscure the path that the disease takes to spread. The influenza data set is the most affected by these issues, as the data is sampled weekly and the serial interval of the disease is 4 days, so it is possible for the weekly data collection to include multiple infections from the same chain of infections.

Both the rubella and influenza data sets have long time series compared to the number of provinces in the data: 673 time points and 195 provinces for rubella, and 365 time points and 15 provinces for influenza. Because we choose the time window width Δ (used for constructing a correlation matrix) to be larger than the number of nodes ($\hat{N}$) in the network, having long time series allows us to create multislice networks that consist of a large number of layers. We can then study the temporal evolution of the disease patterns over a longer time period and potentially in greater depth.

7.2 Rubella

The rubella data set contains weekly new case count data from the 195 provinces of Peru over 13 years between 1997 and 2010 (i.e., 673 weeks). The data show yearly epidemics of

rubella, with especially large epidemics in 2000–2001 and 2005–2006, and a decline in the number of disease cases after 2007 (see Fig. 5.2). We described the data set in detail in Section 5.2.1.2.

For the rubella data set, we use a time-window width of $\Delta = 134$ (after checking that no more than 133 nodes experience disease at once in any of the time windows to comply with the RMT constraint that we discussed in Section 3.1.1), and we let the distance between adjacent time windows be $\upsilon = 4$ to generate a set of 132 static networks; we then take $\Delta = 134$ and $\upsilon = 12$ to generate a multislice network with 44 layers in order to reduce computational complexity. We described the justification for parameter choices and the general properties of the time series and the static networks in Section A.2.

Modularity maximization using the Newman-Girvan and correlation null models allows us to detect spatial partitions in both static and multislice networks. We also find temporal partitions with critical times corresponding to the period between the two large epidemics, and to the fall in the number of disease cases in the last part of the data set. Modularity maximization using gravity and radiation null models yields partitions that are similar to what we observed for dengue: one large community and a small number of singletons corresponding to the highest-populated provinces of the country. However, the spatial null models suggest many more potential temporal partitions of the multislice networks, with very high z-Rand scores.

7.2.1 Modularity maximization using the NG null model

When we apply the Newman-Girvan null model to the static networks constructed from the rubella data set with $\Delta = 134$ and $\upsilon = 4$, we observe spatial partitions (as detected using z-Rand scores versus climate partitions by taking $z_R > 1.96$) for around the first 60 static networks, and with the distance test for the first 30 [see Fig. 7.1(a)-(b)]. The particularly high z-Rand scores and low p-values in the distance test suggest that $\gamma = 1.3$ generates partitions with more spatial organization than other γ values. We select network 9 for $\gamma = 1.3$ for detailed study, as partitions of this network scored high z-Rand scores and significant p-values ($p < 0.05$) in the distance test over the largest fraction out of all the networks for the γ parameter values that we tested.

The partition of network 9 for $\gamma = 1.3$ is composed of 3 communities. Parts of the three communities form spatial clusters [such as the red coastal cluster or the green jungle cluster in Fig. 7.2(a)] but overall each of the communities is distributed across the country. The nodes assigned to the three communities exhibit differences in temporal patterns of disease, with nodes in community 1 starting late and experiencing relatively low levels of disease for the majority of the time period, nodes in community 2 peaking early and nodes in community 3 experiencing large epidemics around networks 200 and 500 (during the large epidemics in years 2000–2001 and 2005–2006) [see Fig. 7.2(b)].

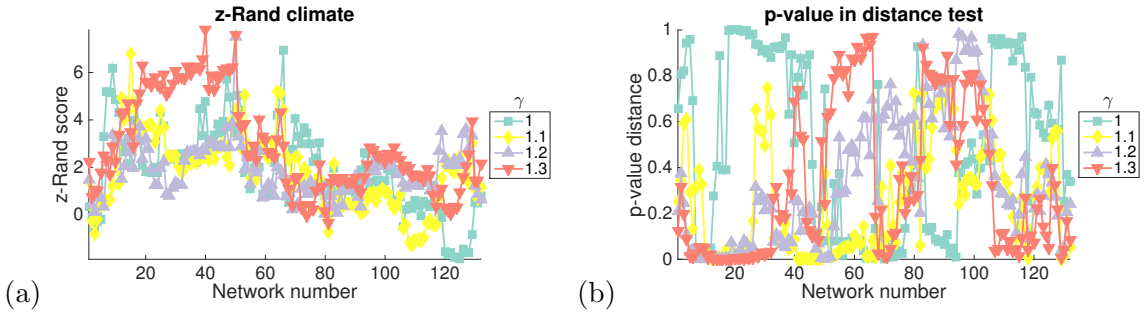


Figure 7.1: Rubella, static networks, NG null model: spatial partitions according to z-Rand scores and the distance test. Properties of the algorithmic community structure for the 132 static networks covering the whole time period (horizontal axis) for $\gamma \in \{1, 1.1, 1.2, 1.3\}$: (a) plot of the z-Rand scores versus the detailed climate partition, and (b) plot of the p-value in the distance test.

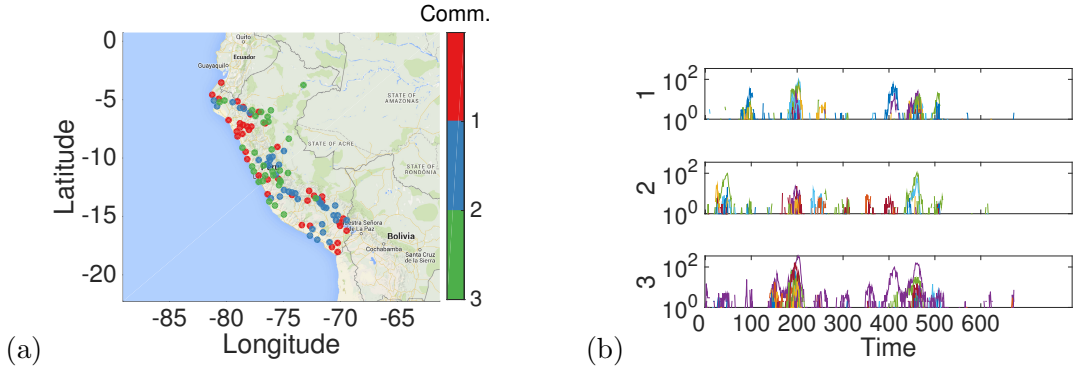


Figure 7.2: Rubella, static networks, NG null model: network 9 at $\gamma = 1$. In panel (a) we show a map of all the nodes in network 9 at $\gamma = 1$ (colored by algorithmically detected community assignment, community number indicated on the color bar), and in panel (b) we show the time series of disease occurrence in the provinces assigned to these communities, with community number indicated on the vertical axis.

When we study the spatial organization of partitions of the rubella multislice networks with the NG null model using z-Rand scores versus climate, we observe that partitions for $\gamma \gtrapprox 1$ are highly spatial; for the partition-wide distance test, most partitions in the parameter regime $1 \lessapprox \gamma \lessapprox 2.1$ are also significantly spatial [see Fig. 7.3(a)–(b)].

We focus on the partition for $\gamma = 1, \omega = 0.1$, as the number of communities increases rapidly for values of γ above it, and higher ω leads to increased prevalence of inter-layer connections in communities. Although the community structure for this pair of parameter values scores as spatial in the partition-wide distance test, the per-layer distance test is only statistically significant for most of the layers covering the first large epidemic of rubella [see Fig. 7.3(e)]. Further, the community assignments do not visually resemble the climate partitions of the country [see Fig. 7.3(c)-(d)]. This suggests that the community partitions have a spatial organization that is related to factors other than climate. This contrasts with the dengue case described in Section 6.2, but it conforms with our expectations of low influence of climate on disease patterns, as rubella is not a vector-borne disease like dengue.

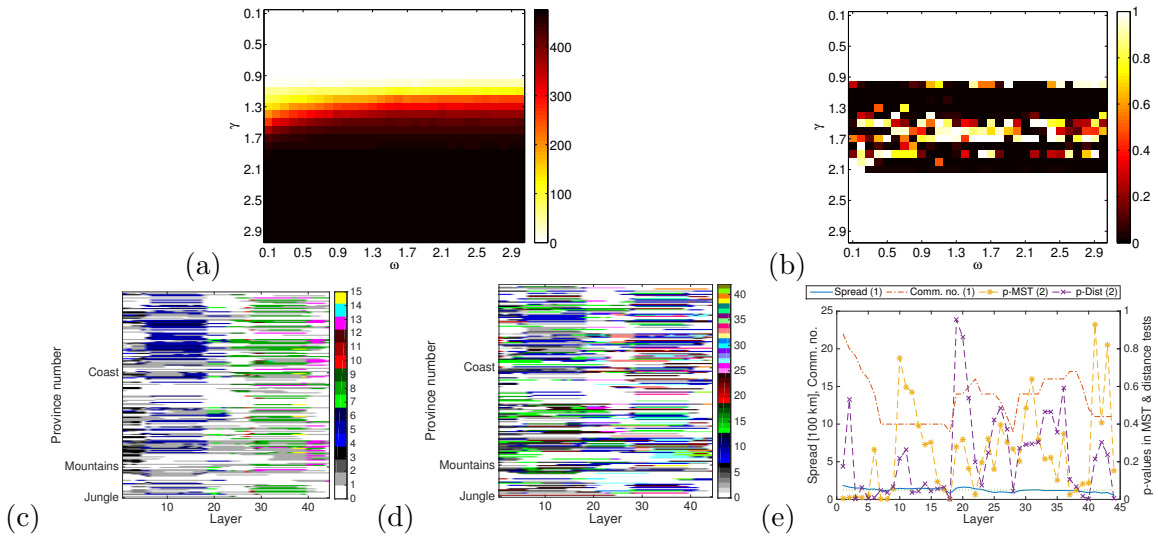


Figure 7.3: Rubella, multislice networks, NG null model: spatial organization of partitions. In parts (a)–(b) we show results of varying the parameters γ and ω: (a) z-Rand scores for similarity to "spatial" partitions by climate, (b) the p-values for distances being smaller than expected at random in the distance test. In (c)–(e) we examine the multislice community structure for select partitions. In (c)–(d) we plot community structure for (c) $\gamma = 1$, $\omega = 0.1$ and (d) $\gamma = 1.1$, $\omega = 0.1$; nodes are on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (e) we plot values for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1.1$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

We also study the temporal organization of the partitions using z-Rand score tests versus temporal partitions. As described in Section 3.4.3, we search for one or two critical time points that mark changes in community structure. Partitions for $\gamma \in \{1, 1.1, 1.2\}$ and $\omega \lessapprox 1$ stand out as scoring comparatively high temporal z-Rand scores, with a critical time point at layer 20 (year 2002) detected by both methods [see Fig. 7.4]. Layers 18–37 roughly correspond to the period with a small number of disease cases between the two large

epidemics, with many provinces not recording any disease cases over several time windows and thus their corresponding nodes having zero strengths. At the critical time point, we see a division of community 5 into two communities [dark blue and green in Fig. 7.3(c)].

On visual inspection of the community structures, we observe a second large change in community structure for $\gamma = 1$, and $\omega = 0.1$ near layer 40 (early 2006). This coincides with a second reduction in disease prevalence across the provinces, shrinking the number of nodes with non-zero strengths in layers 40 onwards. It might be related to the rubella vaccination campaigns that were started in Peru in 2003–2005. This change is not detected by our temporal z-Rand score methodology; perhaps increasing the number of critical time points would allow the methodology to detect it, but this comes at a high increase in computational complexity if one considers all possible combinations of a larger number of critical time points.

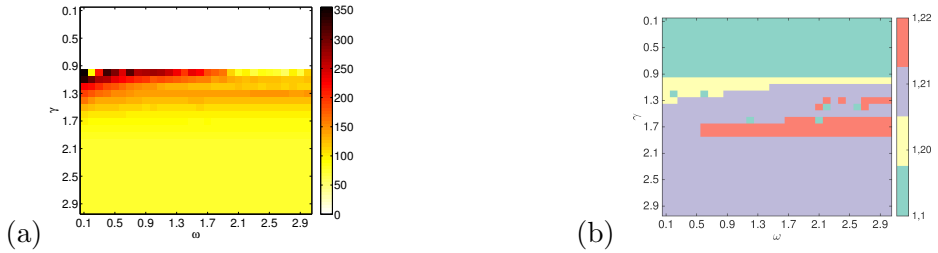


Figure 7.4: Rubella, multislice networks, NG null model — temporal organization of partitions. Results of varying the parameters γ and ω for the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}. (a) The maximum z-Rand score selected out of all t_{c1} and t_{c2} pairs. (b) The pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of their layer numbers (for a comparison versus a partition with two critical time points). The layer numbers of the critical time points are indicated on the colorbar.

7.2.2 Modularity maximization using the correlation null model

If we maximize modularity using the correlation null model on the rubella static networks and use z-Rand scores to study the spatial organization of the networks, we observe strong spatial partitions before 2003. Notably [see Fig. 7.5(a)], we see spatial partitions in layers 1–80, with two peaks in z-Rand scores near layers 10–20 (in year 1998) and 50–70 (around 2002–2003). Both z-Rand scores and distance test p-values are relatively similar for all γ values. Partitions in networks 1–54 consistently score as significantly spatial in both tests [see Fig. 7.5(a)–(b)].

In the absence of large variation with changes in γ, we select the standard $\gamma = 1$ value for detailed examination. We show example community detection results for layer 6 (which consistently exhibits significant p-values in the distance test and high climate z_R). This partition is composed of 4 communities: community 1 is the smallest [see Fig. 7.6(a)] and it experiences small rubella outbreaks in 2000 and 2006 [see Fig. 7.6(b)]. Communities 2 and 4 are also relatively spatially compact, with the exception of a few outlier nodes. They exhibit

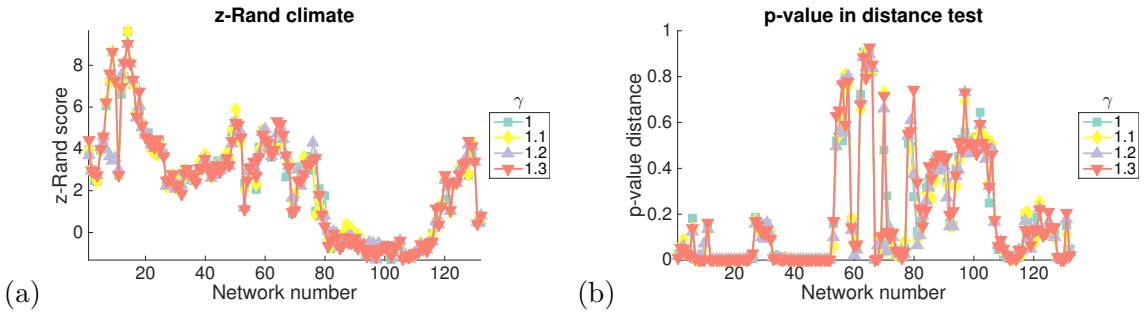


Figure 7.5: Rubella, static networks, correlation null model — spatial partitions. Properties of the algorithmic community structure for the 132 static networks covering the whole time period (horizontal axis) for $\gamma \in \{1, 1.1, 1.2, 1.3\}$: (a) Plot of the z-Rand scores versus the detailed climate partition, and (b) p-value in the distance test.

rather different temporal disease patterns, with nodes in community 2 recording a relatively small number of disease cases which occur during the major epidemic outbreaks. In contrast, communities 3 and 4 experience recurrent disease outbreaks until the countrywide decline in the occurrence of rubella (year 2007).

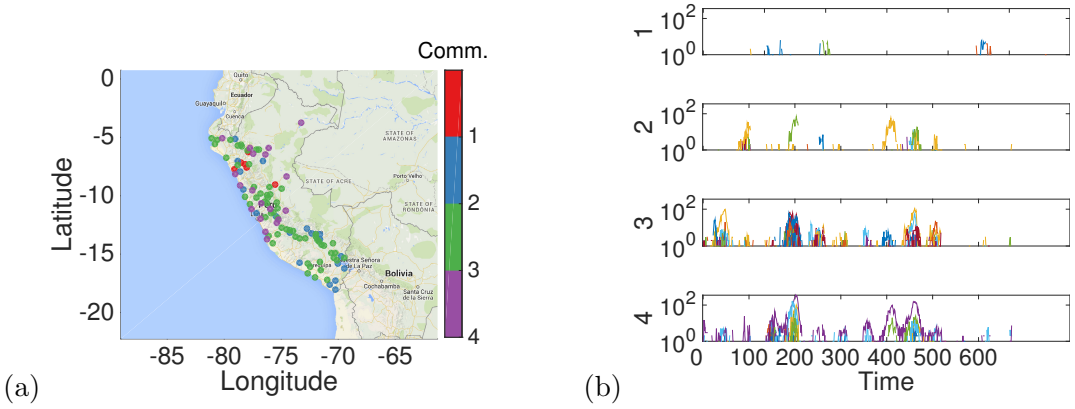


Figure 7.6: Rubella, static networks, correlation null model — network 6 at $\gamma = 1$. In panel (a) we show a map of all the nodes in network 6 at $\gamma = 1$ (colored by algorithmically detected community assignment, community number indicated on the color bar), and in panel (b) we show the time series of disease occurrence in the provinces assigned to these communities, with community number indicated on the vertical axis.

When we study the spatial organization of community structures detected by maximizing multislice modularity with the correlation null model, we see (similarly to the dengue fever case) that partitions for all γ and ω values score as significantly spatial using both the z-Rand scores and the partition-wide distance test (not shown). We study an example partition for $\gamma = 1$, $\omega = 0.1$. As we compare the partition-wide distance test with the per-layer distance test results, we see that the overall significantly spatial partition-wide score is driven by the highly-spatial partitions in layers 1–18 (1997–2002) [see Fig. 7.7(b),(c)]. These partitions have a small number of large communities that do not appear to be related to climate [see Fig. 7.7(a)]. The last spatial partition roughly corresponds to the onset of the period with low case numbers and low numbers of nodes per layer; the partitions for

both this period, and the second large rubella epidemic that follows, do not appear to be significantly spatial as measured by the per-layer distance and MST tests [see Fig. 7.7 (b)].

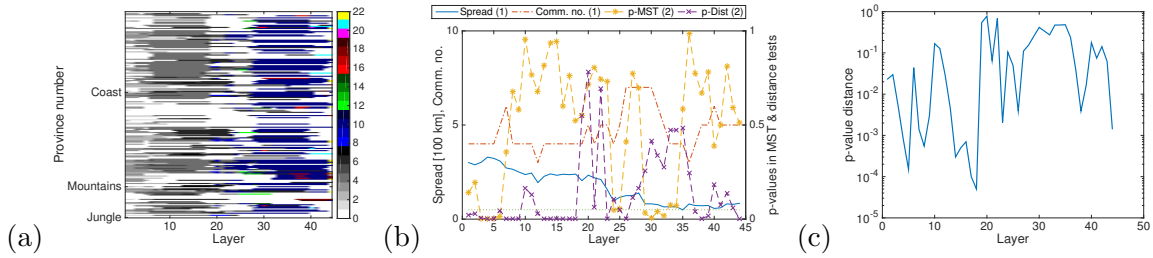


Figure 7.7: Rubella, multislice networks, correlation null model — spatial organization of partitions. We examine the multislice community structure for select partitions. In (a) we plot community structure for $\gamma = 1$, $\omega = 0.1$; nodes are on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (b) we plot values for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1.1$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests. In (c) we show the p-values in the distance test in detail, which we plot on a logarithmic scale to highlight the low p-values in layers 1–18.

When we study the temporal organization of network partitions generated by maximizing multislice modularity using the correlation null model, we observe that structures for all values of ω and γ score high temporal z_R. The critical time point for all parameter values is layer 20, similarly to what we observed on the NG null model partitions (not shown).

7.2.3 Modularity maximization using the gravity null model

If we apply modularity maximization using the gravity null model to the static and multislice rubella networks, we see that similarly to what we found for dengue, partitions obtain lower spatial z-Rand scores than partitions using NG and correlation null models, and they do not appear significant in the distance test (the results are shown in Section B.2). Both static and multislice network partitions tend to consist of one large community and a small number of singleton communities [see static networks in Fig. B.2(a) and multislice in Fig. 7.8(a)–(b)].

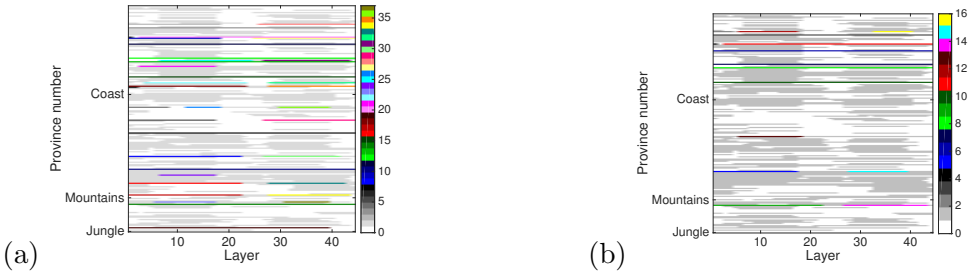


Figure 7.8: Rubella, multislice networks, gravity null model — spatial organization of partitions. We plot community structure for (a) $\gamma = 2.5$, $\omega = 2.9$ and (b) $\gamma = 1$, $\omega = 0.1$; nodes are on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar.

However, when we study the temporal organization of the multislice rubella community structure using modularity maximization with the gravity null model, we see that the z-Rand scores suggest many more different layers than for NG and correlation null models as

potential critical time points, and the z-Rand scores they obtain are very high (see Fig. 7.9). The majority of critical time points with high z-Rand scores correspond to the times when some of the nodes in singleton communities disappear from the network during the period with small numbers of disease cases starting around layer 20. These nodes form their own singleton partitions again when they return to the network around layers 27–28. Finally, layers 40–44 are detected as critical time points for some parameter values; these layers correspond to the drop in rubella case numbers and a fall in the number of nodes with non-zero strength per layer (see example structures in Fig. 7.8). We detected similar changes (visually) in community structures detected using the NG null model in Section 7.2.1, but they were not significant using the temporal z-Rand scores.

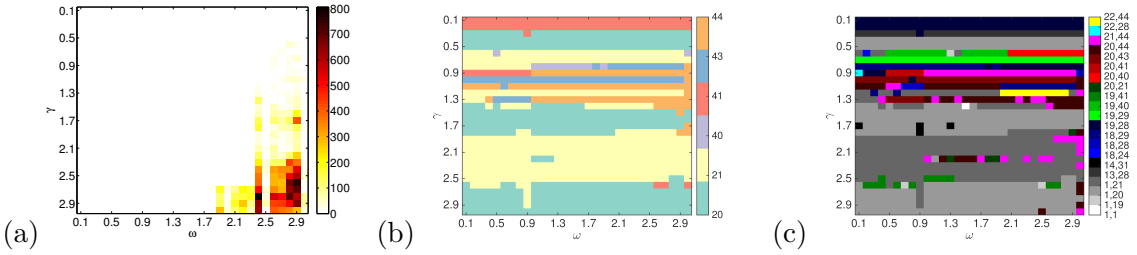


Figure 7.9: Rubella, multislice networks, gravity null model — temporal organization of partitions. In (a)–(c) we show results of varying the parameters γ and ω for: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

7.2.4 Modularity maximization using the radiation null model

Finally, we maximize modularity using the radiation null model on the static networks created from the rubella data set. Similarly to the gravity null model, most of the static partitions score below the significance threshold of 1.96 for spatial z-Rand scores, and the p-values are above 0.05 for the distance test, suggesting that the radiation null model removes most of the spatial variability in the data [see Fig. B.4(a)-(b)].

Similarly to our results for the gravity null model, the structures tend to consist of one large community and the remaining nodes assigned to singleton communities that are spread throughout the country and they contain some of the most populated provinces of Peru. Additionally, for many of the yearly epidemics the singletons experience more disease cases than the nodes in the large community (see Fig. B.5).

The results for modularity maximization on multislice rubella networks using the radiation null model are also very similar to the results of modularity maximization using the gravity null model (see Fig. B.6). The structures show little spatial organization (as detected by z-Rand scores versus climate and administrative partitions, and by the partition-wide

distance test). The z-Rand scores against temporal partitions are again quite high, with many possible critical time points suggested by the method.

7.2.5 Summary of findings for the rubella data set

When we apply our methodology of community detection using modularity maximization with different null models to the rubella data set, we are able to detect spatial communities, especially when using NG and correlation null models. The correlation null model might be slightly better for detecting spatial structures, as the z-Rand scores are higher, and the distance test p-values are lower for this null model than for the NG null model. Both NG and correlation null models detect layers 1–70 as the time period with the strongest spatial organization in the static networks, as shown by high z-Rand scores versus spatial partitions, and significant p-values in the distance test. This period contains the first large epidemic. This result is similar to what we found for dengue, where the partitions with the strongest spatial structure occurred during the largest countrywide disease epidemic. However, the networks containing the second large epidemic of rubella do not appear to possess a significant level of spatial organization. It might be interesting to further study the reasons behind these changes in the context of the vaccination that began in 2003–2005, during the period between the two large epidemics. It is possible that the immunity provided by the vaccination somehow affected the disease patterns to disrupt spatial organization — however, as have been unable to obtain detailed data on dates and locations of vaccination campaigns, we did not pursue this question further.

Spatial partitions. According to our multislice measures of spatial organization (the z-Rand score against administrative and climate partitions, and the partition-wide distance test), community structures found using the NG and correlation null models tend to have a relatively strong spatial organization, especially in the first 18 layers of the multislice networks. However, the community structures are not related to climate or administrative divisions, and we have been unable to find an explanation for the groupings. Perhaps groupings based on sociological data, transportation patterns, or other variables that we do not currently possess the data for, would be able to explain the community structures that we observe.

When we apply modularity maximization with spatial null models to static and multislice rubella networks, we once again obtain partitions with one large community and a small number of singletons (or sometimes, very small communities). This suggests that the spatial null models may remove the majority of the structure present in the rubella disease-correlation networks.

Temporal partitions. We are able to detect two strong temporal divisions in the rubella data, both of which are related to a large fall in the number of disease cases and in the number of provinces affected by the disease. This leads to a reduction in the number of nodes with non-zero strength in the corresponding layers of the multislice network. The first critical time point is related to the period with lower number of disease cases that begins in 2002 (this t_c scores significantly in our temporal z-Rand scores for all null models). The second critical time point corresponds to a drop in the number of new infections after 2008; these kinds of temporal partitions can be visually observed for some parameter regimes using the NG null model but this time point is only detected as a statistically significant possible critical time point by our z-Rand score methodology for partitions found using the gravity and radiation null models.

The temporal z-Rand scores for the gravity and radiation null models are much higher than for the other two null models, and they suggest a larger number of potential critical time points. Thus it appears that at least for this disease, removing the majority of spatial organization allows us to easier detect temporal partitions.

7.3 Seasonal influenza in Chile

7.3.1 Introduction

The Chilean influenza data set contains weekly counts of new disease cases from the 15 provinces of Chile over 7 years (365 weeks) between 1 January 2004 and 31 December 2010. The country forms a thin band on the southwestern side of South America, and the 15 provinces are organized from north to south. Our data set shows yearly countrywide epidemics of seasonal influenza and the 2009 "swine flu" epidemic, which increased the amount of influenza activity in Chile. We describe this data set in detail in Section 5.2.1.3.

For this data set, we use time-window width of $\Delta = 30$ and we let the difference between time window starting points $\upsilon = 4$ to generate a set of 82 static networks. We also use $\Delta = 30$ and $\upsilon = 30$ to generate a multislice network with 11 layers. We describe the justification for parameter choices and the general properties of the time series and the static networks in Section A.2.

7.3.2 Summary of results

The Chilean influenza data set exhibits less spatial organization of communities for the NG and correlation null models than the rubella and dengue data sets. It shows considerable variation in the spatial organization scores of partitions, with partitions for similar parameter values, and partitions corresponding to overlapping static networks showing very different community structures for the same null models. Both static and multislice partitions sometimes show a degree of spatial organization that is related to a north-south

division in node assignment to communities [see Fig. 7.10 for an example of a multislice community structure and the per-layer scores of spatial organization]. This suggests that the north-south location of nodes influences disease patterns (which is expected due to the large north-south climate variability), but the reliability of detailed node assignments to particular communities appears to be low. The results for gravity and radiation null models once again consist of one large community and a small number of singletons (containing the highest-populated nodes).

In multislice networks, we detect relatively strong temporal partitions for some parameter regimes when using the NG and correlation null models [see Fig. 7.10 (a),(c) for example partitions and Figs. B.10 and B.13 for the temporal z-Rand score values and critical time points over various γ and ω parameter values]. The critical time points for the NG null model and correlation null model are layers 6 and 5, corresponding to March 2007 and August 2006. Both of these time points lie during the period with the lowest yearly number of disease cases covered by this data set [see Fig. 5.3 (c)-(d)]. It appears that the temporal partitions may be linked to this low number of disease cases. This is a similar result to the temporal partitions that we detected in the rubella data set, which correspond to an inter-epidemic period. However, in contrast to the rubella data, for chilean influenza the method detects a temporal partition that does not have a corresponding fall in the number of regions that experience disease cases and thus have non-zero strengths.

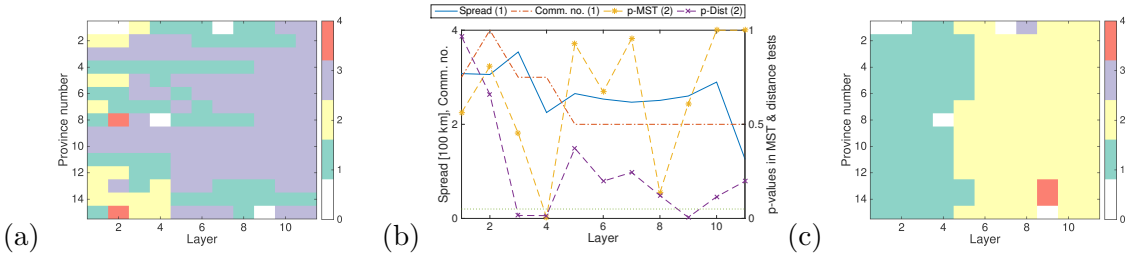


Figure 7.10: Influenza in Chile, multislice networks, NG and correlation null models: spatial organization of partitions. In (a) we plot multislice community structure with nodes ordered by their location (north to south) on the vertical axis and layers on the horizontal axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (b) we plot values for each layer, (left vertical axis) the community spread and the number of communities — both for the NG null model with $\gamma = 1.1$, $\omega = 0.1$. In (c) we plot the multislice community structure for the correlation null model with $\gamma = 0.9$, $\omega = 0.3$.

7.4 Conclusions

In this chapter we applied the network construction and community detection pipeline that we developed in Section 3.5 and previously applied in Chapter 6 for the dengue fever data set to study the patterns of incidence of rubella in Peru and seasonal influenza in Chile, endemic diseases that are established in the local populations and cause repeated (often yearly) epidemics. The results are summarized in Table 7.1.

Table 7.1: Overview of the results of community detection for all endemic disease data sets and all null models.

Data set	Null model	Results
Dengue	NG	• Spatial partitions during epidemics, related to climate • Temporal partitions: 2000–2001 epidemic, start of yearly epidemics • Different patterns of disease occurrence in the whole time series for province-level communities
	Correlation	• Similar as for NG • Little variation with γ
	Grav. & Rad.	• One large community, many singletons with high populations
Rubella	NG	• Spatial partitions during first epidemic, unrelated to climate • Temporal partitions related to inter-epidemic period
	Correlation	• Similar as for NG • Little variation with γ
	Grav. & Rad.	• One large community, many singletons with high populations • Temporal partitions related to inter-epidemic period
Influenza	NG	• Little reliable spatial organization • Strong temporal partition related to the year-long period with low disease case numbers
	Correlation	• Similar as for NG • Little variation with γ
	Grav. & Rad.	• One large community, many singletons with high populations • Temporal partitions related to the year with the lowest number of infections

By maximizing modularity using the NG and correlation null models, we found communities in the rubella networks that have a comparable degree of spatial organization than in the dengue networks. The spatial patterns appeared clearer, and the scores were higher and more repeatable across parameter values when using the correlation null model.

These results suggest that there might be significant local patterns in disease correlation even for diseases that are not explicitly dependent on spatial factors such as climate. This might be due to mechanisms of disease spread spread such as interpersonal contacts, which in turn are affected by transportation networks.

The ability of our methodology to detect spatial communities appears to be better for the rubella data set than for the influenza data set, in which we only found a small degree of spatial organization that is very parameter-dependent. This could be because the spatial patterns that depend on the transport between regions are more visible when the disease

dies out and is being reintroduced to locations, which happens periodically in all regions (and even the whole country) for rubella. We will further study this idea using a spatial model of disease spread in Chapter 9. However, it could also be due to the small number of regions present in the chilean network, or other reasons.

When we take into account the expected influence of distance on community structure through the use of a spatial null model, we once again observe one or two large communities and a small number of singleton communities that contain the highest-populated nodes.

When using multilayer networks, our findings on the spatial organization of partitions are similar as for the respective static networks. Temporal partitions of multilayer networks allow us to detect several important time points in the prevalence history, such as a large lull between epidemics for both rubella and influenza, and the beginning of the decline in the number of new disease cases for rubella. However, most of the critical time points that we detect for rubella are related to a change in the number of provinces that are experiencing disease epidemics in contiguous layers. The NG and correlation null models detect the main changes in disease patters, and the gravity and radiation null models detect more potential critical time points with higher z-Rand scores. For influenza in Chile, in contrast to the rubella data, the method detects a temporal partition that does not have a corresponding fall in the number of regions that experience disease cases, which is a useful demonstration of its ability to detect important temporal change points.

In the last two chapters we used a variety of null models and parameter values to study the spatial spread patterns of three infectious diseases: dengue, rubella and seasonal influenza. We found spatial partitions related to climate for the vector-borne dengue data set, and for the rubella data set we found (yet unexplained) spatial partitions. Both of these diseases die out and reinfect regions periodically. For the seasonal influenza, which is present in the general population at higher levels and does not die out as often, we failed to find spatial or temporal partitions.

In Chapter 8, we will apply the methodology from the previous chapter to case count data for emerging diseases (diseases that have recently entered a susceptible population). These data sets each describe the spread of a disease during one epidemic wave. We will test the ability of our methodology to detect information about the spatial patterns in this type of disease data. We expect that the large influence of transport on the patterns of the dissemination of infections into new regions may influence the spatial organization in the community structures for emerging diseases.

In Chapter 9, we attempt to disentangle the factors that influence the ability of community detection methodology to find spatial communities and planted communities in disease-correlation networks generated from synthetic time series generated from an agent-based model that describes the spread of a both endemic and emerging diseases through a set of interconnected locations.

Chapter 8

Applications to emerging disease epidemics

This chapter consists of original work by MS and M. A. Porter which is not yet published.

8.1 Introduction

In this chapter, we apply the approach of modularity maximization with different null models (that we used for analyzing the spatial spread of endemic diseases in Chapters 6 and 7 and Appendix B) to data sets concerning emerging diseases (diseases that have only entered a susceptible population). These data sets each describe the spread of a disease during one epidemic wave. We use two data sets related to the West African Ebola epidemic (the data is described in Section 5.2.2.1). We also analyze a data set about the 2009 H1N1 influenza in Mexico ("swine flu") described in Section 5.2.2.2. We hope to shed light on the question of whether our methodology can help to gain more information about the patterns of the spread of emerging infections. We expect that the influence of transport on the patterns of the occurrence of infections in new regions may generate a degree of spatial organization in the community structures that we find for emerging diseases.

We compare two data sets for Ebola (see the detailed discussion in Section 5.2.2.1): one of them (which we call "the WHO data set") starts early in the epidemic (5 January 2014) and covers the time when the disease was spreading to new provinces along transport links, which we believe might influence the results of community detection. However, the data are only collected weekly, and as a result the data set is short, with 54 weeks and 63 nodes. The second data set (which we call "the Datamarket data set") is collected daily; it starts after the epidemic reached the majority of provinces but due to the daily collection, the time series are long enough for us to construct multislice networks. Comparing the two data sets allows us to gain some insight into the benefits of having data with higher temporal resolution, and study whether including the early disease phase (when it spreads to new provinces) makes a difference to the community structure.

We present the results of modularity maximization using the Newman-Girvan (NG), correlation, gravity, and radiation null models (described in Section 3.3) for all three data sets. For each of these, we examine spatial organization of community structures that we detect in the static networks. We first measure the extent of spatial organization for different values of the resolution parameter $\gamma \in \{0.1, 0.2, \ldots, 3\}$ and across networks, using the distance test (described in Section 3.4.4). We then select particular parameter values and networks to study in more detail.

For multislice networks, we study the spatial and temporal organization of algorithmically-detected community structures. We use the multislice versions of the climate z-Rand scores and distance test for detecting spatial organization. We search for critical time-points when community structure changes using the temporal z-Rand score methodology that we described in Section 3.4.3. We select interesting parameter values across values of the resolution parameter $\gamma \in \{0.1, 0.2, \ldots, 3\}$ and the inter-layer coupling $\omega \in \{0.1, 0.2, \ldots, 3\}$ to study their community structures in detail.

The Ebola Datamarket data set appears to exhibit some spatial organization that appears to be related to country boundaries. Both NG and correlation null models detect partitions with a mixture of temporal and spatial features in the multislice networks (depending on parameter values). The gravity and radiation null models fail to give useful insights into the factors that affect the spread of disease or provide additional temporal partitions; their results are shown in Appendix C.

The spatial communities that we detect in the Ebola WHO data set using modularity maximization exhibit a pattern in their disease time series, with different times of epidemic onset for nodes in different communities. Further, the partitions for NG and correlation null models show a large degree of spatial organization, which appears to be related to country boundaries, but has a different organization than the structures found in the Datamarket data set.

For the H1N1 data set, the methodology did not find reliable and significantly spatial and informative partitions in the static networks, and thus the majority of results are only included for completeness in Appendix C. In the multislice networks, but NG and correlation null models detect temporal partitions that might correspond to the peak and the end of the epidemic wave of swine flu, and we briefly show these results in this chapter.

8.2 Ebola — Datamarket data set

The Ebola Datamarket data set contains daily new case count data for 105 days starting in August 2014, and originating from 63 provinces of Guinea, Sierra Leone and Liberia. We described the data set in detail in Section 5.2.1.2. We use a time-window width $\Delta = 60$ and a difference between time-window starting points of $\upsilon = 5$ to generate a set of 9 static networks. We use the same parameter values to generate a multislice network with 9 layers.

We describe the justification for parameter choices and the general properties of the time series and the static networks in Appendix A.2.

The Ebola Datamarket data set appears to exhibit some spatial organization. Both NG and correlation null models detect spatial partitions in the static networks, and they detect partitions with a mixture of temporal and spatial features in the multislice networks (depending on the parameter values). The critical time points correspond to the time when the number of disease cases peaked, and the beginning of the decline in new cases.

8.2.1 Modularity maximization using the NG null model

When we maximize modularity for the Ebola Datamarket static networks using the Newman-Girvan null model, we obtain communities that score significantly ($p < 0.05$) in the distance test, with $\gamma \in \{1.1, 1.2, 1.3\}$ giving spatial communities across the largest number of networks. The communities that we find in networks 6–9 appear to have the strongest level of spatial organization according to the test [see Fig. 8.1(a)]. Here we present partitions for $\gamma = 1.1$, as a compromise between high spatial scores and increasing number of communities for increasing γ. For $\gamma = 1.1$ in layer 6, the majority of provinces in Sierra Leone form a separate spatially compact community [yellow in Fig. 8.1(b)], as do several provinces in Liberia [purple and orange in Fig. 8.1(b)]. The community assignment in Guinea appears to be more complicated and does not form an obvious spatial pattern.

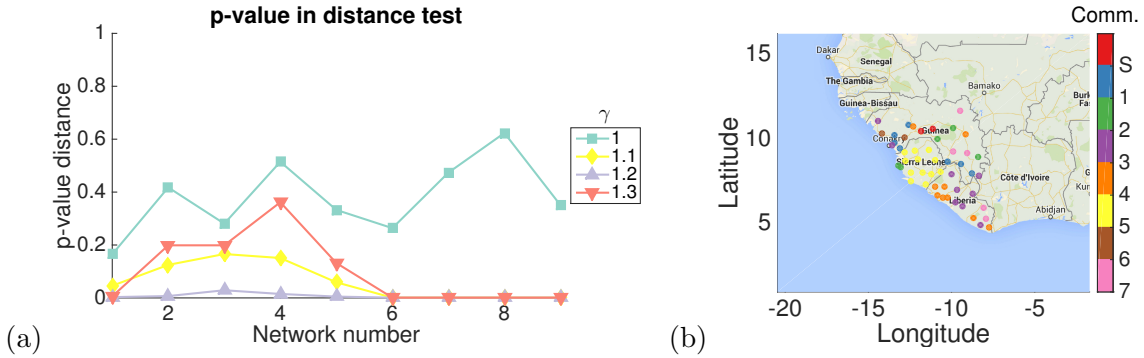


Figure 8.1: Ebola Datamarket data set, static networks, NG null model: spatial partitions. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3\}$. In panel (b) we show a map of all the nodes in network 6 at $\gamma = 1.1$ (colored by algorithmically detected community assignment with singletons grouped into group S, community assignment indicated on the color bar).

For multislice networks, spatial community structures (as detected by the partition-wide distance test) occur at $1 \lessapprox \gamma \lessapprox 2.3$ (see Fig. 8.2). Network partitions change relatively little in time [see Fig. 8.2(b)–(c)]. The community structure for $\gamma = 1, \omega = 0.1$ has 3 communities, and the community number increases to 9 for $\gamma = 1.1, \omega = 0.1$; both structures have a sharp transition at layer 6, at which a new large community forms containing 11 nodes from Sierra Leone for $\gamma = 1$, and for $\gamma = 1.1$ a smaller community grows to include the same group of nodes. This change point could correspond to the rise in the number of disease cases in

Sierra Leone relative to the other two countries, which began to experience large numbers of disease cases later than the other two countries.

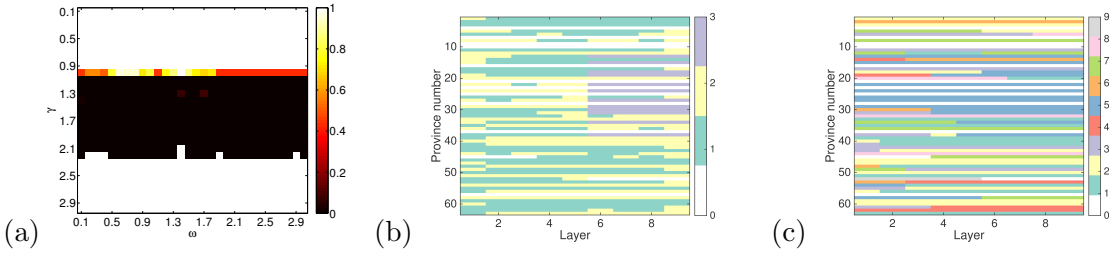


Figure 8.2: Ebola Datamarket data set, multislice networks, NG null model: spatial organization of partitions. In panel (a) we show effects of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b)-(c) we plot community structure with nodes ordered by their location (country-wide then north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for (b) $\gamma = 1$, $\omega = 0.1$ and (c) $\gamma = 1.1$, $\omega = 0.3$. Community number is indicated on the colorbar.

The highest z-Rand scores for temporal partitions of the multislice network occur for $\gamma = 0.9$ and $\gamma = 1$, and $\omega \lessapprox 0.7$. These structures correspond to partitions at layer 6 if one searches for a single critical time t_c or layers 4 and 6 if one seeks two critical times (see Fig. 8.3). Layer 6 corresponds to the sharp growth of the Sierra Leone community that we observed for static networks and for both multislice structures in Fig. 8.2(b)–(c). Layer 4 is the first layer that includes the peak in case numbers, which could be related to the location of this temporal partition.

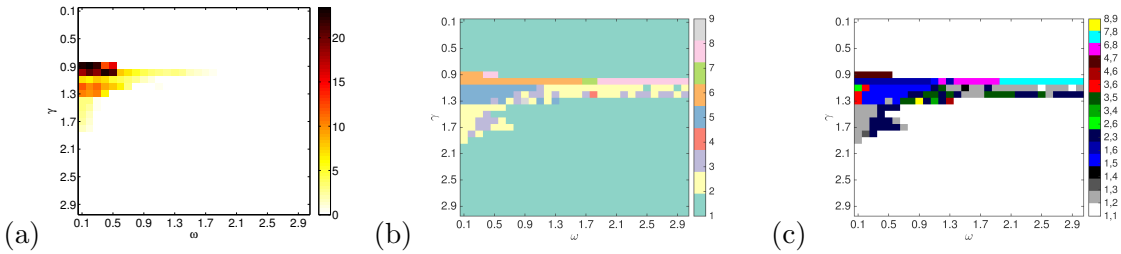


Figure 8.3: Ebola Datamarket data set, multislice networks, NG null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω on: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}, (b) the highest-scoring t_c in terms of the layer number (for a comparison against a single critical time point partition), and (c) pairs of highest-scoring t_{c1} and t_{c2} (for a comparison against a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

8.2.2 Modularity maximization using the correlation null model

By maximizing modularity with the correlation null model, we detect spatial communities (using the distance test) for layers 1–3 and 6–9, and the scores do not vary strongly with the resolution parameter γ. We focus on $\gamma = 1.1$ as it scores lowest p-values in the distance test. We choose network 8 as an example; for this network, most of the provinces of Sierra Leone are assigned to a single community, which also contains some of the bordering nodes

from Guinea and a couple of outlying nodes from Liberia. This could once again be related to the relatively large number of disease cases in this country in the later part of our data set. We also observe another 2 communities that are composed of 2–3 spatial clusters of nodes [see Fig. 8.4(b)].

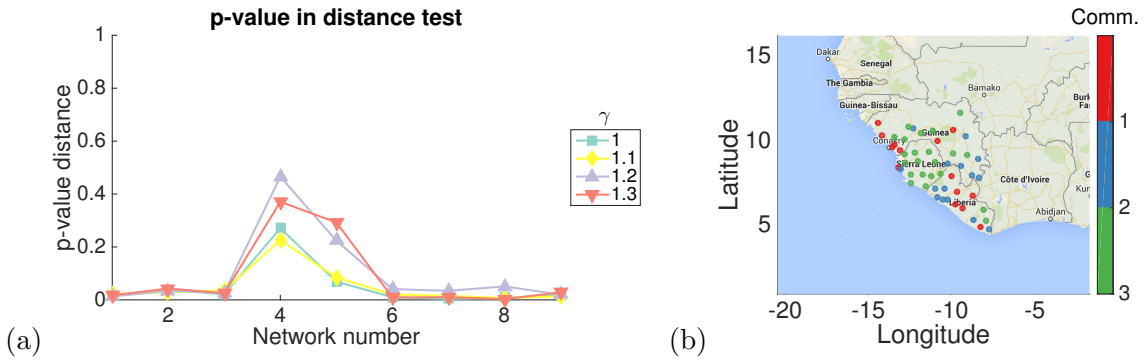


Figure 8.4: Ebola Datamarket data set, static networks, correlation null model: spatial partitions. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3\}$. In panel (b) we show a map of all the nodes in network 8 at $\gamma = 1.1$ (colored by algorithmically detected community assignment, which is indicated on the color bar).

Community structures of the multislice networks that we find by maximizing modularity using the correlation null model once again score high in spatial tests for all parameter pairs that we tested. An example structure for $\gamma = 2.4$ and $\omega = 0.2$ (which had the lowest p-value in the distance test) contains 3 large communities out of a total of 11 [purple, green and yellow in Fig. 8.5(a)]. This multislice community structure contains spatial partitions in layers 1–3 and 6–9 as scored by the per-layer distance test — the layers that contain the largest communities, including a community containing nodes from Sierra Leone that is similar to what we observed using the NG null model.

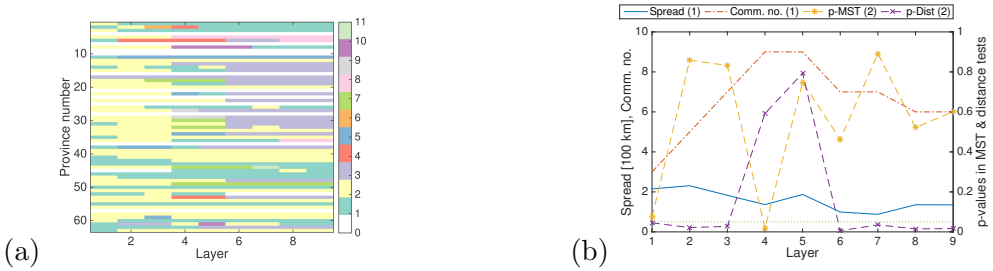


Figure 8.5: Ebola Datamarket data set, multislice networks, correlation null model: spatial organization of partitions. In (a) we plot community structure with nodes ordered by their location (country-wide then north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for $\gamma = 2.4$, $\omega = 0.2$. Community number is indicated on the colorbar. In (b) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 2.4$, $\omega = 0.2$ and (right vertical axis) the p-values for the distance and MST tests.

The partitions that score the highest in the z-Rand score tests against temporal partitions occur for $0.8 \lessapprox \gamma \lessapprox 1.3$ and $\omega \leq 0.5$. They once again correspond to partitions at layer 6 [see Fig. 8.6].

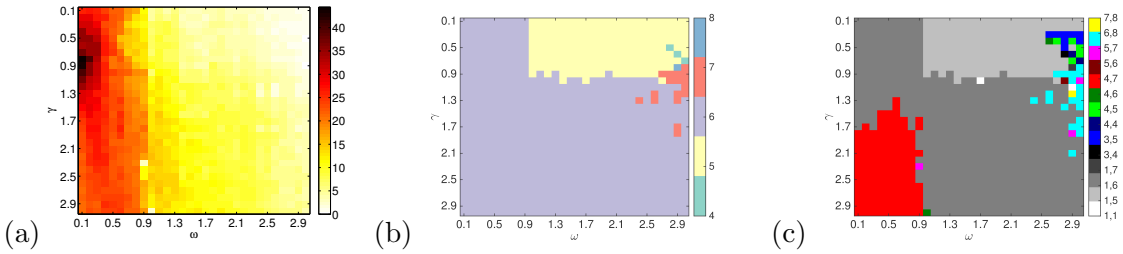


Figure 8.6: Ebola Datamarket data set, multislice networks, correlation null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω on: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}, (b) the highest-scoring t_c in terms of the layer number (for a comparison against a single critical time point partition), and (c) pairs of highest-scoring t_{c1} and t_{c2} (for a comparison against a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

8.2.3 Modularity maximization using the spatial null models

When maximizing modularity for static networks using the gravity and radiation nulls models, community structures again contain one large community and several singleton communities with the highest-populated provinces. The multislice structures have the same basic organization, and they change little in time. These structures tell us little about the patterns of the spread of Ebola in time and space. We therefore do not show them in the main body of the thesis, but they are presented for completeness in Appendix C.2.

8.2.4 Summary of results for the Ebola Datamarket data set

The partitions for the Ebola Datamarket data set appear to exhibit a stronger level of spatial organization than the partitions found in the rubella data, with partitions that both score and visually appear spatial (they appear to be at least partially related to the country boundaries). This strong spatial organization might be related to the fact that the Ebola epidemic is still spreading across space, and the effect of travel on the number of new disease cases is thus relatively strong. In light of this, it would be very useful if the data set that we use to build the disease-correlation networks included the first infection times. For this reason, we hope that we might be able to see interesting effects in the WHO data set which begins much earlier in the epidemic than the Datamarket data set.

For the multislice networks, both NG and correlation null models detect partitions with a mixture of temporal and spatial features (depending on the parameter values). Both static and multislice community structures appear to score the highest in spatial tests for layers 1–2 (August–October, early in the data set but around the peak of the whole epidemic wave) and 6–9 (November–December, later periods in the epidemic when numbers of new cases began to decline).

The gravity and radiation null models fail to give useful insights into the factors that affect the spread of disease. They also do not provide additional temporal partitions for the Ebola Datamarket data.

8.3 Ebola — WHO data set

The WHO data set about the West African Ebola epidemic consists of 54 weeks of data for the 63 provinces of Guinea, Liberia, and Sierra Leone. As we discussed in the introduction to this chapter and in Section 3.1.1, the length of the Ebola WHO data set is shorter than the recommendation for correlation matrices based on random matrix theory ($\Delta > \hat{N}$ where Δ is the length of the time window and $\hat{N}$ is the number of nodes with non-zero strength in the network or layer). However, in an ongoing epidemic situation researchers do not have the ability to wait for more data to become available, and in this context it is interesting to see what can be gleaned from it despite the potential issues. In this section, we attempt to analyze this data set in order to see whether including the times when it was first found in many provinces, enables us to find interesting spatial communities and potentially shed light on the paths that the disease takes to spread across space. We generate a single static network from the whole data set.

Despite the short length of the time series, the data set shows promising results, especially using the NG null model.

8.3.1 Modularity maximization using the NG null model

Modularity maximization using the NG null model detects rather different disease patterns in the WHO network than in the Datamarket network. We present two of the spatial partitions that score as spatial in the distance test. At $\gamma = 1$, the region is divided into two north to south communities of roughly similar size. Note that Sierra Leone is split between the two communities (see Fig. 8.7) — in the Datamarket data set this country formed a highly spatial community for the NG null model. It appears that the communities correspond to regions with slightly earlier and later onset and peak of the main epidemic wave [see Fig. 8.7(b)].

For $\gamma = 1.1$, the number of communities increases to 5, with the southern community largely preserved and the northern community spilt into smaller, roughly spatial, parts, and one singleton. The coastal regions of Guinea and Liberia remain in large communities. The communities appear to show different first infection times and different peak epidemic times [see Fig. 8.7(c)–(d)].

8.3.2 Modularity maximization using the correlation null model

Community structures that we find by maximizing modularity using the correlation null model on the WHO network usually consist of a small community with regions on the border of the three countries (where the epidemic initially developed), some singleton nodes (often with a small number of infections), and the rest of the nodes placed in one large community (see an example partition for $\gamma = 0.6$ in Fig. 8.8). Thus, this methodology may be able to

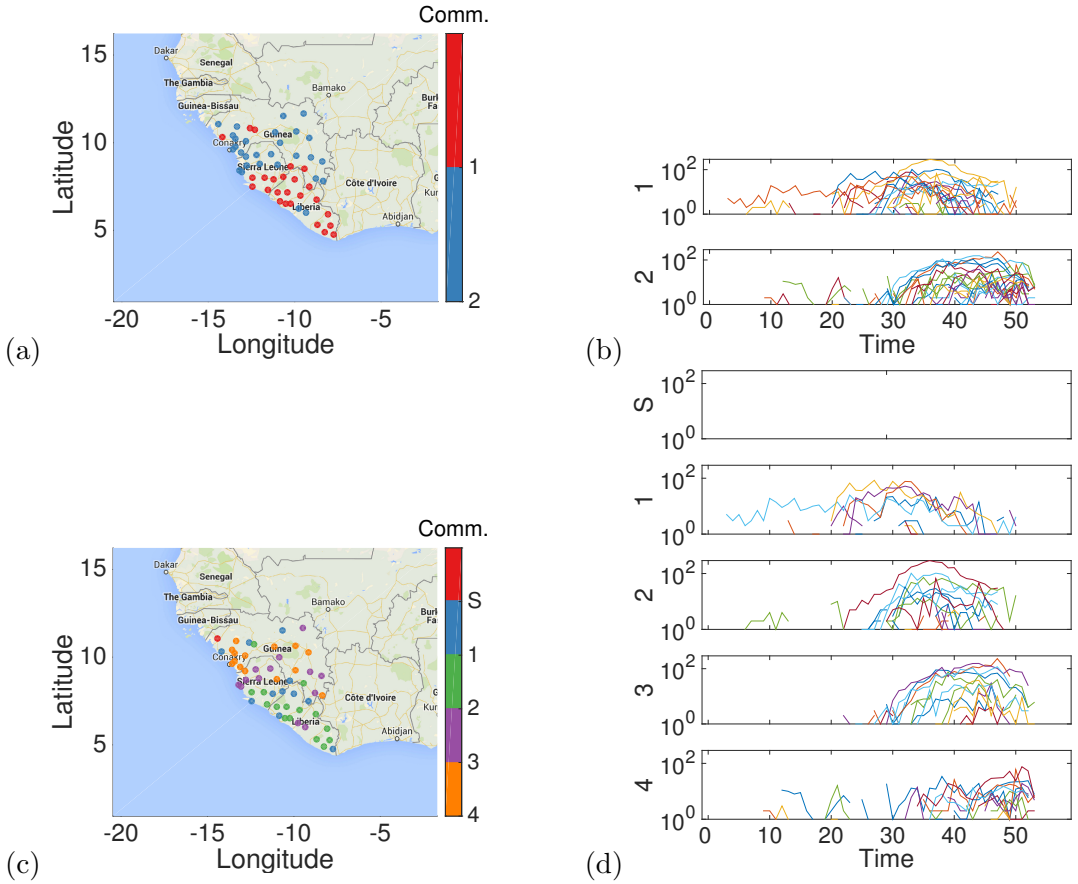


Figure 8.7: Ebola WHO data set, static network, NG null model — example partitions. Example partitions (a)-(b): $\gamma = 1$, (c)-(d): $\gamma = 1.1$). (a,c) Map of all the nodes colored by algorithmically detected communities, (b,d) the time series of disease occurrence for nodes in these communities.

detect the early start of the epidemic in its source nodes. Unfortunately, changing γ does not significantly change the community structure.

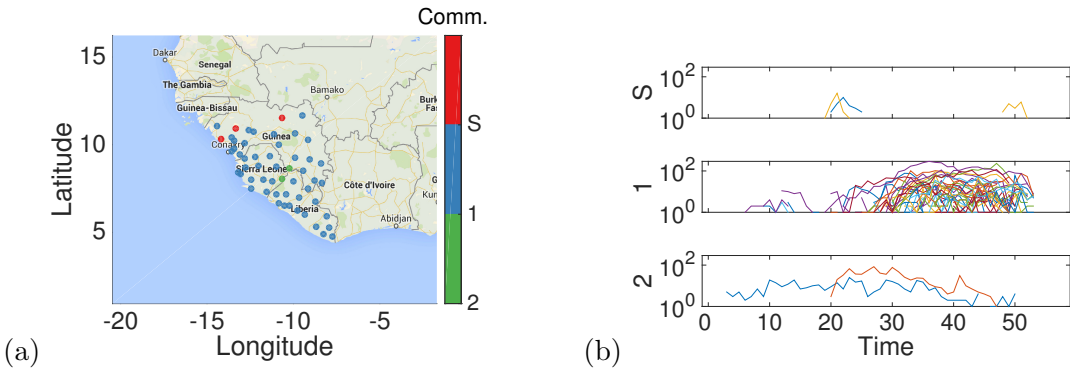


Figure 8.8: Ebola WHO data set, static network, correlation null model — example partitions. Example partition ($\gamma = 0.6$). (a) Map of all the nodes colored by algorithmically detected communities, (b) the time series of disease occurrence for nodes in these communities.

8.3.3 Modularity maximization using the gravity null model

When maximizing modularity using the gravity null model, we see that the majority of provinces in the centre of the affected region, on the border between the three countries (some of which experienced early epidemic onset) are placed in singleton communities [red in the example partition for $\gamma = 1.7$ in Fig. 8.9(a)]. The nodes placed in singletons have larger populations than those in the larger community, similarly to what we found when using this null model in other data sets. There is also one large community that contains the majority of nodes on the outskirts of the epidemic region and a later epidemic onset, but a large number of cases overall [blue in Fig. 8.9(a)] and a small community with a small number of cases in the north-east of Guinea. The general pattern — one large community and many singletons — is similar for all γ values.

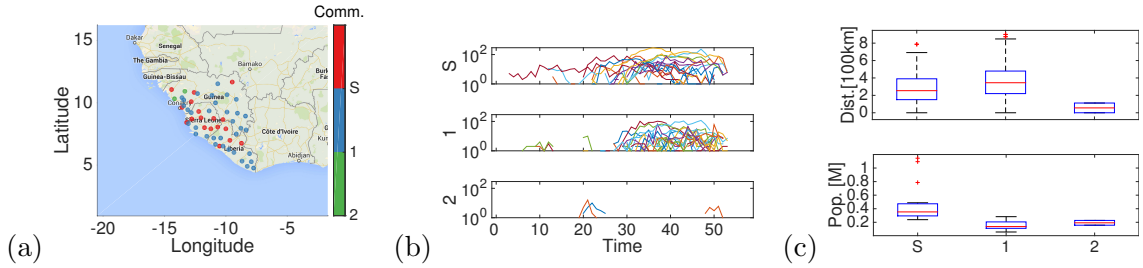


Figure 8.9: Ebola WHO data set, static network, gravity null model — example partitions. Example partition ($\gamma = 1.7$). (a) Map of all the nodes colored by algorithmically detected communities, (b) the time series of disease occurrence for nodes in these communities and (c) community properties: (top) the distances between all pairs of nodes within communities and (bottom) populations of nodes within communities.

8.3.4 Modularity maximization using the radiation null model

When we maximize modularity using the radiation null model, one large community dominates the community structure, and a small number of nodes are again placed into singleton communities (see Fig. 8.10). The singletons once again have larger populations than the nodes in the large community, but there are fewer of them than for the gravity null model, and they are located far from each other. The general pattern — one large community and some singletons — is similar for all γ values.

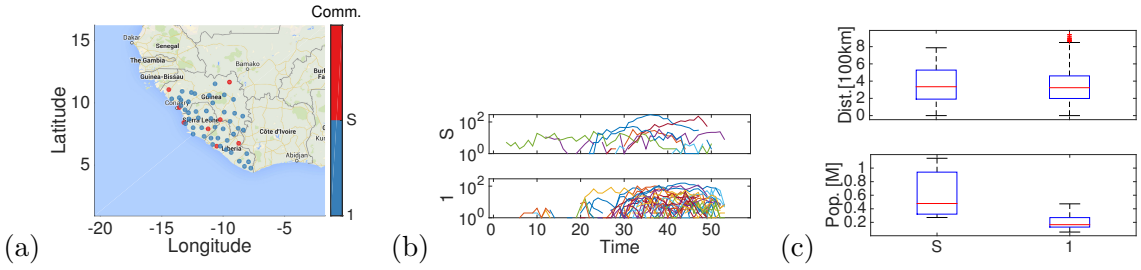


Figure 8.10: Ebola WHO data set, static network, radiation null model — example partitions. Example partition ($\gamma = 1.7$). (a) Map of all the nodes colored by algorithmically detected communities, (b) the time series of disease occurrence for nodes in these communities and (c) community properties: (top) the distances between all pairs of nodes within communities and (bottom) populations of nodes within communities.

The shortness of this data set meant that we chose not to conduct a multislice analysis on this data set due to insufficient number of time points to generate meaningful correlation matrices.

8.3.5 Influence of first infection times on community composition

From plotting the disease time series in communities detected by modularity maximization using the NG and correlation null models on the static network constructed from the Ebola WHO data set, it appears that the algorithm might be grouping provinces that have similar disease time series. In particular, the first infection time (the first time point at which the disease is observed in a region), and the approximate time at which the epidemic wave peaks appear to often differ between communities. Both of these times are affected by transport of new cases into a province, however epidemic peak time is additionally influenced by factors specific to each location, which determine how the epidemic develops once it has reached a particular province. For this reason, we study the first infection times of provinces allocated to different communities for all the null models and γ parameter values that we tested.

We test the network partitions for the four null models with the start time test that we described in Section 3.4.6. This test detects whether the grouping of nodes into communities is statistically significantly associated with the first infection times of these nodes. We find that the p-values are statistically significant ($p < 0.05$ after Bonferroni correction for multiple comparisons) for all null models: NG for $\gamma \in \{1, 1.1, 1.2, 1.3, 1.4, 1.7\}$, correlation for

$\gamma \in \{0.8, 0.9, 1, 1.1, 1.2\} \cup \{2.2, 2.3, \ldots, 3\}$, gravity for $\gamma \in \{0.3, 0.4, \ldots, 2.3\} \cup \{2.5, 2.6, 2.7\}$ and radiation for $\gamma \in \{0.2, 0.3, \ldots, 1\} \cup \{1.5, 1.6, \ldots, 2.3\}$ (see Fig. 8.11).

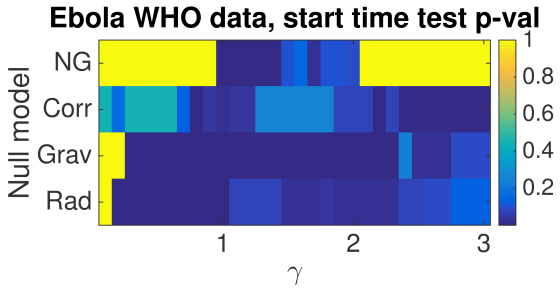


Figure 8.11: Ebola WHO data set, static network, start time test p-values for all null models and γ values. We plot p-values in the start time test for all communities found across $\gamma \in \{0.1, 0.2, \ldots, 3\}$ (horizontal axis) for NG, correlation, gravity and radiation null models.

8.3.6 Summary of the Ebola WHO data set

Despite the short length of the time series, the Ebola WHO data set shows promising results suggesting the presence of strong spatial organization. The communities that we detect using the NG null model exhibit a strong spatial organization and different times of epidemic onset in the disease time series. The results with the gravity and radiation null models are also interesting, with the method again placing highest-populated nodes into singleton communities, which this time are associated with the first infection time as well. The results for the correlation null model appear partially promising, with significant p-values in the start time test and the network partitions grouping the nodes on the border of the three countries where the epidemic initially developed, but the lack of variation across γ parameter values prevented us from exploring these results further. Perhaps using a modification to the correlation null model, as tested on the synthetic time series in Section 9.3.4.2, or using an iterative multiresolution community detection as suggested by the authors in Ref. [171] could allow us to explore these partitions further.

8.4 H1N1 influenza in Mexico

8.4.1 Introduction

The H1N1 data set contains daily new case count data from the 32 provinces of Mexico over 430 days between April 2009 and June 2010. The time series contains the three waves of the initial swine flu epidemic in Mexico, which showed a degree of spatial organization between waves: the first wave mostly affected central regions, the second wave was strongest in the southeast, and the third and fourth waves were geographically widespread [53]. We described the data set in detail in Section 5.2.2.2.

The H1N1 data set contains a large number of time points compared to the number of provinces in Mexico, allowing us to generate a long multislice network and to study the

temporal spread patterns in detail. We use a time-window width $\Delta = 30$ and a difference between the time-window starting points $\upsilon = 7$ to generate a set of 58 static networks. We also use $\Delta = 30$ and $\upsilon = 30$ to generate a multislice network with 14 layers. We described the justification for our parameter choices and the general properties of the time series and the static networks in Section A.2.

8.4.2 Summary of results

Modularity maximization on the H1N1 data set using the NG and correlation null models does not appear to find spatial partitions reliably on either static or multislice networks. The method does not detect the known regional differences in the strength of the three waves of the 2009 epidemic [53]. The gravity and radiation null models once again group all nodes except the highest-populated ones into one large community; these partitions do not appear to give us additional spatial or temporal information.

For the multislice networks, the NG and correlation null models detect temporal partitions [see Fig. 8.12 for example partitions and Figs. C.7 and C.10 for temporal z-Rand scores across a range of γ and ω parameter values]. The z-Rand scores are higher for the correlation null model. All three partitions shown here have critical time points at layers 6 and 9, that roughly correspond to the peak of the main (third) H1N1 epidemic wave and the time when the epidemic wave subsided at the end of the epidemic. This is in line with the findings for the Ebola Datamarket data set where we detected the peak and the beginning of the decline of the epidemic. However, the temporal structure detected in this data set is country-wide rather than regional as for Ebola. The method does not find the earlier dates that mark the first and second wave of the 2009 epidemic.

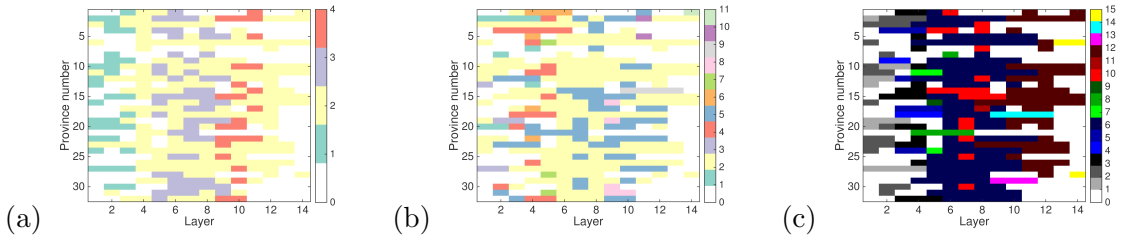


Figure 8.12: H1N1 influenza, multislice networks, NG and correlation null models — spatial and temporal organization of partitions. We plot community structure with nodes ordered by their location (north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for (a) NG null model with for $\gamma = 1$, $\omega = 0.1$, (b) correlation null model with $\gamma = 0.8$, $\omega = 0.1$ and (c) correlation null model with $\gamma = 2.2$, $\omega = 0.3$. Community number is indicated on the colorbar.

8.5 Conclusions

In this chapter, we used our previously described methodology (applying community detection by modularity maximization with different null models to disease-correlation networks

constructed from time series) on data sets relating to the introduction of a new disease into an environment. We developed this methodology on endemic diseases such as dengue fever and seasonal influenza, where we found that spatial communities were often associated with epidemic periods. Further, the spatial organization of partitions was the strongest in networks related to vector-borne dengue, and to rubella — both diseases periodically died out locally during the data sets that we study. In contrast, we did not find strong spatial organization in the seasonal influenza data set, for which the disease is present throughout the data set. For this reason, we were hoping to find strong spatial communities for the emerging diseases, for which the data sets represent the spatial dispersion of one large epidemic each, and the numbers of new disease cases are likely to be more influenced by disease spread along transport links and less influenced by internal population dynamics compared to endemic diseases.

The results of applying our methodology to emerging disease data were mixed (they are summarized in Table 8.1). When using the NG and correlation null models on both Ebola data sets, we observed relatively strong spatial communities. However, we were unable to find interesting spatial structures for the H1N1 data set. The communities that score as "spatial" for both Ebola data sets appear to be composed of provinces that are related to country boundaries. Furthermore, spatial communities that we find in the WHO data set appear to correspond to groups of nodes with different times of first recorded disease case in our disease time series. It is possible that the information about first infection times influences community structure. Thus, network partitions that we detect could be related to transport links. We will begin to explore these ideas using a spatial agent-based model of disease spread in Chapter 9.

For multislice networks, we detect spatial communities using NG and correlation null models for the Ebola Datamarket data set, and we do not find spatial partitions for the H1N1 data. For the Ebola Datamarket data set, we detect a change in community structure that might correspond to the epidemic peak, and a large temporal change corresponding to the change in relative case numbers between Sierra Leone and the other two countries at the end of the data set. We also might be able to detect some of the most important time points in the H1N1 data set (the peak and the end of the epidemic) using both NG and correlation null models.

In Chapter 9, we will explore the conditions that lead to the formation of spatial communities further using a spatially embedded disease model. We will explore the model conditions and parameter regimes for which our methodology of applying community detection by modularity maximization with different null models to disease-correlation networks constructed from disease time series is able to detect spatial communities.

Table 8.1: Overview of the results of community detection for all emerging disease data sets and all null models. "Ebola D" data set refers to the Ebola Datamarket data set.

Data set	Null model	Results
Ebola D	NG	• Spatial partitions related to country boundaries (Sierra Leone) in the later part of the data • Temporal partitions related to the formation of the Sierra Leone community
	Correlation	• Similar as for NG • Little variation with γ
	Grav. & Rad.	• One large community, many singletons with high populations
Ebola WHO	NG	• Spatial partitions related to country boundaries (Guinea and Liberia) • Different first infection times between communities
	Correlation	• Separates central nodes on the border between the three countries • Little variation with γ
	Grav. & Rad.	• One large community, many singletons with high populations • Community assignment related to first infection time
H1N1	NG	• Little reliable spatial organization • Temporal partition related to the peak and end of the epidemic
	Correlation	• Similar as for NG • Little variation with γ
	Grav. & Rad.	• Little reliable spatial organization • Temporal partition related to the peak and end of the epidemic

Chapter 9

Application to time series from an agent-based model

This chapter represents original work. It is the result of a collaboration with A. Elliott (CABDyN, Saïd Business School, University of Oxford) and M. A. Porter that is not published yet. The model development was primarily done by MS and it was implemented in Python by AE and MS. Both parties contributed to the design of the experiments and the analysis of the results.

9.1 Introduction

In this chapter we investigate the ability of the community-detection methodology that we presented in Section 3.5 to detect spatial communities and planted communities in disease-correlation networks that we construct from time series from an agent-based model (ABM) for disease propagation. Our model represents 50 interconnected cities located on a ring, with a known transportation structure that depends on distance along the ring between cities and their community membership. The disease progression in each city assumes full mixing, and the model reduces to a discrete-time SIS model in the limit of large population. We impose a planted community structure by changing an intra-community transport multiplier that governs the proportion of trips that individuals take within and between communities, and we investigate situations with various strengths of planted communities.

We show the results of applying community detection to disease-correlation networks generated from this model using the Newman-Girvan (NG), gravity and correlation null models that we defined in Section 3.3 and used on real disease-correlation networks in Chapters 6, 7 and 8. We do not use the radiation null model, as it is equivalent to the gravity null model for our agent-based model, as we show in Section 9.2.2. We study the algorithmic partitions and we investigate whether our community-detection methodology is able to detect spatial communities and planted communities in this idealized situation with known interaction patterns.

9.2 Model motivation and general definition

Our main consideration for designing the disease model was to choose an approach that is able to reproduce the qualitative behaviors that we see in real disease data sets, which would form the baseline for our investigations of parameter variation. We base our choice of disease model on two observations, which we want to be able to reproduce and test:

1. In Chapter 8, we observed that the spatial organization of communities appears to be strong for emerging diseases, i.e., when a disease initially spreads through a country.

2. In Chapters 6, 7 and Appendix B.3, we observed that the spatial organization of communities for endemic diseases is stronger for dengue fever and rubella, which periodically die out and are reintroduced into provinces (and even an entire country) in occasional epidemics, than for seasonal influenza, which is always present in the country.

Both of these outcomes can in principle be modeled by a single disease model (depending on parameter choices), with (1) representing the initial seeding of the model and (2) representing the situation once the disease has reached all the cities. To simulate the disease emergence described in (1), we seed the disease from one location and study its spread through space. To simulate the endemic disease situation in (2), we want to choose a disease-model design that is capable of reproducing this behavior of dying out and reintroduction for certain parameter regimes, but is capable of simulating a disease that is always present for other parameter regimes, so that we can compare the results of community detection for both cases.

As a simple modelling step, we have attempted to use a differential equation-based metapopulation model with periodic forcing to fit the time series of the occurrence of dengue in Peru, but the outcome is not included in this thesis [236]. We tried both a model containing all provinces and a simplified metapopulation model with three patches corresponding to the three climate zones. Such differential equation models are capable of modelling seasonal outbreaks [195]. However, we struggled to match the model to data due to the very high variation in the data, even after grouping provinces by climate. One disadvantage of differential equation models for modelling the variable long-term disease prevalence of dengue is their deterministic nature, which makes them unable to represent changing sizes and lengths of disease outbreaks. It is possible that a stochastic model would be more successful in matching the qualitative features of the recurring epidemics of dengue.

Further, a large stochastic model with many provinces might be able to match the behavior of the disease dying out and being reintroduced from other regions that we described in (2) above, as stochastic models are more suitable for situations with low numbers of

infected individuals than deterministic ones [40]. This outcome could be achieved by selecting a parameter regime in which the infection rate and the recovery rate for the disease are close to each other, so that stochastic effects are strong for each individual city, and the disease sometimes dies out in individual cities. However, we want to choose the disease and transport-related parameters so that the disease tends to persist in the whole network through the ability of individuals to travel between the model cities.

This kind of model can be implemented through various approaches, such as stochastic differential equations, stochastic difference equations, Markov chains or an agent-based model (ABM) with stochastic behavior [40, 69] as discussed in Chapter 2. Differential-equation and difference-equation models can sometimes be studied analytically, and tend to be less computationally costly than ABMs. However, as we are interested mostly in using the synthetic time series generated by the model for community detection, and we can easiest generate the desired behavior using an ABM, we use this approach to explore the parameter space and get an understanding of whether our method is able to detect spatial communities in synthetic time series.

To simplify the model and reduce computation time, we assume that the population inside each city is perfectly mixed. This is different to the approach that many ABMs take, which often strive for the most realistic implementation possible or focus on the extra heterogeneity generated by the need for close person-to-person contact for disease transmission [49,244]. However, this kind of "hybrid" modelling is a growing field, and it is especially popular with large-scale models as a way of reducing computational complexity while preserving some of the individual-level aspects of the modelling question [34,169,276].

To simplify the spatial aspect of the model, we place the 50 cities on a ring. Each step along the ring has distance 1. The disease model gives each agent a defined probability of changing disease state, and the mobility mechanism gives individuals a defined probability of moving between cities. Individuals can move to any city in one time step, but the likelihood of moving to a city is inversely proportional to the distance from the source city to the target along the ring. It is also influenced by community structure, with the probability of selecting a city as a target for travel increased for cities in the same community as the home city by an intra-community transport multiplier. In the following sections, we formally define the model and present the results of these experiments. We present the basic information about the model in Fig. 9.1.

We implement the model in Python using object-oriented programming to represent the agents and cities. We generate correlation networks from the time series generated by this model and we apply community detection by modularity maximization with different null models. We examine how the results change for different values of disease model parameters (infection rate, recovery rate) and the transport mechanism parameters (transport rate, inter-community mixing).

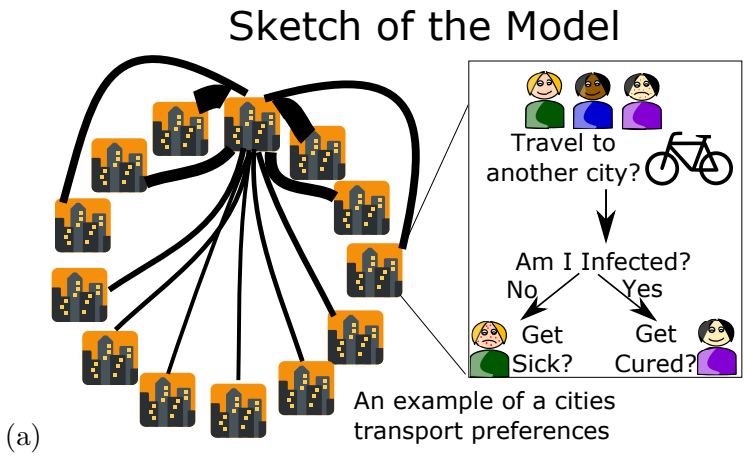


Figure 9.1: The disease model. We place 50 cities on a ring (we show 13 here for simplicity). The transport links between the cities are inversely proportional to distance (signified by edge width; transport links are shown for one city for simplicity). Inset zooms in on the progression of the model in one city at time point t. The transport mechanism is considered first for each individual, after which each individual's health status is separately considered and the disease progress is recorded. Graphics licensed under CC-BY 4.0

9.2.1 Model definition

We place 50 cities on a ring in space. Each city is assigned a population of 100 residents, who can travel between cities but always return to their home city. Each city is connected with transport links to its nearest neighbors along the ring. Individuals can travel to all cities within one time step, but the probability of choosing a particular target city as a travel destination is inversely proportional to the distance along the ring between home city and the target city. We also implement a planted community structure by increasing the probability of choosing cities from the same community by an "intra-community transport multiplier" (we study multiple values of this parameter). We assign cities to one of two communities uniformly at random, with equal probability of assignment to either community.

We model the spread of disease on this set of cities. The model runs in discrete time and for simplicity we will think of the time step as 1 day. At each time step, the model performs two main steps in sequence:

1. The transport mechanism runs: (a) people who were away and are due to come back return to their cities, (b) individuals who are in their home cities are given an opportunity to move with transportation probability ϕ.

2. The SIS infection dynamics occur in the cities.

We record the results of 1000 time steps (about 3 years).

Both the transport mechanism and the SIS infection dynamics contain random aspects which can lead to variation in the overall results of any given simulation. We run the disease model 50 times for each parameter regime in order to achieve representative results for the parameter values.

In the following sections, we specify the details of the model.

9.2.1.1 The disease dynamics

In this section, we define the disease dynamics that we use to simulate the spread of disease within cities. The disease model in each city is a type of an SIS process (Susceptible – Infected – Susceptible, as described in Section 2.2.1); it assumes full mixing but, in contrast to the compartmental models defined there, infections happen on the levels of individual agents. We copy the original differential equation SIS model below for convenience. The SIS model is

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI + \alpha I \\ \frac{dI}{dt} &= \beta SI - \alpha I \,, \end{aligned} \qquad \text{(2.2 revisited)}$$

where

- $S(t)$ is the number of susceptible individuals (who are not infected) at time t;
- $I(t)$ is the number of infected individuals at time t who can transmit the disease by contact with susceptibles.

The SIS model assumes that individuals immediately become susceptible again after recovering. This is a much simpler situation than for the real diseases we study, as they confer immunity to those who recover from them, as discussed in Chapter 5. However, influenza and dengue confer immunity to the specific strain that caused the infection but not to other strains. As our disease data sets do not record the strain involved in infections or the identity of individuals, some individuals might appear in our data sets multiple times due to being infected with different strains over the course of the data set. Further, choosing an SIS model allows us to model the repetitive reintroduction of diseases into cities without the need to model very large numbers of people (which would be computationally prohibitive) or introduce births to ensure a supply of susceptible individuals. An SIRS model (including temporary immunity in the R state) would also be appropriate, although we decided on SIS for simplicity.

In our agent-based model, each individual has a probability $\beta\frac{N_I}{N_T}$ of getting the disease (where N_I is the number of infected individuals and N_T is the total population in each city). Once infected, each individual has a probability α of recovering [we follow the notation used in Eq.(2.2)]. The probability that a susceptible agent becomes infected is

$$P(S \to I) = \beta\frac{N_I}{N_T} \,. \tag{9.1}$$

The probability that an infected agent recovers (and goes back to susceptible state) is

$$P(I \to S) = \alpha \,. \tag{9.2}$$

Therefore, the expected difference in the numbers of susceptible individuals from time t to time $t+1$ is

$$\delta S = N_I P(I \to S) - P(S \to I) N_S = N_I \alpha - \beta \frac{N_S N_I}{N_T}\,, \tag{9.3}$$

where N_S is the number of susceptible individuals. The expected difference in the numbers of infected individuals from time t to time $t+1$ is

$$\delta I = P(S \to I) N_S - N_I P(I \to S) = \beta \frac{N_S N_I}{N_T} - N_I \alpha\,. \tag{9.4}$$

This is a discrete-time version of an SIS model. If we let time go to 0, we recover a process that is related to the SIS process. This model is a type-II model as defined by Colizza et al. in Ref. [60]. However, in practice, the discrete and stochastic nature of the disease model allows us to recover the "spiky" behavior of disease time series for certain parameter values.

For our numerical experiments, we consider $\alpha, \beta \in \{0.1, 0.2, \ldots, 1\}$.

9.2.1.2 The transport mechanism

In this section, we define the transport mechanism that we use to simulate the movement of individuals between cities. To keep the model in line with the countrywide scale of our disease data sets, we want to focus on the effects of long-distance inter-city travel (which we assume to include overnight stay) rather than of daily commuting. Based on reports of mean lengths of stay of visitors in Refs. [104, 210, 216], we select 5 days as the length of a trip to another city in order to simplify the model. (One can also select the trip duration separately for each trip using a suitable distribution. Heterogenous lengths of stay can impact the likelihood of an epidemic reaching all nodes models [216] and they have been shown to influence the spread of real-world diseases over both short and long distances [174, 281].)

We define a parameter ϕ that describes the transportation probability; probability that a person travels from their home city to another city. We use $\phi \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. At each time step, the transport mechanism runs two steps in sequence:

1. It returns all individuals who have been away for exactly 5 days to their home cities.

2. It queries all individuals who are currently located in their home cities (including those who just returned) as to whether they should travel to another city. The probability of an individual being selected for travel is ϕ.

For those individuals that are selected to travel to another city, we select a destination city according to a weighted transport matrix. The probability that an individual travels

from city i to city j at distance d_{ij} along the ring is inversely proportional to the distance between the two cities, and it is influenced by the community structure:

$$p_{ij}^{\text{trav}} \propto \frac{\eta(c_i, c_j)}{Z_t d_{ij}}\,, \tag{9.5}$$

where c_i is the community that contains city i and the function $\eta(c_i, c_j) = \eta$ if nodes c_i and c_j are in the same community and $\eta(c_i, c_j) = 1$ otherwise. The "intra-community transport multiplier" η controls the amount of mixing between communities. When $\eta = 1$, there are no distinct communities. The normalization constant Z_t ensures that the total probability is equal to the transportation probability ϕ ($\sum_{i \neq j} p_{ij}^{\text{trav}} = \phi$).

Therefore the overall probability of traveling from city i to city j in a given time step, taking into account the transportation probability ϕ, is

$$p_{ij}^{\text{trav}} = \frac{\phi\eta(c_i, c_j)}{Z_t d_{ij}}\,. \tag{9.6}$$

9.2.1.3 Updating the system

Synchronous and asynchronous updates are the two main methods to update an ABM in time. For a model with synchronous updates, new infections are recorded all at once for all agents and cities, so the movements and new infections that occur in a given time step are independent of each other. In a model with asynchronous updates, one node is updated at a time, and changes in a given time step can immediately affect the spread of disease and transport for other agents and cities. Asynchronous models are often closer to reality, but synchronous models allow faster simulations. In practice, models in the literature use both designs, depending on the question at hand. In fact, most popular ABM implementations use synchronous updates by default [58]. We implement synchronous updates for simplicity and speed.

9.2.1.4 Seeding infection

We considered two ways to seed infection that could reproduce both emerging and endemic diseases:

1. Seed a single source of infection at time $t = 0$;

2. Implement a random infection mechanism, in which the initial seeding of infection is replaced by a parameter that determines the probability that the disease enters a population from outside (e.g., transported from abroad, or from an animal for a vector-borne disease such as dengue or a zoonotic infection such as Ebola).

We chose the simplest option (1) for our model (following the majority of metapopulation modeling literature [237]), after testing whether "spiky" data, where the disease dies out and is reintroduced to a city, can be generated from our model without introducing random

infection. On a large enough ring it is possible to observe this kind of reinfection behavior without random reinfection when α and β have similar values (so that stochastic effects are strong) and there is enough transport between the cities to cause reinfection. However, the set of disease parameters for which this kind of behavior occurs is larger for a model with a random infection rate than for a model with a single source of infection.

We use 20 infected individuals to initially seed the infection (after testing with values of 5, 10 and 20 and observing no qualitative difference in the results).

9.2.1.5 Endemic versus emerging diseases

We want to model two distinct types of diseases: (1) emerging diseases such as Ebola and H1N1 influenza in Mexico, and (2) endemic diseases such as seasonal influenza and dengue fever and rubella in Peru. The simulation of emerging diseases is the typical purpose of SIS-type epidemic models, and it can be easily simulated by seeding the disease in a single location on the ring. We simulate endemic diseases by introducing a "burn-in period" during which we do not record the disease time series. During this time the disease either dies out or spreads along the ring and becomes endemic in all (or some) of the cities. Essentially, we want the system to "forget" the initial conditions and go to a state in which the dynamics is simply dictated by the parameter choices and not by the starting conditions.

We use a burn-in period length of 9000 time steps, after which we begin to record the 1000 data points. [We examined the disease time series for several parameter regimes the infection spread relatively slowly (where $|\beta - \alpha| \lessapprox 0.1$) over all transport probabilities $\phi \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$, and we observed that the disease either reaches all cities or dies out within 1000 time steps in all the instances that we tested (and we chose 9000 time steps as the "burn-in period" to ensure this for all other parameter regimes).]

9.2.2 Constructing networks and community detection

After we generate the disease time series, we construct one static disease-correlation network from the results of each simulation. We construct the network in the manner that we described in Section 3.1.1.

We then perform community detection using modularity maximization on these networks using the NG, gravity, and correlation null models that we defined in Section 3.3. We will now show that the radiation null model that we introduced in Section 3.3.5 is equivalent to the gravity null model for this ring — which is why we only test the gravity null model.

9.2.2.1 The equivalence of gravity and radiation null models on the ring

For the gravity and radiation null models, we need to define a way of measuring distance in this system. One natural way to think of distance is as distance around the ring: for cities number i and j on a ring of N cities, the distance is $\min(|i - j|, |N - j + i|)$. That is, the

distance from city 1 to city 2 and the distance from city 50 to city 1 are both equal to 1, and so on. One can thus construe the set of cities as lying on a circular road, where it is assumed that all people travel via the road. We define both the radiation model and the gravity model in terms of this "commuting distance", i.e., network distance, rather than physical distance.

For calculating the gravity and radiation null models, we bin the distances into integer-sized bins (of uniform width $b = 1$). The bin with the lowest number of items contains 25 pairs (corresponding to cities that lie opposite each other on the ring), which is more than the minimum of 5 that we used for the disease-correlation networks (see Section A.4).

We now demonstrate that the gravity and radiation null models are equivalent to a mean over the weights of all pairs of nodes at the same binned distance. We first consider the gravity null model with city populations used as node importance,

$$P_{ij}^{\text{grav}} = n_i n_j \frac{\sum_{\{k,l|d_{kl}=d_{ij}\}} W_{kl}}{\sum_{\{k,l|d_{kl}=d_{ij}\}} (n_k n_l)}\,, \qquad \text{(3.21 revisited)}$$

where W_{kl} is the edge weight between nodes k and l, d_{kl} is the distance between them, and n_k is the population of city k. The populations of each of the cities are the same, so we can replace them all with n_1 and simplify to

$$P_{ij}^{\text{grav}} = \frac{\sum_{\{k,l|d_{kl}=d_{ij}\}} W_{kl}}{\sum_{\{k,l|d_{kl}=d_{ij}\}} (1)}\,, \qquad (9.7)$$

which is the mean weight of an edge between nodes that are the same distance apart.

The radiation null model has the form

$$P_{ij}^{\text{rad}} = \hat{T}_{ij} \frac{\sum_{\{k,l|d_{kl}=d_{ij}\}} W_{kl}}{\sum_{\{k,l|d_{kl}=d_{ij}\}} \hat{T}_{kl}}\,, \qquad \text{(3.23 revisited)}$$

where $\hat{T}_{ij} = (T_{ij} + T_{ji})/2$ is the mean flux between i and j predicted by the radiation null model. Therefore, if we can show that fluxes are the same for all cities at the same distance, i.e., $\hat{T}_{ij} = \hat{T}_{ab}$ if $d_{ij} = d_{ab}$, then the $\hat{T}_{ij}$ terms will cancel, leaving the same expression as the gravity null model. To show this, we consider the flux between cities i and j, T_{ij}, assuming the uniform distribution of travelers in cities:

$$T_{ij} = \frac{n_i N_c}{N} \frac{n_i n_j}{(n_i + r_{ij})(n_i + n_j + r_{ij})}\,, \qquad \text{(2.3 revisited)}$$

where N_c is the total number of travelers, N is the total population of the ring and r_{ij} is the population residing in the circle centered in i with radius d_{ij}, minus the populations at the origin i and destination j ($r_{ij} = q_{ij} - (n_i + n_j)$). The N and N_c terms cancel as we know that $\phi N = N_c$.

As the populations in all cities are equal, we can replace all of the n_i terms with n_1. Moreover, as the cities lie on a ring, the population r_{ij} in a circle between the two nodes

$(r_{ij} = n_1(2d_{ij} - 1))$ is simply a function of d_{ij} and n_1. Therefore, for a given n_1, T_{ij} is simply a function of distance as well: $T_{ij} = \tilde{T}(d_{ij})$. Further, the distances are symmetrical (i.e., $d_{ij} = d_{ji}$), so the mean flux between two cities is the same as the flux in either direction: $\hat{T}_{ij} = T_{ij}$. Therefore, for a given distance and a given population n_1, the fluxes between all cities are equal. We can thus substitute all of the fluxes with T_1 and simplify in the same manner as for the gravity null model. This results in the radiation null model reducing to a mean over edge weights at the same distance, which is the same as the expression for the gravity null model in Eq. (9.7).

9.2.3 Examining the algorithmic network partitions

We use three types of tests to examine the results of community detection on the synthetic disease-correlation networks that we produce from the ABM time series. We report a mean of 50 realizations, for one best-scoring resolution parameter γ from $\gamma \in \{0.9, 0.95, \ldots, 1.1\}$, i.e. the γ that generates the largest number of network partitions that have a statistically significant score in a given test.

9.2.3.1 Distance test

We use the distance test that we defined in Section 3.4.4 to quantify the extent of spatial organization in algorithmic partitions. That is, we compute the "total intra-community distance": the sum of the distances between each pair of nodes in the same community. Once again, we compute the p-values of this statistic using a Monte Carlo test, as we do not know of an analytic solution for the general case. We perform 7,500 runs of the Monte Carlo simulation and we use the value $p < 0.05$ after Bonferroni correction for multiple comparisons as the cutoff for assessing the significance of the spatial organization of the community structure.

However, as the geography of the ring is very simple, we can use simple analytical expressions to speed up the calculation of p-values for some (simple) network partitions. Imagine the case in which the community-detection algorithm divides the network "perfectly", i.e., into a set of communities $X = \{X_1, X_2, \ldots, X_K\}$ that are perfectly contiguous in space. As the nodes are on a ring, this will by definition have the lowest possible value of the total intra-community distance for a partition with communities of size $\{|X_1|, |X_2|, \ldots, |X_K|\}$. Therefore, to calculate a p-value in the distance test, we simply need to determine the number of ways that this community assignment is possible and then divide this by the number of possible random rearrangements of the communities that preserve the same community sizes, which is given by the multinomial coefficient

$$\binom{N}{|X_1|, |X_2|, \ldots, |X_K|}. \tag{9.8}$$

The number of possible ways to obtain an assignment of nodes on a ring into K perfectly contiguous communities can be broken into two parts. First, we must consider the number of ways in which the communities can be ordered, giving a contribution of $(K-1)!$. Secondly we need to account for the starting position of the communities (which yields N possibilities). If we assume that the communities then continue in a clockwise (or anti-clockwise) direction around the ring, this is sufficient to uniquely identify each possible network partition with K perfectly contiguous communities. Therefore, the analytical p-value for the distance test is

$$\frac{N(K-1)!}{\binom{N}{|X_1|,|X_2|,\ldots,|X_K|}}\,. \tag{9.9}$$

We can test this approach by comparing an analytical p-value to the distribution of the estimated p-values in multiple Monte Carlo simulations. In an example below, we construct a partition that consists of 5 nodes in one community and 15 nodes in the other. We compute the analytical p-value and 1000 estimated p-values from Monte Carlo simulations. We show the results in Fig. 9.2. As we can see, the Monte Carlo results agree well with the analytical p-value.

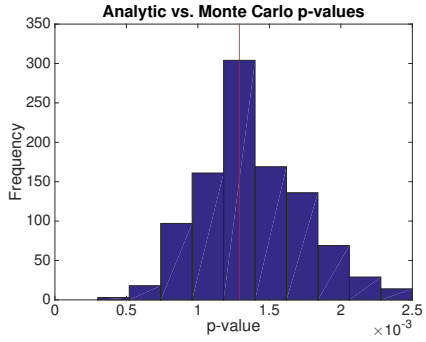


Figure 9.2: Comparison between the analytical p-value (red line) and the estimated p-values from Monte Carlo simulation (histogram) for a partition of a 20-node network with 5 nodes in one community and 15 nodes in a second community.

9.2.3.2 NMI and z-Rand scores versus planted partition

We also assess the similarity of partitions that we detect algorithmically to planted partitions. We use both NMI (see Section 3.4.5) and z-Rand scores (see Section 3.4.3). Recall that NMI is good at detecting close similarity to ground-truth partitions, but it quickly drops off as the partitions diverge from each other. Further, z-Rand scores are good at detecting similarities in coarse structure [268,269] and they are less sensitive to minor changes (such as one node changing community assignment) than NMI. Using both methods on the same set of partitions gives us a more detailed understanding of the performance of our community detection methodology in detecting the planted partitions. NMI detects perfect or near-perfect matches to the planted partitions well, but the scores drop off as partitions

diverge. By additionally using z-Rand scores, we are able to distinguish experimental conditions in which our method detects coarse-structure similarity to a planted partition from the situations in which the method does not work well. Using z-Rand scores is particularly useful for detecting planted partitions in networks generated from the ABM (compared with the spatial benchmarks in Section 4), as the results appear to be more nuanced than on the benchmarks.

We choose a cutoff of NMI > 0.8 to assess similarity between planted partitions and algorithmic partitions (after observing the similarity of partitions at various NMI values on spatial benchmarks and the ABM). We choose a cutoff of $z_R > 2$ for z-Rand scores (which is the 5% statistical significance cutoff — rounded up — assuming a normal distribution).

9.2.3.3 Start-time test

For the emerging disease simulations, we also test whether the algorithmic community assignment by our methodology is a function of the first time that the disease has been observed in each city using the start-time test defined in Section 3.4.6. We calculate the standard deviation of the first infection times (defined as the first time point that the disease is observed in a node) for each community in a partition. We then use the sum of these standard deviations for all communities in a network partition as a test statistic. Once again, we compute the p-values of this statistic using a Monte Carlo test with 7,500 runs, and we use the value $p < 0.05$ after Bonferroni correction as the cutoff for assessing the significance of the association of community structure with start times.

9.3 Results

In this section, we present the results of community detection on the disease-correlation networks that we generate from the ABM time series.

9.3.1 Distance test

First, we study partitions for emerging diseases and the NG and gravity null models. In Figs. 9.3 and 9.4, we plot the mean fraction of realizations that score significantly in the distance test for various values of the transportation probability ϕ and of the ratio β/α of the infection rate and the recovery rate. We select the parameter values where $\beta \geq \alpha$ to ensure the disease spreads in the population. We bin the results by β/α into equal-width bins of width 1 to simplify the assessment of relationships between parameters. We observe communities that score significantly ($p < 0.05$ after Bonferroni correction) in the distance test for some of the parameter regimes. [Note that as we tested values of $\beta, \alpha \in \{0.1, 0.2, \ldots, 1\}$, there are more experimental observations with parameter values β and α corresponding the smaller ratio β/α values than there are experiments with parameter values corresponding to the larger β/α ratios. Therefore, the results in Figs. 9.3– 9.5 for

larger β/α are an average of a smaller number of numerical experiments each (and may be less reliable) than the results for lower β/α].

For the NG null model and no planted partitions (intra-community transport multiplier $\eta = 1$), we observe some spatial organization for all values of β/α when the transportation probability ϕ is relatively low ($\phi \lessapprox 0.05$). For higher transportation probabilities, we are more likely to observe spatial communities for larger values of β/α. When we introduce planted partitions, we observe spatial communities less often as the intra-community transport multiplier η increases and there is less mixing between communities (see Fig. 9.3).

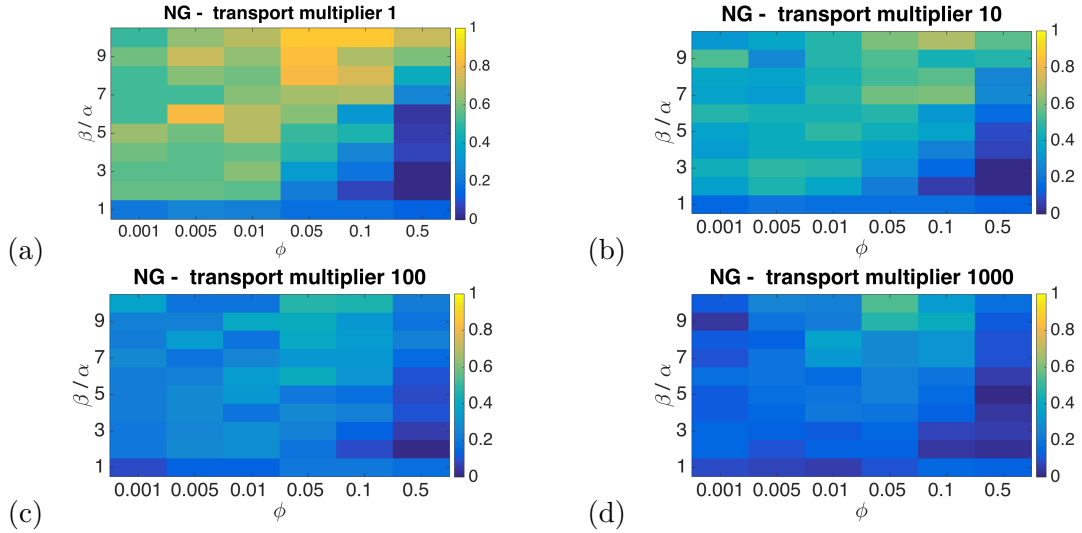


Figure 9.3: Synthetic disease model, emerging diseases, NG null model — spatial partitions according to the distance test. A plot of the mean fraction out of 50 realizations that score significantly ($p < 0.05$) in the distance test for various transportation probabilities ϕ and (binned, with bin lower bounds shown on the axis) ratios between the infection rate and recovery rate β/α for (a) intra-community transport multiplier $\eta = 1$, (b) $\eta = 10$, (c) $\eta = 100$ and (d) $\eta = 1000$.

For the gravity null model and no planted partitions (intra-community transport multiplier $\eta = 1$), we sometimes observe spatial communities for all parameter regimes, except when the infection and recovery rates are too close to each other (for the bin $1 \lessapprox \beta/\alpha \lessapprox 2$), and when the transportation probability is at its maximum tested value ($\phi = 0.5$). This is in contrast to our results on spatial benchmarks and for disease data sets, where the gravity null model was shown to remove spatial elements from community structure and did not detect significantly spatial partitions. See Fig. 9.4 for a comparison of different parameter regimes. The fraction of experimental realizations that score as significantly spatial is lower for the gravity null model than for the NG null model. However, in contrast to NG, the gravity null model allows us to detect spatial partitions for low β/α and high ϕ. Once again, increasing the intra-community transport multiplier η lowers the fraction of networks for which our method detects spatial partitions.

The correlation null model is unable to detect spatial communities (see Fig. 9.5). We omit the cases for $\eta = 10$ and $\eta = 100$, as the results were the same as for $\eta = \{1, 1000\}$.

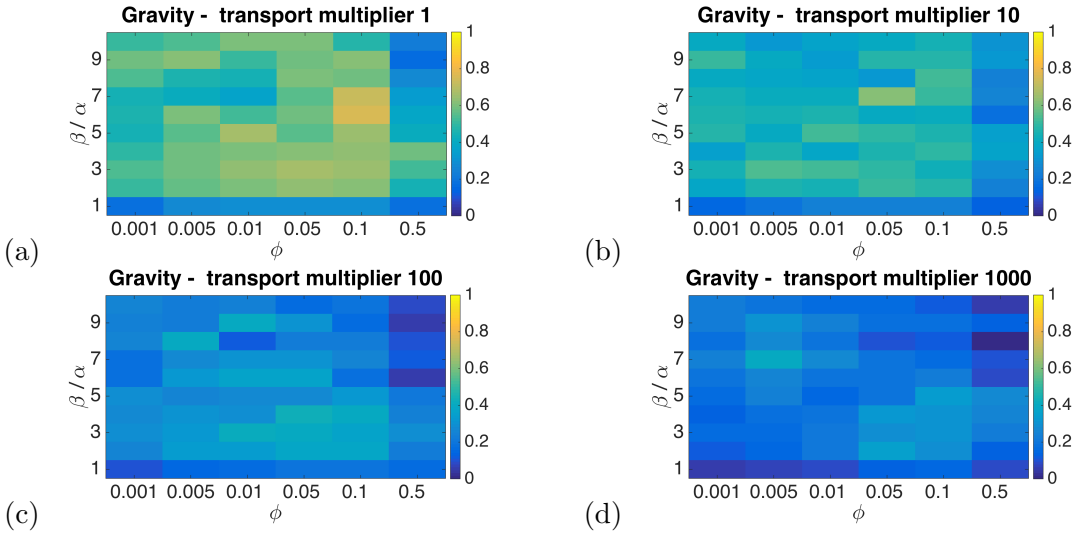


Figure 9.4: Synthetic disease model, emerging diseases, gravity null model — spatial partitions according to the distance test. A plot of the mean fraction out of 50 realizations that score significantly ($p < 0.05$) in the distance test for various transportation probabilities ϕ and (binned, with bin lower bounds shown on the axis) ratios between the infection rate and recovery rate β/α for (a) intra-community transport multiplier $\eta = 1$, (b) $\eta = 10$, (c) $\eta = 100$ and (d) $\eta = 1000$.

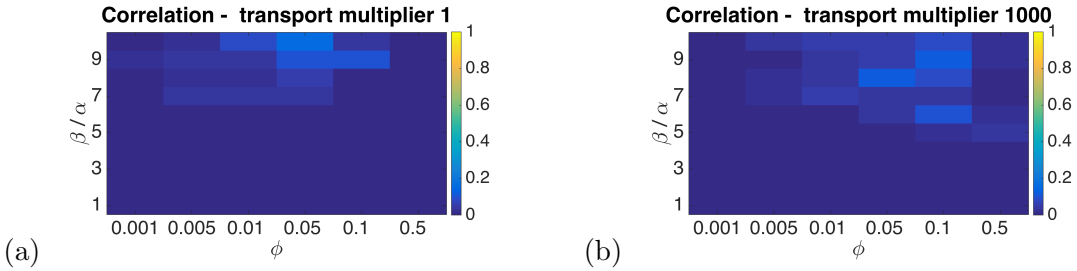


Figure 9.5: Synthetic disease model, emerging diseases, correlation null model — spatial partitions according to the distance test. A plot of the mean fraction out of 50 realizations that score significantly ($p < 0.05$) in the distance test for various transportation probabilities ϕ and (binned, with bin lower bounds shown on the axis) ratios between the infection rate and recovery rate β/α for (a) intra-community transport multiplier $\eta = 1$, (b) $\eta = 1000$.

For endemic diseases, we do not observe spatial communities for any of the ABM parameter combinations that we test, for any of the null models (see Fig. 9.6 for the results at $\eta = 1$, which are qualitatively the same as for other η values). This inability to find spatial partitions for endemic diseases is in contrast to our findings for real disease data sets, where we found communities with a degree of spatial organization in dengue and rubella data.

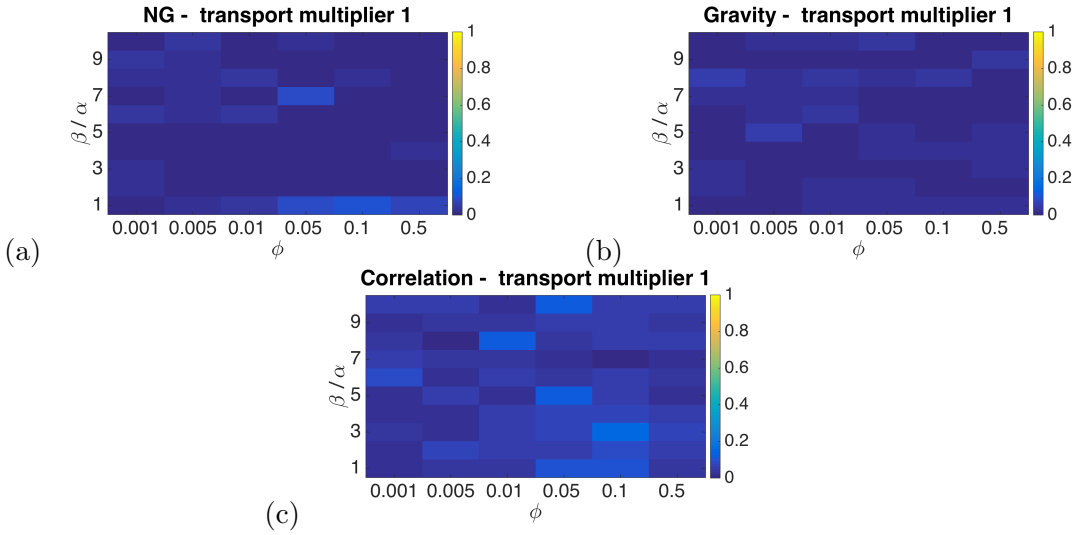


Figure 9.6: Synthetic disease model, endemic diseases — spatial partitions according to the distance test. A plot of the mean fraction out of 50 realizations that score significantly ($p < 0.05$) in the distance test for various transportation probabilities ϕ and (binned, with bin lower bounds shown on the axis) ratios between the infection rate and recovery rate β/α for (a) NG null model, with $\eta = 1$, (b) gravity null model, with $\eta = 1$, and (c) correlation null model, with $\eta = 1$.

9.3.2 Detecting planted communities

We then study the effectiveness of our method (of community detection using modularity maximization with various null models) in detecting planted communities, where the intra-community transport multiplier η describes the increase in a person's probability of travel between cities belonging to the same community over travel between cities that are in two different communities.

We examine the algorithmic network partitions using NMI and z-Rand scores; as discussed in Section 9.2.3.2, we choose a threshold of NMI > 0.8 and $z_R > 2$ to detect algorithmic partitions that are significantly similar to their respective planted partitions. We record the ABM parameter values for which our method succeeds in detecting communities in at least 25 of 50 realizations as "successes". We restrict ourselves to cases in which the disease spreads to all nodes (for simplicity we fix $\beta \geq \alpha$ to ensure this), as the NMI has issues for networks in which the disease does not spread to all nodes.

Detecting successes using NMI. Our community-detection methodology succeeds in detecting planted partitions for at least 25 out of 50 realizations of networks representing

about 8.2% of the parameter regimes that we tested. These tend to be the parameter regimes where the infection rate β is much larger than the recovery rate α, so the disease spreads quickly, and the parameter regimes in which β and α are close, so the epidemics are "explosive" — they last a short time before they die out [see Fig. 9.7(a),(b)]. The reliability of the results for $\beta/\alpha > 6$ is higher of these two detectable groups, with many parameter regimes showing over 90% reliability over the 50 realizations of the numerical experiments (not shown). Further, we are only able to detect planted communities well for emerging diseases [see Fig. 9.7(c)]. Our ability to detect the planted communities appears to sharply increase as the transportation probability ϕ increases above 0.001, but it drops off for $\phi > 0.05$ [see Fig. 9.7(d)]. We also tend to successfully detect planted partitions for networks in which the difference in the likelihood of inter-community travel versus intra-community travel is high ($\eta \in \{100, 1000\}$) [see Fig. 9.7(e)]. This is expected, as it corresponds to situations in which the influence of the planted communities is most pronounced in the transport mechanism, and thus in the disease-spreading model.

The Newman-Girvan null model has the highest fraction of successes of the three null models; the gravity and correlation null models detect only a small fraction of the most pronounced structures (i.e., for high η and ϕ) [see Fig. 9.7(f)]. This is in contrast to what we expected from our studies using benchmark networks in Chapter 4, where the spatial null models were the most effective at recovering planted partitions.

Detecting successes using z-Rand scores. When we examine the results of community detection using z-Rand scores versus the planted partition, we record success in detecting the planted partitions for about 20.6% of the numerical experiments (where we again fix $\beta \geq \alpha$ to ensure that the disease spreads around the ring). Here, we record more successes where the infection rate β is much larger than the recovery rate α [see Figs. 9.8(a)-(b)]. As for NMI, the reliability of the results across experimental realizations is the highest for large β/α. The z-Rand score comparison detects coarse-structure similarities to the planted partition for a small number of endemic disease cases, but the majority of successes come from emerging diseases [see Fig. 9.8(c)]. Furthermore, our ability to detect planted communities appears to increase as the transportation probability ϕ increases above 0.01, and it stays high at high transport rates [see Fig. 9.8(d)]. Using z-Rand scores we are able to successfully detect planted communities at higher levels of mixing between communities (intra-community transport multiplier $\eta = 10$) than if we detect similarity using NMI [compare Fig. 9.8(e) with Fig. 9.7(e)].

The number of parameter values for which we detect planted partitions successfully nearly doubles for the NG null model compared with NMI, but it grows even sharper for the gravity null model, which emerges as the most successful null model in detecting planted partitions when judged using z-Rand scores [see Fig. 9.8(f)]. This, together with the

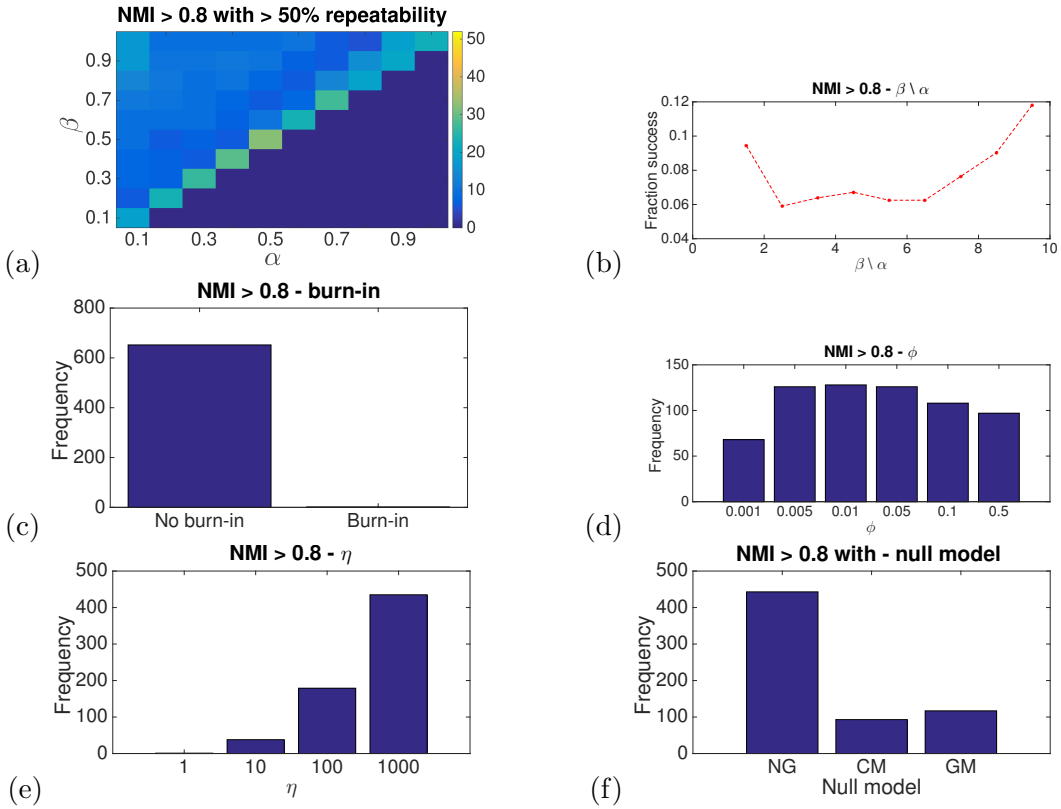


Figure 9.7: Synthetic disease model — detecting planted partitions using NMI. Plot of the counts of the number of conditions in which we detect the planted communities with an NMI score of at least 0.8 (called "successes") for at least 50% of the replications. (a) Number of successes plotted versus infection parameters: infection rate α (horizontal) and recovery rate β (vertical). (b) Binned plot of the fraction of successes versus the ratio of $\frac{\alpha}{\beta}$. (c-f) Frequency of successes plotted versus: (c) the existence of a burn-in period, (d) travel probability ϕ, (e) intra-community transport multiplier η and (f) the null model used for community detection.

NMI results, suggests that the gravity null model is able to detect structures with coarse-structure similarity to the planted partitions, but not the exact planted partitions. In fact, visual examination suggests that the gravity null model usually detects one community perfectly and splits the other community into several small communities. This is still a useful partition, suggesting that the z-Rand score measure is appropriate to judge the success of community detection for the ABM. We will use z-Rand scores in the following section to focus on identifying the parameter values in the disease model for which we are able to detect coarse-structure similarities to planted communities.

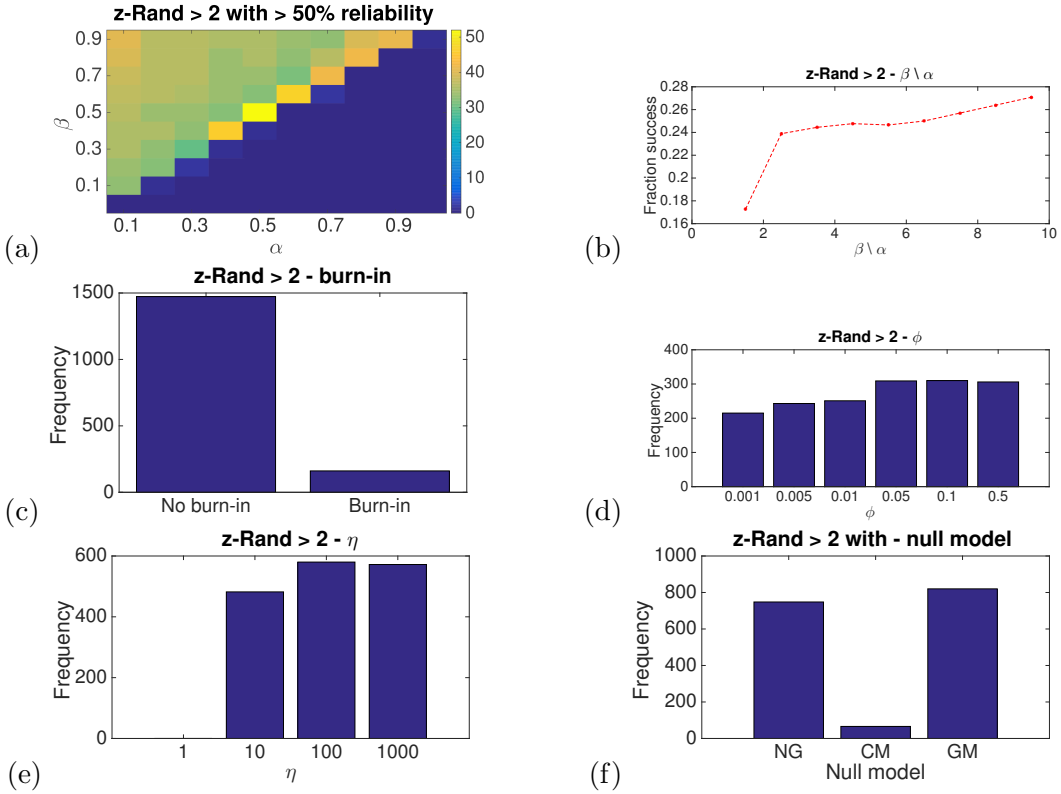


Figure 9.8: Synthetic disease model — detecting planted partitions using z-Rand scores. Plot of the counts of the number of conditions in which we detect the planted communities with an z-Rand score of at least 2 (called "successes") for at least 50% of the replications. (a) Number of successes plotted versus infection parameters: infection rate α (horizontal) and recovery rate β (vertical). (b) Binned plot of the fraction of successes versus the ratio of $\frac{\alpha}{\beta}$. (c)-(f) Frequency of successes plotted versus: (c) the existence of a burn-in period, (d) travel probability ϕ, (e) intra-community transport multiplier η and (f) the null model used for community detection.

Disease parameter regimes. When we focus on identifying the parameter regimes in which we can detect planted communities (as scored by z-Rand scores and NMI) with the NG (see Fig. 9.9) and gravity (see Fig. 9.10) null models for emerging diseases, we observe that we can detect coarse-structure similarity (as scored by $z_R > 2$) to the planted communities with medium reliability (40–80% of the time) for intra-community transport multiplier $\eta = 10$, and very reliably ($> 90\%$ of the time) for higher η values, as long as the ratio $(\beta/\alpha) > 1$. The NMI scores for these regions of the parameter space show that

both null models rarely detect structures that are closely similar to the planted partition (NMI > 0.8), with the exception of the NG null model for low ϕ and high η, where the effect of planted partitions on disease spread is particularly pronounced due to the rarity of travel that is very strongly biased towards travel within communities.

Similarly to what we found for spatial partitions, NG struggles to detect planted partitions for high ϕ and low β/α at $\eta = 10$ (even when scoring using z_R), and the gravity null model is able to detect the planted communities in this region of parameter space. Finally, we are unable to detect planted communities reliably for the endemic diseases.

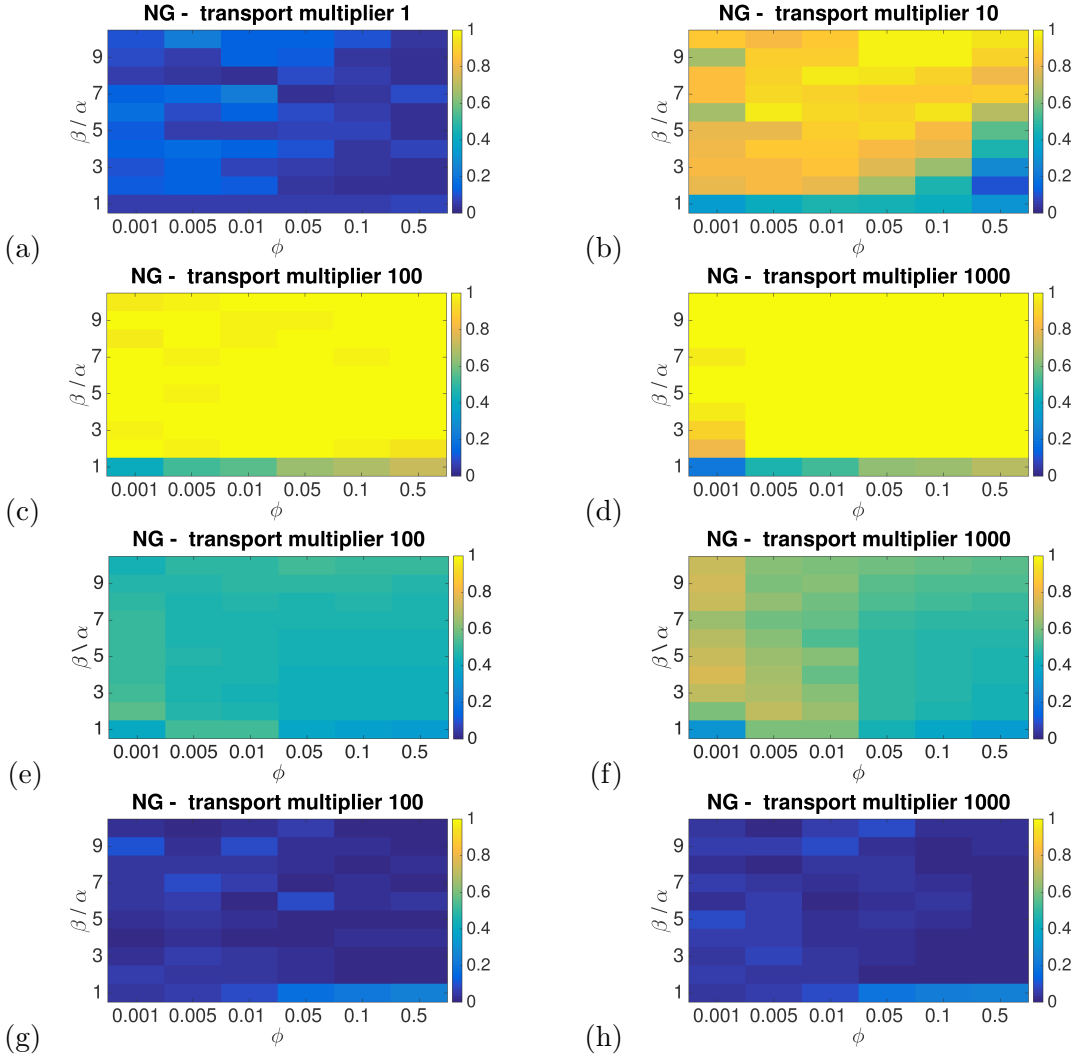


Figure 9.9: Synthetic disease model, emerging and endemic diseases, NG null model — detecting planted partitions using z-Rand scores and NMI. Mean fraction out of 50 realizations that successfully detect planted community structure for various transport probabilities (ϕ) and ratios (β/α) between the infection rate and recovery rate for (a)-(f) emerging diseases and (g)-(h) endemic diseases. Plots (a)-(d) and (g)-(h) use z-Rand scores ($z_R > 2$), and plots (e)-(f) use NMI (NMI > 0.8). The intra-community transport multiplier is (a) $\eta = 1$, (b) $\eta = 10$, (c),(e),(g) $\eta = 100$, and (d),(f),(h) $\eta = 1000$.

For emerging diseases, the correlation null model performs much worse than the NG and gravity null models (see Fig. 9.11). We are unable to detect the planted communities except for some very specific condition sets for which the difference in disease time series is

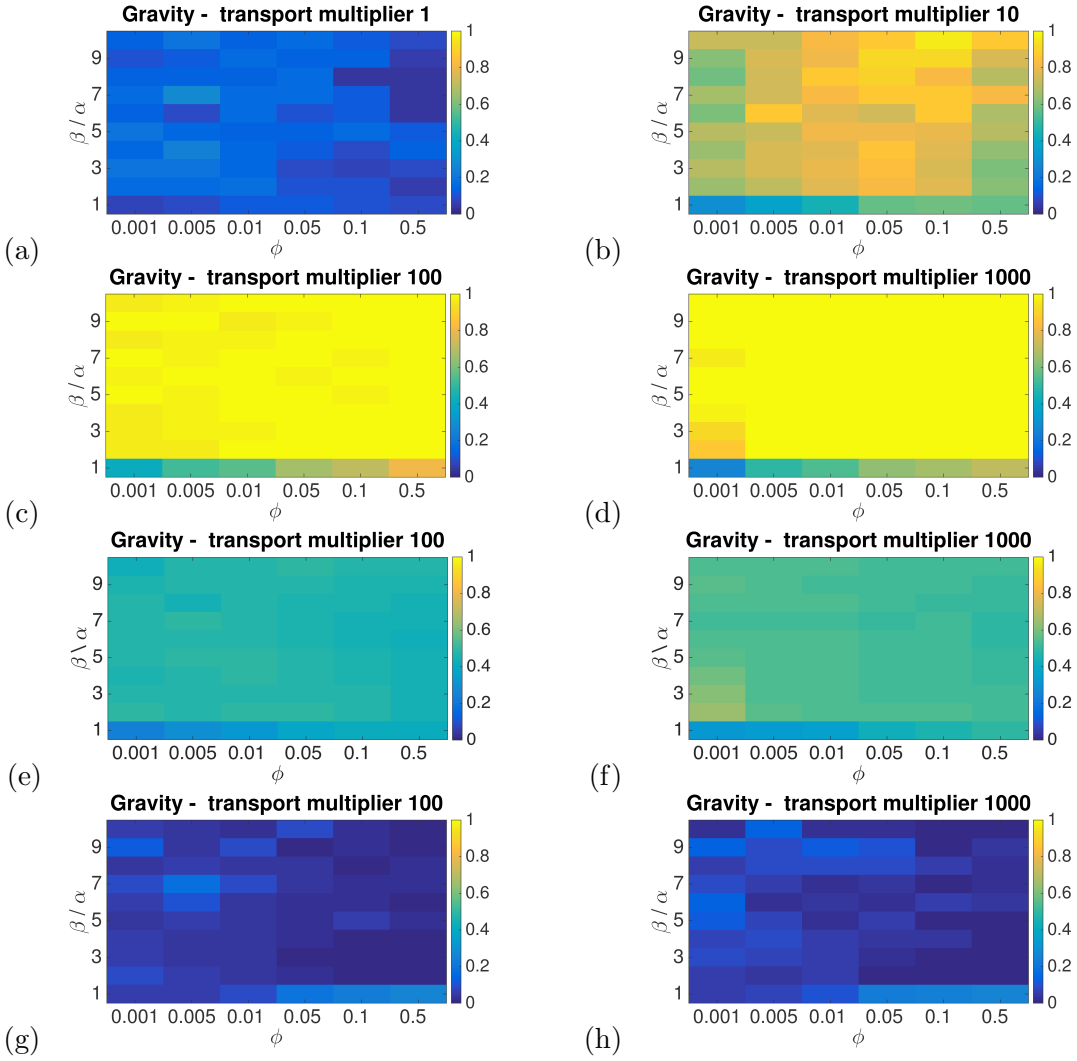


Figure 9.10: Synthetic disease model, emerging and endemic diseases, gravity null model — detecting planted partitions using z-Rand scores and NMI. Mean fraction out of 50 realizations that successfully detect planted community structure for various transport probabilities (ϕ) and ratios (β/α) between the infection rate and recovery rate for (a)-(f) emerging diseases and (g)-(h) endemic diseases. Plots (a)-(d) and (g)-(h) use z-Rand scores ($z_R > 2$), and plots (e)-(f) use NMI (NMI > 0.8). The intra-community transport multiplier is (a) $\eta = 1$, (b) $\eta = 10$, (c),(e),(g) $\eta = 100$, and (d),(f),(h) $\eta = 1000$.

most pronounced between communities: fast-spreading disease ($\beta/\alpha \gtrapprox 7$), high transport ($\phi \gtrapprox 0.1$) and low inter-community mixing ($\eta \gtrapprox 100$). For the endemic diseases the results score as successes in detecting the planted partition for about 30% of the realizations. However, as the success percentage is the same for any η value (even where there is no difference between the two communities at $\eta = 1$), the apparent success does not appear to be a real feature. It might instead be due to a numerical error in subtracting the correlation null model for cases when there are no "non-random" eigenvalues, as defined for the correlation null model in Section 3.3.6. As we will discuss in Section 9.3.4, most of the networks related to endemic diseases contain few "non-random" eigenvalues.

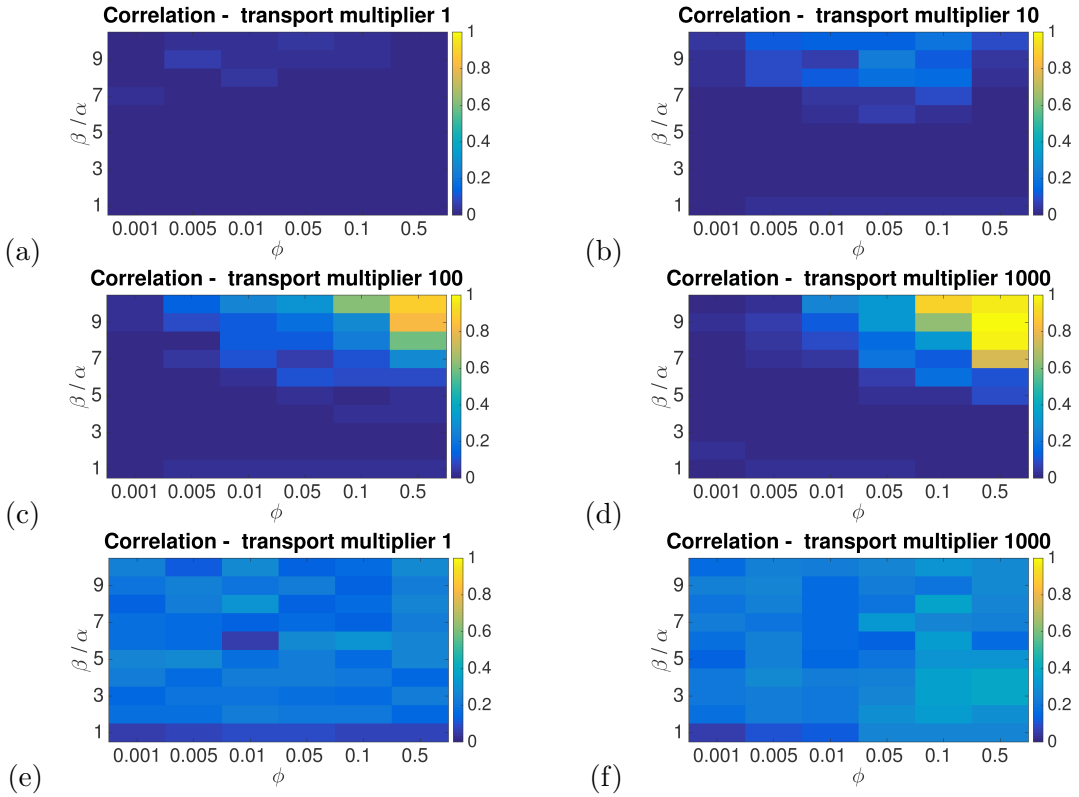


Figure 9.11: Synthetic disease model, emerging and endemic diseases, correlation null model — detecting planted partitions using z-Rand scores. A plot of the average fraction out of 50 repeats that score significantly in the start-time test for various transport probabilities (ϕ) and ratios between the infection rate and recovery rate (β/α) for (a)-(d) emerging diseases and (e)-(f) endemic diseases, for (a,e) transport multiplier $\eta = 1$, (b) $\eta = 10$, (c) $\eta = 100$ and (d),(f) $\eta = 1000$.

In summary, as seen in Fig. 9.12(a), considering all experimental conditions, i.e., parameter values and null models, we are able to detect planted communities well (NMI > 0.8) much more commonly for emerging diseases (i.e., when there is no burn-in period in the model) than for endemic diseases (i.e., ones with a burn-in period). [The results for z-Rand scores in Fig. 9.12(b) are skewed by the results from the correlation null model].

This greater success for emerging diseases may be at least in part due to the fact that as we will show in Section 9.3.1, community assignment is often based on first infection times. Further, it may be that the stochasticity of internal disease dynamics in each city is greater

than the strength of the transport term, so the internal dynamics dominates the pattern of infections inside cities once the disease has reached all of the locations (i.e., for endemic diseases).

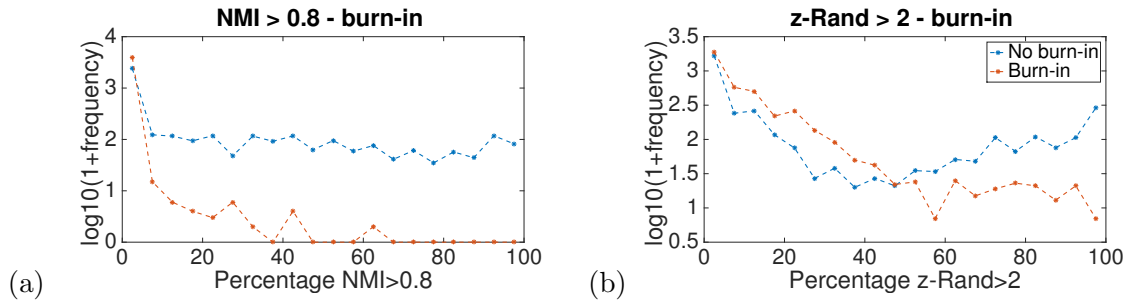


Figure 9.12: Synthetic disease model — detecting planted partitions in burn-in (emerging disease) versus no burn-in (endemic disease) parameter regimes. 1 plus base-10 logarithm of the frequency of observing various percentages of successes, defined as (a) NMI > 0.8 ad (b) $z_R > 2$ against the planted partition. The frequencies are plotted for (red) endemic and (blue) emerging diseases.

9.3.3 Start-time test

When we perform the start-time test on network partitions, we see that the partitions of many networks in the regions where we detect (even a small fraction of) spatial partitions or where we are able to detect planted partitions are statistically significant ($p < 0.05$ after Bonferroni correction) in the start-time test.

The start time is most reliably related to the network partitions for the NG and gravity null models in the emerging disease case for $\eta \in \{100, 1000\}$, where planted partitions strongly influence transportation and disease spread patterns. However, some of the community structures that we detect for $\eta \in \{1, 10\}$ also show a relationship with the first infection times — roughly in the parameter regimes where we are able to detect spatial partitions, with the exception of the region with low β/α and high ϕ. See Fig. 9.13 for the NG null model results and Fig. 9.14 for the gravity null model results. We do not show the results for the correlation null model, which showed no significant results in the start-time test.

The association of spatial partitions with first infection times is in line with our results for both Ebola data sets, which showed spatial partitions, with a degree of dependence on first infection times for the WHO data set.

We do not perform the start-time test for the simulations modeling endemic diseases, as it is not applicable for this case — we use the burn-in period to ensure the disease has spread around the ring and it is already present in all (or almost all) of the cities at the beginning of data collection.

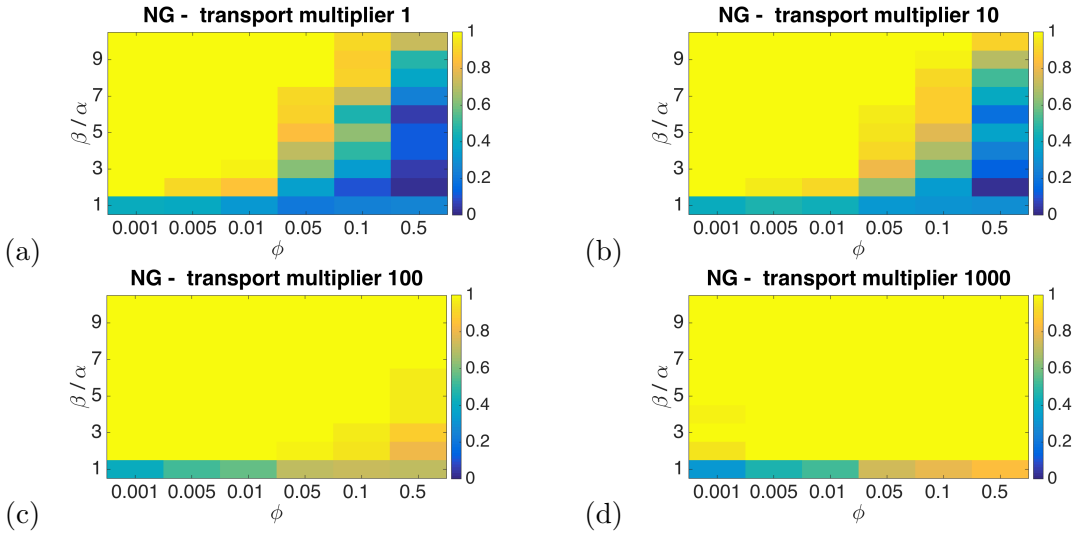


Figure 9.13: Synthetic disease model, emerging diseases, NG null model — partitions that are statistically significant ($p < 0.05$) in the start-time test. A plot of the mean fraction out of 50 realizations that score significantly ($p < 0.05$) in the start-time test for various transportation probabilities ϕ and (binned, with bin lower bounds shown on the axis) ratios between the infection rate and recovery rate β/α for (a) intra-community transport multiplier $\eta = 1$, (b) $\eta = 10$, (c) $\eta = 100$ and (d) $\eta = 1000$.

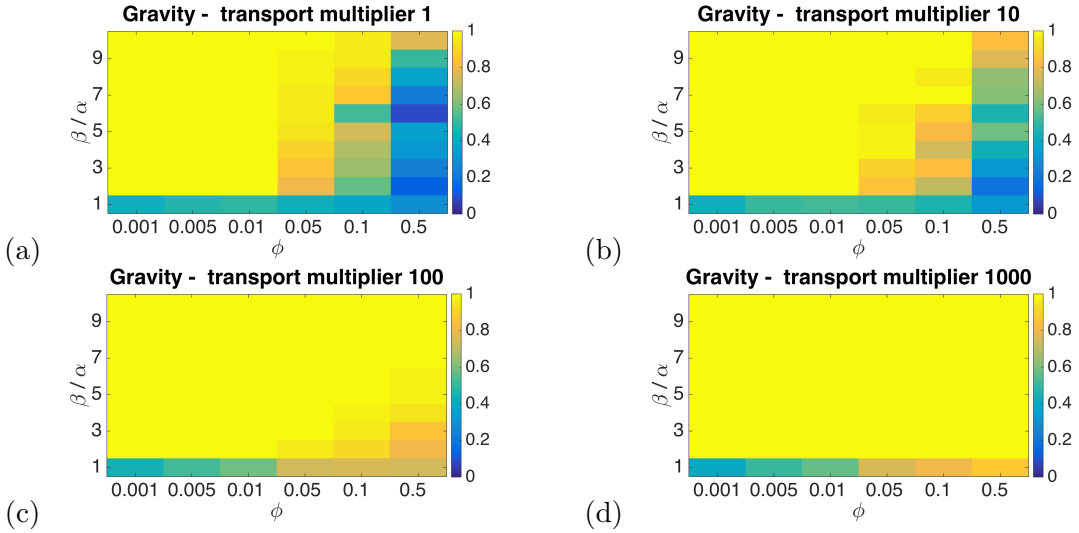


Figure 9.14: Synthetic disease model, emerging diseases, gravity null model — partitions that are statistically significant ($p < 0.05$) in the start-time test. A plot of the mean fraction out of 50 realizations that score significantly ($p < 0.05$) in the start-time test for various transportation probabilities ϕ and (binned, with bin lower bounds shown on the axis) ratios between the infection rate and recovery rate β/α for (a) intra-community transport multiplier $\eta = 1$, (b) $\eta = 10$, (c) $\eta = 100$ and (d) $\eta = 1000$.

9.3.4 Exploring the limitations of the correlation null model

Our experiments suggest that the correlation null model usually performs worse than the other null models on disease-correlation networks constructed from synthetic time series (see Section 9.3.2). This is unlike our findings for real data sets, where the correlation null model was often finding interesting network partitions. To explore the reasons behind the failure of the correlation null model to detect planted and spatial partitions, we examine the number of eigenvalues which are above the RMT threshold that is used to decompose the correlation matrix by the correlation null model into the "group mode" and the "random mode". We search for $\lambda \geq \lambda_+ = (1 + \sqrt{\hat{N}/T})^2$ where $\hat{N}$ is the number of nodes with non-zero strength in the network, and T is the length of the time series, as defined in Section 3.3.6. We show the mean number of eigenvalues that match these criteria for networks corresponding to emerging and endemic diseases in Fig. 9.15. We show results for $\eta = 10$ as this parameter value showed high variability in whether our methodology was able to detect planted communities, spatial communities and communities related to the first infection time.

Observe for the emerging diseases the large section of the parameter space at high transportation probability ϕ and small β/α ratio, in which the correlation matrices seem to contain little "non-random" information, i.e., they have few eigenvalues above the RMT threshold. This section corresponds to the regions of parameter space in which all null models (and NG especially) struggle to find planted communities, and where NG does not find spatial communities. Further, neither null model finds communities related to the first infection time in this parameter region. This parameter region may be worth exploring further.

However, this apparent relationship between the number of "non-random eigenvalues" and the ability to detect communities does not work in the other direction. For endemic diseases, the only region in parameter space where correlation matrices on average contain multiple "non-random" eigenvalues is the region with low β/α ratio. Nevertheless, in our experiments, none of the community-detection methods have been able to detect the planted community structure or spatial organization in any of the endemic disease networks. Further, note that the region with the highest average number of "non-random" eigenvalues for emerging diseases in Fig. 9.15(a) does not correspond to the region where either of the methods performs best. The significant structure contained in the eigenvalue information in both of those regions of the parameter space might correspond to interesting information that our methods are unable to capture, such as a combination between the planted partition and geographical structure — which we may be able to detect using a modified version of our methodology.

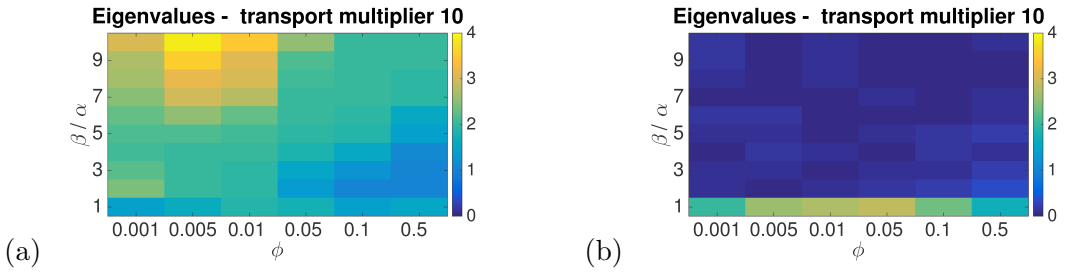


Figure 9.15: Average number of significant eigenvalues (above the RMT cutoff) of the adjacency matrices for emerging and endemic diseases. A plot of the count of the average number of eigenvalues that are significant under the Random Matrix Theory cutoff of significant eigenvalues, and are thus used to generate the modularity matrix B^{cR} in the correlation null model, for (a) emerging and (b) endemic diseases.

9.3.4.1 Problems with the correlation null model

As we saw in the previous section, most disease-correlation networks corresponding to endemic diseases appear to contain only a small number of "non-random" eigenvalues. Thus, the majority of the weight contained in these correlation networks is removed by the correlation null model. This leaves very low values in the modularity matrix B^{cR} and might cause community detection to struggle.

Similar issues are also present for the emerging diseases. We show the original adjacency matrix and the modularity matrix B^{cR} for an emerging disease simulation with $\alpha = 0.5$, $\beta = 0.1$, $\phi = 0.005$, $\eta = 1000$ in Figs 9.16 (a,b). The variation in the modularity matrix is dominated by the difference between node 0 (where the disease was originally seeded) and the rest of the network. For this case, community detection using modularity maximization places all nodes except node 0 in the same community for many resolution parameter γ values.

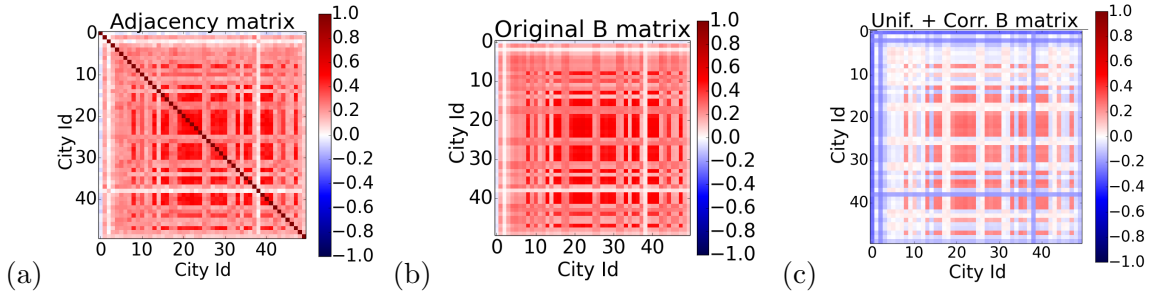


Figure 9.16: The issues with the correlation null model. In (a) we show the correlation matrix A for $\alpha = 0.5$, $\beta = 0.1$, $\phi = 0.005$, $\eta = 1000$, showing correlation between all the cities despite the strong differences in transport depending on community assignment; (b) shows the original modularity matrix B^{cR} generated from the time series of new number of disease cases for this parameter regime, and (c) shows the modularity matrix after subtracting its mean.

9.3.4.2 Modifying the correlation null model

When we examine the community structures generated using the correlation null model for the majority of disease data sets and the agent-based model, we observe little difference

with changes in resolution parameter γ; this parameter does not work in the same manner as for other null models that we test in this thesis, because the unlike them, the correlation null model contains both positive and negative values. This means that the simple relationship between γ and community size does not work for the correlation null model. This has also been noted by the authors of the correlation null model [171]. They propose a different approach to generating a multiresolution community structure by iteratively decomposing communities using the correlation null model. Here, we test a different approach to obtaining meaningful and multi-resolution community structure using the correlation null model.

After removing the "random" elements of the correlation matrix with the correlation null model, we perform a second step in which we apply the uniform null model to the modularity matrix B^{cR}. This allows the community-detection to detect the block-diagonal structure present in the modularity matrix B^{cR} [26, 267]. The final modularity matrix is

$$B_1 = B^{\mathrm{cR}} - \langle B^{\mathrm{cR}} \rangle = A - P^{\mathrm{cR}} - \langle A - P^{\mathrm{cR}} \rangle \,, \tag{9.10}$$

where $\langle B \rangle$ is the mean of a matrix B, and P^{cR} is the correlation null model matrix. This generates more variability in the signs of the new modularity matrix B_1 [see Fig 9.16(c)]. This is a novel use of the uniform null model, in contrast to the previous research, where it is used as a stand-alone null model applied to the adjacency matrix of a network. One could also use a similar approach to generate an "additive resolution parameter", where rather than removing the mean of the modularity matrix, one would remove a constant γ which one could vary over (i.e., let $B_1 = B - \gamma = A - P - \gamma$). This is similar to the approach of Traag et al. [267], who use a constant as their null model ($P = \gamma$).

This approach yields a marked improvement in the ability of the community-detection using the correlation null model to detect planted partitions, as shown in Fig. 9.17. With this additional step, the correlation null model is able to detect the planted communities comparably as well as the NG and gravity null models for most of the parameter regimes, and better than these two null models for high β/α ratio and high ϕ. Further, this correction removes the apparent successes that we observed for endemic diseases using the original correlation null model for all η values (likely due to numerical errors) in Fig. 9.11.

The modified correlation null model performs especially well for $\beta/\alpha \gtrapprox 7$ and $\phi \gtrapprox 0.05$, where it scores significantly on the NMI test for $\eta > 10$. Further, community detection using the modified correlation null model is able to reliably (for over $\geq 80\%$ of realizations) detect a coarse-structure similarity to the planted partition (z-Rand scores above 2) for most of the parameter space for $\eta > 10$, and it has $50 - 80\%$ reliability for $\eta = 10$, only slightly less than the NG and gravity null models. These structures are not detected by NMI because modularity maximization with the modified correlation null model only appears to detect

one of the two planted communities perfectly or near-perfectly, and it splits the nodes that belong to the second planted community into several small communities.

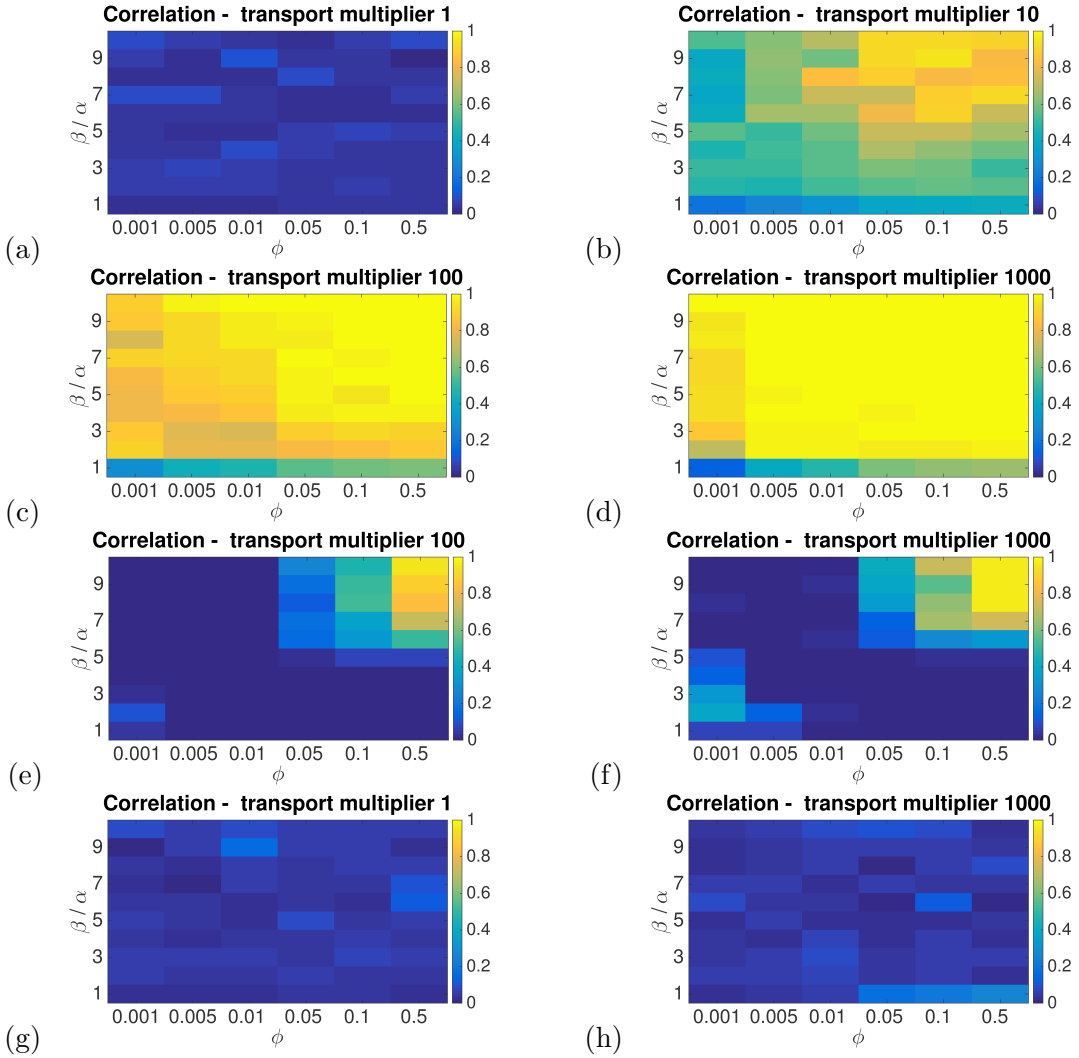


Figure 9.17: Synthetic disease model, emerging and endemic diseases, modified correlation null model — detecting planted partitions using z-Rand scores and NMI. Mean fraction out of 50 realizations that successfully detect planted community structure for various transport probabilities (ϕ) and ratios (β/α) between the infection rate and recovery rate for (a)-(f) emerging diseases and (g)-(h) endemic diseases. Plots (a)-(d) and (g)-(h) use z-Rand scores ($z_R > 2$), and plots (e)-(f) use NMI (NMI > 0.8). The intra-community transport multiplier is (a) $\eta = 1$, (b) $\eta = 10$, (c,e,g) $\eta = 100$, and (d),(f),(h) $\eta = 1000$.

9.4 Conclusions

In this chapter, we presented the results of modelling the progress of a disease in a very simple spatial system and the community detection on disease-correlation networks constructed from the disease time series. We considered the spread of an SIS model of an infectious disease on a ring of 50 cities with a distance-dependent transportation system, which we can influence using a planted community structure. We classify our numerical experiments into emerging and endemic diseases, in line with the classification we used for real disease

data sets in Chapters 7 and 8.

We were able to detect structures that are very close to the planted communities (using NMI) only for emerging diseases. We had the most success in detecting planted partitions for large β/α ratios (corresponding to fast-spreading diseases) and high intra-community transport multiplier η (i.e., when there is little mixing between communities). When studying the coarse-structure similarities in community structure using z-Rand scores, we were able to detect some similarity to planted partitions for parameter regimes for which $\beta/\alpha \approx 1$, so effect of stochastic effects on the spread of disease is larger. The method successfully detects coarse-structure similarity to the planted communities ($z_R > 2$) for 40–80% of the experimental realizations for both null models at $\eta = 10$, and it performs much more reliably for higher η (scoring 80–100% successes over the experimental realizations for both null models). Furthermore, both the NG and gravity null models generate spatial partitions (as scored by the distance test), especially so for experiments with relatively large levels of inter-community mixing (where intra-community transport multiplier $\eta \in \{0, 1\}$). Most of these partitions appear to be related to the first infection times of diseases. This seems consistent with our results for Ebola, especially so for the WHO data set, which also showed spatial communities that were related to the first infection times (see Section 8.3). However, to observe spatial partitions in the ABM, we require a high infection rate β compared with the recovery rate α ($\beta/\alpha \gtrapprox 7$), and high transport ($\phi \gtrapprox 0.01$), which may not be realistic for real situations. Further, we observed spatial partitions using the gravity null model, which is in contrast to our findings for spatial benchmarks (where the spatial null models removed spatial effects from the data allowing the detection of planted communities) and disease data sets, (where the spatial null models appeared to remove the majority of spatial organization from the correlation networks).

For endemic diseases, we observed that the disease-correlation networks for $(\beta/\alpha) > 1$ contain few eigenvalues that are considered non-random by the RMT approach employed by the correlation null model [i.e., $\lambda \geq \lambda_+ = (1 + \sqrt{N/T})^2$]. These networks may contain less "non-random" information than the disease-correlation networks corresponding to emerging diseases. None of the null models that we tested found meaningful communities in this regime (as judged by the degree of spatial organization, start-time test scores, and the ability to detect planted partitions). The findings for endemic diseases are in contrast to our findings for real data sets, as we were able to find strong spatial partitions that appear to be related to climate for dengue fever, and strong (but yet unexplained) spatial partitions for rubella. It might be interesting to modify the agent-based model to include scenarios that are more similar to these diseases to see whether we are able to reproduce these findings. Alternatively, modifying the community-detection methodology might also allow us to detect some meaningful partitions for these networks.

In this chapter, we also experimented with modifying the correlation null model, as the original version of the null model was not detecting spatial and planted communities successfully on our synthetic disease-correlation networks. Further, the resolution parameter γ does not work well for the correlation null model, as suggested by the authors in Ref. [171] and observed in our experiments (for both the disease data sets and the ABM). We found that using the uniform null model on the modularity matrix B^{cR} to generate a second modularity matrix ($B_1 = B^{\mathrm{cR}} - \langle B^{\mathrm{cR}} \rangle = A - P^{\mathrm{cR}} - \langle A - P^{\mathrm{cR}} \rangle$) appeared to mitigate the problem and allowed the correlation null model to perform comparably to the NG and gravity null models in detecting planted partitions for most parameter regimes, and better than these two null models for high β/α ratio and high ϕ. This kind of an additive null model approach with subtracting an arbitrary γ value rather than the mean $\langle B \rangle$ could allow a different version of a resolution parameter that is more suitable for the correlation null model.

Chapter 10

Conclusions

In this thesis, we combined the application of community detection to a new topic (the spatial spread of endemic and emerging diseases) with a detailed study of the results of community detection by modularity maximization using the standard Newman-Girvan (NG) null model as well as spatial and correlation-specific null models. We evaluated the performance of modularity maximization with these null models on synthetic spatially-embedded and temporally-evolving benchmark networks with planted connectivity, and on disease-correlation networks arising from real and synthetic data sets describing the incidence of endemic and emerging diseases. We compared the results of modularity maximization using these null models and sought insights into the ability of this approach to detect meaningful and interesting community structure for these examples.

Generating correlation networks has been a promising approach for applications ranging from gene expression patterns [31,150] to stock market returns [84,171]. Further, community detection is a well established but growing field that has shown many promising results in finding new structures in data sets [91, 219]. Its application to spatially-embedded [22, 47, 82, 119, 242] and temporally-evolving networks [23, 29, 48, 50, 83, 140, 194] are very active fields of research.

Our community-detection approach attempts to find spatiotemporal patterns in the spread of disease. It thus addresses external influences on the spread of disease from factors such as distance, climate and socioeconomic factors, and aims to identify times when the patterns of infection change. Our approach has similar limitations as some of the statistical methods of studying infectious diseases that we reviewed in Chapter 2, as it is only able to detect associations rather than mechanisms and it requires high-quality data, which can be problematic during an emerging epidemic or for epidemics in developing countries. However, the advantages of our approach include the fact that it does not require any knowledge about the disease (although such knowledge can illuminate the analysis of results) and it is applicable (in principle at least) to all kinds of disease data sets without the need for customization.

In practice, we found that while our approach to community-detection in spatial networks is able to relatively reliably detect simple associations on spatial benchmarks, the connectivity structures contained in disease-correlation networks appear to be relatively complex, and we have only begun to explore them. We will now review the main results of each chapter of this thesis.

10.1 Main results of the thesis

Network-science methodology. In Chapter 3, we introduced some vocabulary, reviewed relevant aspects of network science and community detection, and we described several null models for modularity maximization: NG, gravity, and correlation null models. In Eq. (3.23) we developed a novel "radiation null model", which is based on a radiation model for mobility that was proposed relatively recently and appears to match empirical mobility data better than gravity models [245].

We also presented the methods that we used to examine network partitions. We expected the community structure in disease-correlation networks to have strong spatial organization. We thus developed methods of assessing the level of spatial organization in a network partition: in Section 3.4.3 we described using z-Rand scores to compare an algorithmic partition against a manual spatial partition, and in Section 3.4.4 we presented a "distance test" and a "minimum spanning tree test" that measure the extent of spatial clustering in a community structure. Furthermore, for data sets that describe emerging diseases, the time when the infection first reached a node could be important to the algorithmic community assignment. In Section 3.4.6, we developed the "start-time test": a statistic to measure whether community assignments in disease-correlation networks are related to the first infection times for provinces (i.e., the times when the disease first reaches a province). For all three tests, we assessed the significance of the test statistic by Monte Carlo sampling.

We then developed the methodological pipeline that we used for all of the disease-correlation data sets. This consists of:

- Constructing a set of static networks and a multislice network from the disease time series.

- Community detection with each null model (NG, correlation, gravity and radiation null models) on static networks over a variety of γ. Examining the community structures with a focus on assessing the spatial organization of partitions using z-Rand scores (if applicable), distance and MST tests.

- Community detection with each null model (NG, correlation, gravity and radiation null models) on a multislice network over a variety of γ and ω. Examining the community structures with a focus on (1) assessing their spatial organization using distance

and MST tests (both partition-wide and layer-wise) and z-Rand scores if applicable, and (2) assessing their temporal organization using z-Rand scores against manual temporal partitions and the start-time test.

Finally, we discussed alternative approaches to studying the spatial spread of disease, and we evaluated their advantages and disadvantages compared with our methodology.

Testing the null models on spatial benchmarks. In Chapter 4, we tested the usefulness of community detection by modularity maximization using NG, gravity, and radiation null models on synthetic benchmarks with planted connectivity structure. We developed spatial benchmarks in which edge weights between spatially-embedded nodes are based on distance (through relationships based on gravity and flux) and on community membership based on an edge probability distribution generated from a planted partition.

Our results indicate that it is important to incorporate spatial information into null models for community detection (see Fig. 4.6). We also demonstrated that it is not simply a matter of incorporating spatial information in an arbitrary way. It is important to incorporate only relevant information into null models, as extraneous information can decrease the ability to find planted partitions (see Fig. 4.7).

We extended our benchmarks to a multilayer setting, and to make them more realistic, we introduced temporal variation in planted partitions. We implemented this by (1) independently constructing each layer based on the same edge probability distribution and (2) using an evolving planted partition, in which a fraction of nodes change community assignment at each layer. We found that the performance of modularity maximization with different null models on multilayer benchmarks was comparable to their performance for the corresponding static networks. Further, the inter-layer connectivity strength ω only influenced the results of community detection for temporally evolving benchmarks (see Fig. 4.16).

We also tested province-level community detection, in which we aim to recover the most persistent partition in terms of physical nodes by maximizing modularity with the uniform null model on an association matrix representing the frequency of the co-classification of physical nodes in layers of a multislice network. We were able to recover the original planted partition in terms of physical nodes with a similar performance of null models at different mixing levels than for static networks (see Fig. 4.14). Province-level partitions are useful for further analysis of the meaning of community structure in terms of node attributes and locations.

Describing diseases, data sets, and disease-correlation networks. In Chapter 5, we discussed the diseases and data sets that we examined in this thesis in the context of community detection and the expected influences on the spatiotemporal spread of disease.

We divided the diseases into endemic diseases (established in a population at a certain level, with occasional epidemic outbreaks) and emerging diseases (newly-discovered or introduced into a region for the first time). The data sets for the two disease types present different features. Endemic disease data sets cover multiple seasonal epidemics, and the patterns of infection may have changed during the time covered by them. In contrast, emerging disease data sets consist of the time course of only one epidemic that entered a susceptible population. Other aspects that may influence the spatiotemporal patterns of infections are the mode of infection (vector-borne or person-to-person), the degree of contact needed for infection, or the seasonal effects on infectivity.

We also briefly discussed the reasoning behind our parameter choices for constructing disease-correlation networks, taking into account factors such as the serial interval between consecutive infections and the computational complexity of the resulting disease-correlation networks, and the influence of these choices on the disease-correlation networks. We presented comprehensive results of our assessment of various parameter values in Appendix A.

Testing the applicability of our methodology: application to dengue data. In Chapter 6, we applied modularity maximization with different null models to disease-correlation networks constructed from disease time series for dengue in Peru. We performed a detailed analysis of the ability of this approach to deliver reliable and interesting community structures from static and multilayer networks. We also examined the potential for generating informative province-level partitions from multilayer networks and from fully aggregated networks.

We chose this data set for detailed analysis because we expected to find community structure with a strong degree of spatial and temporal organization, as Peru consists of three climatic zones with different disease patterns, and the data set includes several large changes in disease patterns.

We observed that the NG and correlation null models find community structures that are strongly spatial in the static networks, especially during the large epidemic in 2000–2001 and after the onset of yearly epidemics (see Fig. 6.1). The communities are often related to climate. For multilayer networks, both null models produce spatial and temporal partitions, or partitions with both spatial and temporal features (depending on the parameter values); see Fig. 6.4 for an example and spatial scores, and Fig. 6.5 for temporal scores over a variety of parameter values. Temporal partitions successfully find the most important time point in the history of the disease (the introduction of two new disease strains in 2000, which led to a large countrywide epidemic and the following recurrent disease patterns), and several other potentially interesting time points and periods of high spatial correlation.

When studying province-level network partitions, we illustrated that our approach of generating an association matrix from the multilayer network is preferable to complete data

aggregation. Partitions from fully-aggregated networks consist of a single community for most null models and parameter values, and in the few instances where they detect meaningful structures, they are dominated by the single largest event: the 2000–2001 countrywide epidemic. In contrast, the province-level partitions from multislice networks highlight the importance of climate to the spread of dengue, and they yield distinct temporal disease incidence patterns for nodes assigned to different communities (see Fig. 6.14).

When we attempted to remove the influence of space by using the gravity and radiation null models, community structures did not score as spatial in the z-Rand scores and distance and MST tests (see example in Fig. 6.11). The network partitions usually contained one large community with all but the highest-population provinces (which were assigned to singleton communities). This suggests that the spatial null models may be removing the majority of spatial information contained in the disease-correlation networks, with population the only variable remaining.

Application to data on other endemic diseases. In Chapter 7, we applied modularity maximization with different null models to a data set about the incidence of another endemic disease: rubella in Peru. Communities that we found using the NG and correlation null models had a comparable amount of spatial organization to the communities that we found for dengue. These results suggest that there may be significant spatial patterns in disease correlation even for non vector-borne diseases that are not explicitly influenced by climate. Similarly to what we observed for dengue, spatial organization was strong for networks covering the first of the two large epidemics of rubella in Peru; however, the level of spatial organization fell during a large lull between epidemics and did not grow with the onset of the second outbreak. This change in spatial organization might be due to the small-scale vaccination program ran during the period between the two epidemics, however we were unable to verify this intuition due to lack of data.

The temporal partitions of multilayer networks yield several important time points in the local history of disease incidence. This includes a large lull between two major epidemics and a large decline in numbers of disease cases following the introduction of vaccination programs. Partitions with spatial null models contain more potential critical time points with higher temporal z-Rand scores than partitions with the other two null models (see Fig. 7.9). This is in line with our initial expectation that if we removed the spatial organization from a multislice partition, the temporal organization would be highlighted.

We also analyzed a second data set related to an established disease: seasonal influenza in Chile. Our results for that data set did not suggest any clear spatial patterns beyond a general north-south influence that may be due to variability in climate. The spatial organization in the algorithmic network partitions using all null models is is rarely statistically significant in the distance test, and it varies significantly with parameter changes and over

time. We included these results in Appendix B.3 for completeness. However, temporal partitions of the multislice networks of Chilean influenza detect a year-long period with the lowest number of cases in the data set as a highly-scoring and clear-cut temporal partition, which we showed in Fig. 7.10.

Application to data on emerging diseases. In Chapter 8 we applied modularity maximization with different null models to two data sets describing the spread of the 2014 Ebola epidemic in West Africa. The data sets differ in the level of detail: the Datamarket data set is long and contains daily data, but it starts after the epidemic has reached the majority of provinces. The WHO data set contains weekly measurements and despite the fact that is covers a longer time period than the Datamarket data set (stretching back nearly to the origins of the epidemic), it contains fewer time points. We analyzed both data sets in order to test the usefulness of our approach on short data sets — as data constraints are to be expected for urgent investigations into emerging diseases — and to test the intuition that incorporating first infection times into the time series may influence the results of community detection.

Community detection on both Ebola data sets resulted in spatial partitions that roughly correspond to country boundaries. The results for the WHO data set seem the most promising. We have been able to link the community assignments to the first time the infection has reached each province, both visually and using the start-time test (see Fig. 8.7 and Fig. 8.11).

We also applied our methodology to a data set about the H1N1 ("swine flu") epidemic in Mexico in 2009. Our method was unable to find interesting and statistically significant spatial partitions including the known heterogeneity in the three waves of the 2009 epidemic. However, we detected strong temporal partitions which appear to correspond to the peak and the end of the epidemic wave, which we show in Fig. 8.12. For completion, we present the majority of these results in Appendix C.3.

Summary of findings on disease data sets. Here we present a summary of the patterns in our findings for the six disease data sets. We further investigated these patterns using a disease model in Chapter 9.

We have been able to find spatial partitions in endemic diseases (dengue and rubella) and emerging diseases (both Ebola data sets). The spatial partitions tend to correspond to periods with high numbers of disease cases, i.e., large country-wide epidemics, for both types of data sets. However, some large epidemics (notably, the second large epidemic in the rubella data set and the H1N1 epidemic that is contained in the seasonal influenza data set) are not detected by our methodology. Furthermore, the spatial partitions in the Ebola WHO data set are associated with the first infection times of the disease.

The multislice community structures varied between being dominated by spatial and temporal organization as we changed the parameters γ and ω, although we rarely observed clear, smooth progress of variation in the type of structures. The spatial organization of partitions tends to be similar to the structure found in the respective static partitions. We have been able to detect statistically significant temporal partitions in both endemic and emerging disease data sets. For endemic diseases, these partitions often correspond either to a large epidemic with a countrywide change in disease occurrence (dengue), or more commonly, a fall in the number of disease cases (rubella and seasonal influenza). For emerging diseases, the temporal partitions potentially correspond to the epidemic peaks, the beginnings of the fall in the number of disease cases, and to large-scale changes in relative numbers of new cases between regions. The results for disease data sets for the correlation and NG null models are summarized in Table 10.1.

Table 10.1: Overview of the results of community detection for all disease data sets and the NG and correlation null models.

Data set	NG n.m.	Correlation n.m.
Dengue	• Spatial partitions during epidemics, related to climate • Temporal partitions: 2000–2001 epidemic, start of yearly epidemics • Different patterns of disease occurrence in the whole time series for province-level communities	• Similar as for NG • Little variation with γ
Rubella	• Spatial partitions during first epidemic, unrelated to climate • Temporal partitions related to inter-epidemic period	• Similar as for NG • Little variation with γ
Seasonal influenza	• Little reliable spatial organization • Strong temporal partition related to the year-long period with low disease case numbers	• Similar as for NG • Little variation with γ
Ebola Datamarket	• Spatial partitions related to country boundaries (Sierra Leone) in the later part of the data • Temporal partitions related to the formation of the Sierra Leone community	• Similar as for NG • Little variation with γ
Ebola WHO	• Spatial partitions related to country boundaries (Guinea and Liberia) • Different first infection times between communities	• Similar as for NG • Little variation with γ
H1N1	• Little reliable spatial organization • Temporal partition related to the peak and end of the epidemic	• Similar as for NG • Little variation with γ

For the NG null model, the majority of interesting partitions lie in the parameter regime $1 \lessapprox \gamma \lessapprox 2$, and temporal partitions in particular tend to occur most often for $\gamma \approx 1$ and $\omega \lessapprox 0.3$. For the correlation null model, temporal partitions tend to occur for $\omega \lessapprox 1$. For the gravity and radiation null models, the majority of partitions are composed of one large community and several singletons or small communities that contain the highest-populated nodes.

Application to synthetic disease data. In Chapter 9, we applied our methodology of modularity maximization with different null models to the results from an agent-based model of disease spread which we developed. In this model, a hypothetical disease spreads between 50 cities located on a ring through a transport structure. Individuals travel between cities with probability ϕ in each time step. The probability of a traveling individual selecting a particular city as a destination is inversely proportional to network distance between the origin and destination cities. We introduced planted community partitions through an intra-community transport multiplier η that increases the chances of selecting a destination within the same community as the origin. We performed numerical experiments for situations representing both emerging and endemic diseases, in line with the divisions that we used in Chapters 7 and 8.

For emerging diseases, modularity maximization with the NG and gravity null models is able to detect planted communities at a large range of η. Both of these null models also yield spatial partitions, especially when there is extensive inter-community mixing in the model (see Fig. 9.3 and 9.4). This is in contrast to the findings for benchmarks and disease-correlation benchmarks, where the gravity null model was successful in removing spatial organization. The majority of partitions appear to be related to the first times when the infection reached the provinces for all values of η, although there could be other factors affecting community assignment. Furthermore, we should note that the conditions for which we detect spatial and planted communities are not realistic (i.e., the infections are very "aggressive" in terms of having a high ratio of infection rate to recovery rate, and they require a high transport probability to generate spatial partitions). We also found that both null models (and NG in particular) struggle to detect planted and spatial communities for a region in the parameter space with high transport and low ratio of infection rate to recovery rate, in which networks on average contain little "non-random" information (i.e., have few eigenvalues above the RMT threshold that is used to decompose the correlation matrix by the correlation null model into "group mode" and "random mode").

For endemic diseases, most correlation networks appear to contain very little "non-random" information. None of our null models successfully detect the planted partitions, or find any spatial communities in these networks. This is in contrast to our results from real data sets, where we found spatial partitions for dengue and rubella.

However, we also fail to detect meaningful communities in the only region in the parameter space for endemic diseases with on average more than one "non-random" eigenvalue per network. Further, the region with the highest average number of "non-random" eigenvalues for emerging diseases does not correspond to the region where either of the methods performs best. Any structure contained in the eigenvalue information for networks in both of those regions of the parameter space might correspond to information that our methods of detecting and examining communities are unable to capture, such as a combination between the planed partitions and geographical structure.

We also experimented with modifying the correlation null model, as the original version of the null model was not performing well on our synthetic disease-correlation networks. The variability in the correlation null model modularity matrix B^{cR} is dominated by the division between the node where the disease is seeded and the rest of the network, resulting in failure to detect meaningful network partitions. The authors of [171] propose an iterative approach to generating a multiresolution community structure by iteratively decomposing communities using the correlation null model. We tested a different approach: using the uniform null model on the modularity matrix B^{cR} (rather than on the adjacency matrix A directly), to give a new modularity matrix B_1 in Eq. (9.10). This appears to mitigate the problem and allows modularity maximization to detect planted partitions and spatial partitions. One could also use a similar approach to generate an "additive resolution parameter", where rather than removing the mean of the modularity matrix, one would remove the constant γ over which one varies [267].

10.2 Significance and outlook

To our knowledge, no methodology similar to community detection has previously been applied to disease time series. Previous approaches to studying disease time series have tended to focus on specific aspects — e.g., detecting spatial clusters in disease cases [111, 277], detecting cycles in time series [45, 198, 228], or searching for dependence of disease incidence on temporal and climatic variables [76,127,198]. Community detection's flexibility gives it a potential advantage over those approaches, because it makes few assumptions about the structure of the time series. This also allows us to use the same methodology for many different disease networks. We can define an appropriate similarity measure and use it to generate networks that capture desired elements of disease time series, and we can choose null models for modularity maximization to focus on specific aspects of the resulting networks.

Despite its successful track record for financial [84, 171], climate [80, 173] and brain-activity [24] time series, our application of modularity maximization to disease-correlation networks appears to give little additional insight into the mechanisms of disease spread (at least for the null models that we examined). We were largely only able to detect spatial

partitions for diseases with a known strong spatial influence (such as climate for dengue fever), or for emerging diseases (such as Ebola), especially if the initial infection times are part of the time series.

We now discuss potential directions for future research: validating the results of community detection, extending the methodological pipeline, and other network-science research questions that arise from our work. We also discuss the possible impact of such research in the wider context of the fields of spatial epidemiology and community detection in networks.

10.2.1 Validating results

The problem of assessing the reliability and utility of the results of algorithmic community detection is exacerbated by the fact that there is no rigorous definition of a "community". The evaluation of the results of community detection on real data sets can be relatively easy if one finds the structure that one was expecting — for example, the spatial clusters similar to the actual administrative regions in studies of mobility networks in Refs. [222,262] or the language partition of Belgium in the examination of mobile phone data using the gravity null model in Ref. [82]. If the results are not as clear, as is usually the case for most empirical data sets (including our case), there is no consensus as to how to assess their significance and relationship with the data. We took the approach of searching for spatial and temporal signatures in the community structures, but other approaches may also be useful.

Assessing the validity of network partitions. A common approach to assess the reliability of community-detection results is to identify communities using different algorithms and only to consider structures that are similar across multiple methods [219, 264]. In this thesis, we only used a modularity-maximization approach to community detection in order to focus on the effects of using different null models. Moreover, only selecting structures detected by multiple algorithms could lead to discarding structures that are informative but only detected by a small number of methods.

One could also assess the statistical significance of modularity scores achieved in the community detection compared to these expected from suitable random network null models [24]. Such null models are often based on permutation tests, and different aspects of the community structure can be assessed by using different randomizations (e.g., for multislice networks, rewiring edges within a layer, between layers, or simply reordering layers).

Another approach to aid the analysis of real data sets can be through applying community-detection methodology to synthetic data with known properties. This could help to develop better insights as to the types of structures to expect and to the best methodology to use, as we will discuss in more detail in Section 10.2.2.

Further exploration of community structure. One can also further explore the results of community detection in the context of the particular complex system represented by the network. In the context of this thesis, our inability to detect "interesting" community structures in some of our data sets, and in the ABM endemic diseases can arise from multiple sources, which we will discuss now.

First, it is possible that we have indeed not been detecting meaningful community structures for at least some of the diseases. We could test this using some of the aforementioned methods of validating community structure, such as using different algorithms and assessing the statistical significance of community structure.

Second, the problem of validating communities may lie at least in part in our assessment of what are "interesting" communities, rather than in the network partitions themselves. Because we expected strong spatial patterns, we explicitly examined the spatial organization of partitions. We did not explore other factors that could influence disease spread (and thus network partitions), such as socioeconomic differences or transportation. It tends to be difficult to obtain such data, and we do not possess such data for the countries in our data sets. Indeed, there is a dearth of detailed data for these countries compared with those for more developed countries. One interesting extension of our work would be to apply our methodology to a well-known disease in a country for which one possesses a large amount of knowledge and additional data (e.g., influenza in the USA).

Alternatively, internal disease dynamics in the provinces that are driven by interactions within the local population may mask effects that are due to interactions that could be detected by community detection, e.g., transport, or other groupings, e.g., climate. We could explore this possibility by suitably adjusting our ABM. For example, we could explore the relative effects of the internal disease dynamics and the effect of imported cases on endemic diseases. This could then guide modifications to network construction and community-detection methodology that may enable us to detect planted and spatial partitions for endemic diseases in the agent-based model, which could in turn be tested on disease data sets (see Section 10.2.2).

10.2.2 Expanding and improving the methodology

Methodological choices. In this thesis, we made particular choices for network construction, community-detection techniques, null models, and parameter values that can all influence results. It would be interesting to explore the impact that some of these choices have on the disease-correlation networks and community structures both for real data sets and disease models. We will briefly indicate some promising areas for investigation.

There are many different similarity measures for the generation of "similarity networks". Methods that are suitable for disease time series include coherence [23], Granger causality [96], and lagged correlation [239]. These methods might be able to better capture

the intricate relationships between the patterns of disease spread in provinces and even provide information as to the direction of the interactions. Alternatively, using event correlation [173], a measure that searches for patterns between predefined spikes (e.g., epidemics greater than a certain size), could reduce the complexity of the system while preserving its key elements. One could also simplify the system by thresholding the similarity network, for example using false-discovery rate, which only preserves edges that are deemed to be statistically significant [24, 99].

Other community-detection methods than modularity maximization could detect interesting communities on these data sets. For example, approaches based on local community detection [91,135,219] would allow us to focus on the communities with relation to provinces of particular interest (e.g., most-populated ones, or the ones most at risk). However, few such methods have been developed to detect communities in multilayer networks [71, 135]. Further, the ability to easily change null models for community detection is a useful aspect of modularity that many other methods do not possess.

It would also be interesting to use the uniform null model on the original adjacency matrix. It has been common to avoid using the uniform null model in applications because it was considered not to be a good representation of real-world networks [91]. However, as the uniform null model simply emphasizes the block-diagonal structure present in the adjacency matrix, some authors have argued that it can be appropriate for correlation networks [26].

Testing modifications to the methodology on synthetic time series. We could further explore the applicability of our methodology to disease-correlation networks and test any changes to the methodology (e.g., to network creation or community detection, as discussed above) by using more detailed disease model than the agent-based model presented in Chapter 9. Such models also allow one to easily explore the ability to detect communities for different conditions (e.g., different types of diseases or different types of planted partitions). When using synthetic data, it is easy to test whether modifications to the community-detection methodology improve our ability to detect planted partitions (which in real data sets could correspond to features such as socioeconomic factors, climate, or transportation networks) or other spatial partitions. For example, in Section 9.3.4.2, we found that applying the uniform null model to the modularity matrix generated by the correlation null model allowed us to detect planted communities in the ABM time series. We could guide the choice of our approach to network construction and community detection by examining the results of a suitably modified disease model and examining the disease dynamics in detail, as we discussed for the endemic disease example in Section 10.2.1.

We can consider many types of modification to the ABM. For example, it would be interesting to use a more complicated spatial structure — for example, a regular lattice,

random placement similar to the placement in our spatial benchmarks, or a realistic placement of cities in space based on one of the data sets. We could also make the disease model more realistic by introducing an explicit contact step or by increasing the number of individuals, or by some other means. Additionally, we could explore different parameter values and general design choices — including the investigation of different trip lengths, changing the behavior of agents based on their disease status, and planting partitions based on disease-related parameters (to simulate influences of phenomena such as climate). Furthermore, we could redesign our disease-spread model using a different approach from an ABM. For example, we could use stochastic differential equations, which may be able to reproduce features of the disease time series that the current model cannot (e.g., epidemics with different heights and widths of peaks).

Such exploration might direct us to improve the usefulness of our network creation and community detection for analysis of disease time series, and any changes to the methodology can be tested on real disease-correlation networks.

Finally, we have thus far not used the ABM to consider temporal partitions of disease-correlation networks. We could explore such partitions by constructing multislice networks from the synthetic time series. By introducing an explicit change in the model parameters (such as one of the disease dynamics parameters, the intra-community transport multiplier, or the planted partition) at a defined time point, we could plant critical time points at which different aspects of the overall dynamics change. We could then explore the ability of our community-detection methodology to detect these critical time points, either by our temporal z-Rand scores or using a different method to detect change points [20, 214].

Other approaches to studying disease-correlation networks. Other network-science methods beyond community detection can also be used for investigating the structure of disease-correlation networks. One avenue that we briefly examined using the dengue data set was to search for "node roles" — groupings of provinces according to their properties in the disease-correlation networks [84, 115]. This approach has the potential to identify provinces that are important for the spread of disease in their neighborhoods (or their communities). Further, searching for changes in node properties could potentially provide early warning signs before epidemics. Changes in community structure have previously been found to accompany large events such as financial crises in exchange-rate networks [84] and political upheavals in voting networks [172], however, studies to date tended to focus on correlation rather than causation.

Some of the methods used to study the financial-correlation networks (see, e.g., Ref. [171]) might also be applicable to the disease-correlation networks. For example, calculating a minimum spanning tree of such networks could give an intuitive idea as to the path that the disease might take to spread in the country. This may be especially interesting for emerging

diseases, as the community structures we find there are linked to first infection times, which in turn are expected to be related to the number of travelers between regions [237]. In principle, it would also be possible to compare the minimum spanning trees directly to the transportation network of the country in question.

10.2.3 Extensions to the network science aspects of the research

Designing spatial and temporal benchmarks. Using benchmark networks with planted partitions is a well-established approach to analyzing the performance of community-detection algorithms [68,109,157,161]. However, despite a growing interest in community detection in spatially-embedded and temporally-evolving networks, the development of suitable benchmarks has lagged behind the development of algorithms. Many studies seem to be satisfied with applying their methodology to one or a few data sets [14, 82]. This can give the impression that using a spatially-embedded or temporal null model is always preferable to the standard methodology, but our research has demonstrated that this need not be the case. It is important to do additional thorough evaluations of the applicability of spatial and temporal null models in different situations. This can be accomplished by approaches such as building benchmark networks or applying community detection to the time-series output of a dynamical system.

In the context of this thesis, it would be interesting to explore our spatial benchmarks in more detail. In particular, one could further examine the multilayer benchmarks with temporally-evolving planted communities, which we only had time to explore briefly. This would allow one to gain a better understanding of the role of interlayer connectivity strength ω in algorithmic community detection [26]. Furthermore, one could also vary the magnitudes and types of changes in the planted partition between layers to study the ability of community detection to find planted partitions in noisy data [72, 109].

Application to other data sets. While in this thesis we have focused on the application of community detection using modularity maximization with various null models to disease-correlation networks, the methods that we developed here are applicable to a wide range of topics, including correlation networks from finance, gene expression, brain activity, climate, and many other sources — both real and synthetic. Further, many of the insights are not limited to correlation networks. In fact, it is possible that the spatial null models, and our novel radiation null model in particular, would perform better on spatial networks that are related to flow, rather than correlation. Finally, our insights into the importance of testing a variety of null models for community detection, due to their large influence on results, should be applicable to all attempts to detect communities using modularity maximization.

10.3 Final thoughts

The above possibilities represent just a small sample of open questions relating to community detection, disease time series, and the spatiotemporal spread of disease. Many more open questions exist in the fields of network science and epidemiology, and we hope that we showed that the intersection of the two fields is a fascinating subject.

Appendix A

Parameter choices for constructing disease-correlation networks

A.1 Introduction

In this appendix, we present the disease-correlation networks that arise from each of the data sets that we use in this thesis, and we describe the reasoning for choosing parameters for network creation (time window width Δ, the step between time windows υ) and in the binning of distance data for the gravity and radiation null models.

A.2 Influence of time window width on basic network features

In this section, we will study the effect of the choice of time window width Δ on the basic network properties for dengue, rubella, H1N1, seasonal influenza and the Ebola Datamarket data set. Recall the network creation procedure from Section 3.1.1. To inform the choice of Δ, we investigate the effects of varying it on the properties of the time series subsets $E_i^{(s)}$ used for creating the network (i.e, the slices of the time series corresponding to each static network) by examining the mean number of cases in each time series subset. We also study the effects of different Δ choices on aggregate properties of the disease-correlation networks such as mean edge weight, total edge weight and mean node strength.

We want Δ to be high enough that the correlation matrices contain meaningful information, but low enough that the properties of disease time series and disease spread between provinces are smoothed as little as possible. The suggested minimum value for Δ can be based on the random matrix theory constraint described in Section 3.1.1 is that $\Delta > \hat{N}$, where $\hat{N}$ is the number of nodes with non-zero strength in the network or layer. For each of the disease data sets, we present the minimum width of the time window resulting form this constraint, and two longer time window widths. We use these results to inform our parameter choices for the community detection experiments in Chapters 6, 7 and 8.

When we increase time window width, the features of the curve representing the mean number of cases within the time window become smoother, and as a result disease epidemics are increasingly difficult to distinguish. For the dengue fever data set, the yearly epidemics are visible in the aggregated case counts only for $\Delta = 60$ and $\Delta = 80$ out of the values that we tested [Fig. A.1 (a)]. The choice of Δ also affects our the disease-correlation networks; as we increase Δ between 60 and 100 weeks, the total edge weight in the network increases.

The increased mean edge weight and corresponding mean node strength correspond to the time periods with large mean numbers of disease cases [compare Fig. A.1 (b) to Fig. A.1 (a) for each Δ]. We also observe a strong rise in mean correlation within layers during the periods of increasing case numbers before the epidemics, and sharp drops in correlation at the end of the epidemic peaks [see Fig. A.1 (c)]; this is also best visible at Δ=60 and it becomes smoother for wider time windows. It appears that smaller Δ increases our ability to distinguish between epidemic and non-epidemic times. Based on this observation, we choose $\Delta = 80$ (the minimum value for which Δ is always larger than N) in order to preserve as much of the inter-epidemic variability as possible. We use this value for the static networks, with $\upsilon = 4$, giving 175 static networks. Unless stated otherwise, we use $\Delta = 60$ for the multilayer networks (after checking that no layer contains more than 59 nodes) in order to maximize the number of layers to study. This gives a multislice network with 12 non-overlapping layers. We also compare these results to ones with $\Delta = 80$ and $\upsilon = 24$ in order to test the influence of overlapping layers. The choice of $\upsilon = 24$ is motivated by balancing better temporal resolution and the availability to detect changes in the structure of the multislice network with the increased computational complexity resulting from having more layers.

The rubella data set contains two large epidemics. Our ability to detect the division between them worsens with larger time window width (see Fig. A.2). We do not show all of the plots that we did for dengue fever, as the mean edge weight and other network properties follow a similar relation to the number of cases as for dengue fever, with high numbers of disease cases at the same time layers at which we observe high correlations and node strengths. We strive to choose the smallest Δ that satisfies $\Delta > \hat{N}$ for all time windows in order to preserve as much of the inter-epidemic variability as possible. After testing all possible values, we find that to be 134 — for this choice, no more than 133 provinces are affected within one time window. We use $\upsilon = 4$ for the static networks (in order to reduce the number of static networks and thus lower computational complexity), and we use $\upsilon = 12$ for the multislice networks, to increase the number of layers and the temporal resolution of our experiments compared to the non-overlapping approach we took for dengue fever — giving a multislice network with 44 overlapping layers.

The H1N1 influenza data set contains only the 2009 epidemic, but the perceived duration and size of the epidemic depends on the time window width, with the aggregated data

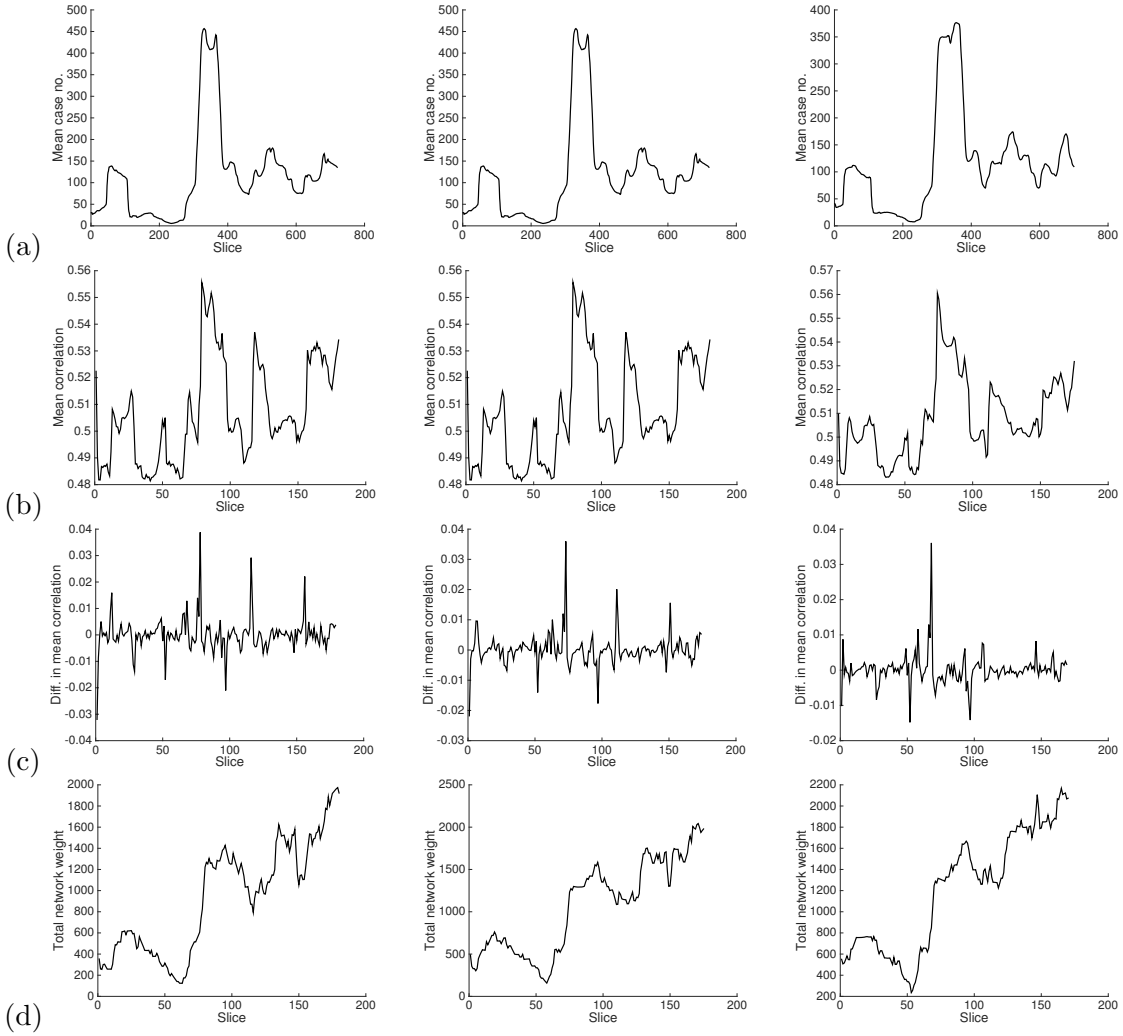


Figure A.1: Influence of time window width Δ on the dengue fever data set and network properties. (a) Mean number of cases in time window, (b) mean edge weight in the static networks, (c) the difference in mean correlation between each pair of neighboring static networks, and (d) the total weight of edges for Δ of (left) 60 weeks, (centre) 80 weeks, and (right) 100 weeks.

showing lower peak number of cases and longer epidemic duration for larger time window widths Δ (see Fig. A.3). Based on this observation, we choose $\Delta = 30$ for the H1N1 data set in order to prevent distorting the shape of the epidemic (after checking that no more than 30 provinces experience the disease at once). We use $\upsilon = 7$ for the static networks, and $\upsilon = 30$ for the multislice networks. For H1N1, we obtain the largest correlation during the very sharp build-up to the main epidemic, and the mean correlation falls after the main epidemic peak. The largest total edge weight in the network and the largest node strengths occur during this period as well, and both values fall near the epidemic peak. There is a very sharp decline in mean correlation at the tail of the epidemic (past the 400th time point in the data set).

The Chilean influenza data set contains yearly disease epidemics, which are only well visible at the $\Delta = 30$ time window width (see Fig. A.4). To preserve the distinction between

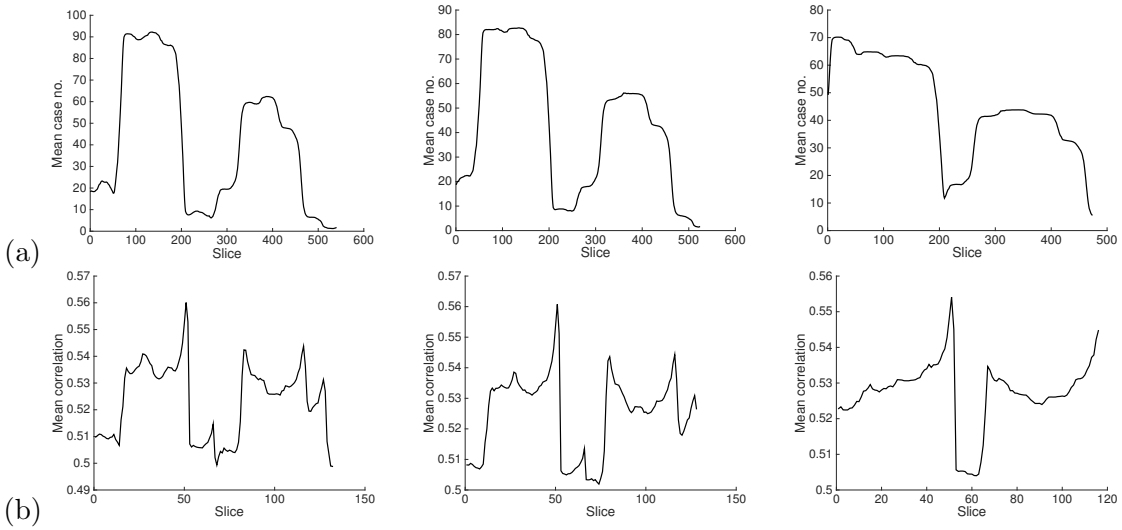


Figure A.2: Influence of time window width Δ on (a) the mean rubella number of cases within the window, and (b) mean edge weight in the static network. We use time window width of (left) 134 weeks, (centre) 150 weeks, and (right) 200 weeks.

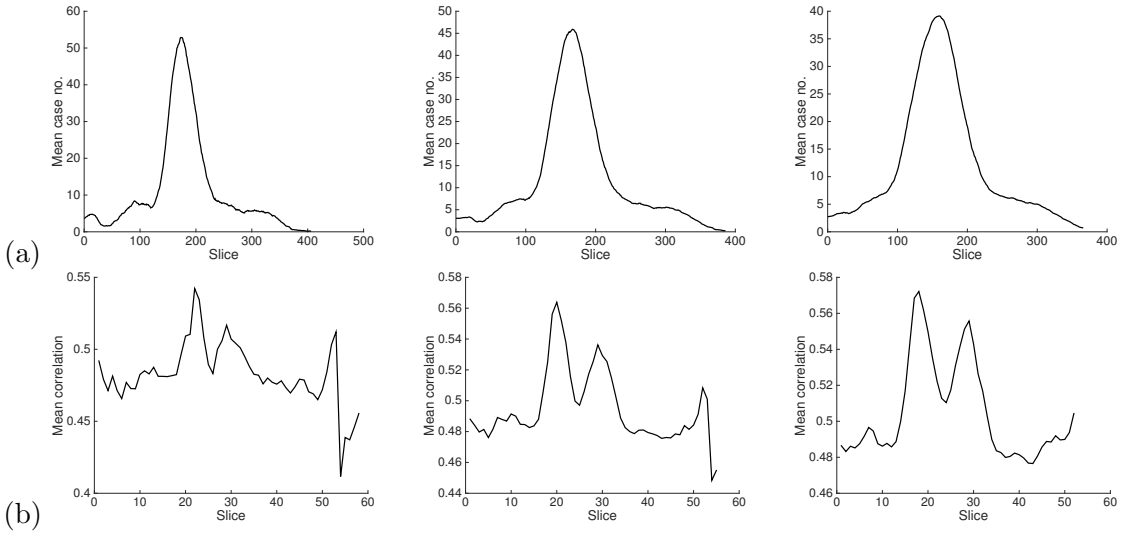


Figure A.3: Influence of time window width Δ on (a) the mean number of H1N1 influenza cases within the window and (b) mean edge weight in the static network. We use time window width of (left) 30 weeks, (centre) 50 weeks, and (right) 70 weeks.

the yearly epidemics, we choose $\Delta = 30$ for the Chile data set. We use $\upsilon = 4$ for the static networks, and $\upsilon = 30$ for the multislice networks. The pattern of mean correlations in the network is once again similar to the pattern of mean numbers of cases, with less variation at higher Δ, and higher node strengths during the influenza epidemics.

For Ebola, we only examine the Datamarket data set (as the WHO data set very is short so we use the whole data set to construct one static network). The Datamarket data set contains a very sharp increase in the number of disease cases that is only visible for time window width Δ=60 (see Fig. A.5). We use $\upsilon = 5$ for both static and multislice networks. Similar to the H1N1 data set, the mean correlation in the network is the strongest at the beginning of the epidemic.

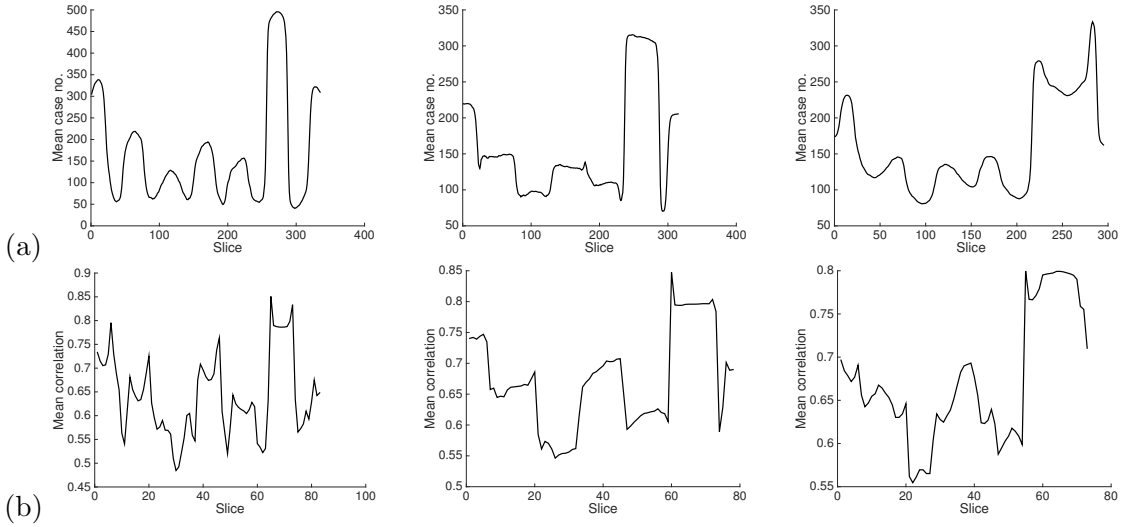


Figure A.4: Influence of time window width Δ on (a) mean number of influenza cases within the time window for the Chilean data set and (b) mean edge weight in the static network. We use time window width of (left) 30 weeks, (centre) 50 weeks and (right) 70 weeks.

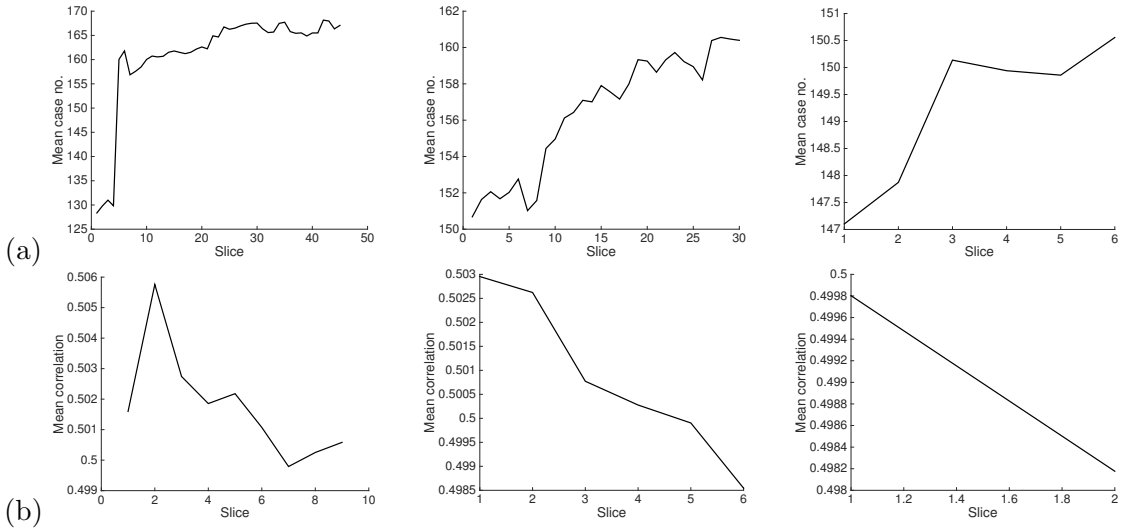


Figure A.5: Influence of time window width Δ on (a) mean number of Ebola cases within the time window and (b) mean edge weight in the static network. We use time window width of (left) 60 weeks, (centre) 75 weeks, and (right) 99 weeks.

A.3 The difference between layer starting points υ

We also choose the values for the difference between layer starting points, υ. This can be as low as 1, which would give us the highest temporal resolution for detecting changes in network structure that occur at particular time points. However, the added benefit of using a lower υ value over a higher one depends on the serial interval of the disease compared to the temporal resolution of the data set; for example, for dengue with a serial interval of 15 days, an υ of 1 day would perhaps include less extra useful information over υ of 2 days than the equivalent change in υ would for influenza (with a serial interval of 3–4 days).

In practice, for most of the disease data sets that we use we are faced with the opposite

resolution problem — data collection being performed not as often as we would prefer. If the data are sampled at a time interval that is larger than the serial interval of the disease in question, the weekly data might include chains of more than one infection, which makes it harder to extract information about the direction of disease spread. The Chilean influenza data set is the most affected by this issue, as the serial interval of the disease is 4 days and the data is sampled weekly. Increasing the υ value reinforces the problem, so if computationally possible, it would be best to choose as low an υ as possible in order to increase temporal resolution of our experiments.

For most of the data sets presented in this thesis we are also constrained in our υ choices by computational complexity. Having a low υ increases computational complexity and makes data analysis difficult, both by generating a large number of structures to study and slowing down calculations. We thus often take a slightly larger υ of at least 4 weeks for the data sets that are collected weekly, and at least 7 days for the data sets that are collected daily. For the multislice networks the computational complexity becomes more of an issue, as the size of the adjacency matrix increases with each layer by the number of nodes present in the layer and it becomes difficult to store and manipulate the matrix on a computer. For this reason, we choose larger values of υ for the multislice networks.

We investigate the influence of having overlapping vs. non-overlapping time windows for the dengue data set in Section 6.6. For the numerical experiments on the remaining disease data sets in Chapters 7 and 8, we decide whether to use overlapping networks on a case-by-case basis depending on the comparison of the serial interval of the disease with the temporal resolution and length of the data set, as well as comparing the data set length with the number of provinces, and thus taking into account the size of the adjacency matrix and the computational complexity of the community detection. Finally, for the Ebola Datamarket data set we choose a low υ of 5 days despite the serial interval of the disease being 21 days, because only lo υ values allow us to generate a multislice network for this short data set. The parameter choices are summarized in Table 5.1.

A.4 Spatial binning for the gravity null model

In this section we investigate the influence of the spatial bin width b on the deterrence function $f(d)$ for the gravity null model (see Chapter 3.3.3). This bin size, which we measure in kilometers for the disease-correlation networks, determines the groupings of nodes for the deterrence function and for the null model. We investigate several options for each data set, and we use them to guide our final parameter choice for bin width b.

Contrary to the mobile phone call data used in the paper that first used the gravity null model for modularity maximization [82], where the deterrence function decreased with distance, we observe that in most of the disease data sets $f(d)$ tends to vary a lot and sometimes increases with distance d. This is largely due to the spread of the population

throughout the countries. See graph of $f(d)$ for 25 km in Fig. A.6(a) to observe the small number of large values that tend to correspond to the interactions between the various larger populations centers and Lima (the highly populous capital of the country). Similar growing trends are observed in all the other diseases (see Fig. A.7).

An additional constraint on the bin width arises from the desire to ensure that the weighted average that is used to calculate the null model [see Eq. (3.20) for the gravity null model and Eq. (3.24) for the radiation null model] contains enough elements for each distance to generate a meaningful null model. As the countries are often shaped in such a manner that there are a small number of far-away nods (Chile being the extreme example), we want to ensure that these nodes are not left in a bin on their own, as this would bias the null model. This places a constraint on the minimum bin size. For this thesis, we (arbitrarily) chose a minimum of 5 region pairs per bin for all diseases. This constrains the lower possible bin size considerably — in case of dengue it forced us to choose a bin size of 400 km [see Fig. A.6].

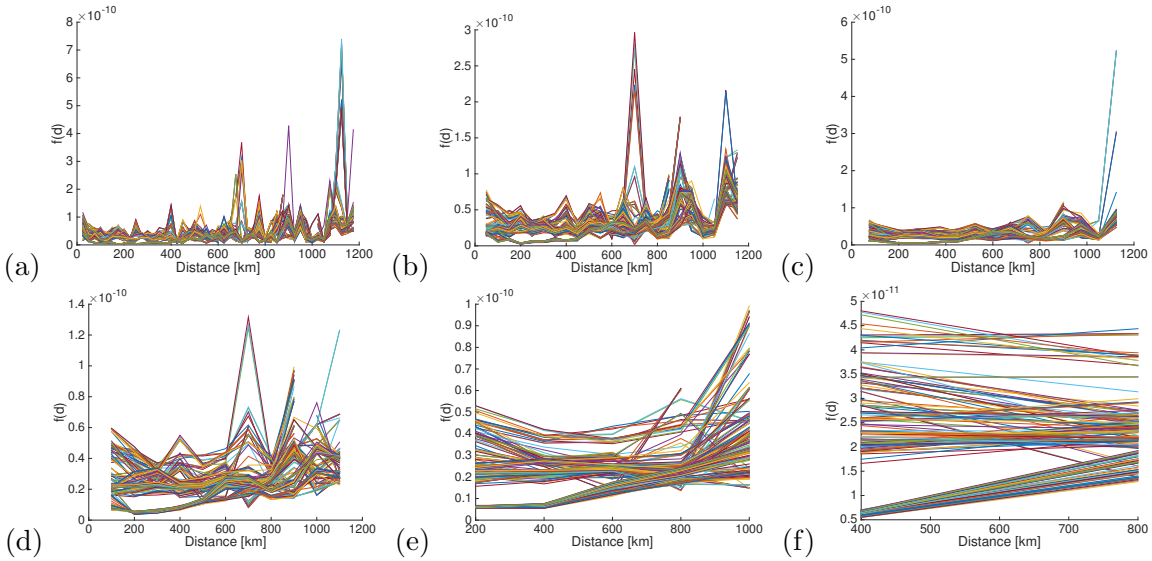


Figure A.6: Influence of bin width b on the shape of the deterrence function $f(d)$ that we use for the gravity null model and radiation null model, versus the binned distance between provinces for the dengue data set; each colorful line corresponds to the shape of the deterrence function for one static network. We use time window width of 60 weeks with bin width (a) 25 km, (b) 50 km, (c) 75 km, (d) 100 km, (e) 200km, and (f) 400km.

We choose the bin width for each data set based on the results from this comparison (see Fig. A.7), combined with the constraint that no bin contains fewer than 5 pairs of nodes to ensure that the calculations of mean correlation values that are used in the gravity null model (and the respective average fluxes for the radiation null model) are sensible. We thus use 400 km for dengue, 100 km for rubella, 600 km for H1N1 influenza in Mexico, 200 km for influenza in Chile, and 100 km for Ebola. We use the same bin sizes for the radiation null model. This limits the level of detail that we are able to see in the null model — for some networks such as dengue, there are only two bins with a much larger number

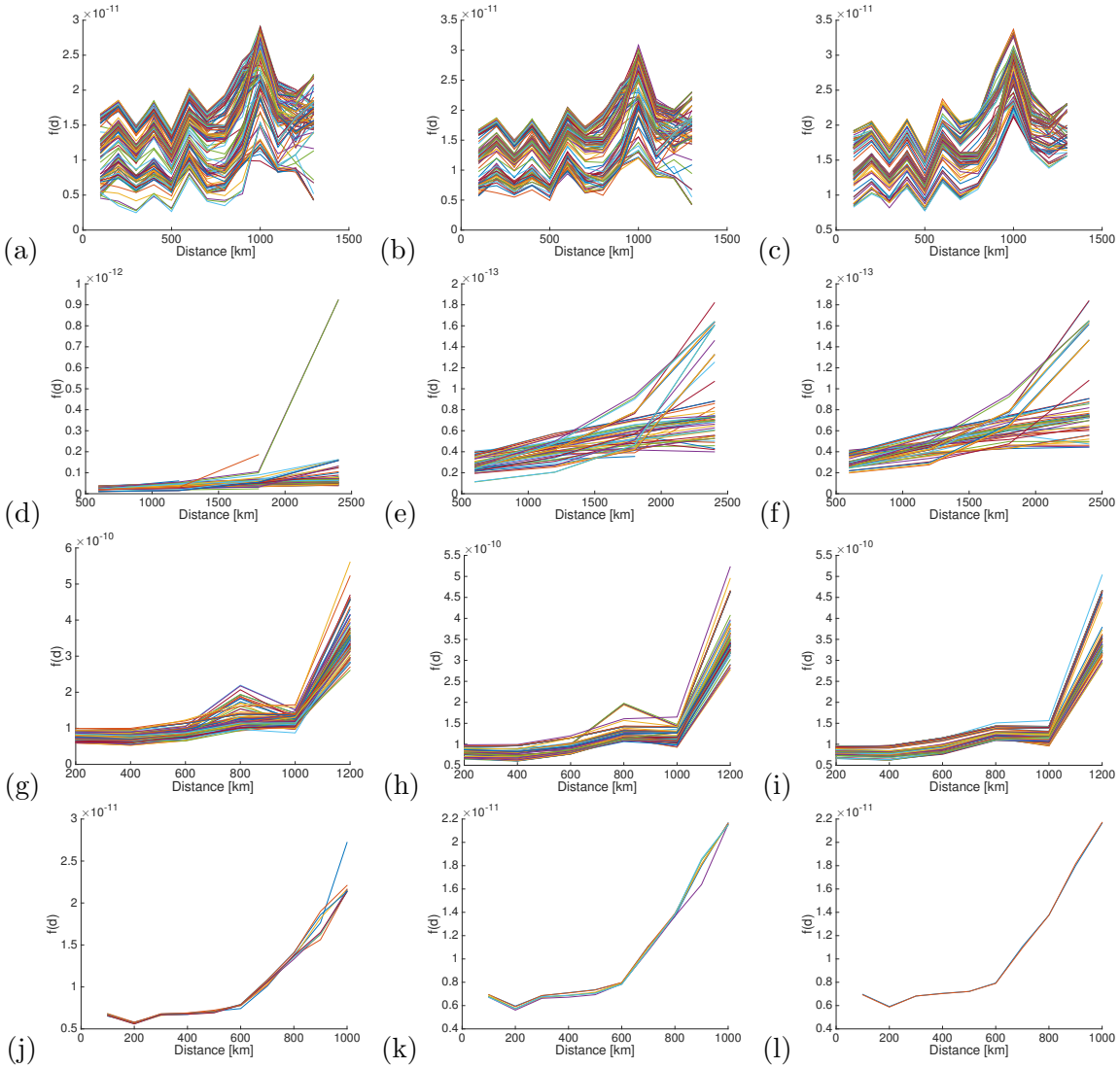


Figure A.7: Influence of bin width b on the shape of the deterrence function $f(d)$ that we use for the gravity null model and radiation null model, versus the binned distance between provinces; each colorful line corresponds to the shape of the deterrence function for one static network. (a)-(c) Rubella network with bin width (a) 50 km, (b) 100 km*, and (c) 200 km; (d)-(f) H1N1 influenza network with bin width (a) 200 km, (b) 400 km, and (c) 600 km*; (g)-(i) Chile influenza network with bin width (a) 100 km, (b) 200 km*, and (c) 300 km; (j)-(l) Ebola network with bin width (a) 50 km, (b) 100 km*, and (c) 200 km. The sign * marks values used in the thesis.

of node pairs in the lower-distance bin and we are unable to distinguish any differences in interaction patterns for provinces at distances lower than the 400 km bin size.

The choice of bin size depends on the number of nodes and the distances between nodes in a network. Because of the larger number of nodes in the rubella network than in the dengue network, we are able to use a finer level of spatial detail for the rubella data set than for the dengue data set. This allows us to include more detail in the spatial null model for rubella than for dengue. However, bin size has a smaller influence on the shape of the deterrence function for rubella and the other diseases than it does for dengue.

It appears that the spatial null model methodology is perhaps best suited to data sets with a large number of nodes, as this allows the use of small bins and detailed binning data, leading to the construction of a more detailed null model. For smaller data sets, one might have to compromise on the number of node pairs in bins or use a small number of (potentially poorly-fitting) bins. Alternatively, one could use a different binning regime, for example binning from the left with a minimum number of items per bin, which would ensure that the structures present in the correlation matrices for nodes at low distances are studied in detail instead of being aggregated into one bin.

Appendix B

Additional results for applications to endemic disease epidemics

This appendix consists of original work by MS and M. A. Porter which is not yet published.

B.1 Introduction

This chapter complements Chapter 7, in which, we apply our community-detection methodology to data sets concerning endemic diseases that are already established in their locations and cause recurrent epidemics. In this appendix, we investigate the patterns of the spatial spread of rubella in Peru and seasonal influenza in Chile, and we present the remaining plots that we did not have space to show in Chapter 7 for completeness.

We use the same approach as we described in Section 3.5 and that we used for analyzing the community structure of correlation networks created from dengue fever data sets in Chapter 6. We present the results of community detection for the NG, correlation, gravity and radiation null models; we analyze the network partitions.

We examine the spatial organization of community structures detected in the static networks, first measuring the degree of spatial organization across values of the resolution parameter $\gamma \in \{0.1, 0.2, \ldots, 3\}$ using z-Rand scores against climate partitions for rubella (defined in Section 3.4.3) and the distance test for both diseases (defined in Section 3.4.4). We then select particular parameter values and networks for which we study the partitions in more detail.

For the multislice networks, we study the spatial and temporal organization of community structures that we detect algorithmically. We use the multislice versions of the z-Rand scores versus a climate partition defined in Section 3.4.3 (for rubella) and the partition-wide distance test defined in Section 3.4.4 for detecting spatial organization (for both diseases). We search for critical time points when community structure changes using the z-Rand score methodology that we described in Section 3.4.3. We use these methods to select interesting partitions for further study from algorithmic partitions generated for parameter values $\gamma \in \{0.1, 0.2, \ldots, 3\}$ and $\omega \in \{0.1, 0.2, \ldots, 3\}$.

The majority of the results for rubella are shown in Chapter 7. This appendix contains the results that we did not have the space to show there — the community structures that do not provide us with extra information. We show the results of community detection on static networks for the gravity null model, and all results for the radiation null model.

We also show results of community detection using all null models for the data set about seasonal influenza in Chile, for which we struggled to find interesting and reliable spatial partitions. However, as mentioned in Chapter 7, we detect strong temporal partitions when using the NG and correlation null models on multislice networks, which may correspond to a year-long period with low case numbers.

B.2 Rubella

Recall that the rubella data set contains weekly new case count data from 175 of the 195 provinces of Peru over 15 years between 1997 and 2003 (i.e., 673 weeks). The data show yearly epidemics of rubella, with especially large epidemics in 2000–2001 and 2005–2006. We described the data set in detail in Section 5.2.1.2.

For the rubella data set, we use a time-window width of $\Delta = 134$, and we let the distance between adjacent time windows to be $\upsilon = 4$ to generate a set of 132 static networks; we then take $\Delta = 134$ and $\upsilon = 12$ to generate a multislice network with 44 layers. We described the justification for parameter choices and the general properties of the time series and the static networks in Section A.2.

B.2.1 Modularity maximization using the gravity null model

If we apply modularity maximization using the gravity null model to the static rubella networks, we see that partitions have lower spatial z-Rand scores than partitions found using NG and correlation null models [compare Fig. B.1(a) with Fig. 7.1(a) and Fig. 7.5(a)], suggesting that the gravity null model removes most of the spatial variability in the data. For the distance test, results also show very little spatial organization [see Fig. B.1(b)]. Some potentially spatial partitions ($p < 0.05$) appear for $\gamma = 1.3$ and $\gamma = 1.9$, for the first 51 networks and networks 81–90. However, visually all partitions for this null model appear very similar: they contain one large community and a small number of singletons (that correspond to highest-populated nodes and are often spread far apart).

We focus on $\gamma = 1.9$ for detailed examination, as partitions for most static networks have the lowest p-values in the distance test at this γ [see Fig. B.2(a)]. For this partition, the singleton communities are spread throughout the country; they contain some of the most populated provinces of Peru. For many of the yearly epidemics, they have more disease cases than the nodes in the large community.

We now consider multilayer networks and examine the spatial organization of community structures that we detect using modularity maximization with the gravity null model

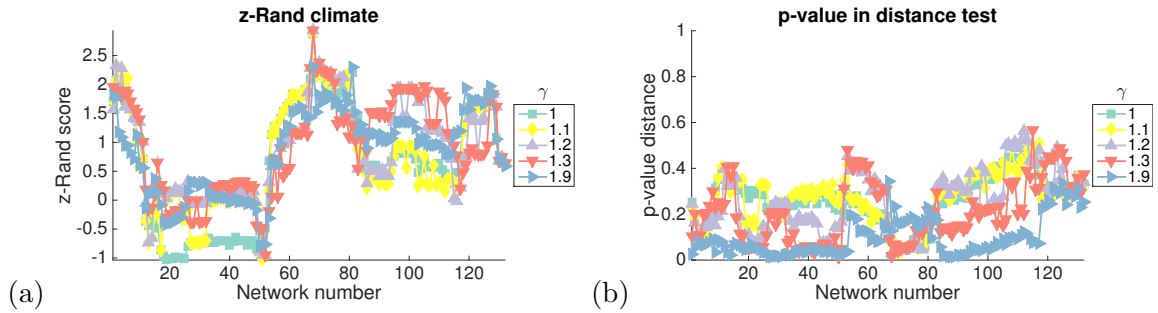


Figure B.1: Rubella, static networks, gravity null model — spatial partitions. Properties of the algorithmic community structure for the 132 static networks covering the whole time period (horizontal axis): (a) Plot of the z-Rand scores versus the detailed climate partition and (b) p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3, 1.9\}$.

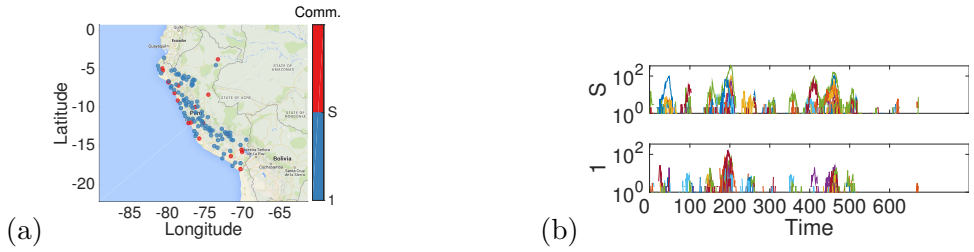


Figure B.2: Rubella, static networks, gravity null model — example partitions. In panel (a) we show a map of all the nodes in network 89 at $\gamma = 1.9$ (colored by algorithmically detected community assignment with singletons grouped into group S, community assignment indicated on the color bar), and in panel (b) we show the time series of disease occurrence in the provinces assigned to these communities, with community number (or S for singletons) indicated on the vertical axis.

for various values of the parameters γ and ω. We observe that all partitions for $\gamma \gtrapprox 0.4$ are statistically significant using the spatial z-Rand test [see Fig. B.3(a)], and they score as significant in the partition-wide distance test [see Fig. B.3(b)]. However, when we study these structures in detail in Figs. 7.8(a)-(b), we observe that once again they consist of one large community that contains the majority of nodes and a small number of singleton communities that connect provinces to their counterparts across layers. These kinds of partitions are not very informative. This highlights the difference between detecting partitions that are statistically significant for spatial organizations and detecting partitions that are informative for examination of infectious disease.

The scores for temporal partitions in the multislice rubella networks are presented with the main results in Chapter 7, as they contain interesting temporal partitions.

B.2.2 Modularity maximization using the radiation null model

We maximize modularity using the radiation null model on the static networks created from the rubella data set. Similarly to the gravity null model, most of the static partitions score below the significance threshold of 1.96 for spatial z-Rand scores, suggesting that they contain little spatial organization (i.e., the it appears that the radiation null model removes most of the spatial variability in the data).

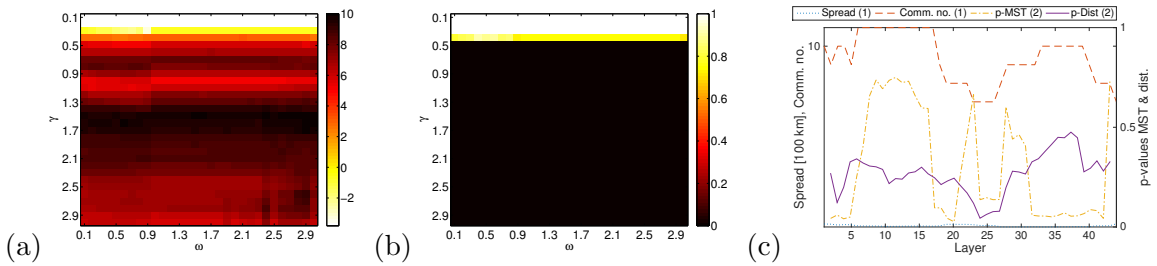


Figure B.3: Rubella, multislice networks, gravity null model — spatial organization of partitions. In parts (a)–(b) we show results of varying the parameters γ and ω: (a) z-Rand scores for similarity to "spatial" partitions by climate, (b) the p-values for distances being smaller than expected at random in the distance test. In (c) we plot values for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

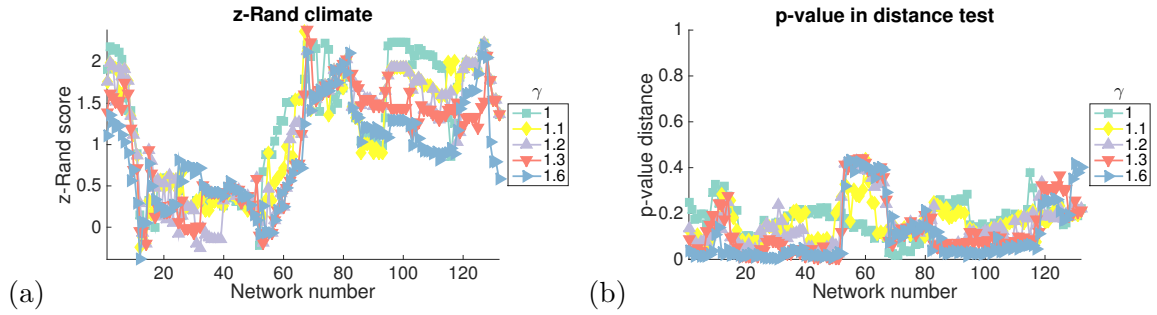


Figure B.4: Rubella, static networks, radiation null model — spatial partitions. Properties of the algorithmic community structure for the 132 static networks covering the whole time period (horizontal axis): (a) Plot of the z-Rand scores versus the detailed climate partition and (b) p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3, 1.6\}$.

However, the distance test scores are statistically significant for networks that contain both of the large rubella epidemics for some parameter values, e.g. $\gamma = 1.6$, as we illustrate in Fig. B.4(b). We focus on $\gamma = 1.6$ for a detailed examination, as partitions for this resolution parameter are most commonly classified as spatial across time. Similarly to our results for the gravity null model, the structures tend to consist of one large community and the remaining nodes assigned to singleton communities that are spread throughout the country and contain some of the highest-populated provinces of Peru (see Fig. B.5).

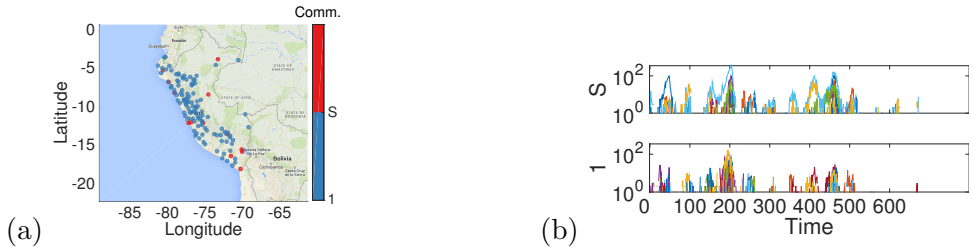


Figure B.5: Rubella, static networks, radiation null model — example partitions. In panel (a) we show a map of all the nodes in network 31 at $\gamma = 1.6$ (colored by algorithmically detected community assignment with singletons grouped into group S, community assignment indicated on the color bar), and in panel (b) we show the time series of disease occurrence in the provinces assigned to these communities, with community number (or S for singletons) indicated on the vertical axis.

The results for modularity maximization on multislice rubella networks using the radiation null model are also similar to the results of modularity maximization using the gravity null model (see Fig. B.6). Structures for $\gamma \gtrapprox 0.1$ score as spatial using the z-Rand scores and the partition-wide distance tests. Once again, the multislice network partitions are composed of one large community and several temporal singleton partitions that connect highly populated nodes with their copies across layers [see example in Fig. B.6(c)] and thus do not provide us with additional spatial information.

The z-Rand scores for temporal organization of the partitions of the multislice networks for different γ and ω parameter values yield very high temporal z-Rand scores. Similarly to the results for the gravity null model, the method detects the beginning and end of the period with few disease cases and few nodes (approximately layers 18 and 28), and the large drop in the number of cases after layer 40.

B.2.3 Summary of additional findings for the rubella data set

The results for using community detection by modularity maximization with the gravity and radiation null models on static rubella networks suggest that the null models succeed (at least partially) in removing spatial organization from the correlation networks. This is shown by the low spatial organization in the partitions, as measured by the z-Rand scores against climate and administrative partitions, and the distance test.

The multislice partitions contain multiple possible critical time points with high z-Rand scores at which the structure of the network significantly changed, corresponding to large

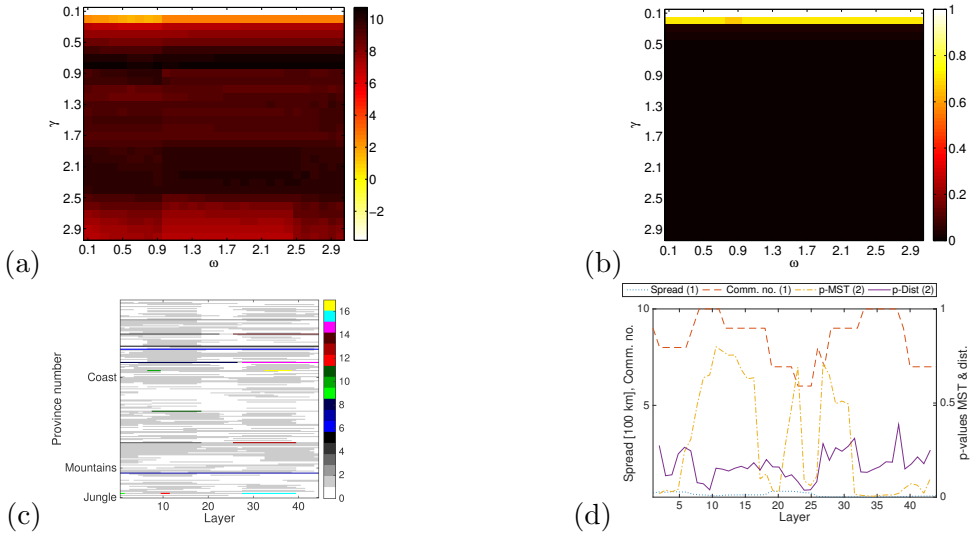


Figure B.6: Rubella, multislice networks, radiation null model: spatial organization of partitions. In parts (a)–(b) we show results of varying the parameters γ and ω: (a) z-Rand scores for similarity to "spatial" partitions by climate, (b) the p-values for distances being smaller than expected at random in the distance test. In (c) we plot community structure for $\gamma = 1$, $\omega = 0.1$; nodes are on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (d) we plot values for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

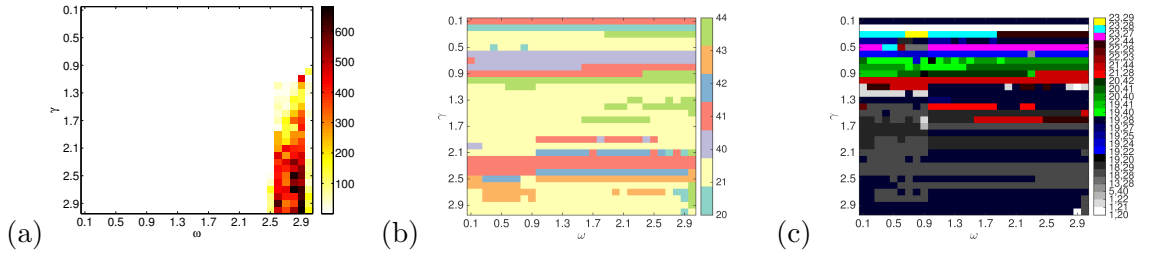


Figure B.7: Rubella, multislice networks, radiation null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω for: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

changes in network structure such as a drop in the number of nodes with non-zero strength that corresponds to a period with low numbers of new disease cases.

B.3 Seasonal influenza in Chile

B.3.1 Introduction

The Chilean influenza data set contains weekly counts of new disease cases from the 15 provinces of Chile over 7 years (365 weeks) between 1 January 2004 and 31 December 2010. Chile covers a long, narrow strip in the southwestern part of South America, and the 15 provinces are organized from north to south. Our data set shows yearly countrywide epidemics of seasonal influenza and the 2009 "swine flu" epidemic, which increased the amount of influenza activity in Chile. We describe this data set in detail in Section 5.2.1.3.

For this data set, we use time-window width of $\Delta = 30$ and we let the difference between time window starting points $\upsilon = 4$ to generate a set of 82 static networks. We also use $\Delta = 30$ and $\upsilon = 30$ to generate a multislice network with 11 layers. We describe the justification for parameter choices and the general properties of the time series and the static networks in Section A.2.

Modularity maximization using the Newman-Girvan (NG) and correlation null models allows us to detect partitions that appear somewhat spatial in both static and multislice networks. However, the results of community detection appear to be very parameter-dependent, and change quickly in time. The results for gravity and radiation null models once again consist of one large community and a small number of singletons (containing the most-populated nodes). However, we detect strong temporal partitions using the NG and correlation null models, which may correspond to a year-long period with low case numbers.

B.3.2 Modularity maximization using the NG null model

For the static networks, the extent of spatial organization that we observe in communities detected by maximizing modularity using the NG null model appears to vary a lot across time. When we use the distance test to detect spatial organization of partitions, we sometimes observe rapid changes from a significantly spatial partition ($p < 0.05$) to a partition more dispersed than expected at random ($p = 1$) in the next (partially-overlapping) network. The community structures also appear to be very sensitive to the values of resolution parameter γ [see Fig. B.8(a)].

We examine partitions for γ between 1 and 1.3 for layers 40–45, and 66–71 which might have significant levels of spatial organization. We plot a representative partition in layer 67 (April 2009) for $\gamma = 1.1$. It contains a north-south division of the country into roughly equal-sized communities [see Fig. B.8(b)].

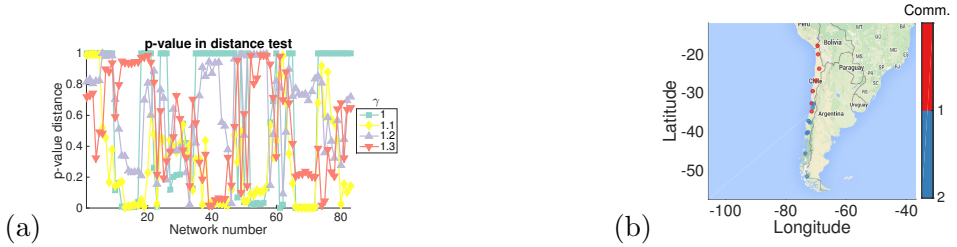


Figure B.8: Influenza in Chile, static networks, NG null model: spatial partitions. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3\}$.In panel (b) we show a map of all the nodes in network 67 at $\gamma = 1.1$ (colored by algorithmically detected community assignment, community assignment indicated on the color bar).

We then study the spatial organization of community structures of the multislice networks found using the NG null model for different values of parameters γ and ω. We observe a similar pattern as for rubella and dengue: there is a band for $1 \lessapprox \gamma \lessapprox 1.5$ that includes all of the spatial partitions — although the total number of parameter regimes with spatial partitions is smaller than for either of the other two diseases. We examine the structures that score as spatial in the partition-wide distance test, and we show partitions for $\gamma = 1.1$ and $\omega \in \{0.1, 1.3\}$ in Figs. B.9 (b)-(c). They both consist of 4 communities and score $p < 0.05$ in the partition-wide distance test. However, the temporal patterns of their per-layer distance and MST test scores are different [see Figs. B.9 (d)-(e)]. The community assignments of most physical nodes in the $\omega = 0.1$ partition change several times. The whole network partition scores as spatial overall due to layers 3, 4, and 9, which have very low p-values in the per-layer distance tests. They have very low total intra-community distance each, and thus they bring the overall total intra-community distance down for the multislice network enough for it to appear significant. The $\omega = 1.3$ partition does not change much in time, and it is significant in per-layer distance tests across layers 1–8, perhaps due to spatially clustered nature of community 4 [see Fig. B.9 (c)].

When we calculate the z-Rand scores of partitions across the γ and ω parameter range against temporal partitions, we find that $\gamma = 1.1$ and $\omega = 0.1$ scores the highest, and several parameter values in the neighboring region of parameter space also score relatively high (see Fig. B.10). For these parameter values, layer 6 (June 2007) is the critical time point. As briefly discussed in Chapter 7, this time point corresponds to a year-long period with a low number of disease cases, which is a similar result to the temporal partition that we detected in the rubella multislice networks.

B.3.3 Modularity maximization using the correlation null model

When we maximize modularity using the correlation null model on the static chilean influenza networks, we find little dependence of the level of spatial organization in the partitions as detected by the distance test on the resolution parameter γ. The partitions once again change from strongly spatially clustered to strongly spatially dispersed in as little as

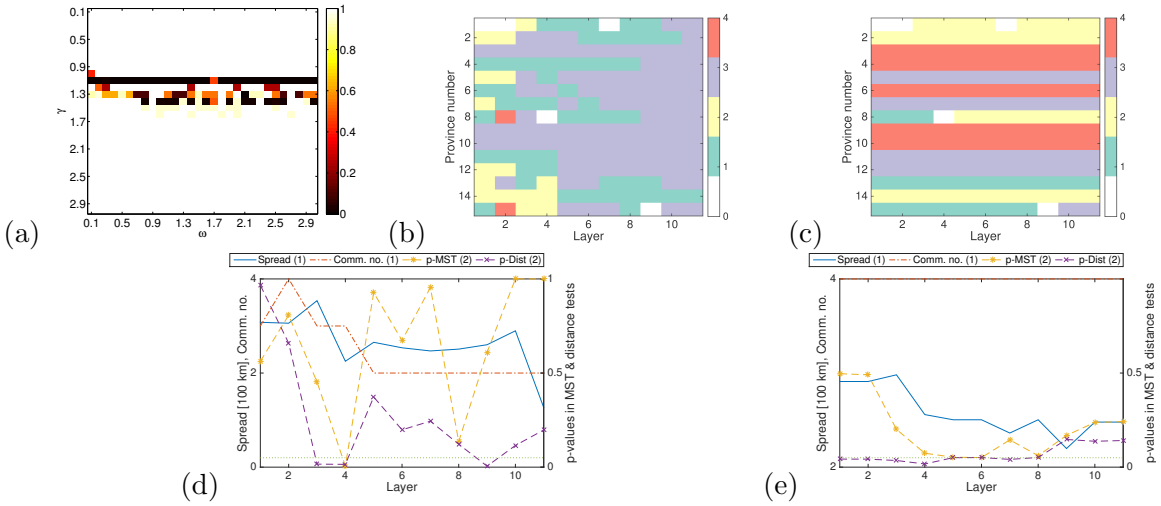


Figure B.9: Influenza in Chile, multislice networks, NG null model: spatial organization of partitions. In part (a) we show effect of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b)-(c) we plot multislice community structure with nodes ordered by their location (north to south) on the vertical axis and layers on the horizontal axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for (b) $\gamma = 1.1$, $\omega = 0.1$ and (c) $\gamma = 1.1$, $\omega = 1.3$. Community number is indicated on the colorbar. In (d)-(e) we plot values for each layer, (left vertical axis) the community spread and the number of communities and (right vertical axis) the p-values for the distance and MST tests, for (d) $\gamma = 1.1$, $\omega = 0.1$, and (e) $\gamma = 1.1$, $\omega = 1.3$.

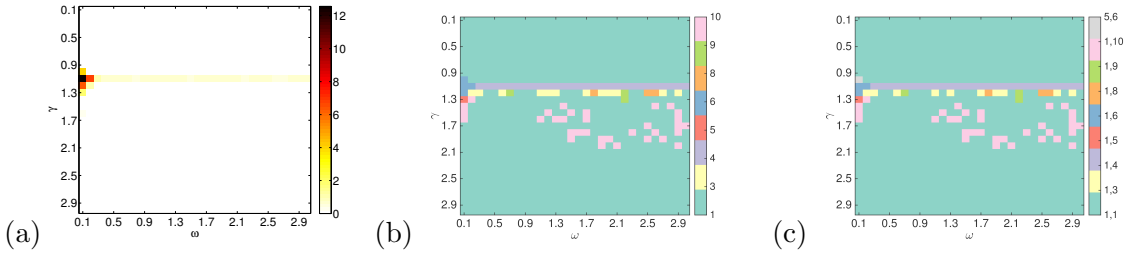


Figure B.10: Influenza in Chile, multislice networks, NG null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω on: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

one time step [see Fig. B.11(a)]. Networks 13–20 and 50–58 score as strongly spatial in the distance test. We select layer 16 for further study; at $\gamma = 3$, the provinces are divided into two communities, with the northernmost two nodes forming their own community [see Fig. B.11(b)].

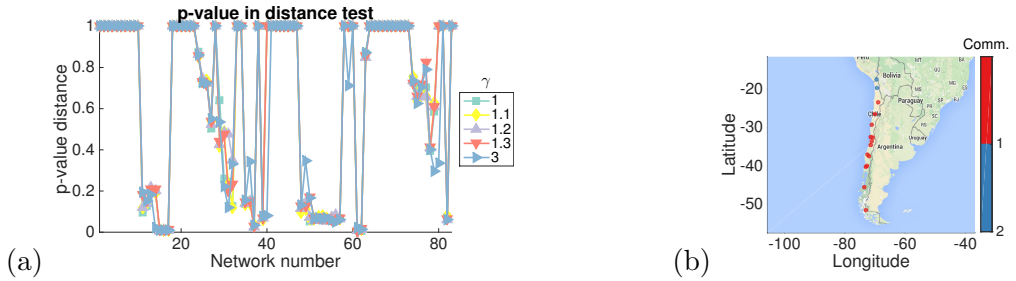


Figure B.11: Influenza in Chile, static networks, correlation null model: spatial partitions according to distance test. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3, 3\}$. In panel (b) we show a map of all the nodes in network 16 at $\gamma = 3$ (colored by algorithmically detected community assignment, community assignment indicated on the color bar).

When we study the effects of varying the parameters γ and ω on the spatial organization of community structures detected by maximizing multislice modularity with the correlation null model, we see that strongly spatial partitions occur for all γ, for $\omega \lessapprox 0.4$ (with more ω values yielding spatial partitions for higher γ). The partitions are dominated by one large community grouping the majority of the multislice nodes; they also contain a small number of other communities that are present in 1–2 layers only [see a representative partition for $\gamma = 1$, $\omega = 0.1$ in Fig. B.12(b)]. The layers with small communities formed by nodes on the northern or southern ends of the country are the ones that score as significantly spatial in the per-layer distance test [see Fig. B.12(c)]. However, as these groupings do not persist or reappear in time, we gain little knowledge about the disease patterns from them. Other strongly spatial partitions additionally possess temporal features, as we show in Fig. B.12(d) for $\gamma = 0.9$, $\omega = 0.3$.

Across a variation of ω and γ parameters, $(\gamma, \omega) \in \{(0.7, 0.3), (0.8, 0.3), (0.9, 0.3), (0.8, 0.4)\}$ appear to have strong temporal structure, as scored using the z-Rand score test against temporal partitions [Fig. B.13(a)]. Once again, as with the NG null model, layer 6 (March 2007) appears to be a highly scoring critical time point [Figs. B.13(b)-(c)]. The partitions closely resemble the idealized temporal partitions, with a large community before and after the change, and two communities of comparable size in layer 6 [see partition for $\gamma = 0.9$, $\omega = 0.3$ in Fig. B.12(d)]. As briefly discussed in Chapter 7, this time point corresponds to a year-long period with a low number of disease cases, which is a similar result to the temporal partition that we detected in the rubella multislice networks.

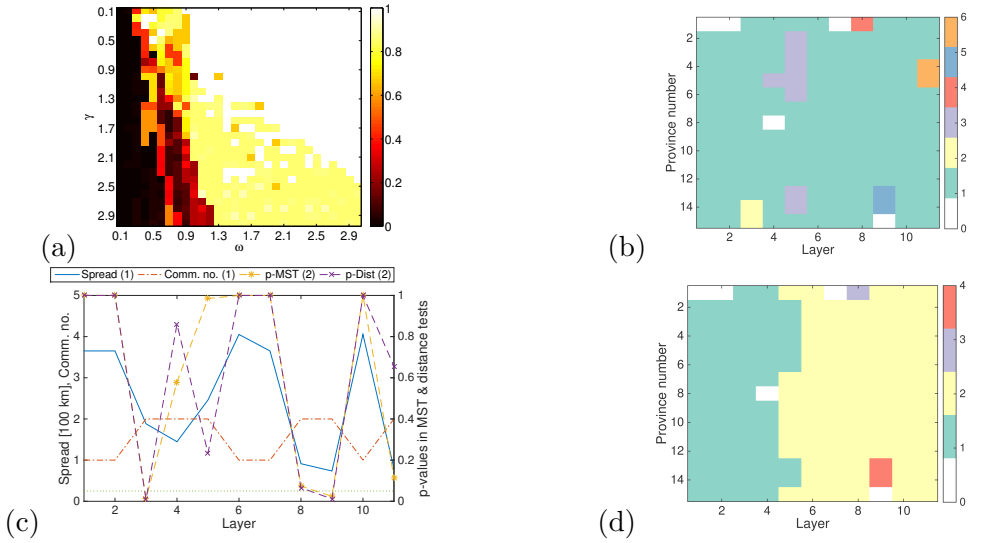


Figure B.12: Influenza in Chile, multislice networks, correlation null model: spatial organization of partitions. In parts (a) we show effects of varying the parameters γ and ω: (a) the p-values for distances being smaller than expected at random in the distance test. In (b) and (d) we plot community structure with nodes ordered by their location (north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for (b) $\gamma = 1$, $\omega = 0.1$ and (d) $\gamma = 0.9$, $\omega = 0.3$. Community number is indicated on the colorbar. In (c) we plot values for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

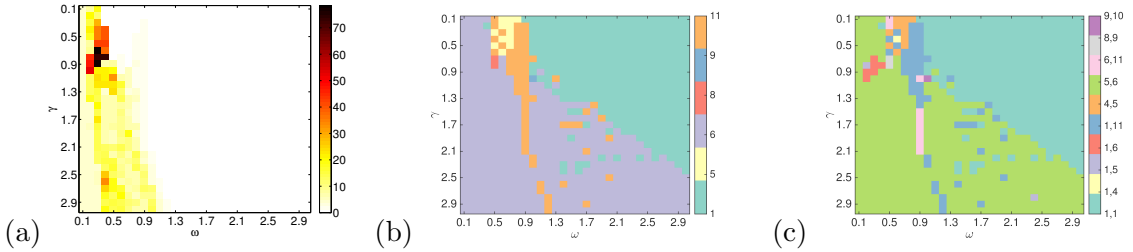


Figure B.13: Influenza in Chile, multislice networks, correlation null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω on: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

B.3.4 Modularity maximization using spatial null models on Chilean influenza data

In this section we present the results of modularity maximization using the spatial null models on the Chilean influenza data. These structures do not show strong spatial features when scored using the distance test. The structures for the gravity null model contain one large community and a small number of singletons. The pattern is similar for partitions found using the radiation null model [Fig. B.14(c)-(e)], including the partitions nominally scoring as spatial in the distance test, such as network 15 at $\gamma = 0.8$, which consists of the country capital as a singleton, the southernmost and northernmost nodes in a two-node community, and all other nodes in a large community.

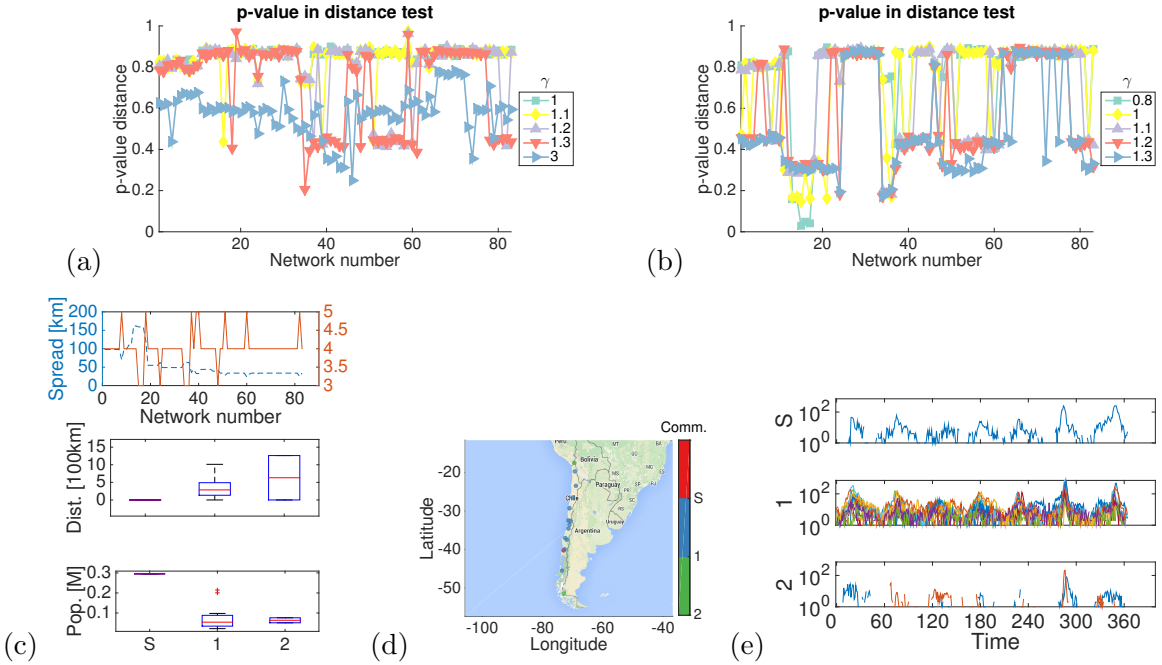


Figure B.14: Influenza in Chile, static networks, gravity and radiation null models — spatial partitions according to the distance test. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3, 3\}$ and the gravity null model, and in panel (b), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{0.8, 1, 1.1, 1.2, 1.3, 3\}$ and the radiation null model. In panels (c)–(e) we focus on partitions using the radiation null model. In panel (c), we plot (top): community spread and number of communities for community structures of all static networks at $\gamma = 0.8$, (middle): box plot of intra-community distance for community structure of network 15 (corresponding to January 2002) at $\gamma = 0.8$, (bottom): box plot of populations of communities in the same partition. In panel (d) we show a map of all the nodes in network 15 at $\gamma = 0.8$ (colored by algorithmically detected community assignment with singletons grouped into group S, community assignment indicated on the color bar), and in panel (e) we show the time series of disease occurrence in the provinces assigned to these communities, with community number (or S for singletons) indicated on the vertical axis.

By maximizing multislice modularity using the gravity null model for many values of ω and γ parameters, we observe that none of the structures score as significantly spatial in the partition-wide distance test except a small parameter range near $\gamma = 1.6$, $\omega = 3$. For all values including this one, we obtain one large community and a few singleton communities

that persist through time (see Fig. B.15). We study the per-layer scores in Fig. B.15(c) and we observe that none of the layers scores significantly in the per-layer distance or MST tests. This is a similar result to what we observed in the dengue and rubella data sets for the gravity null model.

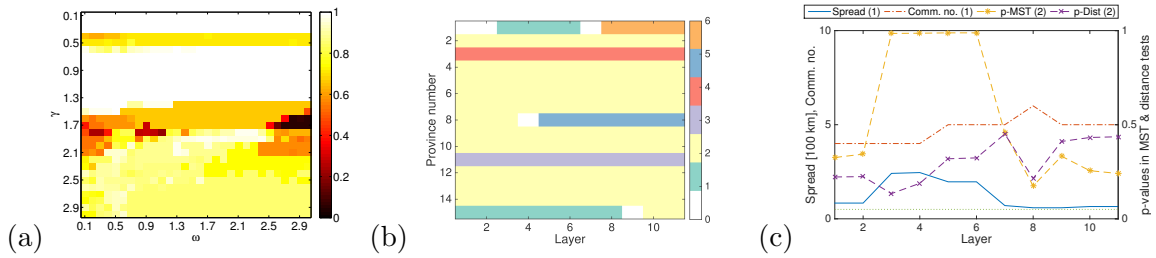


Figure B.15: Influenza in Chile, multislice networks, gravity null model: spatial organization of partitions. In panel (a) we show influence of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b) we plot community structure for $\gamma = 1.6$, $\omega = 3$ with nodes ordered by their location (north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (c) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1.6$, $\omega = 3$ and (right vertical axis) the p-values for the distance and MST tests.

For all parameter values, partitions using the gravity null model score low against temporal partitions [see Fig. B.16(a)]. The community structures for the gravity null model suggest a large range of potential critical time points, but their significance is disputable as the corresponding z-Rand scores are low [see Fig. B.16(b)-(c)].

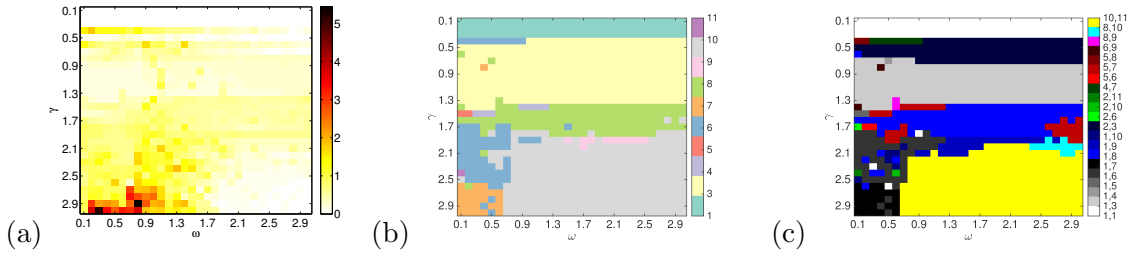


Figure B.16: Influenza in Chile, multislice networks, gravity null model — temporal organization of partitions. In (a)–(c) we show results of varying the parameters γ and ω for: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

The results of maximizing multislice modularity using the radiation null model are similar to what we have observed using the gravity null model. We obtain one large community and a small number of temporal singleton communities that persist through time [see Fig. B.17(b)]. None of the structures score as statistically significant in the per-layer distance or MST tests [see Fig. B.17(c)].

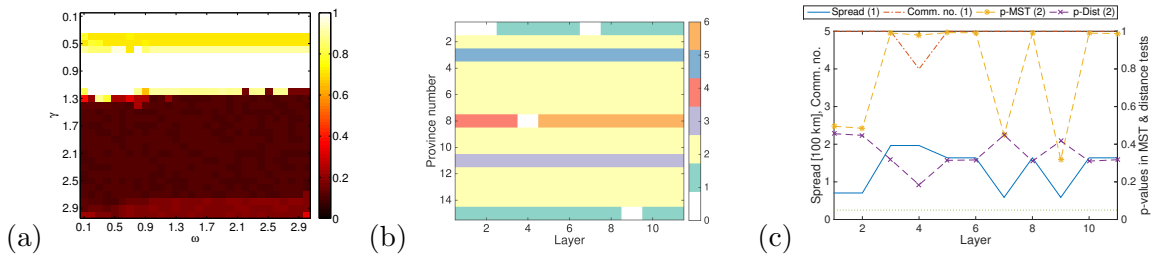


Figure B.17: Influenza in Chile, multislice networks, radiation null model: spatial organization of partitions. In panel (a) we show influence of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b) we plot community structure for $\gamma = 2.5$, $\omega = 0.1$ with nodes ordered by their location (north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (c) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 2.5$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

Similarly, very few partitions found using the radiation null model score as statistically significant in z-Rand scores for temporal partitions [see Fig. B.18(a)].

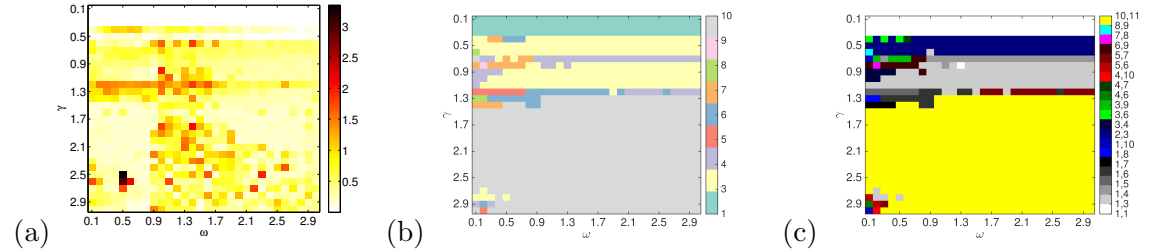


Figure B.18: Influenza in Chile, multislice networks, radiation null model — temporal organization of partitions. In (a)–(c) we show results of varying the parameters γ and ω for: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

B.3.5 Summary of findings for the Chilean influenza data set

The Chilean influenza data set exhibits less spatial organization of communities for the NG and correlation null models than the rubella and dengue data sets. The spatial organization scores of partitions are highly parameter-dependent and they vary across time.

Both NG and correlation null models often detect north–south partitions of the network into two spatially contiguous communities. The partitions generated using the correlation null model appear to be more stable across different values of the resolution parameter γ than the NG partitions. These results might give us an indication of a degree of difference in the spatial patterns of disease between the northern and southern provinces, but they do not provide informative groupings.

For the multislice networks, we detect some strong temporal partitions that appear to correspond to a period of low numbers of disease cases. This is a similar result to the temporal partitions of the rubella data set, and it is made more interesting by the fact that the strength of structural change in the networks is lower for the chilean data set (in contras to rubella, it does not involve an increase nodes with non-zero strength in the corresponding layers).

Appendix C

Additional results for applications to emerging disease epidemics

This chapter consists of original work by MS and M. A. Porter which is not yet published.

C.1 Introduction

This appendix complements Chapter 8, in which we applied our community-detection methodology to data sets concerning emerging diseases, i.e., diseases that are newly discovered or are introduced into a new population. In this appendix, we investigate the patterns of the spatial spread of Ebola in West Africa and H1N1 influenza in Mexico, and we present the plots that we did not have the space to show in Chapter 8 for completeness.

We use the same approach as we described in Section 3.5 and that we used for analyzing the community structure of correlation networks created from dengue fever data sets in Chapter 6 and to endemic diseases in Chapter 7 and Appendix B. We present the results of community detection for the NG, correlation, gravity and radiation null models and we analyze the network partitions.

We examine the spatial organization of community structures detected in the static networks using the four null models for values of the resolution parameter $\gamma \in \{0.1, 0.2, \ldots, 3\}$ using the distance test (defined in Section 3.4.4). We then select particular parameter values and networks to study in more detail.

For multislice networks, we study spatial and temporal organization of algorithmically-detected community structures. We use the multislice partition-wide version of the distance test defined in Section 3.4.4 to detect spatial organization. We search for critical time points at which community structure changes using the z-Rand score methodology that we described in Section 3.4.3. We select interesting parameter values across various values of the resolution parameter $\gamma \in \{0.1, 0.2, \ldots, 3\}$ and the inter-layer coupling $\omega \in \{0.1, 0.2, \ldots, 3\}$, to study their community structures in detail.

The majority of the results for the Ebola Datamarket data set are shown in Chapter 8. This appendix contains the results that we did not have the space to show for the spatial

null models, for which the algorithm finds community structures that do not provide us with extra information about the patterns of disease infections.

We also show results of community detection using all null models for the data set about H1N1 influenza in Mexico, in which we did not find strong and reliable spatial partitions. However, when using the NG and correlation null models on the multislice networks, we find temporal partitions that might correspond to the peak and the end of the epidemic wave, which we briefly showed in Chapter 8 and we describe in more detail here.

C.2 Ebola — Datamarket data set

C.2.1 Modularity maximization using the gravity null model

When maximizing modularity for static networks using the gravity null model, none of the networks score as significantly spatial ($p < 0.05$) in the distance test. For most networks, several provinces are placed in singleton communities (see example for network 6 at $\gamma = 2.4$ in Fig. C.1). These provinces tend to have larger populations than the nodes that are placed in the one large community. As γ increases, so does the number of singletons. These structures tell us little about the patterns of disease spread in time and space.

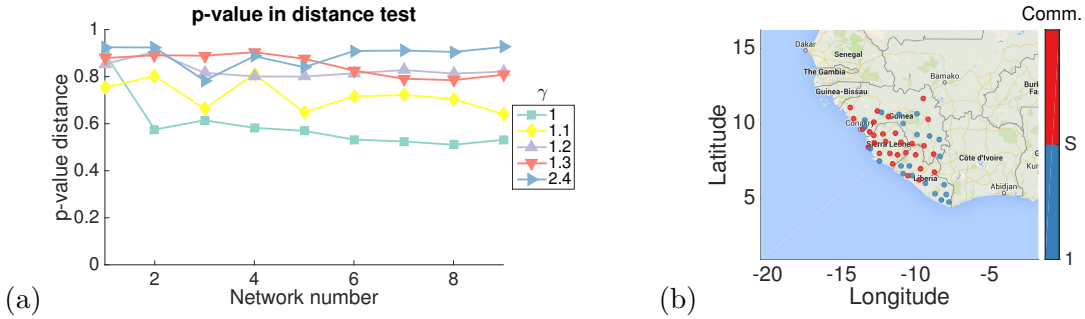


Figure C.1: Ebola Datamarket data set, static networks, gravity null model — spatial partitions. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3, 2.4\}$, and in panel (b) we show a map of all the nodes in network 6 at $\gamma = 2.4$ (colored by algorithmically detected community assignment with singletons grouped into group S, community assignment indicated on the color bar).

The typical multislice community structure that we find using the gravity null model contains one large community and a small number of temporal singleton communities [see example for $\gamma = 0.4, \omega = 0.1$ in Fig. C.2(b)]. This structure has one of the lower p-values in the partition-wide distance test for this null model. However, it is not significantly spatial in the partition-wide or layer-wise distance tests. This result is similar to the results that we found when using the gravity null model for all other diseases.

None of the partitions of the multislice network found using the gravity null model score significantly high ($z_R > 1.96$) in comparisons versus manual temporal partitions. This seems in accordance with the low number of changes in the community structures in time.

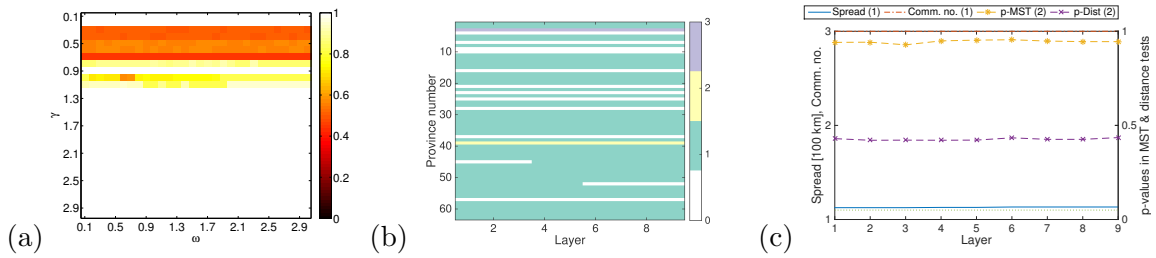


Figure C.2: Ebola Datamarket data set, multislice networks, gravity null model — spatial organization of partitions. In panel (a) we show the influence of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b) we plot community structure with nodes ordered by their location (country-wide then north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for $\gamma = 0.4$, $\omega = 0.1$. Community number is indicated on the colorbar. In (c) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 0.4$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

C.2.2 Modularity maximization using the radiation null model

When we maximize modularity for static Ebola Datamarket networks using the radiation null model, we find that similarly to what we observed when using the gravity null model, highly-populated provinces are placed in singleton communities – see Fig. C.3. However, the number of singletons is lower than for the gravity null model, and some of the network partitions score as significantly spatial in the distance test. However, we failed to find patterns to their composition beyond the aforementioned division by population.

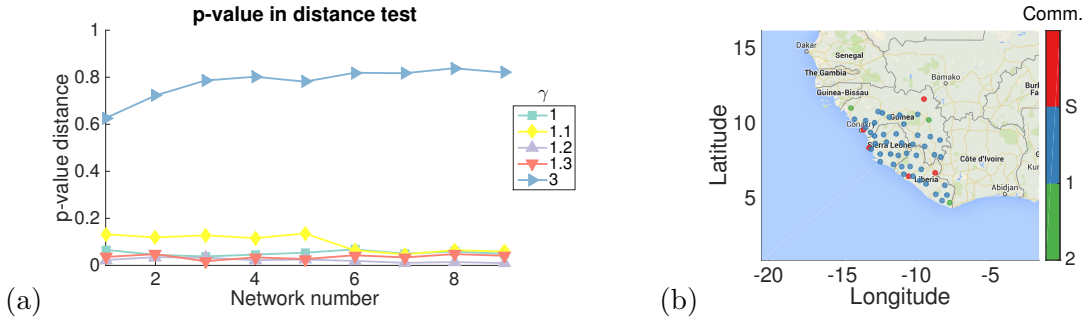


Figure C.3: Ebola Datamarket data set, static networks, radiation null model — spatial partitions. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3, 3\}$, and in panel (b) we show a map of all the nodes in network 7 at $\gamma = 1.2$ (colored by algorithmically detected community assignment with singletons grouped into group S, community assignment indicated on the color bar).

Despite the fact that some of the structures appear to contain significant spatial organization when studied using the partition-wide distance test, the multislice community structures that we detect using the radiation null model visually resemble the structures found using the gravity null model: they contain one large community that persists with a very small number of changes across time, and a small number of temporal singletons [see example for $\gamma = 1.2$, $\omega = 1.4$ in Fig. C.4 (b)].

None of the community structures that we detect in the multislice networks using various ω and γ parameter values with the radiation null model score significantly high ($z_R > 1.96$)

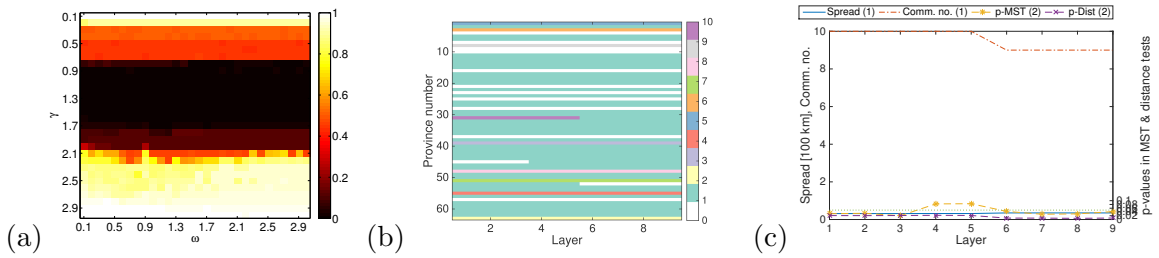


Figure C.4: Ebola Datamarket data set, multislice networks, radiation null model — spatial organization of partitions. In panel (a) we show the influence of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b) we plot community structure with nodes ordered by their location (country-wide then north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for $\gamma = 1.2$, $\omega = 1.4$. Community number is indicated on the colorbar. In (c) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1.2$, $\omega = 1.4$ and (right vertical axis) the p-values for the distance and MST tests.

in comparisons versus manual temporal partitions.

C.2.3 Summary of additional findings for the Ebola Datamarket data set

The gravity and radiation null models fail to give useful insights into the factors that affect the spread of disease. The two null models appear to succeed in removing the majority of spatial structure that is present in the disease-correlation networks. The remaining partitions appear only to be influenced by province population. The multislice partitions do not vary significantly in time for any of the parameter regimes, and thus they do not provide additional temporal partitions for the Ebola Datamarket data set.

C.3 H1N1 influenza in Mexico

The H1N1 data set contains daily new case count data from the 32 provinces of Mexico over 430 days between April 2009 and June 2010. The data set contains the three waves of the initial swine flu epidemic in Mexico. We described it in detail in Section 5.2.2.2.

The H1N1 data set contains a large number of time points compared to the number of provinces in Mexico, allowing us to generate a long multislice network and to study the temporal spread patterns in detail. We use a time-window width $\Delta = 30$ and a difference between the time-window starting points $\upsilon = 7$ to generate a set of 58 static networks. We also use $\Delta = 30$ and $\upsilon = 30$ to generate a multislice network with 14 layers. We described the justification for our parameter choices and the general properties of the time series and the static networks in Section A.2.

Modularity maximization on the H1N1 data set does not appear to find spatial partitions reliably. The Newman-Girvan (NG) null model appears to detect the strongest temporal partitions, including the peak of the main H1N1 epidemic wave and the time when the wave subsided at the end of the epidemic. The gravity and radiation null models do not appear to give us additional spatial or temporal information.

C.3.1 Modularity maximization using the NG null model

When we maximize modularity using the Newman-Girvan null model on the set of static H1N1 networks, the spatial organization scores in the distance test appear to be very sensitive to changes in resolution parameter γ, and to strongly vary across time, including large changes in scores from one network to the next (partially-overlapping) network [see Fig. C.5(a)]. The partitions for $\gamma \in \{1, 1.2, 1.3, 1.4\}$ appear to contain the largest amount of spatial organization, but even $\gamma = 1.3$ [the parameter for which the largest fraction of static networks score as significantly spatial ($p < 0.05$)] generates spatial partitions for only less than 10% of the networks. Further, we do not see clear patterns regarding which γ values give spatial communities most often, or during which time periods spatial communities exist. When we focus on a partition for the standard value of $\gamma = 1$ for network 43 (one of the significantly spatial partitions) in more detail, we notice a community composed mainly of central nodes [red in Fig. C.5(b)] and another community with eastern and northern nodes [blue in Fig. C.5(b)].

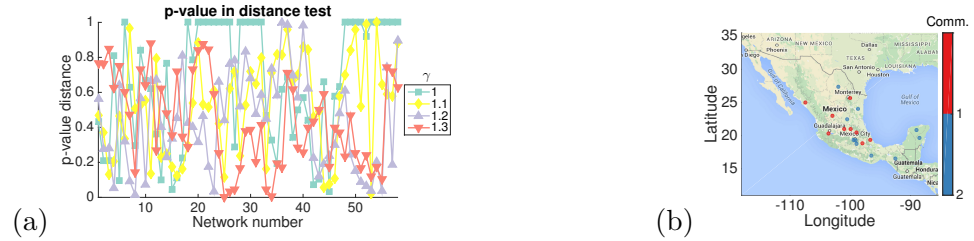


Figure C.5: H1N1 influenza, static networks, NG null model — spatial partitions. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3\}$. In panel (b) we show a map of all the nodes in network 43 at $\gamma = 1$ (colored by algorithmically detected community assignment, which is indicated on the color bar).

When we examine the spatial organization of partitions of the multislice network found using using the NG null model for a variety of values of parameters γ and ω, the majority of partitions do not appear to contain significant spatial organization ($p > 0.05$ in the partition-wide distance test) [see Fig. C.6(a)]. The partition at $\gamma = 1$, $\omega = 0.1$ is one of the few that appear significant in the partition-wide distance test. We show it in Fig. C.6(b). The spatial structure is only statistically significantly spatial in the per-layer distance and MST tests for layer 11, which contains 2 communities roughly divided into northern and southern nodes. This drives the apparent significance in the partition-wide tests.

When we study the temporal z-Rand scores, partitions for $\gamma \in \{1, 1.1\}$ and $\omega \in \{0.1, 0.2\}$ achieve the highest z-Rand scores. These partitions represent a partition at layers 6 and 9 [see Fig. C.7(c)]. These time points correspond to the birth of communities 3 and 4, and might be related respectively to the peak of the main epidemic wave and the end of it.

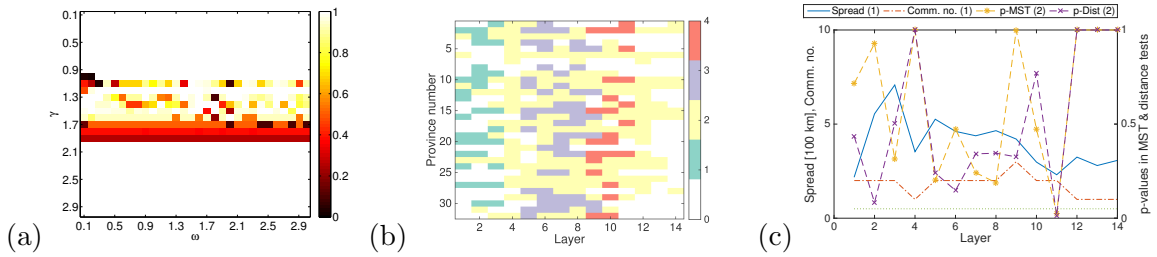


Figure C.6: H1N1 influenza, multislice networks, NG null model — spatial organization of partitions. In panel (a) we show influence of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b) we plot community structure for $\gamma = 1$, $\omega = 0.1$ with nodes ordered by their location (north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window. Community number is indicated on the colorbar. In (c) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

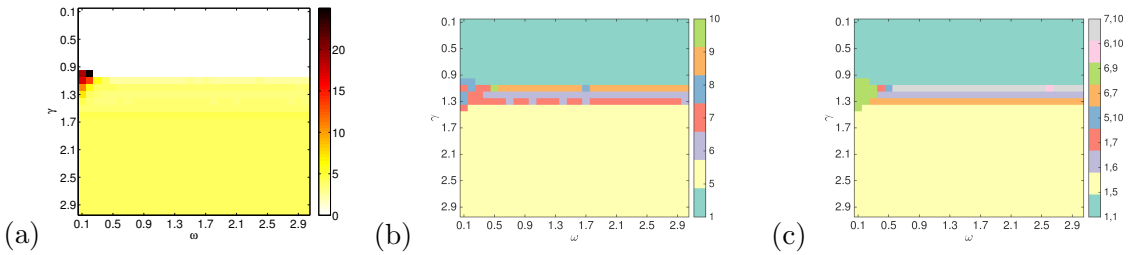


Figure C.7: H1N1 influenza, multislice networks, NG null model — temporal organization of partitions. In (a)–(c) we show results of varying the parameters γ and ω for: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

C.3.2 Modularity maximization using the correlation null model

When we use modularity maximization with the correlation null model on the static H1N1 networks, once again the pattern of spatial organization in partitions is not very clear. The number of spatial partitions across the data set is low, and there is no clear pattern as to which time periods generate spatial structures. As shown in Fig. C.8(a), we detect some spatial communities for a much larger parameter range than for any other null model–data set combination that we tested to date: the top 5 values of γ that generate the largest numbers of spatial partitions across the set of networks are (in order) $0.8, 0.9, 0.5, 0.2$ and 1.7. This variation between γ values is in contrast to the previous results using this null model, where it showed relatively little variation. The community structures that score as significantly spatial in the distance test are often dominated by one–two large communities [see example for network 43 at $\gamma = 0.8$ in Fig. C.8(b)].

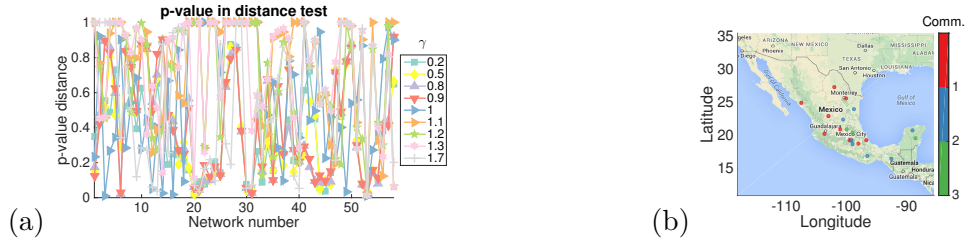


Figure C.8: H1N1 influenza, static networks, correlation null model — spatial partitions according to distance and MST measures. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{0.2, 0.5, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.7\}$. In panel (b) we show a map of all the nodes in network 43 at $\gamma = 0.8$ (colored by algorithmically detected community assignment, which is indicated on the color bar).

The multislice network partitions found using the correlation null model for almost all (γ, ω) parameter values that we test score as significantly spatial in the partition-wide distance test ($p < 0.05$). We show two representative partitions in Fig. C.9(b)-(c): the partition for $(\gamma, \omega) = (1, 0.1)$ has many changes in community structure between layers, and contains at least 2 large communities for most layers, and the partition for $(\gamma, \omega) = (2.3, 3)$ contains a small number of temporal singleton communities. The two partitions appear to contain spatial communities at different points in time: layers 8 and 10 for the former, and 1, 6, and 8 for the latter. The structures for other (γ, ω) values vary between these two extremes, as we change the parameters. Larger ω tends to lead to more temporal singletons, and lower γ tend to increase the total number of communities in the partition, and they introduce potential temporal partitions. Both of these effects are as expected from the definition of these two parameters, but they are best visible on this data set compared to other data sets, where the changes in community structure with changes in parameter values are not as gradual.

When studying the temporal organization of the multislice network partitions found using the correlation null model by using temporal z-Rand scores, we find that many of the

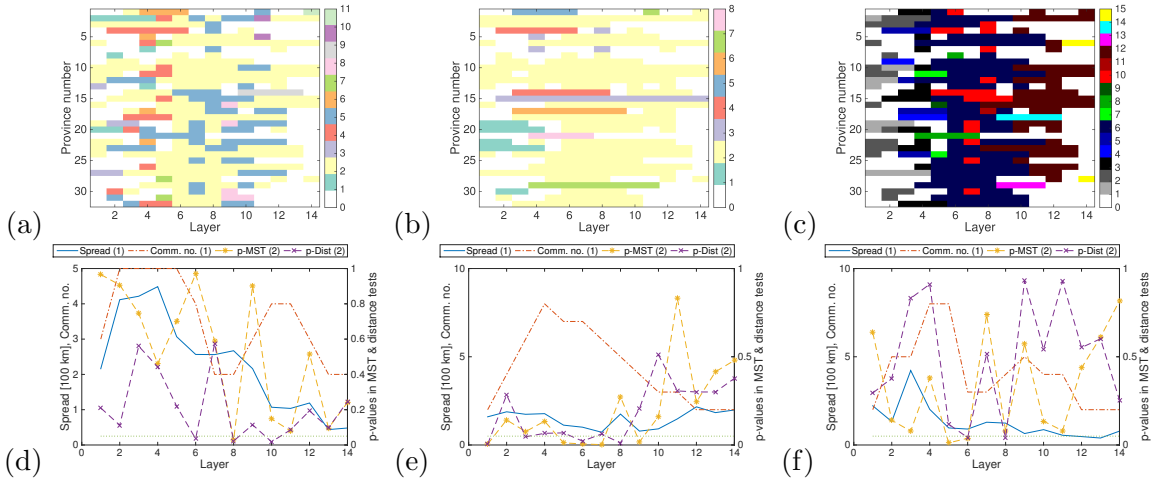


Figure C.9: H1N1 influenza, multislice networks, correlation null model — spatial organization of partitions. In (a)-(c) we plot community structure with nodes ordered by their location (north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for (a) $\gamma = 0.8$, $\omega = 0.1$, (b) $\gamma = 2.3$, $\omega = 3$ and (c) $\gamma = 2.2$, $\omega = 0.3$. Community number is indicated on the colorbar. In (d)-(f) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests, for (d) $\gamma = 0.8$, $\omega = 0.1$, (e) $\gamma = 2.3$, $\omega = 3$ and (f) $\gamma = 2.2$, $\omega = 0.3$.

partitions for $\omega \lessapprox 0.9$ except those for γ between approximately 0.9 and 1.2 have significant values of $z_R > 1.96$. The partition for for $(\gamma, \omega) = (0.8, 0.1)$ in Fig. C.10(a) has a critical time point at layers 6 and 9, similarly to the results for the NG null model. The partition for $(\gamma, \omega) = (2.2, 0.3)$ scores higher z_R than the $(\gamma, \omega) = (0.8, 0.1)$ partition versus the same pair of critical times [see Fig. C.10(c)]. These high temporal z_R scores are interesting. As mentioned in Chapter 8, these critical time points may correspond to the peak of the epidemic wave and the end of it. This is similar to our findings for the Ebola Datamarket data set using NG and correlation null models.

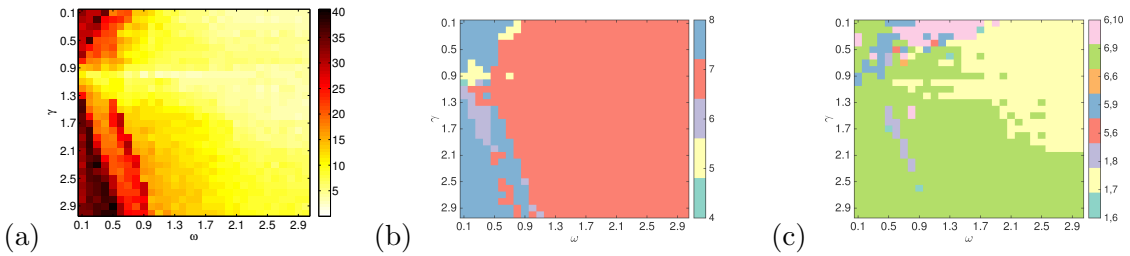


Figure C.10: H1N1 influenza, multislice networks, correlation null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω for: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

C.3.3 Modularity maximization using the gravity null model

Community structures for the static networks from the H1N1 data set that we find by maximizing modularity using the gravity null model tend to be dominated by one–two large communities, and contain several singleton communities. The structures for most values of γ do not score as significantly spatial on the distance test. The partitions contain many singleton communities that are often spread around the country (see Fig. C.11). Similarly to the results for other diseases, the nodes assigned to singleton communities when using the gravity null model tend to have larger populations than the nodes in the larger communities.

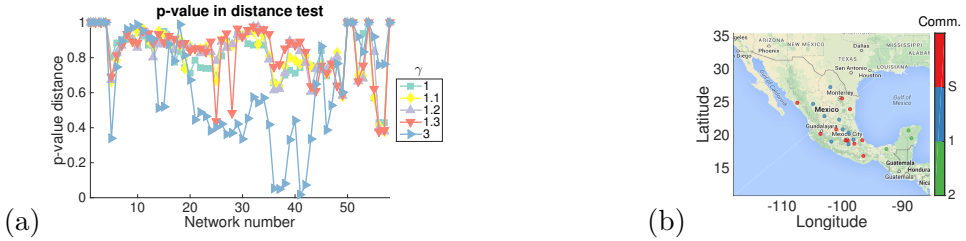


Figure C.11: H1N1 influenza, static networks, gravity null model — spatial partitions according to distance and MST measures. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{1, 1.1, 1.2, 1.3, 3\}$. In panel (b) we show a map of all the nodes in network 41 at $\gamma = 3$ (colored by algorithmically detected community assignment with singletons grouped into group S, community assignment indicated on the color bar).

For community detection using the gravity null model on the multislice networks, none of the values of γ and ω that we examine give partitions that score as significantly spatial ($p < 0.05$) in the distance test. The partitions for the gravity null model consist of one large community that persists through time and several small communities (containing 1–3 nodes). See Fig. C.12(b) for an example of such a partition.

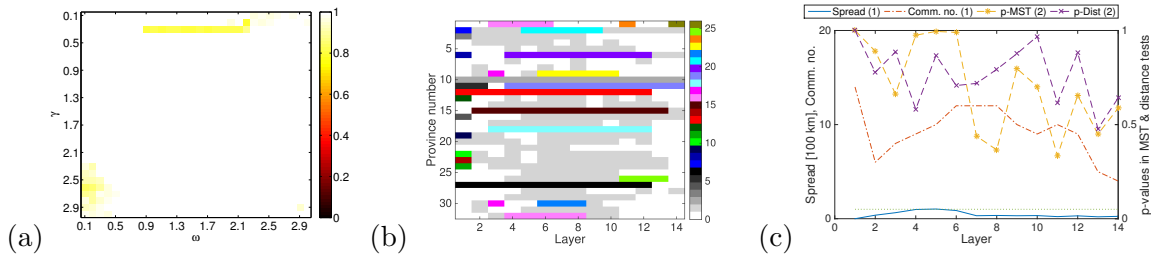


Figure C.12: H1N1 influenza, multislice networks, gravity null model — spatial organization of partitions. In panel (a) we show the influence of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b) we plot community structure with nodes ordered by their location (north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for $\gamma = 1.8$, $\omega = 0.1$. Community number is indicated on the colorbar. In (c) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 1.8$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

The multislice network partitions that score the highest z-Rand scores versus temporal partitions are all for $\gamma \lessapprox 0.4$ (see Fig. C.13). These structures correspond to divisions after layer 2 in the search for a single critical time, and as divisions at layers 2 and 14 in

the search for a pair of critical times. This temporal partition appears to be due to many singleton communities in the first layer [see structure in Fig. C.12(b)]. The method detects two other large groupings of critical time point pairs: layers 3 and 8 [see the dark blue region in Fig. C.13(c)] and layers 4 and 13 [see the red region in Fig C.13(c)]. Both of these critical times correspond to times when the number of nodes with non-zero strengths in the corresponding layers decreases (similarly to our previous findings for rubella), and they do not provide us with additional information about the disease patterns.

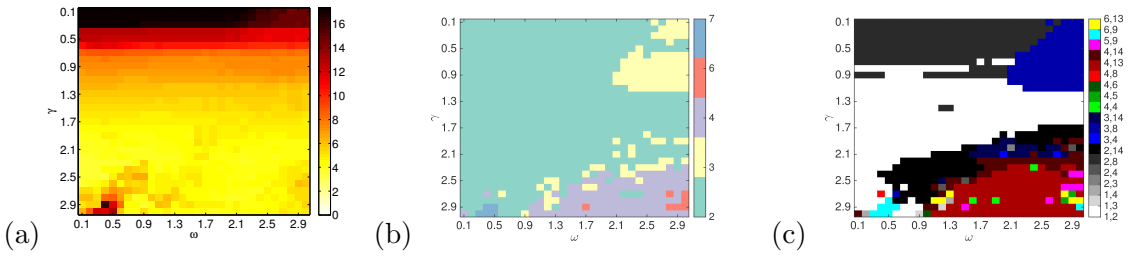


Figure C.13: H1N1 influenza, multislice networks, gravity null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω for: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

C.3.4 Modularity maximization using the radiation null model

The static network partitions that we obtain by maximizing modularity using the radiation null model do not appear to have a strong spatial organization. They do not score significantly in the distance test [see Fig. C.14(a)]. The representative structure for $\gamma = 0.6$, layer 23 has one large community and two singleton communities that are far apart from each other in space and have relatively large populations [see Fig. C.14(b)].

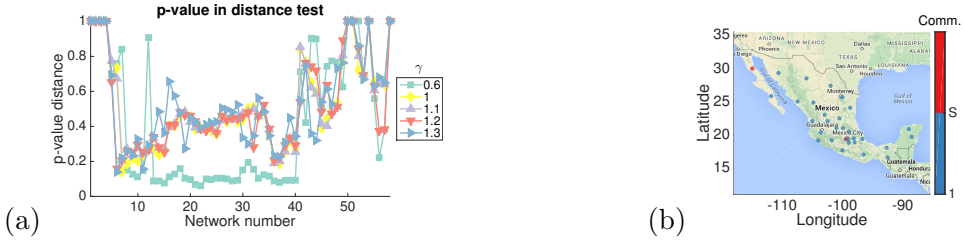


Figure C.14: H1N1 influenza, static networks, radiation null model — spatial partitions according to distance and MST measures. In panel (a), we plot the p-value in the distance test for community structures of all static networks for $\gamma \in \{0.6, 1, 1.1, 1.2, 1.3\}$. In panel (b) we show a map of all the nodes in network 23 at $\gamma = 0.6$ (colored by algorithmically detected community assignment with singletons grouped into group S, community assignment indicated on the color bar).

When we study the spatial organization of the partitions of the multislice networks found using the radiation null model for a variety of ω and γ parameters, we see that partitions for

$\gamma \in \{0.5, 0.6, 0.7\}$ appear to score significantly in the distance test. Once again, however, all of the network partitions for this null model tend to consist of one large community and several small (often singleton) communities, and none of the layers has a statistically significant p-value in distance and MST tests on their own. We show a representative partition for $\gamma = 0.8, \omega = 0.1$ and its summary statistics in Figs. C.15(b)-(c).

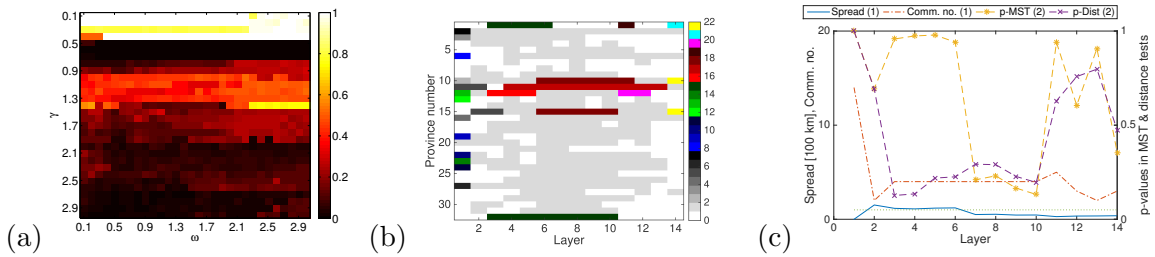


Figure C.15: H1N1 influenza, multislice networks, radiation null model — spatial organization of partitions. In panel (a) we show the influence of varying the parameters γ and ω on the p-values for distances being smaller than expected at random in the distance test. In (b) we plot community structure with nodes ordered by their location (north to south) on the vertical axis, with node community membership indicated by color, and 0 (white) indicating no disease in a given time window, for $\gamma = 0.8$, $\omega = 0.1$. Community number is indicated on the colorbar. In (c) we plot statistics for each layer, (left vertical axis) the community spread and the number of communities for $\gamma = 0.8$, $\omega = 0.1$ and (right vertical axis) the p-values for the distance and MST tests.

The multislice network partitions for $\gamma \lessapprox 0.3$ and $\omega \lessapprox 2$ appear to score highly significantly in the temporal z-Rand score test. Partitions for $0.4 \lessapprox \gamma \lessapprox 0.8$, and $\gamma \gtrapprox 2.7$ and $\omega \lessapprox 0.9$ [see the top stripe and bottom left corner of Fig. C.16(a)] also score statistically significant temporal z-Rand scores. The majority of these partitions correspond to layer 2 for the single critical time point test, and layers 2 and 8 for the search for two critical time points. However, the detected temporal partitions appear to be due to the presence of many singleton communities in the first layer, similarly to what we found for the gravity null model [see an example structure in Fig. C.15(b)].

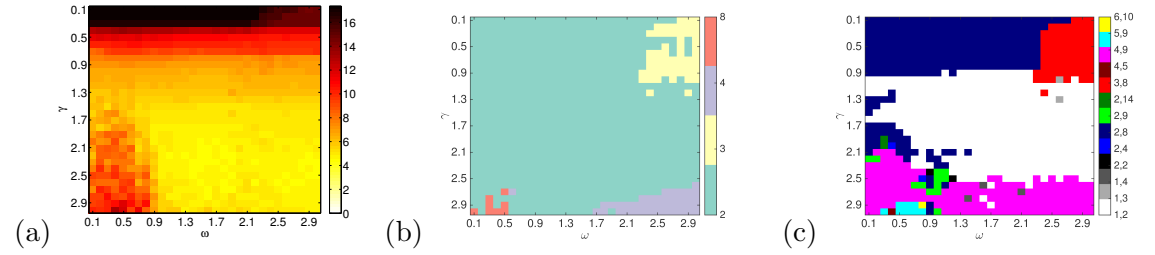


Figure C.16: H1N1 influenza, multislice networks, radiation null model — temporal organization of partitions. In (a)-(c) we show results of varying the parameters γ and ω for: (a) the z-Rand scores for similarity to "temporal" partitions before and after a pair of critical time points t_{c1} and t_{c2}; we plot maximum selected out of all t_{c1} and t_{c2} pairs. In (b) we plot the single critical time point t_c corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with a single critical time point partition), and (c) we plot the pairs of highest-scoring critical times (t_{c1}, t_{c2}) corresponding to the maximum z-Rand score in terms of its layer number (for a comparison versus a partition with two critical time points). For (b)-(c), the layer numbers of the critical time points are indicated on the colorbar.

C.3.5 Summary of the results for the H1N1 influenza data set

Our community-detection results for the H1N1 data set have considerable variation in time and with respect to parameter values. We fail to find reliable spatial partitions in static and multislice networks.

The correlation and NG null models appear to detect significant temporal partitions of the multislice networks, that may be related to the peak of the main H1N1 epidemic wave and the time when the wave subsided at the end of the epidemic. The gravity and radiation null models do not appear to give us additional spatial or temporal information.

Bibliography

[1] Global Administrative Areas, 2015. http://www.gadm.org.

[2] A E Aiello, G F Murray, V Perez, R M Coulborn, B M Davis, M Uddin, D K Shay, S H Waterman, and A S Monto. Mask use, hand hygiene, and seasonal influenza-like illness among young adults: A randomized intervention trial. *J. Infect. Dis.*, 201(4):491–498, 2010.

[3] R Albert and A-L Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, 2002.

[4] J Aldstadt, I-K Yoon, D Tannitisupawong, R G Jarman, S J Thomas, R V Gibbons, A Uppapong, S Iamsirithaworn, A L Rothman, T W Scott, and T Endy. Space-time analysis of hospitalised dengue patients in rural Thailand reveals important temporal intervals in the pattern of dengue virus transmission. *Trop. Med. Int. Health*, 17:1076–1085, 2012.

[5] K S Alexander, C E Sanderson, M Marathe, B L Lewis, Rivers C M, Shaman J, J M Drake, E Lofgren, V M Dato, M C Eisenberg, and S Eubank. What factors might have led to the emergence of Ebola in West Africa?, 2014. http://blogs.plos.org/speakingofmedicine/2014/11/11/factors-might-led-emergence-ebola-west-africa/.

[6] R Allard. Use of time-series analysis in infectious disease surveillance. *Bull. WHO*, 76(4):327–333, 1998.

[7] L A N Amaral and J M Ottino. Complex networks. *Eur. Phys. J. B*, 38(2):147–162, 2004.

[8] K B Anderson, R V Gibbons, D A T Cummings, A Nisalak, S Green, D H Libraty, R G Jarman, A Srikiatkhachorn, M P Mammen, Buddhari D, et al. A shorter time interval between first and second dengue infections is associated with protection from clinical illness in a school-based cohort in Thailand. *J. Infect. Dis.*, 209(3):360–368, 2014.

[9] R M Anderson and R M May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[10] L Anselin. Local Indicators of Spatial Association—LISA. *Geogr. Anal.*, 27(2):93–115, 1995.

[11] A Arenas, J Duch, A Fernández, and S Gómez. Size reduction of complex networks preserving modularity. *New J. Phys.*, 9(6):176, 2007.

[12] A Arenas, A Fernández, and S Gómez. Analysis of the structure of complex networks at different resolution levels. *New J. Phys.*, 10(5):53039, 2008.

[13] A H Auchincloss and A V Diez Roux. A new tool for epidemiology: The usefulness of dynamic-agent models in understanding place effects on health. *Am. J. Epidemiol.*, 168(1):1–8, 2008.

[14] M Z Austwick, O O Brien, E Strano, and M Viana. The structure of spatial networks and communities in bicycle sharing systems. *PloS One*, 8(9):e74685, 2013.

[15] D Balcan, V Colizza, B Gonçalves, H Hu, J J Ramasco, and A Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA*, 106(51):21484–21489, 2009.

[16] D Balcan, H Hu, B Gonçalves, P Bajardi, C Poletto, J J Ramasco, D Paolotti, N Perra, M Tizzoni, W Van den Broeck, V Colizza, and A Vespignani. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Med.*, 7(1):45, 2009.

[17] J E Banatvala and D W G Brown. Rubella. *Lancet*, 363(9415):1127–37, 2004.

[18] Z Bar-Yehuda. plot_google_map [Computer program], 2010. `http://www.mathworks.com/matlabcentral/fileexchange/27627-zoharby-plot-google-map`.

[19] M J Barber. Modularity and community detection in bipartite networks. *Phys. Rev. E*, 76(6):066102, 2007.

[20] I Barnett and J P Onnela. Change point detection in correlation networks. arXiv:1410.0761, 2014.

[21] M Barthélemy. Spatial networks. *Phys. Rep.*, 499(1-3):1–101, 2011.

[22] D S Bassett, E T Owens, M A Porter, M L Manning, and K E Daniels. Extraction of force-chain network architecture in granular materials using community detection. *Soft Matter*, 11:2731–2744, 2015.

[23] D S Bassett, M A Porter, N F Wymbs, S T Grafton, J M Carlson, and P J Mucha. Robust detection of dynamic community structure in networks. *Chaos*, 23(1):013142, 2013.

[24] D S Bassett, N F Wymbs, M A Porter, P J Mucha, J M Carlson, and S T Grafton. Dynamic reconfiguration of human brain networks during learning. *Proc. Natl. Acad. Sci. USA*, 108(18):7641–7646, 2011.

[25] D G Bausch and L Schwarz. Outbreak of Ebola virus disease in Guinea: Where ecology meets economy. *PLoS Negl. Trop. Dis.*, 8(7):e3056, 2014.

[26] M Bazzi, M A Porter, M McDonald, D J Fenn, S Williams, and S D Howison. Community structure in multilayer networks of financial-asset correlations. arXiv:1501.00040, 2014.

[27] N J Beeching, M Fenech, and C F Houlihan. Ebola virus disease. *Brit. Med. J.*, 349:1756–1833, 2014.

[28] V Belik, T Geisel, and D Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Phys. Rev. X*, 1(1):011001, 2011.

[29] T Y Berger-Wolf and J Saia. A framework for analysis of dynamic social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528, 2006.

[30] J M Best. Rubella. *Seminars in Fetal & Neonatal Medicine*, 12(3):182–192, 2007.

[31] A Bhan, D J Galas, and T G Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493, 2002.

[32] S Bhatt, P W Gething, O J Brady, J P Messina, A W Farlow, C L Moyes, J M Drake, J S Brownstein, A G Hoen, O Sankoh, et al. The global distribution and burden of dengue. *Nature*, 496:504–507, 2013.

[33] V D Blondel, J-L Guillaume, R Lambiotte, and E Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008:P10008, 2008.

[34] G V Bobashev, D M Goedecke, Feng Yu, and J M Epstein. A hybrid epidemic model: Combining the advantages of agent-based and equation-based approaches. In *Simulation Conference, 2007 Winter*, pages 1532–1537, 2007.

[35] N Boccara. *Modeling Complex Systems*. Springer, 2003.

[36] P-Y Boëlle, S Ansart, A Cori, and A-J Valleron. Transmission parameters of the A/H1N1 (2009) influenza virus pandemic: A review. *Influenza Other Respir. Viruses*, 5(5):306–316, 2011.

[37] B Bollobás. *Modern Graph Theory*. Springer, 1998.

[38] B Bollobás. *Random Graphs*. Springer, 1998.

[39] U Brandes, D Delling, M Gaertler, R Gorke, M Hoefer, Z Nikoloski, and D Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20:172–188, 2008.

[40] F Brauer and C Castillo-Chávez. *Mathematical Models in Population Biology and Epidemiology*. Springer, 2011.

[41] D Brockmann, L Hufnagel, and T Geisel. Dynamics of modern epidemics. In A R McLean, R M May, J Pattison, and R Weiss, editors, *SARS: A case study in emerging infections*, pages 81–91. Oxford University Press, 2005.

[42] D S Burke. Computational modeling and simulation of epidemic infectious diseases. In Smolinski M S, Hamburg M A, and Lederberg J, editors, *Microbial threats to health: emergence, detection, and response*, pages 335–342. National Academies Press, Washington (DC), 2003.

[43] G Caldarelli, S Battiston, D Garlaschelli, and M Catanzaro. Emergence of complexity in financial networks. In E Ben-Naim, H Frauenfelder, and Z Toroczkai, editors, *Complex Networks*, volume 650 of *Lecture Notes in Physics*, pages 399–423. Springer, 2004.

[44] P H Cao, X Wang, S S Fang, X W Cheng, K P Chan, X L Wang, X Lu, C L Wu, X J Tang, R L Zhang, H W Ma, J Q Cheng, C M Wong, and L Yang. Forecasting influenza epidemics from multi-stream surveillance data in a subtropical city of China. *PLoS One*, 9(3):e92945, 2014.

[45] B Cazelles, M Chavez, G C de Magny, J F Guégan, and S Hales. Time-dependent spectral analysis of epidemiological time-series with wavelets. *J. R. Soc. Interface*, 4(15):625–636, 2007.

[46] Centers for Disease Control and Prevention. Dengue, 2011–2014. `http://www.cdc.gov/NCIDOD/DVBID/DENGUE`.

[47] F Cerina, V De Leo, M Barthelemy, and A Chessa. Spatial correlations in attribute communities. *PLoS One*, 7(5):e37507, 2012.

[48] S Y Chan, P Hui, and K Xu. Community detection of time-varying mobile social networks. In J Zhou, editor, *Complex Sciences*, volume 4 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 1154–1159. Springer Berlin Heidelberg, 2009.

[49] D Chen. Modeling the spread of infectious diseases: A review. In D Chen, B Moulin, and J Wu, editors, *Analyzing and modeling spatial and temporal dynamics of infectious diseases*, pages 19–43. John Wiley & Sons, Hoboken, New Jersey, 2014.

[50] Y Chen, V Kawadia, and R Urgaonkar. Detecting overlapping temporal community structure in time-evolving networks. arXiv:1303.7226, 2013.

[51] G Chowell, B Cazelles, H Broutin, and C V Munayco. The influence of geographic and climate factors on the timing of dengue epidemics in Perú, 1994–2008. *BMC Infect. Dis.*, 11(1):164, 2011.

[52] G Chowell, S Echevarría-Zuno, C Viboud, L Simonsen, C Grajales Muñiz, R A Rascón Pacheco, M González León, and V H Borja-Aburto. Recrudescent wave of pandemic A/H1N1 influenza in Mexico, winter 2011–2012: Age shift and severity. *PLoS Curr.*, 4:RRN1306, 2012.

[53] G Chowell, S Echevarría-Zuno, C Viboud, L Simonsen, J Tamerius, M A Miller, and V H Borja-Aburto. Characterizing the epidemiology of the 2009 influenza A/H1N1 pandemic in Mexico. *PLoS Medicine*, 8(5):e1000436, 2011.

[54] G Chowell and F Sanchez. Climate-based descriptive models of dengue fever: The 2002 epidemic in Colima, Mexico. *J. Environ. Health*, 68(10):40–44, 55, 2006.

[55] G Chowell, C A Torre, C Munayco-Escate, L Suárez-Ognio, R López-Cruz, J M Hyman, and C Castillo-Chavez. Spatial and temporal dynamics of dengue fever in Peru: 1994–2006. *Epidemiol. Infect.*, 136(12):1667–1677, 2008.

[56] G Chowell, S Towers, CG Viboud, R Fuentes, and Sotomayor V. Rates of influenza-like illness and winter school breaks, Chile, 2004–2010. *Emerg. Infect. Dis*, 20, 2014.

[57] A D Cliff and P Haggett. Statistical modelling of measles and influenza outbreaks. *Stat. Methods Med. Res.*, 2(1):43–73, 1993.

[58] S Coakley, M Gheorghe, M Holcombe, S Chin, D Worth, and C Greenough. Exploitation of high performance computing in the flame agent-based simulation framework. In *IEEE 14th International Conference on High Performance Computing and Communications*, pages 538–545, 2012.

[59] V Colizza, A Barrat, M Barthélemy, and A Vespignani. Predictability and epidemic pathways in global outbreaks of infectious diseases: The SARS case study. *BMC Med.*, 5(1):34, 2007.

[60] V Colizza, R Pastor-Satorras, and A Vespignani. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(4):276–282, 2007.

[61] V Colizza and A Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations. *J Theor Biol*, 251(3):450–67, 2008.

[62] F J Colón-González, I R Lake, and G Bentham. Climate variability and dengue fever in warm and humid Mexico. *Am. J. Trop. Med. Hyg.*, 84(5):757–763, 2011.

[63] L Coudeville and G P Garnett. Transmission dynamics of the four dengue serotypes in southern Vietnam and the potential impact of vaccination. *PLoS One*, 7(12):e51244, 2012.

[64] A Cruz Marques. Human migration and the spread of malaria in Brazil. *Parasitol. Today*, 3(6):166–170, 1987.

[65] P Csermely, A London, L Y Wu, and B Uzzi. Structure and dynamics of core/periphery networks. *J. Complex Networks*, 1:93–123, 2013.

[66] D A T Cummings, R A Irizarry, N E Huang, T P Endy, A Nisalak, K Ungchusak, and D S Burke. Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. *Nature*, 427(6972):344–347, 2004.

[67] B D Dalziel, B Pourbohloul, and S P Ellner. Human mobility patterns predict divergent epidemic dynamics among cities. *Proc. R. Soc. B*, 280(1766):20130763, 2013.

[68] L Danon, J Duch, A Díaz-Guilera, and A Arenas. Comparing community structure identification. *J. Stat. Mech.*, 2005(09):10, 2005.

[69] L Danon, A P Ford, T House, C T Jewell, M J Keeling, G O Roberts, J V Ross, and M C Vernon. Networks and the epidemiology of infectious disease. *Interdiscip. Perspect. Infect. Dis.*, 2011:284909, 2011.

[70] Datamarket, 2014–2015. `https://datamarket.com/data/set/4spl/sub-national-time-series-data-on-ebola-cases-and-deaths-in-guinea-liberia-sierra-leone-nigeria-and-senegal-since-march-2014#!ds=4spl!88d0=1x.r.20.9.1o.g.f.1s.h.13.1z.1j.2e.a.7.6.b.1n.1w.c.2g.y.2f.2j.2i.1y.e.2h.21.14.15.16.p.17.18.i.1k.19.j.o.1v.k.l.1r.m.1u.n.1a.1b.1c.s.w.1e.x.q.z.t.u.10.v.1t.2d.11.1f.1g.12.1h.1i:88d1=1&display=line`.

[71] M De Domenico, A Lancichinetti, A Arenas, and M Rosvall. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys. Rev. X*, 5:011027, 2015.

[72] M De Domenico, V Nicosia, A Arenas, and V Latora. Structural reducibility of multilayer networks. *Nat. Commun.*, 6:6864, 2015.

[73] M De Domenico, A Solé-Ribalta, E Cozzo, M Kivelä, Y Moreno, M A Porter, S Gómez, and A Arenas. Mathematical formulation of multilayer networks. *Phys. Rev. X*, 3:041022, 2013.

[74] J-C Delvenne, S N Yaliraki, and M Barahona. Stability of graph communities across time scales. *Proc. Natl. Acad. Sci. USA*, 107(29):12755–60, 2010.

[75] C Depradine and E Lovell. Climatological variables and the incidence of Dengue fever in Barbados. *Int. J. Environ. Health Res.*, 6:429–441, 2004.

[76] E Descloux, M Mangeas, C E Menkes, M Lengaigne, A Leroy, T Tehei, L Guillaumot, M Teurlai, A-C Gourinat, J Benzler, et al. Climate-based models for understanding and forecasting dengue epidemics. *PLoS Negl. Trop. Dis.*, 6(2):e1470, 2012.

[77] O Diekmann, H Heesterbeek, and T Britton. *Mathematical Tools for Understanding Infectious Disease Dynamics:*. Princeton Series in Theoretical and Computational Biology. Princeton University Press, 2012.

[78] K Dietz and J A P Heesterbeek. Bernoulli was ahead of modern epidemiology. *Nature*, 408(6812):513–514, 2000.

[79] A-C Disdier and K Head. The puzzling persistence of the distance effect on bilateral trade. *Rev Econ. Stat.*, 90:37–48, 2008.

[80] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal Special Topics*, 174(1):157–179, 2009.

[81] J Duch and A Arenas. Community detection in complex networks using extremal optimization. *Phys. Rev. E*, 72(2):027104, 2005.

[82] P Expert, T S Evans, V D Blondel, and R Lambiotte. Uncovering space-independent communities in spatial networks. *Proc. Natl. Acad. Sci. USA*, 108(19):7663–7668, 2011.

[83] D J Fenn, M A Porter, M McDonald, S Williams, N F Johnson, and N S Jones. Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007–2008 credit crisis. *Chaos*, 19(3):033119, 2009.

[84] D J. Fenn, M A. Porter, P J. Mucha, M Mcdonald, S Williams, N F. Johnson, and N S. Jones. Dynamical clustering of exchange rates. *Quantitative Finance*, 12:1493–1520, 2012.

[85] N M Ferguson, D A T Cummings, S Cauchemez, C Fraser, S Riley, A Meeyai, S Iamsirithaworn, and D S Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214, 2005.

[86] D A Focks and R Barrera. Dengue transmission dynamics: assessment and implications for control. In *WHO Report of the Scientific Working Group meeting on Dengue, Geneva*, 2006.

[87] Centres for Disease Control and Prevention, 2015. `http://www.cdc.gov/flu/index.htm`.

[88] Office for the Coordination of Humanitarian Affairs in West & Central Africa, 2014–2015. `http://www.unocha.org/rowca/`.

[89] A Fornito and E T Bullmore. Connectomics: A new paradigm for understanding brain disease. *Europ. Neuropsychopharm.*, 25(5):733–748, 2014.

[90] B M Forshey, A C Morrison, C Cruz, C Rocha, S Vilcarromero, C Guevara, D E Camacho, A Alava, C Madrid, L Beingolea, et al. Dengue virus serotype 4, northeastern Peru, 2008. *Emerg. Infect. Dis.*, 15(11):1815–1818, 2009.

[91] S Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3-5):75–174, 2010.

[92] S Fortunato and M Barthélemy. Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA*, 104(1):36–41, 2007.

[93] C Fraser, C A Donnelly, S Cauchemez, W P Hanage, M D Van Kerkhove, T D Hollingsworth, J Griffin, R F Baggaley, H E Jenkins, E J Lyons, T Jombart, W R Hinsley, N C Grassly, F Balloux, A C Ghani, N M Ferguson, A Rambaut, O G Pybus, H Lopez-Gatell, C M Alpuche-Aranda, I B Chapela, E P Zavala, D Guevara, F Checchi, E Garcia, S Hugonnet, C Roth, and The WHO Rapid Pandemic Assessment Collaboration. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324(5934):1557–1561, 2009.

[94] L C Freeman. Some antecedents of social network analysis. *Connections*, 19(1):39–42, 1996.

[95] C Fuhrmann. The effects of weather and climate on the seasonality of influenza: what we know and what we need to know. *Geogr. Compass*, 4(7):718–730, 2010.

[96] A Fujita, P Severino, J R Sato, and S Miyano. Granger causality in systems biology: Modeling gene networks in time series microarray data using vector autoregressive models. In C E Ferreira, S Miyano, and P F Stadler, editors, *Advances in bioinformatics and computational biology*, volume 6268 of *Lecture Notes in Computer Science*, pages 13–24. Springer, 2010.

[97] K L Gage, T R Burkot, R J Eisen, and E B Hayes. Climate and vectorborne diseases. *Am. J. Prev. Med.*, 35(5):436–450, 2008.

[98] R C Geary. The contiguity ratio and statistical mapping. *Incorp. Stat.*, 5(3):115–127+129–146, 1954.

[99] C R Genovese, N A Lazar, and T Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

[100] Geonames.org, 2011–2014. `http://www.geonames.org/`.

[101] M P Girard, J S Tam, O M Assossou, and M P Kieny. The 2009 A (H1N1) influenza virus pandemic: A review. *Vaccine*, 28(31):4895–902, 2010.

[102] M Girvan and M E J Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002.

[103] S Goh, K Lee, J S Park, and M Y Choi. Modification of the gravity model and application to the metropolitan Seoul subway system. *Phys. Rev. E*, 86:026102, 2012.

[104] U Gokovali, O Bahar, and M Kozak. Determinants of length of stay: A practical use of survival analysis. *Tour. Manag.*, 28(3):736–746, 2007.

[105] A Goldenberg, A X Zheng, S E Fienberg, and E M Airoldi. A survey of statistical network models. *Found.Trends Mach. Learn.*, 2(2):129–233, 2010.

[106] M F C Gomes, Pastore y Piontti A, L Rossi, D Chao, I Longini, M E Halloran, and A Vespignani. Assessing the international spreading risk associated with the 2014 West African Ebola outbreak. *PLOS Curr. Out.*, 1, 2014.

[107] Gómez, S and Jensen, P and Arenas, A. Analysis of community structure in networks of correlated data. *Phys. Rev. E*, 80(1):016114, 2009.

[108] B H Good, Y A de Montjoye, and A Clauset. Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81:046106, 2010.

[109] C Granell, R K Darst, A Arenas, S Fortunato, and S Gómez. A benchmark model to assess community structure in evolving networks. arXiv:1501.05808, 2015.

[110] N G Gratz. Critical review of the vector status of *Aedes albopictus*. *Med. Vet. Entomol.*, 18(3):215–227, 2004.

[111] C Green, D Krause, and J Wylie. Spatial analysis of campylobacter infection in the canadian province of manitoba. *Int. J. Health Geogr.*, 5(1):2, 2006.

[112] S K Greene, E L Ionides, and M L Wilson. Patterns of influenza-associated mortality among us elderly by geographic region and virus subtype, 1968–1998. *Am. J. Epidemiol.*, 163(4):316–326, 2006.

[113] D J Gubler. Dengue and Dengue Hemorrhagic Fever. *Clin. Microbiol. Rev.*, 11(3):480–496, 1998.

[114] D J Gubler and M Meltzer. Impact of dengue/dengue hemorrhagic fever on the developing world. *Adv. Vir. Res.*, 53:35–70, 1999.

[115] R Guimerà and L A Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[116] R Guimerà, M Sales-Pardo, and L A N Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70(2):025101, 2004.

[117] M G Guzman, S B Halstead, H Artsob, P Buchy, J Farrar, D J Gubler, E Hunsperger, A Kroeger, H S Margolis, E Martínez, M B Nathan, JL Pelegrino, C Simmons, S Yoksan, and R W Peeling. Dengue: a continuing global threat. *Nat. Rev. Microbiol.*, 8(12 Suppl):S7–16, 2010.

[118] S Hales, N De Wet, J Maindonald, and A Woodward. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *Lancet*, 360(9336):830–834, 2002.

[119] J Hannigan, G Hernandez, R M. Medina, P Roos, and P Shakarian. Mining for spatially-near communities in geo-located social networks. arXiv:1311.1924, 2013.

[120] L C Harrington, T W Scott, K Lerdthusnee, R C Coleman, A Costero, G G Clark, J J Jones, S Kitthawee, P Kittayapong, R Sithiprasasna, et al. Dispersal of the dengue vector *Aedes aegypti* within and between rural communities. *Am. J. Trop. Med. Hyg.*, 72(2):209–220, 2005.

[121] M B Hastings. Community detection as an inference problem. *Phys. Rev. E*, 74:035102, 2006.

[122] D M Hawkins and K D Zamba. A change-point model for a shift in variance. *J. Quality Technology*, 37(1):21–31, 2005.

[123] W A Hawley, P Reiter, R S Copeland, C B Pumpuni, and George B Craig. *Aedes albopictus* in North America: probable introduction in used tires from northern Asia. *Science*, 236(4805):1114–1116, 1987.

[124] S I Hay, D J Rogers, S E Randolph, D I Stern, J Cox, G D Shanks, and R W Snow. Hot topic or hot air? Climate change and malaria resurgence in East African highlands. *Trends Parasitol.*, 18(12):530–534, 2002.

[125] C G Hayes, I A Phillips, J D Callahan, W F Griebenow, K C Hyams, S J Wu, and D M Watts. The epidemiology of dengue virus infection among urban, jungle, and rural populations in the Amazon region of Peru. *Am. J. Trop. Med. Hyg.*, 55(4):459–463, 1996.

[126] K Heath, H A Peck, K L Laurie, P Wu, H Nishiura, and B J Cowling. The age-specific cumulative incidence of infection with pandemic influenza H1N1 2009 was similar in various countries prior to vaccination. *PLoS One*, 6(8):e21828, 2011.

[127] U Helfenstein. Box-Jenkins modelling of some viral infectious diseases. *Stat. Med.*, 5(1):37–47, 1986.

[128] J H Holland. *Emergence: From chaos to order*. Da Capo Press, 1999.

[129] P Holme and J Saramäki. Temporal networks. *Phys. Rep.*, 519(3):97–125, 2012.

[130] P Holme and J Saramäki, editors. *Temporal Networks*. Springer, 2013.

[131] L Hufnagel, D Brockmann, and T Geisel. Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci. USA*, 101(42):15124–15129, 2004.

[132] Instituto Nacional de Estadística e Informática (INEI). `http://www.inei.gob.pe/`, 2011–2014.

[133] K Jaakkola, A Saukkoriipi, J Jokelainen, R Juvonen, J Kauppila, O Vainio, T Ziegler, E Ronkko, J Jaakkola, T Ikaheimo, and the KIAS-Study Group. Decline in temperature and humidity increases the occurrence of influenza in cold climate. *Environ. Health*, 13(1):22, 2014.

[134] C C Jansen and N W Beebe. The dengue vector *Aedes aegypti*: what comes next. *Microbes Infect.*, 12(4):272–279, 2010.

[135] L G S Jeub, P Balachandran, M A Porter, P J Mucha, and M W Mahoney. Think locally, act locally: Detection of small, medium-sized, and large communities in large networks. *Phys. Rev. E*, 91:012821, 2015.

[136] M A Johansson, F Dominici, and G E Glass. Local and global effects of climate on dengue transmission in Puerto Rico. *PLoS Negl. Trop. Dis.*, 3(2):e382, 2009.

[137] N P S Johnson and J Mueller. Updating the accounts: Global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bull. Hist. Med.*, 76:105–115, 2002.

[138] K E Jones, N G Patel, M A Levy, A Storeygard, D Balk, J L Gittleman, and P Daszak. Global trends in emerging infectious diseases. *Nature*, 451:990–993, 2008.

[139] I S Jutla, P J Mucha, and L Jeub. GenLouvain: A generalized Louvain method for community detection implemented in MATLAB, version 2.0, 2011–2014. `http://netwiki.amath.unc.edu/GenLouvain`.

[140] V Kawadia and S Sreenivasan. Sequential detection of temporal communities by estrangement confinement. *Sci. Rep.*, 2:794, 2012.

[141] J Keating. An investigation into the cyclical incidence of dengue fever. *Soc. Sci. Med.*, 53:1587–1597, 2001.

[142] M J Keeling and K T D Eames. Networks and epidemic models. *J. R. Soc. Interface*, 2(4):295–307, 2005.

[143] K Khan, J Arino, W Hu, P Raposo, J Sears, F Calderon, C Heidebrecht, M Macdonald, J Liauw, A Chan, and M Gardam. Spread of a novel influenza A (H1N1) virus via global airline transportation. *New Engl. J. Med.*, 361(2):212–214, 2009. PMID: 19564630.

[144] A M Kilpatrick, A A Chmura, D W Gibbons, R C Fleischer, P P Marra, and P Daszak. Predicting the global spread of H5N1 avian influenza. *Proc. Natl. Acad. Sci. USA*, 103(51):19368–19373, 2006.

[145] Y Kim, S-W Son, and H Jeong. Finding communities in directed networks. *Phys. Rev. E*, 81(1), 2010.

[146] M Kivelä, A Arenas, M Barthelemy, J P Gleeson, Y Moreno, and M A Porter. Multilayer Networks. *J. Complex Networks*, 2:203–271, 2014.

[147] A B Knudsen and R Slooff. Vector-borne disease problems in rapid urbanization: new approaches to vector control. *B. World Health Organ.*, 70(1):1–6, 1992.

[148] T Kochel, P Aguilar, V Felices, G Comach, C Cruz, A Alava, J Vargas, J Olson, and P Blair. Molecular epidemiology of dengue virus type 3 in Northern South America: 2000–2005. *Inf. Genet. Evol.*, 8(5):682–688, 2008.

[149] T J Kochel, D M Watts, S B Halstead, C G Hayes, A Espinoza, V Felices, R Caceda, C T Bautista, Y Montoya, S Douglas, and K L Russell. Effect of dengue-1 antibodies on American dengue-2 viral infection and dengue haemorrhagic fever. *Lancet*, 360(9329):310–312, 2002.

[150] F Kose, W Weckwerth, T Linke, and O Fiehn. Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics*, 17(12):1198–1208, 2001.

[151] A Kraskov, H Stägbauer, R G Andrzejak, and P Grassberger. Hierarchical clustering using mutual information. *Europhys. Lett.*, 70(2):278–284, 2005.

[152] M Kulldorff. A spatial scan statistic. *Commun. Stat. A–Theor*, 26(6):1481–1496, 1997.

[153] G Kuno. Factors influencing the transmission of dengue viruses. In D J Gubler and Kuno G, editors, *Dengue and Dengue Hemorrhagic Fever*, pages 61–88. CAB International, Wallingford, UK, 1997.

[154] J L Kyle and E Harris. Global spread and persistence of dengue. *Annu. Rev. Microbiol.*, 62(1):71–92, 2008.

[155] R Lambiotte, J-C Delvenne, and M Barahona. Laplacian dynamics and multiscale modular structure in networks. arXiv:0812.1770, 2009.

[156] R Lambiotte, J-C Delvenne, and M Barahona. Random walks, Markov processes and the multiscale modular organization of complex networks. *IEEE Trans. Net. Sci. Eng.*, 1:76–90, 2015.

[157] A Lancichinetti and S Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80(1):016118, 2009.

[158] A Lancichinetti and S Fortunato. Community detection algorithms: a comparative analysis. *Phys. Rev. E*, 80:056117, 2009.

[159] A Lancichinetti and S Fortunato. Limits of modularity maximization in community detection. *Phys. Rev. E*, 84:066122, 2011.

[160] A Lancichinetti and S Fortunato. Consensus clustering in complex networks. *Sci. Rep.*, 2(336):794, 2012.

[161] A Lancichinetti, S Fortunato, and F Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78(4):6, 2008.

[162] A Lancichinetti, F Radicchi, J J Ramasco, and S Fortunato. Finding statistically significant communities in networks. *PLoS One*, 6(4):e18961, 2011.

[163] S Lee, L E C Rocha, F Liljeros, and P Holme. Exploiting temporal network structures of human interaction to effectively immunize populations. *PLoS One*, 7(5):e36439, 2012.

[164] S H Lee, R Ffrancon, D M Abrams, B J Kim, and M A. Porter. Matchmaker, matchmaker, make me a match: Migration of populations via marriages in the past. *Phys. Rev. X*, 4:041009, 2014.

[165] E Leicht and M Newman. Community structure in directed networks. *Phys. Rev. Lett.*, 100(11), 2008.

[166] A C F Lewis, N S Jones, M A Porter, and C M Deane. The function of communities in protein interaction networks at multiple scales. *BMC Sys. Biol.*, 4(100), 2010.

[167] X Li, H Tian, D Lai, and Z Zhang. Validation of the gravity model in predicting the global spread of influenza. *Int. J. Environ. Res. Public Health*, 8(8):3134–3143, 2011.

[168] S W Lindsay and W J Martens. Malaria in the African highlands: past, present and future. *B. World Health Organ.*, 76(1):33–45, 1998.

[169] I M Longini, A Nizam, S Xu, K Ungchusak, W Hanshaoworakul, D A T Cummings, and M E Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

[170] A C Lowen and J Steel. Roles of humidity and temperature in shaping influenza seasonality. *J. Virol.*, 88(14):7692–7695, 2014.

[171] M MacMahon and D Garlaschelli. Community detection for correlation matrices. *Phys. Rev. X*, 5:021006, 2015.

[172] K T Macon, P J Mucha, and M A Porter. Community structure in the United Nations General Assembly. *Physica A*, 391(1–2):343–361, 2012.

[173] N Malik, B Bookhagen, N Marwan, and J Kurths. Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. *Clim. Dyn.*, 39(3-4):971–987, 2012.

[174] J S Malik Peiris, L L M Poon, and Y Guan. Emergence of a novel swine-origin influenza A virus (S-OIV) H1N1 virus in humans. *J. Clin. Virol.*, 45(3):169–173, 2009.

[175] R N Mantegna. Hierarchical structure in financial markets. *Eur. Phys. J. B*, 11(1):193–197, 1999.

[176] A Marí Saéz, S Weiss, K Nowak, V Lapeyre, F Zimmermann, A Dux, H S Kuhl, M Kaba, S Regnaut, K Merkel, A Sachse, U Thiesen, L Villanyi, C Boesch, P W Dabrowski, A Radonicá, A Nitsche, S A J Leendertz, S Petterson, S Becker, V Krahling, E Couacy-Hymann, C Akoua-Koffi, N Weber, L Schaade, J Fahr, M Borchert, J F Gogarten, S Calvignac-Spencer, and F H Leendertz. Investigating the zoonotic origin of the west african ebola epidemic. *EMBO Molecular Medicine*, 7(1):17–23, 2015.

[177] P Martens and L Hall. Malaria on the move: human population movement and malaria transmission. *Emerging Infect Dis.*, 6:103–109, 2000.

[178] A P Masucci, J Serras, A Johansson, and M Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Phys. Rev. E*, 88:022812, 2013.

[179] J McDevitt, S Rudnick, M First, and Spengler J. Role of absolute humidity in the inactivation of influenza viruses on stainless steel surfaces at elevated temperatures. *Appl. Environ. Microb.*, 76(12):3943–3947, 2010.

[180] M L Mehta. *Random Matrices*. Academic Press, San Diego, California, third edition, 2004.

[181] M Meilá. Comparing clusterings – an information based distance. *J. Multivariate Anal.*, 98(5):873–895, 2007.

[182] M Meltzer, C Y Atkins, S Santibanez, B Knust, B W Petersen, E D Ervin, S T Nichol, I K Damon, and M L Washington. Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014-2015. *MMWR Surveill Summ.*, 63:1–14, 2014.

[183] M I Meltzer, N J Cox, and K Fukuda. The economic impact of pandemic influenza in the United States: Priorities for intervention. *Emerg. Infect. Dis.*, 5:659–671, 1999.

[184] R A Meyers, editor. *Complex Systems in Finance and Econometrics*. Springer, 2011.

[185] Ministerio de Salud de Chile Departamento de Epidemiologia. Vigilancia epidemiológica, investigación y control de brotes. `http://epi.minsal.cl/epi/html/normas/circul/CircularInfluenzaESTACIONALyPANDEMICA.pdf`, 2013.

[186] N A M Molinari, I R Ortega-Sanchez, M L Messonnier, W W Thompson, P M Wortley, E Weintraub, and C B Bridges. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, 25(27):5086–5096, 2007.

[187] Y Montoya, S Holechek, O Caceres, A Palacios, J Burans, C Guevara, F Quintana, V Herrera, E Pozo, and E Anaya. Circulation of dengue viruses in north-western Peru, 2000–2001. *Dengue Bulletin*, 27:52–62, 2003.

[188] C G Moore and J E Freier. Geospatial technologies and spatial data analysis part 1: Geographic information system approaches to data analysis. In N M Míkanatha, R Lynfield, and B C A Van, editors, *Infectious disease surveillance*. John Wiley & Sons, Somerset, NJ, USA, 2013.

[189] P A P Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[190] D M Morens, G K Folkers, and A S Fauci. What is a pandemic? *J. Infect. Dis.*, 200(7):1018–1021, 2009.

[191] C W Morin, Comrie A C, and Ernst K C. Climate and dengue transmission: evidence and implications. *Environ. Health. Perspect.*, 121:1264–1272, 2013.

[192] A C Morrison, S L Minnick, C Rocha, B M Forshey, S T Stoddard, A Getis, D A Focks, K L Russell, J G Olson, P J Blair, et al. Epidemiology of dengue virus in Iquitos, Peru 1999 to 2005: interepidemic and epidemic patterns of transmission. *PLoS Negl. Trop. Dis.*, 4(5):e670, 2010.

[193] P J Mucha and M A Porter. Communities in multislice voting networks. *Chaos*, 20(4):041108, 2010.

[194] P J Mucha, T Richardson, K Macon, M A Porter, and J-P Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[195] J D Murray. *Mathematical Biology*. Springer, 3rd edition, 2003.

[196] J D Murray, E A Stanley, and D L Brown. On the spatial spread of rabies among foxes. *P. Roy. Soc. Lond. B Bio.*, 229(1255):111–150, 1986.

[197] R R Nadakuditi and M E J Newman. Graph spectra and the detectability of community structure in networks. *Phys. Rev. Lett.*, 108:188701, 2012.

[198] S Naish, P Dale, J Mackenzie, J McBride, K Mengersen, and S Tong. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC Infect. Dis.*, 14(1):167, 2014.

[199] M Newman and M Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[200] M E J Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003.

[201] M E J Newman. Analysis of weighted networks. *Phys. Rev. E*, 70:056131, 2004.

[202] M E J Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.

[203] M E J Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582, 2006.

[204] M E J Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[205] M E J Newman. Communities, modules and large-scale structure in networks. *Nat. Phys.*, 8(1):25–31, 2012.

[206] S Ng and B J Cowling. Association between temperature, humidity and Ebola virus disease outbreaks in Africa, 1976 to 2014. *Eurosurveillance*, 19:20892, 2014.

[207] L Osorio, J Todd, and D J Bradley. Travel histories as risk factors in the analysis of urban malaria in Colombia. *Am. J. Trop. Med. Hyg.*, 71(4):380–386, 2004.

[208] G Palla, A-L Barabási, and T Vicsek. Quantifying social group evolution. *Nature*, 446:664–667, 2007.

[209] G Palla, I Derényi, I Farkas, and T Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[210] N Parikh, M Youssef, S Swarup, and S Eubank. Modeling the effect of transient populations on epidemics in Washington DC. *Sci. Rep.*, 3:3152, 2013.

[211] M Pascual, J A Ahumada, L F Chaves, X Rodo, and M Bouma. Malaria resurgence in the East African highlands: temperature trends revisited. *Proc. Natl. Acad. Sci. USA*, 103(15):5829–5834, 2006.

[212] R Pastor-Satorras, C Castellano, P Van Mieghem, and A Vespignani. Epidemic processes in complex networks. arXiv:1408.2701, 2014.

[213] R Pastor-Satorras and A Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001.

[214] L Peel and A Clauset. Detecting change points in the large-scale structure of evolving networks. arXiv:1403.0989, 2015.

[215] J E Pinzon, J M Wilson, C J Tucker, R Arthur, P B Jahrling, and P Formenty. Trigger events: Enviroclimatic coupling of Ebola hemorrhagic fever outbreaks. *Am. J. Trop. Med. Hyg.*, 71(5):664–674, 2004.

[216] C Poletto, M Tizzoni, and V Colizza. Heterogeneous length of stay of hosts movements and spatial epidemic spread. *Sci. Rep.*, 2:476, 2012.

[217] M S Porta, S Greenland, and J M Last. *A dictionary of epidemiology.* Oxford University Press, 2008.

[218] M A Porter and J Gleeson. Dynamical systems on networks: A tutorial. arXiv:1403.7663, 2015.

[219] M A Porter, J-P Onnela, and P J Mucha. Communities in networks. *Notices Amer. Math. Soc.*, 56(9):1082–1097, 1164–1166, 2009.

[220] R C Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36:1389–1401, 1957.

[221] J Raghwani, A Rambaut, E C Holmes, V T Hang, T T Hien, J Farrar, B Wills, N J Lennon, B W Birren, M R Henn, et al. Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission. *PLoS Pathog.*, 7(6):e1002064, 2011.

[222] C Ratti, S Sobolevsky, F Calabrese, C Andris, J Reades, M Martino, R Claxton, and S H Strogatz. Redrawing the map of Great Britain from a network of human interactions. *PLoS One*, 5(12):e14248, 2010.

[223] N G Reich, S Shrestha, A A King, P Rohani, J Lessler, S Kalayanarooj, I-K Yoon, R V Gibbons, D S Burke, and D A T Cummings. Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *J. R. Soc. Interface*, 10(86):20130414, 2013.

[224] J Reichardt and S Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, 2006.

[225] P Reiter. Climate change and mosquito-borne disease. *Environ. Health Persp.*, 109(Suppl 1):141–161, 2001.

[226] P Reiter, S Lathrop, M Bunning, B Biggerstaff, D Singer, T Tiwari, L Baber, M Amador, J Thirion, J Hayes, et al. Texas lifestyle limits transmission of dengue virus. *Emerg. Infect. Dis.*, 9(1):86–89, 2003.

[227] L E C Rocha, F Liljeros, and P Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comp. Bio.*, 7(3):e1001109, 2011.

[228] D J Rogers, S E Randolph, R W Snow, and S I Hay. Satellite imagery in the study and forecast of malaria. *Nature*, 6872:710–715, 2002.

[229] M P Rombach, M A Porter, J H Fowler, and P J Mucha. Core-periphery structure in networks. *SIAM J. Applied Mathematics*, 74(1):167–190, 2014.

[230] M Rosvall and C T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA*, 105(4):1118–1123, 2008.

[231] L A Rvachev and I M Jr. Longini. A mathematical model for the global spread of influenza. *Math. Biosci.*, 75(1):3–22, 1985.

[232] Eubank S, Guclu H, Anil Kumar V S, Marathe M, Srinivasan V, Toroczkai Z, and Wang N. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, 2004.

[233] L K Saker, B Cannito, A Gilmore, and D Campbell-Lendrum. *Globalization and infectious diseases: A review of the linkages.* Economic and Behavioural Research. Special Topics No.3. World Health Organization, Geneva, 2004.

[234] Sanofi Pasteur, 2014. `http://www.sanofipasteur.com/en/articles/the-new-england-journal-of-medicine-publishes-results-of-final-landmark-phase-III-efficacy-clinical-study-of-sanofi-pasteur-s-dengue-vaccine-candidate.aspx`.

[235] M Sarzynska, E Leicht, G Chowell, and M A Porter. Null models for community detection in spatially-embedded, temporal networks. arXiv:1407.6297, 2014.

[236] M Sarzynska, O Udiani, and N Zhang. A study of gravity-linked metapopulation models for the spatial spread of dengue fever. arXiv:1308.4589, 2013.

[237] L Sattenspiel. *The Geographic Spread of Infectious Diseases: Models and Applications.* Princeton University Press, 2009.

[238] R Schäfer and T Guhr. Local normalization: Uncovering correlations in non-stationary financial time series. *Physica A*, 389(18):3856–3865, 2010.

[239] W A Schmitt, R M Raab, and G Stephanopoulos. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res.*, 14(8):1654–1663, 2004.

[240] A M Sengupta and P P Mitra. Distributions of singular values for some random matrices. *Phys. Rev. E*, 60:3389–3392, 1999.

[241] R E Serfling. Historical review of epidemic theory. *Human Biology*, 24(3):145–166, 1952.

[242] P Shakarian, P Roos, D Callahan, and C Kirk. Mining for geographically disperse communities in social networks by leveraging distance modularity. arXiv:1305.3668, 2013.

[243] C R Shalizi. Methods and techniques of complex systems science: An overview. In *Complex systems science in biomedicine*, pages 33–114. Springer, 2006.

[244] C I Siettos and L Russo. Mathematical modeling of infectious disease dynamics. *Virulence*, 4(4):295–306, 2013.

[245] F Simini, M C González, A Maritan, and A-L Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[246] F Simini, A Maritan, and Z Néda. Human mobility in a continuum approach. *PLoS One*, 8(3):e60069, 2013.

[247] C P Simmons, J J Farrar, N van Vinh Chau, and B Wills. Dengue. *New Engl. J. Med.*, 366(15):1423–1432, 2012.

[248] H A Simon. *The architecture of complexity.* Springer, 1991.

[249] S Sinha, A Chatterjee, A Chakraborti, and B K Chakrabarti. *Econophysics.* Wiley-VCH, 2011.

[250] S M Smith, K L Miller, G Salimi-Khorshidi, M Webster, C F Beckmann, T E Nichols, J D Ramsey, and M W Woolrich. Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891, 2011.

[251] J Q Stewart. Empirical mathematical rules concerning the distribution and equilibrium of population. *Geogr. Rev.*, 37:461–485, 1947.

[252] J Q Stewart and W Warntz. Physics of population distribution. *J. Reg. Sci.*, 1:99–121, 1958.

[253] S T Stoddard, A C Morrison, Gonzalo M Vazquez-Prokopec, Valerie Paz Soldan, T J Kochel, Uriel Kitron, J P Elder, and T W Scott. The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl. Trop. Dis.*, 3(7):e481, 2009.

[254] S A Stouffer. Intervening opportunities: A theory relating mobility and distance. *Am. Soc. Rev.*, 5(6):845–867, 1940.

[255] A Strehl, J Ghosh, and C Cardie. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, 2002.

[256] J D Tamerius, J Shaman, W J Alonso, K Bloom-Feshbach, C K Uejio, A Comrie, and C Viboud. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS Pathogens*, 9(3):e1003194, 2013.

[257] J W Tang, N Shetty, T T Y Lam, and K L E Hon. Emerging, novel, and known influenza virus infections in humans. *Inf. Dis. Clint. N. Am.*, 24(3):603–17, 2010.

[258] T Tango and K Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *Int. J. Health Geogr.*, 4(1):11, 2005.

[259] A J Tatem, S I Hay, and D J Rogers. Global traffic and disease vector dispersal. *Proc. Natl. Acad. Sci. USA*, 103(16):6242–6247, 2006.

[260] A J Tatem, D J Rogers, and S I Hay. Global Transport Networks and Infectious Disease Spread. In A Graham S I. Hay and D J Rogers, editors, *Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications*, volume 62 of *Advances in Parasitology*, pages 293–343. Academic Press, 2006.

[261] WHO Ebola Response Team. Ebola virus disease in west africa – the first 9 months of the epidemic and forward projections. *New Engl. J. Med.*, 371(16):1481–1495, 2014.

[262] C Thiemann, F Theis, D Grady, R Brune, and D Brockmann. The structure of borders in a small world. *PLoS One*, 5(11):e15422, 2010.

[263] W W Thompson, D K Shay, E Weintraub, L Brammer, N Cox, L J Anderson, and K Fukuda. Mortality associated with influenza and respiratory syncytial virus in the United States. *J. Am. Med. Assoc.*, 289(2):179–186, 2003.

[264] G Tibély, L Kovanen, M Karsai, K Kaski, J Kertész, and J Saramäki. Communities and beyond: Mesoscopic analysis of a large social network with complementary methods. *Phys. Rev. E*, 83:056125, 2011.

[265] M Tizzoni, P Bajardi, A Decuyper, G Kon Kam King, C M Schneider, V Blondel, Z Smoreda, M C González, and V Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS Comput. Biol.*, 10(7), 2014.

[266] C A Torre. *Deterministic and Stochastic Metapopulation Models for Dengue Fever.* PhD thesis, Arizona State University, 2009.

[267] V A Traag, P Van Dooren, and Y Nesterov. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E*, 84:016114, 2011.

[268] A L Traud, E D Kelsic, P J Mucha, and M A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM Review*, 53(3):526–543, 2011.

[269] A L Traud, P J Mucha, and M A Porter. Social structure of Facebook networks. *Physica A*, 391(16):4165–4180, 2012.

[270] J Truscott and N M Ferguson. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS Comput. Biol.*, 8(10):e1002699, 2012.

[271] J Truscott and N M Ferguson. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS Comput. Biol.*, 8(10):e1002699, 2012.

[272] A A Tsonis, G Wang, K L Swanson, F A Rodrigues, and L F Costa. Community structure and dynamics in climate networks. *Climate Dynamics*, 37(5–6):933–940, 2011.

[273] C Viboud, O N Bjørnstad, D L Smith, L Simonsen, M A Miller, and B T Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451, 2006.

[274] C J Villabona-Arenas and P M de Andrade Zanotto. Worldwide spread of dengue virus type 1. *PLoS One*, 8(5):e62649, 2013.

[275] L Villar, G H Dayan, J L Arredondo-García, D M Rivera, R Cunha, C Deseda, H Reynales, M S Costa, J O Morales-Ramírez, G Carrasquilla, L C Rey, R Dietze, K Luz, E Rivas, M Montoya, M Consuelo, M Cortés Supelano, B Zambrano, E Langevin, M Boaz, N Tornieporth, M Saville, and F Noriega. Efficacy of a tetravalent dengue vaccine in children in Latin America. *New Engl. J. Med.*, 372(2):113–123, 2015.

[276] C E Vincenot and K Moriya. Impact of the topology of metapopulations on the resurgence of epidemics rendered by a new multiscale hybrid modeling approach. *Ecol. Inform.*, 6(3–4):177–186, 2011.

[277] L A Waller and C A Gotway. *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Incorporated, Hoboken, NJ, USA, 2004.

[278] S Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[279] A S Waugh, L Pei, J H Fowler, P J Mucha, and M A Porter. Party polarization in Congress: A network science approach. arXiv:0907.3509, 2012.

[280] R G Webster and E A Govorkova. H5N1 Influenza — continuing evolution and spread. *New Engl. J. Med.*, 355(21):2174–2177, 2006.

[281] A Wilder-Smith and D O Freedman. Confronting the new challenge in travel medicine: SARS. *J. Travel Med.*, 10(5):257–258, 2003.

[282] A G Wilson. A statistical theory of spatial distribution models. *Transport. Res.*, 1(3):253–269, 1958.

[283] World Health Organisation Global Health Observatory Data Repository. Cause - specific mortality, 2008, 2008. `http://www.who.int/gho/mortality_burden_disease/global_burden_disease_DTHInc_2008.xls`.

[284] Y Xia, O N Bjornstad, and B T Grenfell. Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *Am. Nat.*, 164(2):267–281, 2004.

[285] Z Yao, J Tang, and F B Zhan. Detection of arbitrarily-shaped clusters using a neighbor-expanding approach: A case study on murine typhus in south texas. *Int. J. Health Geogr.*, 10(1):23, 2011.

[286] A Yeh, D Lin, H Zhou, and C Venkataramani. A multivariate exponentially weighted moving average control chart for monitoring process variability. *J. Applied Statistics*, 30(5):507–536, 2003.

[287] A Zalesky, A Fornito, and E T Bullmore. On the use of correlation as a measure of network connectivity. *NeuroImage*, 60(4):2096–2106, 2012.

[288] X Zhang, T Zhang, A A Young, and X Li. Applications and comparisons of four time series models in epidemiological surveillance data. *PLoS One*, 9(2):e88075, 2014.

[289] G K Zipf. The P1 P2/D hypothesis: On the intercity movement of persons. *Am. Sociol. Rev.*, 11:677–686, 1946.