Identifying Unbiased Meso-scale Structures in Spatial Networks

Alison Peard

St Peter's College

University of Oxford

A thesis submitted for the degree of

MSc Mathematical Modelling and Scientific Computing

September 2021

Firstly, I would like to thank my supervisors Prof. Jim Hall and Prof. Renaud Lambiotte, for helping me to formulate the subject of this dissertation and for all of their guidance and interesting discussions throughout. I would also like to express my deep gratitude to Jasper and Rodrigo, whose input was invaluable and helped to shape this dissertation from start to finish.

Secondly, I would like to thank our course director, Dr Kathryn Gillow, for the tremendous amount of work she has done for us this year, which has helped to make the MMSC a very welcoming and pleasant community to be a part of.

On a more personal note, I would like to thank my parents for their continued support and encouragement during an exceptional year, and finally, I would like to thank the MMSC cohort, along with my housemates Nelli, Miriam, and Ida for making this year one to remember despite the best efforts of a global pandemic.

The identification of *meso-scale structures* such as *community* or *core-periphery* structures in spatial networks can assist researchers in areas such as assessing the efficiency and resilience of transport networks, understanding the origins of uneven development in trade networks, and identifying evidence-based cultural and economic boundaries in human mobility networks. When classical network science methods are used to identify meso-scale structures in spatial networks, the results are often heavily biased by space. The result is a 'masking' effect, and other useful information contained in the network remains undetected. As new technologies contribute to the global bank of geolocation data, there is a growing interest in the development of algorithms that are suitable for spatial networks.

This dissertation makes two main contributions to this field. First, we address the community detection algorithm, which identifies groups of densely connected nodes in an otherwise sparse network. We extend this measure to directed networks, and also incorporate a 'fine-tuning' step that we find improves prediction accuracy on test networks. We also study a pre-processing step that allows one to remove spatial bias from a network *before* meso-scale methods are applied, and apply this to both the community detection and the core-periphery detection problems. Core-periphery structures consist of a densely connected core and a sparse periphery, and have not yet been addressed to any significant extent in spatial networks. We apply these methods to a maritime container ship network of 1433 ports, which makes an excellent candidate for this analysis due to its strong spatial element. We find that spatial-correction in the community detection algorithm uncovers significant trade routes which were not identified by the standard algorithm, and spatial-correction in core-periphery detection is able to highlight the more regional roles played by some ports, while the uncorrected method takes a broader perspective.

Contents

Chapter 1

Introduction

1.1 Motivation and problem statement

A *graph* is a mathematical object, which can be used to make sense of complex data by abstracting it into a set of pairwise interactions, denoted by *nodes* and *edges*. Spatial graphs are equipped with an additional layer of information; each node is embedded in a particular location. Many familiar networks, such as transport networks, human mobility, and social networks, and even neural networks, are spatial networks [5, 16], and often space has warped the topology of these networks in some way. In transport networks, the cost of fuel and equipment discourages long connections. This results in edge length having a truncated distribution. A similar logic applies to the brain, regions that are closer spatially are more likely to be well-connected due to the cost associated with axon length [5]. When the propensity of a pair of nodes to form a link is affected by the distance between them, the bias that this introduces to classical network science methods often renders their results as trivial [16].

A well-studied example is that of the community detection problem. Community is a *meso-* or intermediate-scale structure, where densely connected groups of nodes exist in an otherwise sparsely connected network [39]. In spatial networks, classical community detection algorithms, such as Newman-Girvan modularity optimisation [17], regularly isolate groups of spatially-proximate nodes, and other information contained in the network structure goes undetected [16]. Several papers have proposed solutions to correct this spatial bias, most using variations of the popular gravity law [16, 9, 42, 40, 37, 49], often used in geography and economics. An optimal method to correct for space, however, has not yet been agreed upon. Or indeed, as in classical community detection [39], an optimal method may not exist, and correcting for spatial bias in a network may need to be carried out on a more case-specific basis, carefully

considering the specific spatial identity of a network before choosing a suitable spatial correction method.

Most of the literature on spatial networks tends to focus exclusively on community structure [16, 9, 37, 42, 40], and how space affects other meso-scale structures such as *core-periphery* structures has not been explored in much depth. Existing methods have been applied to a wide variety of networks, primarily a Belgian mobile phone network [16], dengue fever in Peru [37], a retail network [10], census data [40], and a terrorist network [42]. However, to date, their implementation on a spatial transport network has not been studied. In this dissertation, we review, synthesise, and extend existing methods for the identification of meso-scale structures in spatial networks. We use a transport network of containerships with a strong spatial element as an example throughout, but the technique can also be easily applied to other spatial networks and applications.

State of the art A wide variety of meso-scale structures can be found in networks, but by far the most popular, and the most researched, is that of *community structure*. Classical community detection seeks to divide a network of undirected edges into communities such that the maximum number of intra-group edges is obtained [39]. Networks with community structure are also referred to as *modular*, or *assortative*. Community detection is performed by (approximately) optimising a quality function known as the *modularity function*, where the significance of community structure is assessed via comparison against some unstructured random graph, known as a *null model*. The standard (Newman-Girvan) function uses the configuration model as a null model [31]. Directed extensions to the modularity problem exist [27], and the two most popular methods to optimise this function are the Louvain algorithm [6] and spectral optimisation [35]. These methods are effective for aspatial graphs [6, 35]. In 2011, Expert *et al.* proposed a spatial-modification of these problems which introduced a variation of the gravity law [10, 49], into the modularity function. There are now myriad papers exploring implementations and extensions of Expert *et al.*'s method. In 2016, Sarzynska *et al.* modified the problem by replacing the gravity law with the radiation model of Simini *et al.* [37, 43]. Leal Cervantes developed a set of procedures to *constrain* spatial models such that they are *dimensionally consistent*, as per Wilson's 1970 definitions of constrained gravity models. Namely, they share more physical characteristics with the empirical data [10, 49, 29]. Liu, Murata, and Wakita [29] incorporated these models into the modularity function in an undirected setting. While community is the most researched meso-scale structure for spatial

networks, all variations of the methods have not been exhausted, and in particular, many choices for a suitable spatial null model have yet to be considered.

Classical core-periphery detection is a lesser-explored class of meso-scale structure, where an undirected network is divided into a densely interconnected core and a sparsely connected periphery. The network science definition and algorithm was first formalised by Borgatti and Everett in 2000 [7]. In 2018, Kojaku and Masuda proved that it is impossible to detect a discrete core-periphery structure consisting of a single core and periphery when using the classical configuration model as a null model, and proposed an algorithm to detect *multiple* core-periphery pairs using the Erős-Rényi null model [21]. This paper was followed up by a variation that was able to detect multiple core-periphery pairs using the configuration model [22]. In 2019, Elliot *et al.* proposed a directed extension of the aspatial core-periphery model, where both the core and periphery were divided into nodes that predominantly received or sent out links. This provided new insights into the different structural roles that nodes may play within the core or the periphery.

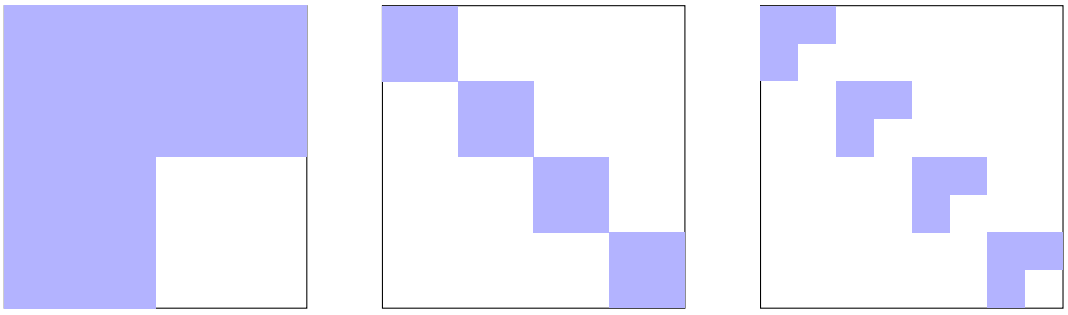


Figure 1.1: **Adjacency matrices exhibiting different meso-scale structure**. In (a) we see global core-periphery structure, in (b) global community structure, in (c) global community structure with local core-periphery structure.

Instead of modifying a particular community detection method, Leal Cervantes proposed a pre-processing method to remove spatial bias from a network by extracting only the edges whose occurrence cannot be explained by the spatial organisation of a network. These are called *spatial backbones*. Modifying algorithms to apply them to spatial backbones is a simpler problem than spatially correcting entire algorithms. Spatial backbones are *signed networks*, where edges can take negative values. Thus, modifying any algorithm to work with spatial networks is then sufficient to use it in a spatially corrected setting. This is shown using stochastic block models in [10],

where stochastic block models are used with spatial backbones to unveil a range of meso-scale structures for a U.K. retail network. Stochastic block models divide nodes into a wide range of assortative and disassortative groupings but these spatial backbones have not previously been used for algorithms that seek more specific meso-scale structures. They have not yet been tested with the classical modularity function or used in any form of core-periphery detection.

Contributions of this dissertation The developments in the current literature are by no means extensive and the majority of work focuses exclusively on community structure. In this dissertation, we aim to synthesise many of the advances in the literature and build a broader picture of modern spatially-corrected methods for both community and core-periphery structures. The dissertation makes the following contributions:

We use the developments of Leal Cervantes in the construction of dimensionally consistent gravity and radiation null models [10], and apply these directly to the modularity optimisation problem. To the best of our knowledge, dimensionally constrained models have only been used in this way by Liu, Murata, and Wakita [29], and with a slightly different formulation to what is used here. We find using these dimensionally constrained null models improves on the results of Expert *et al.* in [16].

Spatial correction by directly modifying an algorithm can be a laborious task, and successful results are not guaranteed. Spatially correcting the Louvain method [40, 42], for example, involves modifying multiple steps of the algorithm, and computational constraints reduce the applicability from networks of 118 million nodes [6] to networks of only 100 thousand nodes. The pre-processing step developed by Leal Cervantes allows the spatial-correction framework to be generalised to different methods in a more straightforward manner. In this dissertation, we consider a methodology to do this for both community detection and core-periphery detection by using signed modifications of existing algorithms [44, 13, 27]. For community detection, we find these methods to be effective, but with limitations for sparser networks. As an exercise, we also study the construction of a novel spatial null model, inspired by the work of Kosowska-Stamirowska and Zusanna in [23] on shipping networks, which we call the common neighbours+sea distance null model.

The core-periphery method we modify is a recently published extension of the classical problem for directed networks [13] based on maximising the likelihood of a discrete block structure. We account for spatial bias in this algorithm by implementing it on the spatial backbones. To verify our methods, we extend the directed core-

periphery synthetic network proposed by Elliot *et al.* [13] to include spatial effects, and use this for testing our methods before analysing their results on the maritime shipping network.

Dissertation outline The rest of this dissertation is organised as follows: the remainder of this chapter will formalise the notation to be used throughout the dissertation, cover classical community detection theory, and introduce the maritime shipping dataset to which we apply our methods. In Chapter 2 we discuss modifying the modularity method for spatial bias. We also cover the theory of constructing dimensionally consistent null models and incorporating them into the modularity function or using them to extract spatial backbones. We assess the performance of these methods using two types of synthetic spatial networks with a known partition, known as *benchmarking networks*, before implementing them on the maritime shipping network. In Chapter 3, we study undirected and directed core-periphery detection methods applied to the spatial backbones and test these on benchmarking networks and the container ship network. We compare these results to those of existing algorithms [7, 36, 13] on our container ship network.

1.2 Mathematical background and notation

To begin, we lay out some of the notation that will be used throughout this dissertation. In network science, there is an important distinction between networks with directed and undirected edges, and some metrics are defined differently between the two. In the context of transport networks, such as flight or shipping networks, it is preferable to use directed networks as flows cannot be assumed to be bi-directional. Throughout this dissertation, we predominantly focus on directed graphs but may occasionally need to derive a measure from the undirected setting.

Undirected graphs Formally, a graph is defined as a tuple $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of vertices or nodes and $E = \{(v_1, v_1), (v_1, v_2), ..., (v_n, v_n)\}$ is the set of observed links or edges between them. We define the cardinality of set V as $|V| = n$, namely the total number of nodes in the network. We let e_{ij} represent the number of edges between nodes v_i and v_j, where we allow for multi-edges and so $e_{ij} \in \mathbb{Z}^+ \cup \{0\}$. For G undirected $e_{ij} = e_{ji}$. We let m represent the total *flow* in the network, $m = \sum_{i,j=1}^{n} e_{ij}$. We can define the set of neighbours of a node v_i by

$$\eta_i = \{v_j | \exists (v_i, v_j) \in E\}.$$

The degree of node v_i is given by k_i where

$$k_i = \sum_{j \,|\, v_j \in \eta_i} e_{ij},$$

and self-loops are counted twice. For an undirected network, summing the degrees of each node v_i yields $\sum_i k_i = 2m$. In the undirected case, this sum is known as the handshaking lemma [26]. It is convenient to represent a graph in the form of an adjacency matrix A, where each edge is represented by assigning $A_{ij} = e_{ij}$. In undirected networks, this matrix is symmetric. This allows many standard results from linear algebra relating to symmetric postive semidefinite matrices to be utilised.

For a given graph, we define a partition $\mathcal{P} = \{G_1, G_2, ..., G_k\}$ of its vertices, where each G_i is a group of nodes. When all nodes are assigned to one group $G = V$, this is the *trivial partition* $\mathcal{P}_0$, and when each node is assigned its own individual group, $v_1 \in G_1, v_2 \in G_2, ..., v_n \in G_n$ this is the *singleton partition* $\mathcal{P}_s$. We will also frequently use the notation g_i, which represents the group assignment of node v_i, i.e., $g_i = G$ if $v_i \in G$.

Directed graphs For a directed graph $e_{ij} \neq e_{ji}$, and we must explicitly define out- and in-neighbours as

$$\eta_i^{\text{out}} = \{v_j | \exists (v_i, v_j) \in E\} \qquad \text{and} \qquad \eta_i^{\text{in}} = \{v_j | \exists (v_j, v_i) \in E\}.$$

The total neighbours η_i for a node v_i in a directed graph is the union of these two sets. For a directed graph,

$$k_i^{\text{out}} = \sum_{j \,|\, v_j \in \eta_i^{\text{out}}} e_{ij} \qquad \text{and} \qquad k_i^{\text{in}} = \sum_{j \,|\, v_j \in \eta_i^{\text{in}}} e_{ij}$$

and the total degree is defined as $k_i = k_i^{\text{out}} + k_i^{\text{in}}$. The total flow for a directed graph is then $\sum_i k_i^{\text{out}} = \sum_i k_i^{\text{in}} = m$. The adjacency matrix of a directed graph is no longer symmetric and so the task of extending derivations from undirected to directed graphs is an active field of research [27, 13, 38]. In the case where we wish to apply undirected methods to directed graphs, we may symmetrise the graphs by redefining the adjacency matrix as $\bar{A} = \frac{1}{2}\left(A + A^\top\right)$. This naturally leads to a loss of information but will suffice in many cases.

Spatial graphs A spatial graph is a graph where each node is embedded into a spatial location, geographic or otherwise. This is usually represented by a co-ordinate vector allowing for the pairwise distance between nodes to be calculated. The notation varies slightly in this context and is more specific to what different spatial metrics represent. The adjacency matrix, in particular, may also be called a flow matrix or an origin-destination (OD) matrix and represented by T_{ij}. The out-and in-degrees for a node v_i, usually denoted by k_i^{out} and k_j^{in} are called out- and in-flows, O_i and D_i. We will use this notation of T_{ij}, O_i and D_j when we are discussing a result specific to spatial networks, but otherwise we will adhere to the more general network science notation of A_{ij}, k_i^{out} and k_j^{in}.

1.3 Methods for classical community detection

1.3.1 Newman-Girvan modularity

The classic modularity optimisation problem aims to find groups of nodes such that the nodes within the groups are most densely connected and the connections between different groups are sparse [26, 39]. This measure is classically defined for an undirected graph, but the extension to a directed graph is discussed later in this section. The modularity statistic, Q, sums the number of intra-group links and calculates their significance compared to the expected number of such links under a random graph without community structure. In the random graph, links are random variables with probabilities distributed according to some generative model. This random graph is referred to as the null model, as we are testing a null hypothesis that our empirical graph is also such a random graph. The standard random graphs are the Erdős-Rényi and configuration random graphs [14, 31].

The classical Newman-Girvan modularity uses the configuration random graph model [17]. According to this, the expected number of edges between two nodes, v_i and v_j, with respective degrees k_i and k_j, is $\frac{k_i k_j}{2m}$ [26]. Considering a partition $\mathcal{P}$ of our network, we arrive at the classic modularity equation

$$Q = \frac{1}{2m} \sum_{G \in \mathcal{P}} \sum_{i,\, j \in G} \left(A_{ij} - \frac{k_i k_j}{2m} \right).$$

The null model preserves the total flow of the original network, so for the trivial partition $\mathcal{P}_0$, $Q = 0$. There is also the option to include a resolution parameter, γ in the form

$$Q = \frac{1}{2m} \sum_{G \in \mathcal{P}} \sum_{i,\, j \in G} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right). \tag{1.1}$$

Decreasing γ will prefer fewer, larger communities and vice versa. The directed extension of Q is given by

$$Q = \frac{1}{m} \sum_{G \in \mathcal{P}} \sum_{i,\, j \in G} \left(A_{ij} - \gamma \frac{k_i^{\text{out}} k_j^{\text{in}}}{m} \right). \tag{1.2}$$

The modularity measure can be generalised by considering different null models. We denote a general null model representing the expected number of the links between nodes as $\hat{A}_{ij}$. From this, we can write a generalised modularity measure as

$$Q = \frac{1}{m} \sum_{G \in \mathcal{P}} \sum_{i,\, j \in G} \left(A_{ij} - \gamma \hat{A}_{ij} \right). \tag{1.3}$$

Equation 1.3 is used to introduce spatial-correction to the modularity problem in Chapter 2.

Optimisation of the modularity function Optimising the modularity function has been proven to be NP-hard [26] and many algorithms for finding approximate solutions have been developed. For our purposes, we will use spectral partitioning, which relies on the symmetry of the modularity matrix $B = A - \gamma \hat{A}$. The form of spectral partitioning appflied to the modularity matrices in this dissertation is the bi- and tripartitioning methods developed by Richardson, Mucha, and Porter [35], and we give an overview of some of the key concepts of these methods in Appendix A.1.

To extend the optimisation problem to directed networks, where the asymmetry of the modularity matrix causes technical issues, we follow the procedure proposed by Leicht and Newman [27], who restore symmetry to the modularity matrix by noting that since Q is a scalar, it is equal to its transpose. Thus, $Q = \frac{1}{2}(Q + Q^\top) = \frac{1}{4m} s^\top (B + B^\top) s$ and the new quantity to be optimised is the symmetric matrix $(B + B^\top)$. Unlike simply symmetrising the adjacency matrix, this approach does not discard information about edge directions [27].

1.3.2 Motivating example

Throughout this report, we implement our methods on a network of container ship journeys. The network contains data for the years 2019 and 2020 obtained from Automatic Identification System (AIS) data, which records the dynamics (e.g. location, speed, direction) and statics (e.g. type of vessel, length) of all ongoing maritime vessels above 300 GT. Nowadays more than 100,000 maritime vessels, covering the vast majority of the maritime fleet in terms of tonnage, have an AIS transponder, and

hence are included in the data. The years 2019 and 2020 are particularly interesting to study due to disruption suffered as a consequence of Covid-19 [48, 45]. The container network also has a strong directional component [18], where routes taken by vessels tend to be circular, rather than 'back-and-forth'.

The processed network consists of 1433 ports and their pairwise flow counts. The approximate sea distances between ports were calculated over a network of possible routes (Figure 1.2), weighted by distance in kilometers, using the Djikstra shortest path length [12]. The data is separated into networks for 2019 and 2020 and any nodes with zero incoming or outgoing flows are removed, leaving 950 nodes in the 2019 network and 1006 nodes in the 2020 network.

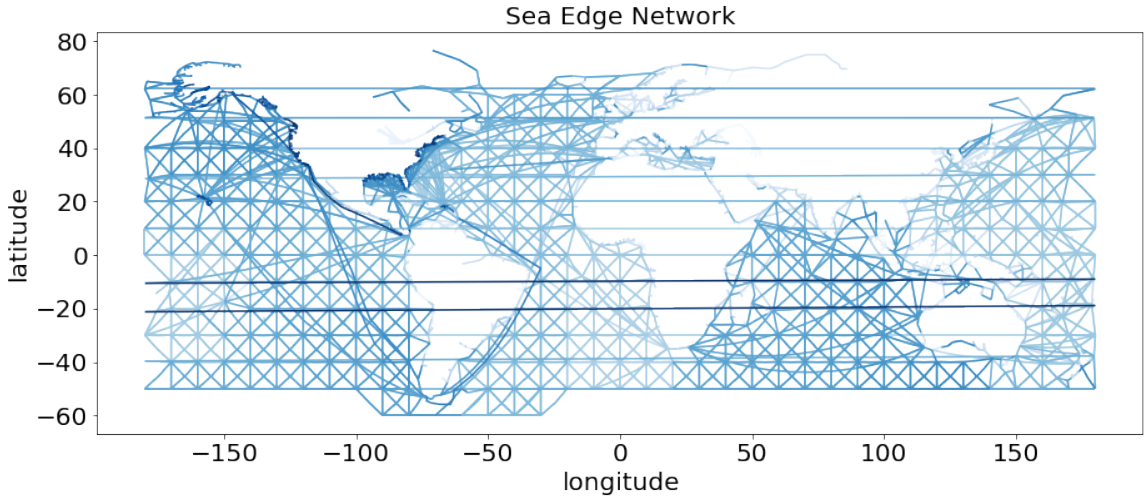
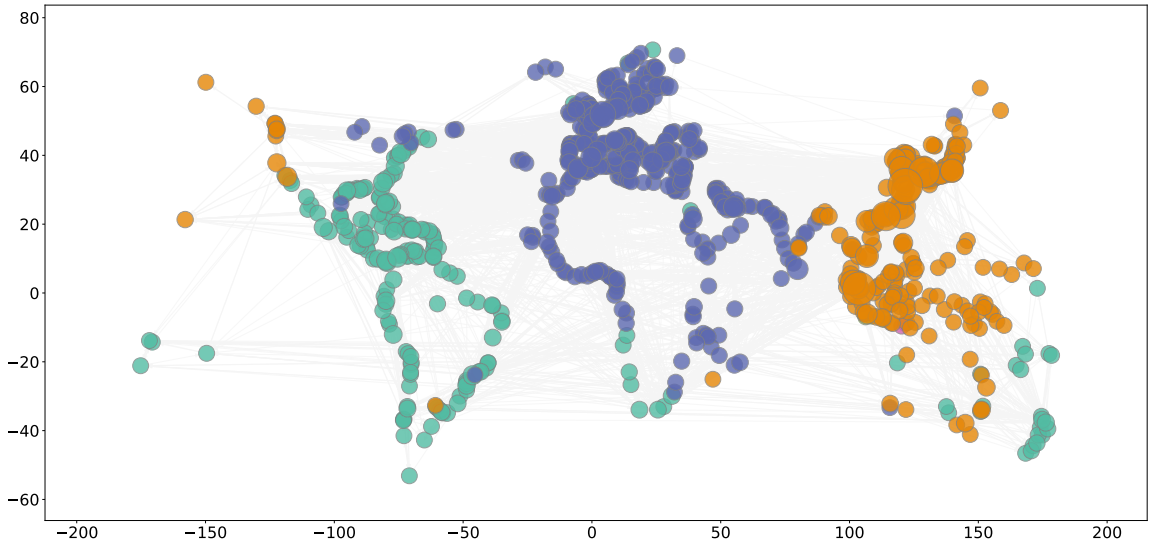


Figure 1.2: **Network of sea edges** over which the Djikstra shortest path length was calculated to find approximate sea distances between all pairs of ports.

Results Changing the resolution parameter γ in (1.3) allows one to 'zoom' through different granularities of communities detected [26]. For higher values of γ, more communities will be identified and at lower values these communities will be collected into larger groups, giving a broad overview of a network's structure. In the main text, we focus on the resolution $\gamma = 0.2$ as this produces sufficiently large communities to enable us to conduct a broad-scale analysis, we also limit our focus to the 2019 network but some supplementary figures, results, and comments for different resolutions and years are included in Appendix A.2.

Figure A.5 shows the results of classical community detection using the Newman-Girvan modularity function on the 2019 container ship network. Nodes are plotted in their spatial locations and community assignment is denoted by colour.



(a) 2019, $\gamma = 0.2$

Figure 1.3: **Visualisation of communities detected using Newman-Girvan modularity in 2019 shipping network.** Ports are shown in their spatial locations with community assignment denoted by colour. Community detection is implemented using Newman-Girvan modularity with resolution $\gamma = 0.2$ and optimised using spectral partitioning [35].

The Newman-Girvan modularity primarily assigns nodes to communities based on continent. There were just five communities found in the 2019 network with 99% of nodes divided between three major groups, North and South America, Europe and Africa, and Asia and Oceania. This pattern is consistent across multiple resolutions. The only deviation we see from this occurs for some South African ports and ports from Oceania which are grouped with the American continent. South-South and intra-Asian containerised flows accounted for 39.9% of containerised trade in 2019 [45], which is possibly what the model is picking up here. Just six Indonesian ports were divided between two groups. Of these, four are the Indonesian ports of Cirebon, Pelabuhan Ratu Coal Power Plant, Probolinggo and Tanah Merah. The other two ports are Ende and Waingapu.

Chapter 2

Spatially-Corrected Community Detection

In this chapter, we consider methods by which we can adapt the classical community detection problem to a spatial network in an unbiased manner. Many community detection algorithms are built on the assumption that the growth of a graph was determined by some generative or null model, such as the configuration model in the Newman-Girvan modularity function [31, 17]. Spatially corrected community detection involves altering this null model to incorporate space, i.e. incorporating a *spatial null* or *mobility* model.

Throughout this section, we focus on directed, spatial networks, and we use the conventional notation for spatial networks. That is, the matrix T used to represent the origin-destination (OD) matrix for the network, and out- and in-flows for a node v_i are represented by the vectors O_i and D_i. We default to this notation forthwith, unless we are explicitly deriving a more general network science result.

For a great deal of this chapter, we make use of the work by Leal Cervantes in [10], which collects a number of useful results for spatial networks from different sources, including a procedure for tuning the parameters of the gravity model, and proposes a methodology for constructing spatial null models which share certain dimensional characteristics with the empirical network. We use these null models to adapt the classical modularity community detection algorithm. Similar work has been done in the past by Liu, Murata, and Wakita [29], but for undirected networks, and with a slightly different constrained gravity model. As we shall see shortly, tuning the parameters of the gravity model improves on the results of Expert *et al.*

In Section 2.1, we give a detailed overview of Leal Cervantes' construction of the gravity model and how it may be modified such that the out- and in-flows are preserved by the model, creating a *dimensionally-constrained* model. We then more

briefly discuss the extension of this to the radiation model in Section 2.1.2, though a more thorough discussion can be found in the original paper [10]. In Section 2.1.3 we present our proposed mobility model, the common neighbours+sea distance model, which is inspired by maritime shipping networks, and discuss how it may be dimensionally constrained using the same methodology. We use these three models as null models for all subsequent developments in this dissertation.

In Section 2.2 we cover the *spatial backbone* methodology developed by Leal Cervantes for removing edges whose occurrence in a network is not statistically significant relative to some spatial null model [10]. These spatial backbones can then be used for community detection as part of a two-step procedure. In the Results 2.3 the methods covered in this section will be tested using two synthetic spatial networks proposed by Expert *et al.* and Cerina *et al.* [16, 9] before we use them to perform community detection on the container ship network.

2.1 Dimensionally-constrained mobility models

General set-up We begin by presenting the models which will be used as spatial null models for all spatially-corrected meso-scale structure detection in this dissertation. We utilise the following set up throughout: we assume that the number of trips T_{ij} between nodes v_i and v_j follows a binomial distribution $T_{ij} \sim \text{Bin}(X, p_{ij})$. Here, X is either the total number of flows originating from node v_i, O_i, if we are in a production-constrained setting, or else the total number of flows arriving at v_j, D_j if we are in an attraction-constrained setting. The probabilities p_{ij} are generated using a suitable mobility model. The expected flow between v_i and v_j is thus given by the expectation of the binomial distribution: $\hat{T}_{ij} = Xp_{ij}$. In this section, we discuss three mobility models, the gravity model in Section 2.1.1, the radiation model in Section 2.1.2 and our common neighbours+sea distance model in Section 2.1.3, though the logic of this formulation can be extended to any reasonable mobility model.

2.1.1 The gravity model

The gravity model has been used for decades in the social sciences [8], geography [15] and economics [2] to model the intensity of interaction between two entities, separated by some measure of distance. In 1970, a paper by Wilson [49] showed how the gravity model, which gets its name from the Newtonian principles upon which it is based, is not a single model, but actually a whole family of spatial interaction models. Wilson introduces additional constraints to the classical model in order to ensure either the

out- or in-flows of the observed data, or both, are conserved. This yields four members of the gravity model family, the *unconstrained* (standard) model, the *production*- and *attraction*-constrained models, and the *doubly*-constrained model.

In its simplest form, the gravity model predicts the volume of interaction between two vertices v_i and v_j with the *affinity function*

$$\psi_{ij} = O_i^{\alpha} D_j^{\beta} f(d_{ij}) \tag{2.1}$$

where d_{ij} is the distance between the two vertices, α and β are parameters which may be chosen arbitrarily, or optimised, and $f(\cdot)$ is a distance decay function which can take various forms, which we discuss later. The out- and in-flows O_i and D_j are used to measure the importance of the nodes v_i and v_j. Alternative measures are often used in place of O_i and D_j, such as trade volume in economic settings [23], or population in geographic settings [16]. As in [10], we use out- and in-flows, which keeps things simple. Using (2.1), a matrix of pairwise affinities between nodes Ψ with entries $(\Psi)_{ij} = \psi_{ij}$ may be generated. This matrix is symmetric in the undirected case if, and only if, $\alpha = \beta$.

The one-step method for community detection In the past [16, 42, 40, 9], spatially-corrected community detection has been performed by replacing the null model, denoted $\hat{T}$ in the spatial context, in the modularity function (1.3) with a variation of the gravity model affinity function (2.1). Various forms for the distance decay function $f(\cdot)$ have been considered. Expert *et al.* compute a weighted average for the probability that a link exists based on the data [16]

$$f(d) = \frac{\sum_{i,j \,|\, d_{ij}=d} T_{ij}}{\sum_{i,j \,|\, d_{ij}=d} k_i k_j}. \tag{2.2}$$

where k_i is the degree of node v_i. In this dissertation, we use the simple inverse power function

$$f(d) = d^{-\ell} \tag{2.3}$$

where ℓ is a tunable parameter, which is used by [10].

Inserting (2.1) or any of its variations into (1.3), the final form of the gravity modularity measure becomes

$$Q^{\text{gravity}} = \frac{1}{m} \sum_{C \in \mathcal{P}} \sum_{i,\, j \in C} \left(T_{ij} - \gamma \hat{T}_{ij}) \right) \tag{2.4}$$

where $\hat{T}$ is the expected network under one of our gravity models. The same logic may be used to construct a spatially-corrected modularity function using any other mobility model, such as the radiation or common neighbours+sea distance models.

Constrained gravity models

In [10], Leal Cervantes develops a rigorous framework for constructing null models which are dimensionally consistent as per the 1970 definitions of Wilson [49]. The rest of this section and Section 2.1.2 are adapted directly from this work.

The unconstrained gravity model The unconstrained gravity model is constructed such that total expected flow is equal to that in the observed network. This is achieved by introducing a weighting Z such that $Z\sum_{i,j}\hat{T}_{ij} = m$, i.e. $Z = m/\sum_{i,j}\hat{T}_{ij}$, where m is the toal flow in the network. Thus the unconstrained gravity model $\hat{T}^{\mathrm{UC}}$ takes the form

$$\hat{T}_{ij}^{\mathrm{UC}} = Z\frac{O_i^{\alpha}D_J^{\beta}}{d_{ij}^{\ell}}. \tag{2.5}$$

Production and attraction-constrained gravity models For a null model $\hat{T}$ to be dimensionally-consistent in a *production*-constrained setting then the sum of the out-flows $\sum_j \hat{T}_{ij}^{\mathrm{PC}}$ must satisfy

$$\sum_j \hat{T}_{ij}^{\mathrm{PC}} = O_i$$

for all nodes v_i. In an *attraction*-constrained setting, the in-flows must satisfy

$$\sum_i \hat{T}_{ij}^{\mathrm{AC}} = D_j$$

for all nodes v_i. In the *doubly*-constrained setting, both of these conditions must hold.

To create a production-constrained gravity null model, we construct the probabilities $p_{ij}^{\mathrm{grav}} = \psi_{ij}/\sum_k \psi_{ik}$. The rows of the probability matrix (p_{ij}^{grav}) sum to unity so the matrix (p_{ij}^{grav}) is row-stochastic. If we set $\hat{T}_{ij} = O_i p_{ij}^{\mathrm{grav}}$ then the production condition is satisfied. Substituting the gravity affinity function, ψ_{ij} given by Equations (2.1) and (2.3) into this allows us to cancel O_i^{α}, and we obtain the production-constrained model $\hat{T}^{\mathrm{PC}}$

$$\hat{T}_{ij}^{\mathrm{PC}} = O_i\frac{D_j^{\beta}/d_{ij}^{\ell}}{\sum_k D_k^{\beta}/d_{ik}^{\ell}}. \tag{2.6}$$

Analogously, to create an attraction-constrained gravity null model, we construct the column-stochastic probability matrix (p_{ij}^{grav}), $p_{ij} = \psi_{ij}/\sum_k \psi_{kj}$. If we set $\hat{T}_{ij} = p_{ij}^{\mathrm{grav}} D_j$ then the attraction condition is satisfied. Substituting the gravity affinity function, Equation (2.1) into this allows us to cancel D_j^{β} and we obtain the attraction-constrained model $\hat{T}^{\mathrm{AC}}$

$$\hat{T}_{ij}^{\mathrm{AC}} = D_j\frac{O_i^{\alpha}/d_{ij}^{\ell}}{\sum_k O_k^{\alpha}/d_{kj}^{\ell}}. \tag{2.7}$$

The doubly-constrained gravity model The doubly-constrained gravity model T^{DC} takes the form [10, 49]

$$\hat{T}_{ij}^{\mathrm{DC}} = (A_i O_i) \times \psi_{ij} \times (B_j D_j). \tag{2.8}$$

This formulation has the additional requirement that we solve iteratively for two additional balancing factors A_i and B_j. With a small bit of algebra it can be shown that

$$A_i = \frac{1}{\sum_j \psi_{ij}(B_j D_j)} \qquad B_j = \frac{1}{\sum_i (A_i O_i)\psi_{ij}}. \tag{2.9}$$

The balancing parameters, A_i and B_j are solved using the iterative proportional fitting procedure, which has guaranteed convergence for closed systems [10]. Hence, for this constrained setting we require that there be no zero entries of O_i or D_j. From (2.9), it can be shown that multiplying the rows or columns of $\hat{T}_{ij}^{\mathrm{DC}}$ by a constant will not affect its final form, hence the O_i^{α} and D_j^{β} terms can be neglected and the only parameter which needs to be specified is ℓ.

Constraint type	Relevant params.	Affinity function	Expected matrix
Unconstrained (UC)	(α, β, ℓ)	$\psi_{ij} = {O_i}^{\alpha} {D_j}^{\beta} {d_{ij}}^{-\ell}$	$\hat{T}_{ij} = Z\psi_{ij}$
Production (PC)	(β, ℓ)	$\psi_{ij} = {D_j}^{\beta} {d_{ij}}^{-\ell}$	$\hat{T}_{ij} = O_i \left(\psi_{ij} / \sum_k \psi_{ik}\right)$
Attraction (AC)	(α, ℓ)	$\psi_{ij} = {O_i}^{\alpha} d_{ij}^{-\ell}$	$\hat{T}_{ij} = \left(\psi_{ij} / \sum_k \psi_{kj}\right) D_j$
Doubly (DC)	ℓ	$\psi_{ij} = {d_{ij}}^{-\ell}$	$\hat{T}_{ij} = A_i O_i \psi_{ij} B_j D_j$

Table 2.1: Summary of the different constrained model types and their relevant parameters. Table adapted with permission directly from [10].

The iterative fitting procedure is inherently asymmetric and produces asymmetric matrices even for undirected networks. Thus the doubly-constrained gravity model is unsuitable for undirected networks. Symmetric fitting procedures exist [25], but for now, we consider this beyond the scope of the dissertation. The production- and attraction-constrained models are asymmetric by construction and these are also only appropriate for directed networks. For this dissertation, we restrict our focus to directed networks or occasionally symmetrise a null model where appropriate. For a more rigorous treatment of undirected networks, we refer the interested reader to the work of Liu, Murata, and Wakita [29].

Calculating the parameters Finding the optimal parameters to use for a network is an optimisation problem, where we seek to minimise a cost function of the difference between the empirical and predicted adjacency matrices. The parameters for the unconstrained problem can be found by solving the linear least squares problem

$$K + \alpha \log O_i + \beta \log D_j - \ell \log d_{ij} = \log T_{ij}$$

where we solve the overdetermined system of $|\Omega| = |\{(i,j)|T_{ij} > 0\}|$ equations for four unknowns (K, α, β, ℓ). In the production- or attraction-constrained setting, this is no longer possible due to the terms in the denominators of (2.6) and (2.7). Instead, the loss function

$$D(\beta, \ell) = \frac{1}{2} \sum_{i,j \in \Omega} \left(\hat{T}_{ij} - T_{ij} \right)^2. \tag{2.10}$$

can be minimised.

For the doubly-constrained model, the calculation of the balancing factors is included as part of the cost-function and the parameter ℓ is optimised while A_i and B_j are also solved for at each iteration.

2.1.2 The radiation model

The radiation model is a member of the intervening-opportunities (IO) family of mobility models [10, 43]. Unlike the gravity model, the link probability from node v_i to v_j does not depend explicitly on the distance d_{ij} but rather on the number of alternative, closer destinations that are available to a journey leaving v_i. The more of these that exist, the less likely a journey is to continue all the way to node v_j. The authors of the radiation model claim it has a number of advantages over, and resolves a number of the limitations of the gravity model, which are listed in [43]. In particular, it has no free parameters that need to be arbitrarily chosen or fitted to the data. It also only depends implicitly on the distance between nodes d_{ij}, removing the need to explicitly chose a decay function $f(\cdot)$ as in the gravity model. In [43], when applied to US census data, it far better predicts the commuter flux between Alabama and Utah, while the gravity model is wrong by an order of magnitude.

The model is formulated as follows: consider two nodes, v_i and v_j with distance between them d_{ij}. For simplicity, we consider v_i to be located at the origin. The radiation model usually uses a measure of population to represent the importance of each node, but here, we replace this with in-flows D_i at each node, as in [10]. We construct a circle of radius d_{ij} with center $(0, 0)$ and consider all other nodes v_k within

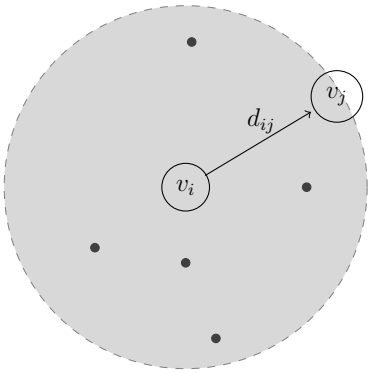


Figure 2.1: **Schematic of the radiation intervening-opportunities model.** The probability of an edge existing from node v_i to v_j depends on the number of and populations of other nodes (black dots) within the circle of radius d_{ij} that is centered at v_i.

this circle. The measure $s_{ij}^{\mathrm{IO}} = \sum_{k \neq i,j} D_k$ gives the total in-flow within this circle. The radiation model is formed as

$$p_{ij}^{\mathrm{rad}} = \frac{D_i D_j}{(D_i + s_{ij}^{\mathrm{IO}})(D_i + D_j + s_{ij}^{\mathrm{IO}})}, \tag{2.11}$$

and we let the matrix S^{IO} be the IO matrix where $(S^{\mathrm{IO}})_{ij} = s_{ij}^{\mathrm{IO}}$. In [10], these probabilities are normalised such that the matrix (p_{ij}^{rad}) is row-stochastic by dividing by $(1 - D_i/M)$. We obtain the production-constrained expectations

$$\hat{T}_{ij} = O_i \frac{D_i D_j}{(1 - D_i/M)(D_i + s_{ij}^{\mathrm{IO}})(D_i + D_j + s_{ij}^{\mathrm{IO}})}. \tag{2.12}$$

These dimensionally-constrained mobility models have been used as successful spatial null models in [10] to detect meso-scale structures by use of stochastic block models.

2.1.3 The common neighbours model

Kosowska-Stamirowska and Zusanna used machine learning techniques to study the performance of different network measures in link prediction for maritime shipping networks [23]. Among these, a gravity model correctly predicted on average 14-20% of links while a parameter-free model using only the number of common neighbours between ports was found to correctly predict 19-23% of links. The model of Kosowska-Stamirowska and Zusanna combined the common neighbours measure with sea distances and slightly improved on this, correctly predicting up to 24% of links. The improvement was observed for container carriers, bulk carriers, general cargo ships,

and petroleum tankers. The authors suggest that this phenomenon could be justified by the tendency to 'shortcut routes' in the shipping industry. If there is a high volume of trade between two ports, creating a direct route will increase efficiency. Motivated by this result, we suggest a novel spatial null model for maritime networks, the common neighbouts+sea distance null model, where the probability of a link between two nodes depends on the number of flows taking indirect paths between them, augmented by their pairwise (sea)-distance.

For a directed spatial network with OD matrix T, the number of two-step paths between two nodes, v_i and v_j, is given by T^2. If the intermediate node, v_k, has a very high degree, then its appearance as a common neighbour is less remarkable. To correct for this in undirected networks, the Adamic-Adar index is often used [1]. The Adamic-Adar index between nodes v_i and v_j is given by

$$s_{ij}^{\mathrm{AA}} = \sum_{v_k \in \eta_i \cap \eta_j} \frac{1}{\log k_k}$$

where k_k is the degree of node v_k and η_i is the set of neighbours of node v_i. To extend this to spatial, directed networks we use

$$s_{ij}^{\mathrm{AA}} = \sum_{v_k \in \eta_i^{\mathrm{out}} \cap \eta_j^{\mathrm{in}}} \frac{1}{\log(O_k + D_k)}, \tag{2.13}$$

where O_k and D_k represent the out- and in-flows for node v_k. We let the matrix S^{AA} be the Adamic-Adar matrix where $(S^{\mathrm{AA}})_{ij} = s_{ij}^{\mathrm{AA}}$. This can be calculated in matrix form as

$$S^{\mathrm{AA}} = T\tilde{D}T$$

where $\tilde{D}$ is a diagonal matrix with elements

$$(D)_{ii} = \frac{1}{\log(O_i + D_i)}.$$

To introduce distances, we construct a new affinity function

$$\psi_{ij} = \frac{(s_{ij}^{\mathrm{AA}})^{\alpha}}{d_{ij}^{\ell}} \tag{2.14}$$

where α is a tunable parameter, and once again are faced with the task of constructing dimensionally-consistent null models.

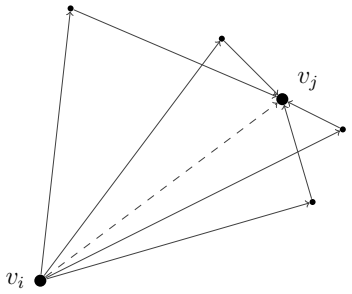


Figure 2.2: **Common neighbours intuition**: if there are many links from v_i to v_j, the model expects there to also be direct links from v_i to v_j, especially if the other nodes have low total out- and in-flows.

Dimensionally-consistent null models The construction of the unconstrained common neighbours model remains the same as for the gravity and radiation model and we find the expected flow $\hat{T}$ to be given by

$$\hat{T}_{ij}^{\mathrm{UC}} = Z\psi_{ij} = Z\frac{(s_{ij}^{\mathrm{AA}})^{\alpha}}{d_{ij}^{\ell}}$$

where $Z = m/\sum_{i,j}\psi_{ij}$ and $m = \sum_{i,j}T_{ij}$. To formulate the production-constrained model, we construct the row stochastic matrix of probabilities with entries $p_{ij}^{\mathrm{CN}} = \psi_{ij}/\sum_k \psi_{ik}$ where ψ_{ij} is as defined in Equation (2.14). Then

$$\hat{T}_{ij}^{\mathrm{PC}} = \frac{O_i(s_{ij}^{\mathrm{AA}})^{\alpha}d_{ij}^{-\ell}}{\sum_k (s_{ik}^{\mathrm{AA}})^{\alpha}d_{ik}^{-\ell}}.$$

Likewise, for the attraction-constrained model

$$\hat{T}_{ij}^{\mathrm{AC}} = \frac{(s_{ij}^{\mathrm{AA}})^{\alpha}d_{ij}^{-\ell}D_j}{\sum_k (s_{kj}^{\mathrm{AA}})^{\alpha}d_{kj}^{-\ell}}.$$

The doubly-constrained model is also formulated in the usual way [10], where

$$\hat{T}_{ij}^{\mathrm{DC}} = (A_iO_i) \times \frac{(s_{ij}^{\mathrm{AA}})^{\alpha}}{d_{ij}^{\ell}} \times (B_jD_j)$$

and A_i and B_j are balancing factors.

Fitting the parameters Since the terms O_i and D_j don't appear in the affinity matrix ψ_{ij}, we cannot perform the elegant substitutions that were used for the gravity and radiation models in [10], however, the optimisation procedure remains the same and we solve for a different set of parameters. The doubly constrained problem is now

optimised using nonlinear least squares and a proportional fitting procedure, where we solve for (α, ℓ) rather than just ℓ at each step.

To visualise how the model may be fit to the shipping data, we show the optimisation landscapes for the unconstrained (left panel) and attraction-constrained (right panel) models on container ship data for three different metrics in Figure 2.3, the loss function (2.10), a logarithmic variation of (2.10) given as

$$LD(\beta, \ell) = \frac{1}{|\Omega|} \sum_{i,j \in \Omega} (\log \hat{T}_{ij} - \log T_{ij})^2 \tag{2.15}$$

and the common part of commuters (CPC) index. The CPC index is widely used in Applied Mathematics to quantify the similarity between two sets [28]. It returns a score of 1 for two sets that match perfectly and a score of 0 for two sets that do not match at all. Minima of the loss functions are indicated by pink dots, minima of a logarithmic variation of (2.10) are indicated by turquoise triangles, and maxima of the CPC index are indicated by black squares. In Appendix B.1.1, the full set of optimisation landscapes for all three metrics is included.

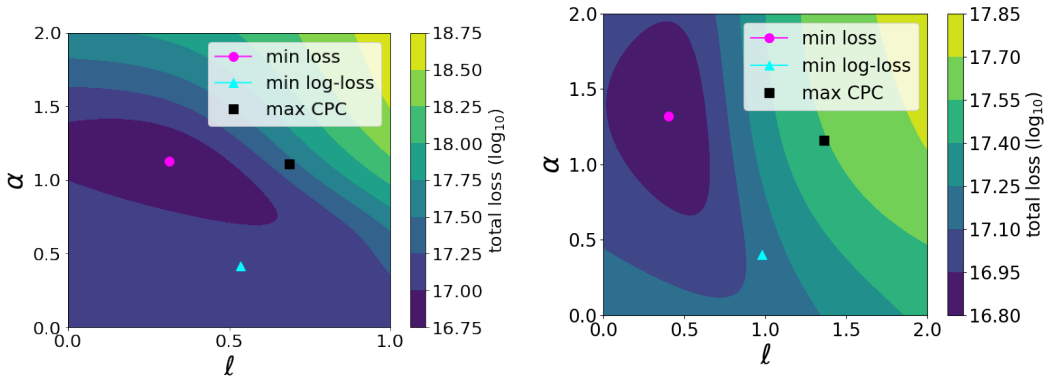


Figure 2.3: The optimisation landscapes for the unconstrained (left panel) and attraction-constrained (right panel) for the loss function (2.10). Minima of the loss functions are indicated by pink dots. Also shown are minima of (2.15) (turquoise triangles), and maxima of the common part of commuters index (black squares).

2.2 The spatial backbone problem

The first method of spatial-correction uses the spatial null models presented in Sections 2.1 to directly modify the general modularity function (1.3) in a one-step process. The spatial backbone problem presents us with an opportunity to generalise these spatially corrected community detection methods. Spatial backbone extraction is a

preprocessing step that removes spatial bias from the network *before* any community detection methods are applied. The result is a signed network, for which it is often more straightforward to modify existing meso-scale structure detection algorithms.

The goal of extracting the spatial backbone from a network is to remove those edges from the network whose occurrence can be explained by our spatial null model, be that the gravity model, radiation model, or something else. In this section, we closely follow the developments of Leal Cervantes in [10] for extracting spatial backbones from networks. If there are no other factors at play (such as meso-scale structures) in the growth of a network, we expect the edges in the spatially corrected graph to be arranged randomly. If so, we expect the graph to be largely explained by the configuration [31] or the Erdős-Rényi random graphs [13].

2.2.1 The general backbone extraction problem

The spatial backbone extraction problem begins with the null hypothesis H_0 that the observed network, T, has edges generated according to the rules of a spatial null model. Thus, with one of the mobility models from Section 2.1.1-2.1.3 as the spatial null model, the distribution of flows across the edges is given by a binomial distribution, i.e.

$$\mathrm{H}_0 : T_{ij} \sim \mathrm{Bin}(X, p_{ij}). \tag{2.16}$$

The probability mass function (pmf) for each entry of the flow matrix T taking some value $w > 0$ is

$$\mathbb{P}(T_{ij} = w) = \binom{X}{w} p_{ij}^{w}(1 - p_{ij})^{w}, \tag{2.17}$$

where p_{ij} is the probability of an edge occurring according to our spatial null model. Here, X is either the production O_i or attraction D_j vector, dependent on whether we are in a production or an attraction setting. We seek entries of the empirical flow matrix T which are significantly high or low with respect to these probability distributions. Left and right-tailed statistical tests are constructed for this. The right-tailed test is used to find edges with significantly larger flows than could be explained by our null model. The probability of observing an edge with a flow count greater than or equal to t in a graph generated by our null model is given by the p-value

$$p = \mathbb{P}(T_{ij} \geq t | H_0) = \sum_{w=t}^{X} \mathbb{P}(T_{ij} = w). \tag{2.18}$$

We extract the significant edges with p-values larger than some constant α (usually $\alpha = 0.01$) and form the network which consists of just these edges: the positive backbone Φ^+. For the left-tailed test, the probability of observing an edge of value less than t is

$$p = \mathbb{P}(T_{ij} < t | H_0) = \sum_{w=0}^{t} \mathbb{P}(T_{ij} = w), \tag{2.19}$$

and this forms the negative backbones Φ^-.

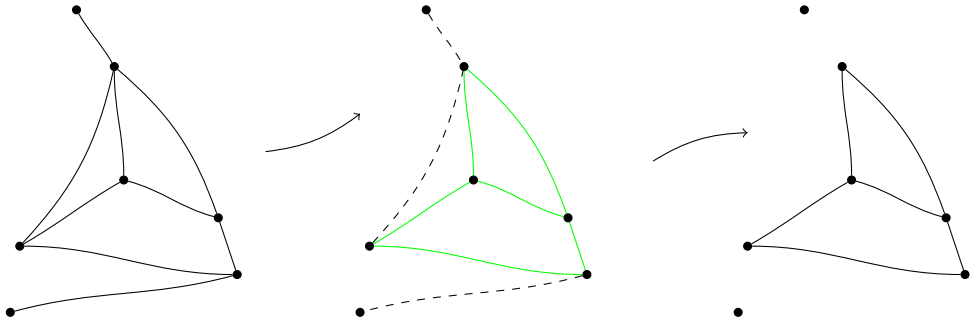


Figure 2.4: **Visualisation of the spatial backbone extraction process for the positive backbone** Green links denote positive edges which are extracted to form the positive spatial backbone (final panel).

We can combine both backbones into a single backbone

$$\Phi = \Phi^+ - \Phi^- = \begin{cases} 1 & \text{if } T_{ij} \text{ is a positive edge} \\ -1 & \text{if } T_{ij} \text{ is a negative edge} \\ 0 & \text{else} \end{cases}$$

Intuitively, the positive backbone can be understood as the set of edges in which there is a lot more flow than our mobility model predicts. This could be indicative of some auxiliary positive relationship between two nodes. In the case of a network of ports, for example, this could represent a close trading partnership due to cultural or historical ties between regions. By similar logic, the edges in the negative backbone suggest a possible negative relationship between two regions, perhaps strong cultural differences, sanctions, or historical animosity.

2.2.2 Two-step community detection

In this subsection, we consider calculating the modularity using the extracted spatial backbones of a network, a methodology not yet explored in the current literature. In this case, we need only use a signed modification of a more classical null model in our modularity function, such as the usual Newman-Girvan or the Erdős-Rényi null

models. In [44], Traag proposes that instead of directly calculating the modularity of a signed network, we split the network's adjacency matrix into one for its positive edges A^+ and one for its negative edges A^- as follows

$$A^+_{ij} = \begin{cases} A_{ij} & \text{if } A_{ij} \geq 0, \\ 0 & \text{else.} \end{cases} \qquad A^-_{ij} = \begin{cases} |A_{ij}| & \text{if } A_{ij} < 0, \\ 0 & \text{else.} \end{cases} \tag{2.20}$$

These can be used to calculate the modularities Q^+ and Q^-, respectively for each matrix. We wish to minimise the number of negative links within communities, so Traag proposes we maximise Q^+ whilst minimising Q^-. Thus, a new objective function for signed networks is formed as $Q = Q^+ - Q^-$. For a partition $\mathcal{P} = \{G_1, G_2, \ldots, G_k\}$, where $g_i = G$ if $v_i \in G$, we obtain

$$Q^{\text{signed}} = \frac{1}{m} \sum_{i,j=1}^{n} \left(A_{ij} - \left(\hat{A}^+_{ij} - \hat{A}^-_{ij} \right) \right) \delta_{g_i g_j} \tag{2.21}$$

where

$$\delta_{g_i g_j} = \begin{cases} 1 & \text{if } g_i = g_j, \\ 0 & g_i \neq g_j. \end{cases}$$

The formulation of the Newman-Girvan null model is provided in [44] so we only include the formulation for the Erdős-Rényi (ER) null model here. Letting m^+ be the total flow in the positive network and m^- the total flow in the negative network, and noting that the number of nodes is the same for both networks, we have signed ER-modularity

$$Q^{\text{signed}} = \frac{1}{m} \sum_{i,j=1}^{n} \left(A_{ij} - \frac{m^+ - m^-}{n^2} \right) \delta_{g_i g_j}. \tag{2.22}$$

From now on we refer to this method, which involves extracting the spatial backbones before optimising the modularity as the *two-step* method, and we refer to the method which involved incorporating space directly into the modularity function as the *one-step* method.

2.3 Results

2.3.1 Synthetic networks

Before we apply our methods to an empirical dataset, we must verify that they achieve their purpose. In community detection, validation is usually performed using a synthetic benchmarking network where a known community structure affects link formation. An algorithm's accuracy may be assessed by comparing its predicted

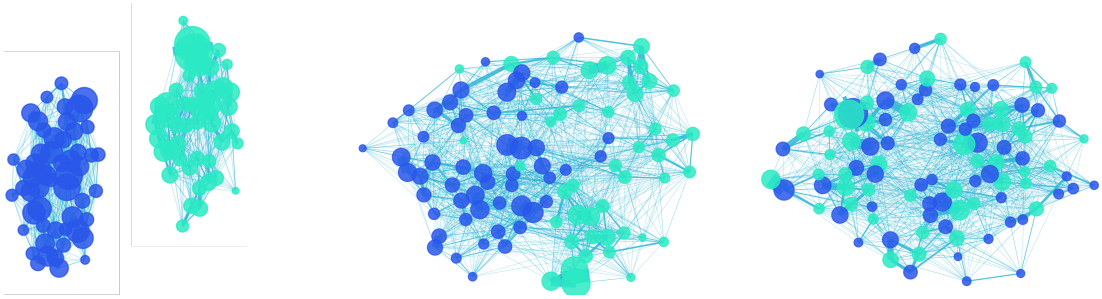
partition with the true partition [9, 16, 37]. The benchmarking methodology has the additional advantage that it allows one to explore parameter spaces in great detail, building an understanding as to what type of network an algorithm is most suited to. To quantify the similarity between the predicted partition and the true partition, we use the normalised mutual information (NMI) score, which is based on the concepts of entropy and mutual information [13, 11]. This score ranges between 0 when the partitions contain completely different information, and 1, when they are identical. A formal definition is included in Appendix B.2.

Two benchmarking networks are used by Leal Cervantes for community detection in [10]. The first, originally proposed by Expert *et al.* generates a random network where nodes are embedded in space and have random binary group assignments determined by a uniform distribution, we refer to this henceforth as the *uniform model*. In the second benchmark, proposed by Cerina *et al.* the level of correlation between space and community assignment is a tunable parameter, and we refer to this as the *correlated group membership model*. We give some details of the construction of these networks in Appendix B.3.1 and B.3.2, which are adapted from [10, 16] and [9], where more thorough discussions can be found.

The uniform model

The uniform model by Expert *et al.* [16] produces networks where attributes are randomly assigned, and *edge density* and graph assortativity are tunable parameters. Here, edge density refers to the total number of edges in the network relative to the number of nodes. Assortativity is connected to the type of community structure a graph displays. If a graph is modular, then nodes within communities tend to be densely connected, whilst connections between communities are sparse. Conversely, a bipartite graph [3], where nodes are predominantly connected to nodes from other groups, is *disassortative*. We show this type of modular community detection, naturally, only performs well for assortative graphs.

The parameters used for assortativity and edge density are, respectively, λ and ρ. For $\lambda = 0$, a graph consists of a set of fully disconnected communities with dense internal connectivity. For $\lambda < 1$ graphs are assortative, and for $\lambda > 1$ they are disassortative. Synthetic networks generated using the uniform model with $n = 100$ nodes for $\lambda = 0$, $\lambda < 1$ and $\lambda > 1$ are shown in Figure 2.5 (a)-(c). The parameter ρ controls edge density in that the number of edges in the graph is $\rho n(n-1)$, where n is the number of nodes.



(a) (b) (c)

Figure 2.5: **Three undirected synthetic networks** of $n = 100$ nodes placed in a 10×10 square with link probabilities determined by Expert *et al.*'s uniform model [16]. The nodes are not plotted in their spatial locations but according to Python library NetworkX's `spring_layout` to exhibit the effects of assortativity. Attribute communities are denoted by colour and node degree is represented by size. Figure (a) shows a network of disconnected communities produced when $\lambda = 0$. Figure (b) shows a highly assortative graph produced when $\lambda = 0.1$ and Figure (c) shows a primarily disassortative graph produced for $\lambda = 20$.

Results: the one-step method

We begin by studying the results of the one-step method, where the modularity function is directly modified to incorporate a spatial null model. We run parameter searches over $\lambda \in [0, 2]$ and $\rho \in [1, 100]$, to test a wide range of assortative and disassortative structures and edge densities. Expert *et al.* and Leal Cervantes use similar search domains, so it is straightforward to compare method performances.

Undirected networks The Expert *et al.* and classical Newman-Girvan modularity functions are defined for the undirected setting so we take a brief departure to consider undirected networks. Since the constrained models are asymmetric by construction, only the unconstrained models are suitable for this purpose. We use the unconstrained gravity model, with parameters (α, β, ℓ) tuned to the data as described in Section 2.1, and perform 100×100 grid parameter searches, constructing uniform networks of $n = 20$ nodes, with $\lambda \in [0, 2]$ and $\rho \in [1, 100]$. For each (ρ, λ)-pair, the modularity function is optimised using spectral methods [35], and the NMI between the predicted and true partition is calculated for the unconstrained gravity, Expert *et al.*[1], and Newman-Girvan modularity[2]. Heatmaps of the results are shown in Figure 2.6, where

[1] A bin size of two is chosen for the Expert *et al.* binning procedure (2.2), and this choice is discussed further in Appendix B.4.1.

[2] We use the original `MATLAB` code from [16] to predict the Newman-Girvan and Expert *et al.* partitions, and the code of Leal Cervantes in [10] to construct the constrained gravity and radiation models, but the rest of this code was developed for this dissertation in Python.

we see that tuning to the parameters to the data has a positive effect on predictive performance, even for the unconstrained gravity model.

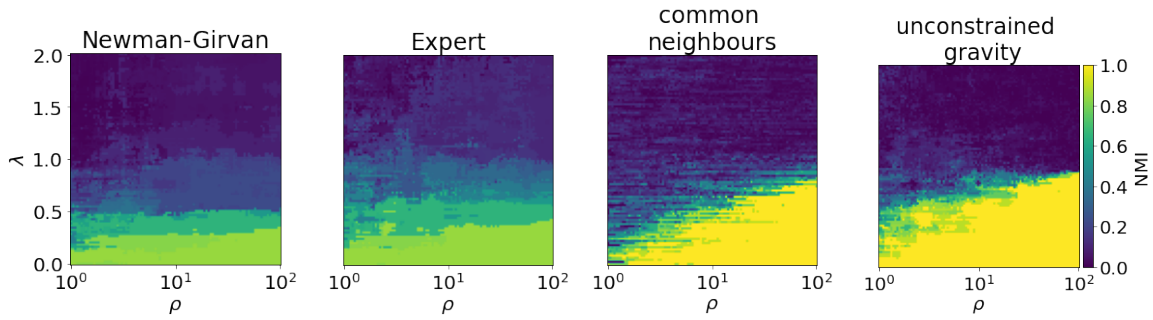


Figure 2.6: **Comparison of NMI scores across (ρ, λ) parameter space for undirected uniform benchmarking networks**. Parameter searches are performed across a 100×100 grid with $\lambda \in [0, 2]$ and $\rho \in [1, 100]$. For each (ρ, λ)-pair a synthetic network of 20 nodes is generated, and community detection is performed using one of (a) Newman-Girvan modularity, (b) the method of Expert *et al.* [16], or the one-step method with (c) the common neighbours or (d) the unconstrained gravity model. The y-axes and colourbar scale are the same for all heatmaps so have each only been included once.

As can be seen in Table 2.2, the differences in performance between methods overall (ρ, λ) combinations are on the scale of $\mathcal{O}(10^{-1})$, but the tuned, unconstrained gravity model performs best. If we limit our scope to assortative graphs, the average NMI for λ in the $[0, 1]$-range is 0.7859 for the unconstrained gravity model, 0.6327 for the unconstrained common neighbours model, 0.5935 for the Expert *et al.* model and 0.47816 for the Newman-Girvan modularity. We, therefore, conclude that the unconstrained gravity and common neighbours models yield the best predictive performances for assortative community structures in these networks. The average modularity for the tuned, unconstrained gravity model is also lower, indicating that the null model more closely resembles the data [16].

Due to time constraints, we do not consider the common neighbours null model any further after this point, but leave it as an illustrative example of how new spatial null models might be developed.

Null Model	Avg. Time	Avg. Modularity	Avg. NMI
Newman-Girvan	0.2687	0.2692	0.2687
Expert	0.0567	0.2251	0.3541
Gravity	0.0646	0.1330	0.4125
Common Neighbours	0.0633	0.0574	0.3469

Table 2.2: **Comparison of averaged scores across (ρ, λ) parameter space for undirected synthetic uniform networks**. Parameter searches are performed with $\lambda \in [0, 2]$ and $\rho \in [1, 100]$. Results show the average calculation time, modularity and NMI scores for each method.

Directed networks We now turn our attention to directed graphs and limit our parameter domain to the assortative regime of $\lambda \in [0, 1]$ to compare the performance of the unconstrained model with the three types of constrained models.

We generate directed versions of the synthetic spatial benchmarks, where there is a directed net flow from the first to the second community. This is described in more detail in Appendix B.3.1 and [10]. Over a 100×100 grid of parameter values, we generate a single network of $n = 20$ nodes for each (ρ, λ)-pair, optimise the modified modularity function using spectral methods [35], then calculate the NMI between the true and predicted predictions. Heatmaps for the results for both the gravity and radiation models are included in Appendix B.4.1. The average NMI scores, calculation time, and modularity across the domain are shown for the gravity model family in Table B.1 and this is shown for the radiation model family in Appendix B.4.1. The radiation model did not perform as well as the gravity model overall, and the averaged NMI score for the doubly-constrained radiation model was 0.4016 in contrast to 0.8545 for the gravity model

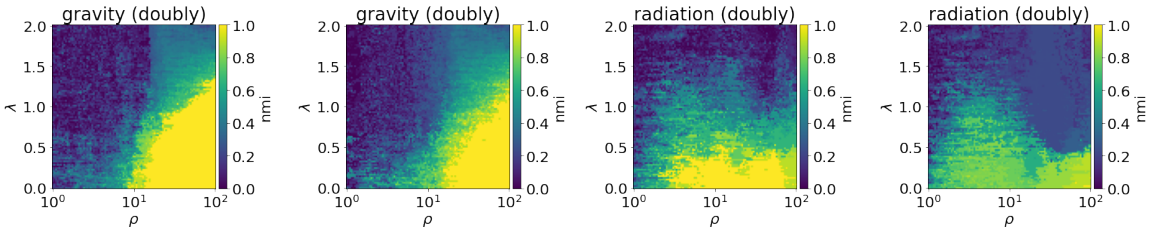
Null Model	Avg. Time	Avg. Modularity	Avg. NMI
Unconstrained	0.0628	0.3545	0.8253
Production	0.0626	0.3395	0.8419
Attraction	0.0603	0.3422	0.8489
Doubly	0.0833	0.3322	0.8545

Table 2.3: **Averaged results for one-step community detection using the gravity model family on directed, uniform benchmarking networks.** Parameter searches were run over with $\lambda \in [0, 1]$ and $\rho \in [1, 100]$. Results show the average calculation time, modularity and NMI scores across the entire (ρ, λ) domain.

Results: the two-step method

We now compare these results to those of the two-step community detection algorithm described in Section 2.2.2. First, we extract the spatial backbone, using the doubly-constrained gravity or radiation model, then we perform classical community detection on the backbones, using the Newman-Girvan or Erdős-Rényi random graph as a null model in the modularity function (1.3) where we use signed extensions as described in Section 2.2.2.

We again generate 100×100 directed and assortative synthetic spatial graphs, with $n = 20$ nodes, $\lambda \in [0, 2]$, and $\rho \in [1, 100]$. Heatmaps show the resulting NMIs for the doubly-constrained gravity (left) and radiation (right) models with Newman-Girvan or Erdős-Rényi null models in Figure 2.7, and the findings are summarised in Table 2.4. We note that the gravity model outperforms the radiation model on the scale of $\mathcal{O}(10^{-1})$.



(a) Newman-Girvan (b) Erdős-Rényi (c) Newman-Girvan (d) Erdős-Rényi

Figure 2.7: **Normalised mutual information scores for two-step spatial community detection using the doubly-constrained gravity(top) and radiation (bottom) null models.** The left panel shows the results using a Newman-Girvan null model and the right panel shows the results using an Erdős-Rényi null model.

Null Model	Backbone	Benchmark	Avg. Time	Avg. Modularity	Avg. NMI
Newman-Girvan	Gravity	Uniform	0.065	0.0194	0.6315
Erdős-Rényi	Gravity	Uniform	0.0607	0.0206	0.6776
Newman-Girvan	Radiation	Uniform	0.0548	0.0348	0.6811
Erdős-Rényi	Radiation	Uniform	0.0525	0.0385	0.5711

Table 2.4: Averaged NMI scores, modularity scores and calculation times of the two step method as in Section 2.2.2 across (ρ, λ) parameter space using the signed Newman-Girvan or Erdős-Rényi modularities with doubly-constrained gravity or radiation spatial backbones.

Comment on results For the gravity model, performance for the two-step method suffers compared to the one-step method. This is particularly pronounced for sparse graphs, which is likely a result of the backbone extraction. For very sparse graphs, not enough edges will be extracted to produce meaningful results. For dense, assortative graphs, however, the performance is good, with NMI scores near unity. The performance of the radiation model actually improves for the two-step procedure, and it performs better than the gravity model in sparse regimes. For higher edge densities $\rho > 10$, we see this performance drop as the gravity model's performance improves.

This decrease in performance of the radiation model for high edge densities is also observed in [10] for stochastic block model methods.

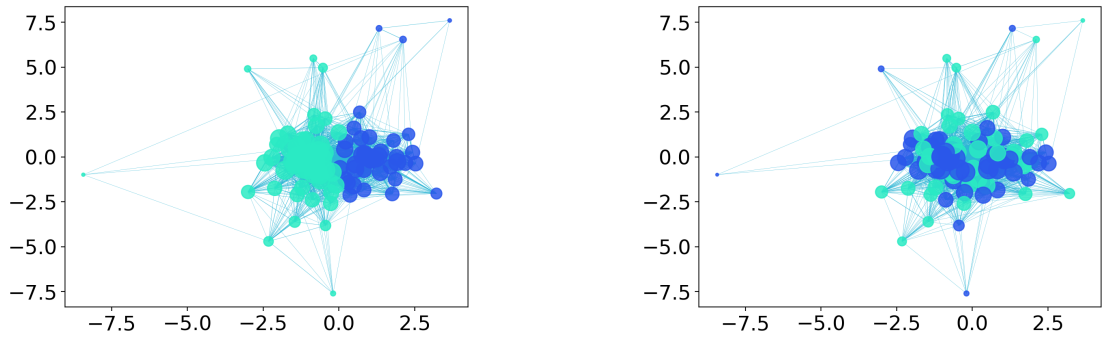
We note from Table 2.4 that the radiation model performs better with the Newman-Girvan null model and the gravity model performs better with the Erdős-Rényi null model. Like the configuration model, the gravity null model 'controls' for degree, so in effect, this is being done twice when the gravity model is combined with the Newman-Girvan null model. This possibly explains why we see the gravity model's performance drop for the Newman-Girvan null model compared to the Erdős-Rényi null model.

For both the one and two-step methods, the gravity model outperforms the radiation model over the entire (ρ, λ)-space. However, we note that the uniform model is based on the gravity model, so the gravity model shares more commonalities with this data. It is possible that some overfitting is at play here, and we cannot assume that the gravity model will generalise better without testing these models on more synthetic networks. We turn to a different synthetic model, the correlated group membership model of Cerina *et al.* to further study the performance of these methods.

The correlated group membership model

The uniform model constructs a network where space and other attributes are uncorrelated; node attributes and locations are randomly assigned, and it is necessary to correct for space in order to uncover attribute communities. Cerina *et al.* [9] highlight that if there exists some unknown degree of correlation between space and attributes then this must be accounted for in our methods. In fact, if space and attributes are strongly correlated, then the removal of spatial effects can result in less accurate results [9, 10]. Cerina *et al.* propose a model where the degree of correlation between some binary attribute and space is a tunable parameter, thus allowing us to explore the effectiveness of our methods on different degrees of spatial correlation in attribute communities.

The parameter ϵ is used to control how correlated space and community are, where $\epsilon = 0.5$ corresponds to the fully random case and $\epsilon = 0$ corresponds to community being completely determined by space. The parameter β controls the role of space in link formation. The $\beta \gg 1$ regime corresponds to space having no impact on link formation while the $\beta \ll 1$ regime corresponds to space being the main factor. Figure 2.8 shows two extremes of this model and a more thorough discussion of the construction of these networks is given in Appendix B.3.2 and in the papers [10, 9].



(a) (b)

Figure 2.8: Two extremes of the correlated group membership model by Cerina *et al.* [9, 10] generated according to B.3. Setting (a) $\epsilon = 0.0$ results in attribute assignment being fully dependent on space, while (b) setting $\epsilon = 0.5$ generates a random network where space and attributes are entirely uncorrelated[1].

Results

Leal Cervantes uses the benchmarks of Cerina *et al.* with $\epsilon = \{0.0, 0.5\}$ and β varied logarithmically in the range $[0.1, 10]$ to benchmark the performance of stochastic block models applied to the spatial backbones.

For each parameter search in this section, we adapt code from [10] to generate the correlated synthetic networks of size $n = 20$ nodes, with a high edge density of $\rho = 100$ and a binary partition of nodes, and perform similar parameter searches for the one- and two-step methods. Figures 2.9 (a)-(d) show the average NMI for varying $\beta \in [10^{-1}, 10]$ for a fixed ϵ, with error bars denoting one standard deviation.

[1]The code used to generate these graphs is adapted from code by Leal Cervantes `https://github.com/rodrigolece/spatial-nets`.

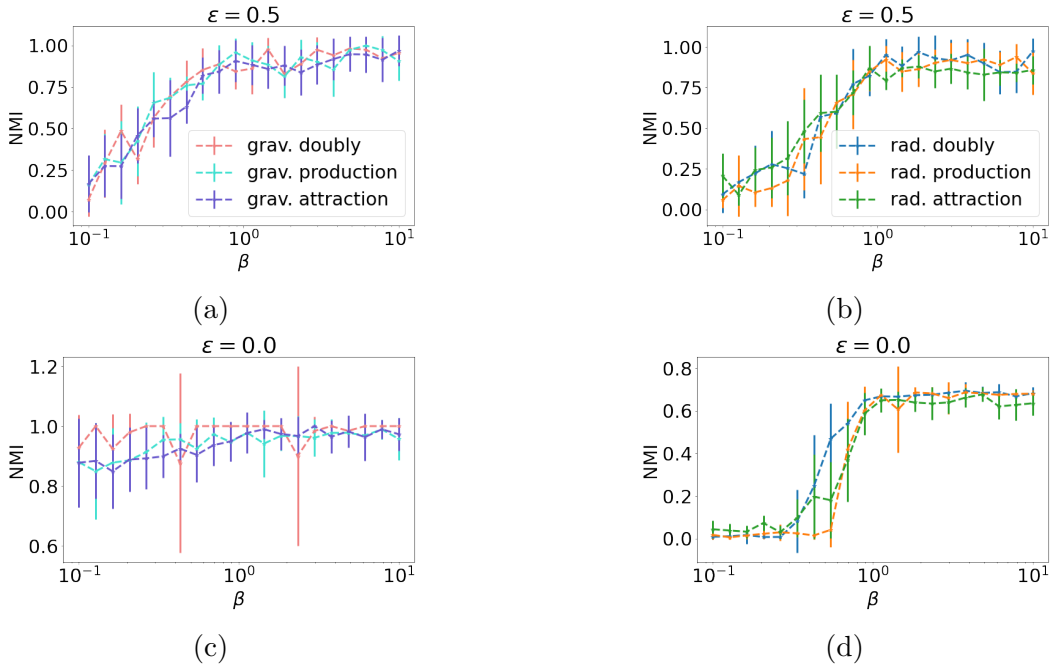


Figure 2.9: **Error bar plots of average NMI for different values β and ϵ in the correlated group membership model**. Results for the correlated group membership model (Appendix B.3.2 or [9]) when performing modularity community detection on gravity (left) and radiation (right) backbones. For each spatial null model and constraint, 10 directed networks were constructed with 20 nodes, edge density $\rho = 100$ and $\ell = 1$, and β was varied on a logarithmic scale in $[10^{-1}, 10^{1}]$. The top row uses $\epsilon = 0.5$ which creates networks where space and attributes are completely uncorrelated and the bottom row shows results for $\epsilon = 0.0$ where space and attributes are fully correlated. The $\beta \gg 1$ regime corrresponds to space having no impact on link formation while the $\beta \ll 1$ regime corresponds to space being the main factor. The error bars here represent one standard deviation.

The trends we observe follow a similar pattern to those in [10]. The gravity and radiation models perform similarly for the uncorrelated case of $\epsilon = 0.5$ while the gravity models perform better than the radiation models for the fully-correlated case. From these results, we understand that in the completely uncorrelated case ($\epsilon = 0.5$), the one-step methods work best when space does not play a significant role in link formation. In the fully-correlated case, the gravity model works well whether or not space plays a significant role in link formation, though it works best when it does not. For the radiation model, there is a clear distinction for β-values below or above 1, though it observes a far smoother transition in the $\epsilon = 0.5$ case than for the SBMs in [10]. When space has an impact on link formation, the radiation model does not perform well, but performance increases rapidly as β increases. Observing the scale on the y-axis, we note that for the fully-correlated case the radiation model produces

partitions with lower NMI scores in general.

Results: the two-step method

We repeat the same parameter search for the two-step procedure, and the results are similar to those in Figure 2.9. These are shown in Appendix B.4.1. A table summarising the average NMI for each (ϵ, null model)-combination is provided in Table 2.5.

	$\epsilon = 0.0$		$\epsilon = 0.5$	
	gravity	radiation	gravity	radiation
ER	0.859299	0.645152	0.679714	0.705310
NG	0.904397	0.589243	0.711061	0.672457

Table 2.5: **Average NMI for different two-step methods**. The doubly-constrained gravity model performs better with the Newman-Girvan modularity, while the doubly-constrained radiation model performs better with the Erdős-Rényi modularity.

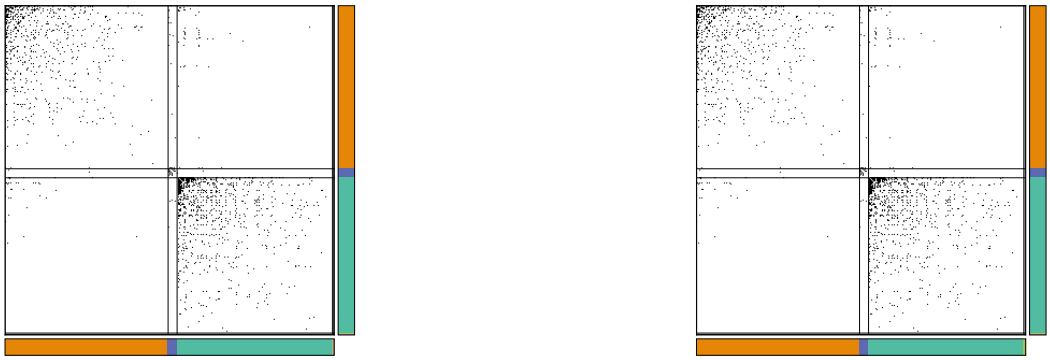
Comment on results For the correlated group membership model, the gravity model performs well, except in the case where space has a strong impact on link formation and space and attributes are correlated ($\epsilon = 0, \beta \ll 1$). In this case, removing space will also remove information about community structure. The radiation model performs slightly better for the two-step method though results are not on par with those of the gravity model.

2.3.2 The maritime shipping network

Next, we apply these methods to the container ship network and compare them to the results of Newman-Girvan community detection in Chapter 1.

The one-step method For the one-step method we detect communities by using the doubly-constrained gravity null model (2.8) in the modularity function (2.4). Figure 2.10 shows the adjacency matrices for the (a) 2019 and (b) 2020 networks, where nodes are ordered first by community assignment and secondly by degree. Figure 2.11 shows ports plotted spatially for the 2019 network, with community assignment denoted by colour. All community detection in this section is implemented using a resolution parameter of $\gamma = 0.2$, which we found gave a small number of communities

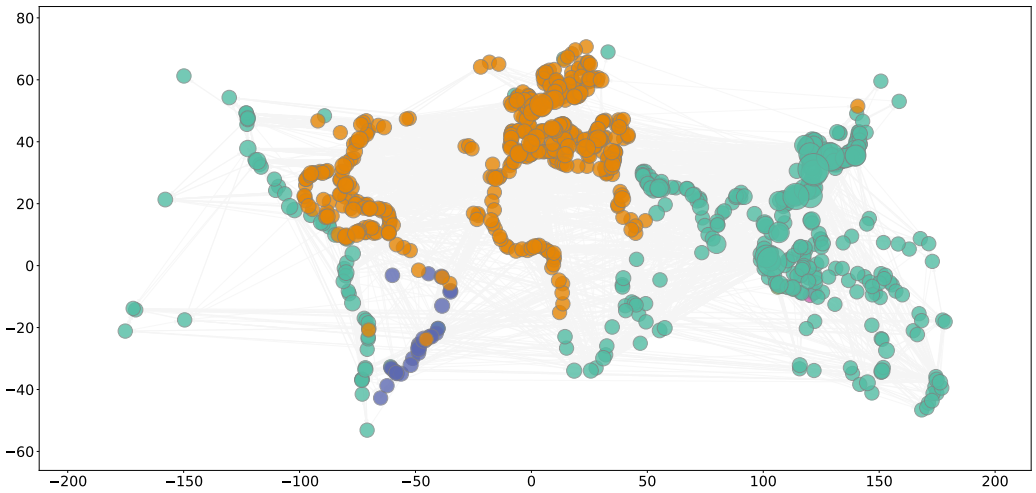
across methods, which facilitated interpretation of results. Further results for the 2020 network are included in Appendix B.4.2.



(a) 2019 (b) 2020

Figure 2.10: **Adjacency matrices for spatially-corrected community detection on the 2019/2020 shipping network.** Here, black points indicate the presence of an edge, and nodes are ordered by community and by degree within communities. Colour bars denoting community are included below and to the right of each matrix. Communities are detected using the one-step method with the doubly-constrained gravity model.

On a broad level, comparing the communities in the spatially-corrected 2019 network, Figure 2.11 to the Newman-Girvan results for the same resolution in Figure A.5, where communities appear to be mostly determined by continent, we see there a clear shift. The gravity model gives more weight to trading routes across major oceans, and two clear groups of ports involved in trans-Pacific trade, or transatlantic trade, emerge. This grouping did not appear in the Newman-Girvan results for any resolution values. We consider this is a sensible result. In 2019, mainlane East-West containerised trade routes, namely Asia-Europe, transatlantic and trans-Pacific routes, dominated the market, handling 39.1% of the market share of globalised trade [45]. We see a similar grouping in the 2020 network and note that the Latin American group, which is isolated in Figure 2.11, rejoins the Atlantic group while a community consisting of 14 ports in the Great Lakes Maritime System becomes isolated.



(a) 2019

Figure 2.11: **Visualisation of spatially-corrected communities detected using the one-step method with the doubly-constrained gravity model on the container ship network.** Ports are shown in their spatial locations and groupings are denoted by colour. At resolution $\gamma = 0.2$ five communities are found, 47% of ports are placed in a group which encompasses mostly trans-Pacific routes (teal), and 49% of ports are placed in a group which contains mostly transatlantic routes (orange). A group of 26 Latin American ports is shown in purple. Not visible in this figure, six Indonesian communities are also divided into two separate groups.

Since the gravity model is explicitly distance-based, it is not surprising that it favours lengthy transatlantic and trans-Pacific crossings. However, we must also consider that this gravity model is now attributing inflated importance to journeys across major oceans due to the huge geographic distances associated with them. In this respect, aspatial and spatially corrected results are best viewed in conjunction, as we are then able to identify bias in either model by considering it in comparison to the other. In general, however, we see that the gravity model can effectively remove the spatial bias which grouped ports by continent, and uncovers a more global perspective of containerised trade relationships which better aligns with observations based on market shares and trade volume in [45].

The radiation model is an intervening-opportunities model, it is based on the number of intermediate nodes between ports rather than the geographic distance between them. Thus we do not expect trans-oceanic voyages to be attributed quite as much importance. Across most resolutions, the results of the doubly-constrained radiation model were somewhat less structured than those of the gravity model. In light of this, and since the radiation model did not perform particularly well on the

benchmarking networks, we have included the full results in Appendix B.4.2 and only summarise them here.

Using resolutions of $\gamma \leq 0.5$, the doubly constrained radiation-based modularity groups all the ports into a single group. For $\gamma = 0.8$, both the 2019 and 2020 groups are split into two communities that exhibit very little logical organisation. One group has Shanghai, Pusan (South Korea), and Hong Kong as its ports with the largest degree, while the other has Singapore, Ningbo (China), and Rotterdam. We leave addressing the interpretation of this to future work and now consider the performance of the radiation model using the two-step method.

The two-step method The results for the two-step method using the doubly-constrained radiation for the 2019 shipping network are shown in Figure 2.12, where communities are denoted by colour. The two-step method detected 14 communities for 2019 and 15 communities for 2020. The communities detected differ in structure from any detected by the one-step method or the Newman-Girvan method. Ports in communities of size less than 30 are shown in white for visual clarity, which reduces the number of visible communities to ten. The groups of size less than 30 in the 2019 network include a group of 22 ports from Brazil, Argentina, Japan, Uruguay, and the Democratic Republic of Congo and two small communities of Indonesian ports.

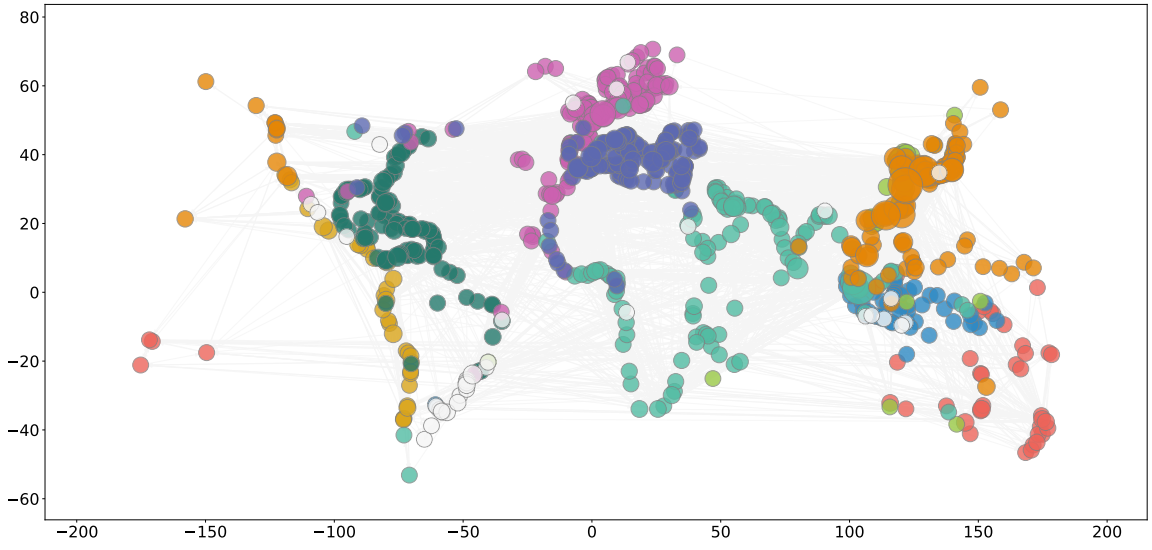


Figure 2.12: **Visualisation of spatially-corrected communities detected using the two-step method with the doubly-constrained radiation model and Erdős-Rényi modularity [13] on the 2019 container ship network.** Ports are plotted in their spatial locations with community denoted by colour. The methods detected 14 communities $\gamma = 0.2$. Ports in communites of size less than 30 are shown in white for visual clarity, which reduces the number of communities to ten.

The community pattern in Figure 2.12 most closely resembles the Newman-Girvan partition for $\gamma = 0.8$ (shown in Appendix A.3b). The main differences are the East-West division of the American continent (both coasts are of North America are grouped together for all resolutions when using the Newman-Girvan model), Singapore is no longer grouped with the Southeast Asian nations but with the larger Middle Eastern and Southern Asian group, which better reflects its major role in the maritime trade system. Additionally, the communities appear slightly more dispersed than in the Newman-Girvan model results, the predominantly Northern European group has a number of North and West African members, as well as five North and one South American member.

In general, the radiation model attributes less import to transatlantic and trans-Pacific crossings than the radiation model but still identifies a different community structure to the Newman-Girvan model.

Chapter 3

Spatially-Corrected Core-Periphery Detection

Core-periphery (CP) structures are meso-scale structures that contain both assortative and disassortative substructures. Classically defined for undirected networks, core-periphery pairs consist of an assortative and densely connected core and a disassortative and sparsely connected periphery. In many definitions [13, 36], the connections between core and peripheral nodes are also required to be dense. Nodes in the core are considered to be well-integrated in the network whilst nodes in the periphery are regarded as more isolated and as performing less important roles in the functioning of the network.

There is a multitude of papers across the disciplines of economics, neuroscience, trade theory, and geography concerned with the existence, identification, and implications of core-periphery structures [47, 24, 21, 7, 4, 41]. In transportation networks, core-periphery structures may have economic implications which contribute to uneven development. They can be used to understand the level of integration of regions across trade networks [24, 22], or determine the resilience of a network to different forms of failure. A network with a pronounced core-periphery structure has been found to be robust against random failure but sensitive to targeted attacks [21].

Core-periphery is, however, an aspatial concept: algorithms determine the core-periphery status of a node based solely on the topology of the network. In spatial networks where space plays a role in link-formation, this may lead to trivial results. Groups of nodes lying in close spatial proximity may be identified as playing artificially inflated core roles in the network. Similarly, a node that is geographically isolated but relatively well-connected, considering its spatial location, may be identified as less integrated with the system than it truly is.

A second limitation of classical core-periphery detection is that it is defined for undirected networks. Potentially useful information contained in the directionality of the network is not considered when implementing core-periphery algorithms. In trade networks, the differentiation between manufacturing and consuming economies is particularly important and the ability to classify a port as a major exporter or importer is useful and relevant.

In this section, we synthesise the work of three papers [10], [13] and [44]. We utilise the spatial backbones developed by Leal Cervantes [10] which we covered in Chapter 2, this time to perform spatially-corrected core-periphery detection. This we integrate with the work of Elliot *et al.* [13] who propose a method for core-periphery detection in directed networks involving the optimisation of a directed core-periphery (DCP) modularity function. We return to the logic of Traag in [44] to extend this measure to signed networks, which allows for its direct application to the spatial backbones. We develop a novel extension of the directed core-periphery benchmarks used by Elliot *et al.* to a spatial setting by including space as a contributing factor in link-formation. Once again, these developments are applied to the maritime container ship network. The resulting directed and spatially-unbiased core-periphery partitions are then compared to the results of the classical Borgatti-Everett [7] and Rombach [36] core-periphery detection methods, which are applied to the symmetrised backbones.

3.0.1 Directed core-periphery detection

According to the definition of [13], an idealised directed core-periphery structure occurs when the vertices in a network $v_i \in V$ can be divided into four groups. First, there is an out-periphery $\mathcal{P}_{\text{out}}$: peripheral nodes that only have outgoing edges. These outgoing edges connect to the in-core $\mathcal{C}_{\text{in}}$ which has only incoming edges and links between different $\mathcal{C}_{\text{in}}$ nodes. Next there is an out-core, $\mathcal{C}_{\text{out}}$ consisting of only outgoing edges and links between different $\mathcal{C}_{\text{out}}$ nodes. Lastly, there is an in-periphery $\mathcal{P}_{\text{out}}$ which only has incoming edges (from $\mathcal{C}_{\text{out}}$). The partition of V into $\mathcal{P} = \{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$ creates an adjacency matrix with the following block structure:

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \mathcal{P}_{\text{out}} \\ \mathcal{C}_{\text{in}} \\ \mathcal{C}_{\text{out}} \\ \mathcal{P}_{\text{in}} \end{matrix}$$

where each entry M_{ij} represents a group of nodes.

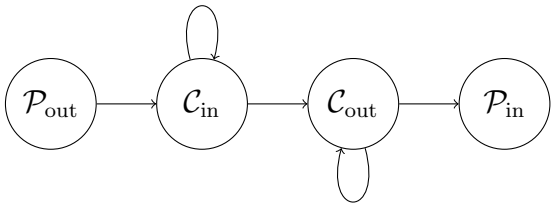


Figure 3.1: Schematic of flows in the idealised directed core-periphery structure.

In Elliot *et al.* [13], an adaptation of the classical modularity function [30] is used, where the community indicator $\delta_{g_i g_j}$ is replaced by the block matrix $M_{g_i g_j}$. They define the directed core-periphery modularity as

$$\text{DCPM}(g) = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(A_{ij} - \hat{A}_{ij} \right) M_{g_i g_j} \tag{3.1}$$

where $g_i \in \{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$ is the set which node v_i is assigned to [13]. This measure lies in $(-1, 1)$ and, just like classical modularity, has a value of 0 for the trivial partition into one community. The null model used in this case $\hat{A} = m/n^2$, is the Erdős-Rényi null model. In core-periphery detection the core nodes are expected to have higher degrees than the peripheral nodes, so using null models that control for degree such as the configuration model, may actually obscure core-periphery structures. This is proved by Kojaku *et al.* [21] who show that is impossible to detect a single core-periphery structure using the configuration model, unless we incorporate more blocks, be they other core-periphery structures, other community structures or a set of nodes without structure.

3.0.2 DCP detection on the spatial backbone

Since we are now dealing with a signed network things become a little more complicated. We cannot directly apply the methodologies described in Elliot *et al.* to the spatial backbones as the signed network would return meaningless results. We revisit Traag's discussion of modularity clustering on signed networks for inspiration [44].

In Traag's discussion of partitioning signed networks [44], it is shown that calculating the modularity using the Newman-Girvan null model for a signed network yields meaningless results. We include here a similar constructive example for an undirected core-periphery structure, where calculating the undirected core-periphery modularity (CPM) using the Erdős-Rényi null model also yields unhelpful results. We will assume this proof is sufficient to conclude that the same limitation applies

in the directed core-periphery modularity (DCPM). The sample signed network with our definition of signed core-periphery structure is shown in Figure 3.2 and has $n_c = 4$ nodes in its core, which is a complete graph of positive links. It has a periphery of $n_p = 4$ nodes which is also a complete graph, but of negative links. Each node in the periphery is connected to a different node in the core by two positive links.

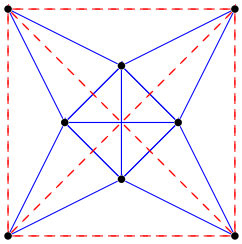


Figure 3.2: **Example undirected core-periphery structure in a signed network.** Positive links are represented by blue edges and negative links are represented by red edges.

Using a discrete block-structure definition of core-periphery structure [13, 7], the undirected core periphery structure for an unsigned network here is given as

$$\text{CPM} = \frac{1}{2m} \sum_{i,j=1}^{n} \left(A_{ij} - \frac{m}{n^2} \right) M_{g_i g_j} \tag{3.2}$$

The total flow m of the network is $m = \binom{4}{2} - \binom{4}{2} + 8 = 8$. Thus the contribution to the CPM of a link between core nodes is

$$A_{ij} - \frac{m}{n^2} = 1 - \frac{1}{8} = \frac{7}{8}.$$

For two nodes in the periphery

$$A_{ij} - \frac{m}{n^2} = -1 - \frac{1}{8} = \frac{-9}{8}$$

and for links between the core and the periphery, we obtain

$$A_{ij} - \frac{m}{n^2} = 1 - \frac{1}{8} = \frac{7}{8}.$$

Since the periphery is disassortative the modularity function is increased by placing peripheral nodes in separate groups. Also, the positive links between the core and peripheral nodes suggests core and peripheral nodes should be grouped together. Altering the block matrix $M = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ to $M = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ would resolve the negative contribution to the CPM for negative links between the peripheral nodes but does not

address the positive contribution from core-to-periphery links. The total flow $m = 8$ of the network is simply to small due to the presence of negative links. Thus, we see that we cannot simply extend the DCPM definition provided in Elliot *et al.* [13] to a signed network.

We propose to extend the logic of Traag [44] to the case of the directed core-periphery detection on signed networks as follows: as usual, we seek to maximise the number of positive edges within the core and maximise the number of positive edges between core and peripheral nodes. In our signed network, negative edges indicate that there is less connection between a pair of nodes than would be expected under our spatial null model. In this case it makes sense that we do not want there to be many negative edges within the core but we do want there to be plenty of negative edges between peripheral nodes. Thus, with the positive and negative spatial backbones denoted as Φ^+ and Φ^-, respectively, we define the objective function, which we wish to maximise, for the positive backbone as

$$\text{DCPM}^+ = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\Phi^+_{ij} - \frac{m^+}{[n^+]^2}\right) M_{g_i g_j}, \tag{3.3}$$

where we have dropped the leading coefficient, and n^+ and m^+ are the number of nodes and edges in the positive backbone, respectively. Analogously, we define the objective function for the negative backbone as

$$\text{DCPM}^- = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\Phi^-_{ij} - \frac{m^-}{[n^-]^2}\right) M_{g_i g_j} \tag{3.4}$$

and in this case we seek to minimise this quantity. Here we use the Erdős-Rényi null model. Like in [44], this is equivalent to maximising the negative of DCPM^-, and we can combine the maximisation of DCPM^+ and DCPM^- into one objective function as

$$\text{DCPM} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\Phi_{ij} - \left(\frac{m^+}{[n^+]^2} - \frac{m^-}{[n^-]^2}\right)\right) M_{g_i g_j} \tag{3.5}$$

where $\Phi = \Phi^+ - \Phi^-$. Additionally, the number of nodes remains the same in both backbones $n^+ = n^- = n$, so we obtain

$$\text{DCPM} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\Phi_{ij} - \left(\frac{m^+ - m^-}{n^2}\right)\right) M_{g_i g_j} \tag{3.6}$$

Thus, the extension of the directed core-periphery detection to signed networks may be boiled down to simply replacing the null model $\hat{A}$ with $\hat{A}^+ - \hat{A}^-$ in the directed core-periphery modularity function (3.1).

Elliot *et al.* compare a number of optimisation algorithms to optimise the DCPM [13][1]. Of these methods, we choose to modify the Advanced HITS (AdvHITS) algorithm, an extension of the Hyperlink-Induced Topic Search (HITS) algorithm [20]. We chose AdvHITS it performs well with respect to both speed and accuracy. An overview of the AdvHITS algorithm and some further details regarding the optimisation of the DCPM are given in Appendix C.0.1 and C.0.2.

3.1 Results

3.1.1 Synthetic networks

We construct a simple spatial extension of the directed core-periphery networks described in [13], where a more complete discussion of the network is provided. As in Elliot *et al.* the network is partitioned into an out-periphery, an in-core, an out-core and an in-periphery, $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$ and the idealised block structure of this partition has the 'L' shape.

$$M = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} \mathcal{P}_{\text{out}} \\ \mathcal{C}_{\text{in}} \\ \mathcal{C}_{\text{out}} \\ \mathcal{P}_{\text{in}} \end{matrix}$$

In [13], links in the synthetic networks are distributed according to

$$M = \begin{pmatrix} p_2 & p_1 & p_2 & p_2 \\ p_2 & p_1 & p_2 & p_2 \\ p_2 & p_1 & p_1 & p_1 \\ p_2 & p_2 & p_2 & p_2 \end{pmatrix} \begin{matrix} \mathcal{P}_{\text{out}} \\ \mathcal{C}_{\text{in}} \\ \mathcal{C}_{\text{out}} \\ P_{\text{in}} \end{matrix}.$$

We mimic the first form of this benchmark proposed in [13] and use only one probability $p \in [0, 0.5]$ such that

$$(p_1, p_2) = (0.5 + p, 0.5 - p). \tag{3.7}$$

To extend this to a spatial setting, we make the following adjustments: each node is randomly assigned to a spatial location and a pairwise distance matrix (d_{ij}) is constructed from the coordinates. Each entry of the probability matrix p_{ij} is then multiplied by $d_{ij}^{-\ell}$, where ℓ is a parameter which can be specified, and the probability of an edge e_{ij} becomes

[1]Original code is available at: https://github.com/alan-turing-institute/directedCorePeripheryPaper

$$p_{ij} = \frac{1}{Z}\left(\frac{M_{g_i g_j}}{d_{ij}^{-\ell}}\right) \tag{3.8}$$

where Z is a normalisation constant such that $\sum_{i \neq j} p_{ij} = 1$. The edge density may specified be specified using the parameter $\rho > 0$ in the same way as the uniform [16, 10] and correlated group membership models [9, 10], and $\rho n(n-1)$ edges will be placed in the network. In this way, link probabilities in the synthetic benchmark are controlled by both the core-periphery structure and the pariwise distances between nodes. [2]

Figure 3.3: **Synthetic spatial network with ideal directed core-periphery structure** Network with link probabilities determined by (3.8) with $p = 0.5$, $\ell = 2$ and $\rho = 1$. Purple represents the in-core, teal the out-core, green the in-periphery and orange the out-periphery. Nodes are plotted spatially on the left and according to NetworkX's spring-block layout on the right.

The heatmaps in Figure 3.4 (a)-(b) visualise the adjacency matrix of a spatial DCP benchmark with a near-ideal block structure with $p = 0.48$, with its corresponding positive and negative backbones, extracted using the doubly-constrained gravity (blue) and the doubly-constrained radiation (red) null models.

[2] More sophisticated benchmarks, say, by introducing parameters controlling the strength of the spatial effects or the degree of correlation between directed core-periphery assignment and space, similar to the work of Cerina *et al.* [9], would be useful to fully validate the performance of this algorithm, but we will consider that beyond the scope of this dissertation.

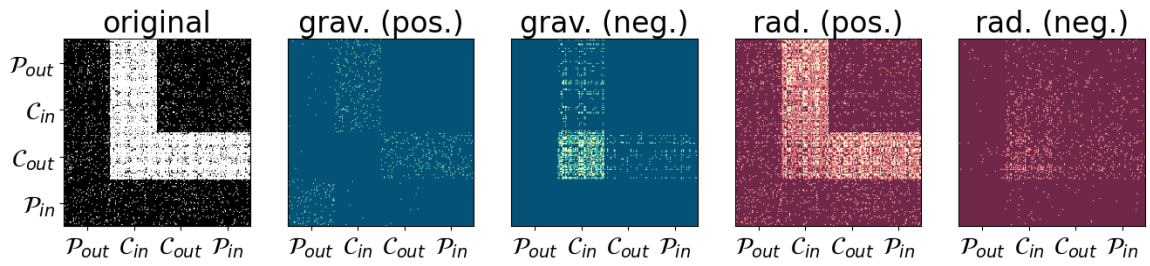


Figure 3.4: **Backbones of synthetic spatial directed core-periphery networks with an almost-ideal directed core-periphery structure** ($p = 0.48$). Backbones are extracted from a network of 200 nodes using the doubly-constrained gravity null model (blue), and the doubly-constrained radiation null model (red). All matrices are binary and lighter colours correspond 1s and darker colours correspond to 0s.

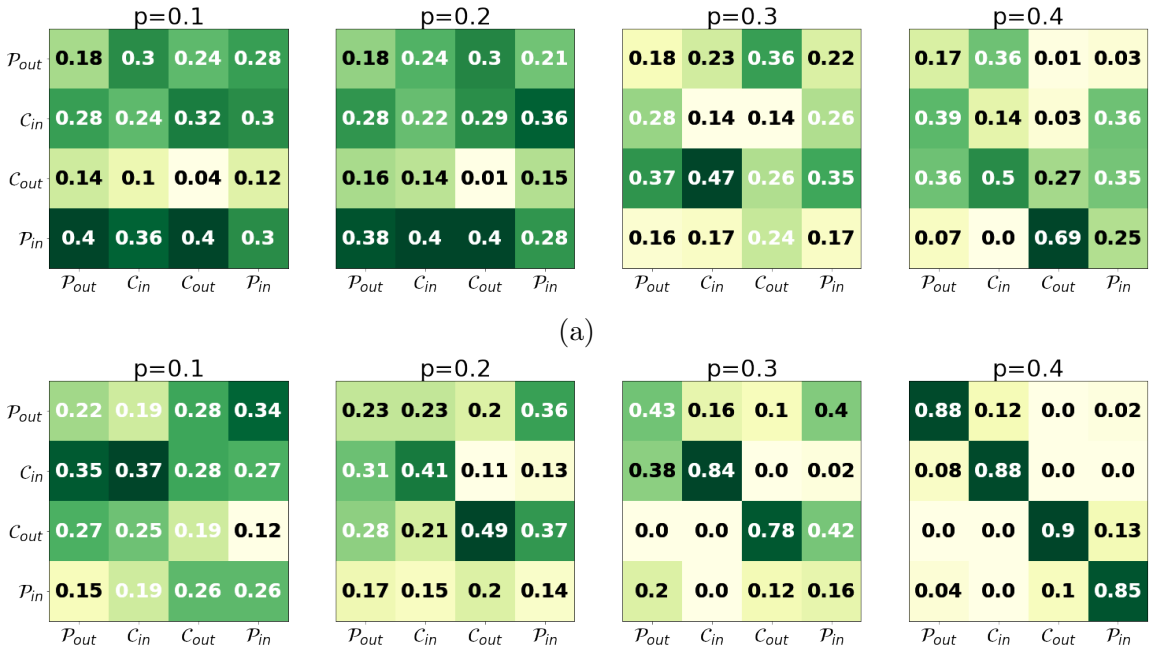
Both positive backbones generally identify the full DCP block structure, but the gravity model misses the bend of the 'L' structure ($\mathcal{C}_{\text{out}} \to \mathcal{C}_{\text{in}}$). A possible explanation for this is the high out-degree of nodes in $\mathcal{C}_{\text{out}}$ and the high in-degree of nodes in $\mathcal{C}_{\text{in}}$, which causes the gravity null to estimate very high predicted fluxes between these groups. Kojaku, Sadamori, and Masuda prove in [21] that degree-controlled methods often 'mask' core-periphery structure and, since the gravity model controls for degree, this is a possible limitation of the gravity model in this context. The positive radiation backbone, however, performs better in this respect. Both of the negative backbones mistakenly identify negative edges within the 'L' structure and miss the blocks below and to the left of the main 'L' structure.[3]

Applying our modification of the AdvHits algorithm to the positive gravity and radiation backbones returns NMI scores of 0.6265 and 0.9696 for the predicted partitions, respectively, compared to the known partitions. Thus, the radiation-based backbone performs better here in the almost ideal block-structure case.

Directed confusion matrices To further investigate this, we generate confusion matrices where each column shows the percentage of times nodes in a network generated by (3.8) with true assignments in $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$ are assigned to each group. We visualise the results obtained using the doubly-constrained gravity backbone for p-values in $[0.1, 0.2, 0.3, 0.4]$ in Figure 3.5 (a) with and (b) without the negative backbone. In this case, the algorithm actually performs better when the negative backbone

[3]The author of [10] has previously noted that the methodology for extracting the negative backbone has some remaining limitations. The negative backbone has a tendency to be either too sparse or too dense relative to the positive backbone. This may be due to increased noise associated with smaller entries, or the fact that the algorithm only considers nonzero entries of the adjacency matrix, i.e., a non-existing edge cannot become an edge in the negative backbone.

is ignored. Heatmaps showing similar results for the radiation backbone are included in Appendix C.1.



(a)

(b)

Figure 3.5: **Confusion matrices for the directed core-periphery algorithm using the doubly-constrained gravity backbone of a directed graph with known core-periphery structure.** For each p-value, 20 random synthetic networks with known core-periphery structure are constructed according to (3.8) and the AdvHits [13] algorithm applied to them. The columns of the confusion matrices show the percentage of times core and periphery nodes were assigned to each of $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$.

3.1.2 The maritime shipping network

In this section, we present a selection of results produced when our methods for spatial connection are applied to the 2019 container shipping network. Ports are assigned to the out-periphery $\mathcal{P}_{\text{out}}$, in-core $\mathcal{C}_{\text{in}}$, out-core $\mathcal{C}_{\text{out}}$, or in-periphery $\mathcal{P}_{\text{in}}$ using first the unmodified DCP method of Elliot *et al.* [13], then incorporating spatial effects by extracting the doubly-constrained gravity and radiation spatial backbones, using the methods of Leal Cervantes [10], and applying the modified DCP method to these backbones.

We find that the gravity model uncovers regional roles played by ports that were masked by the broader spatial structure of the network, while the radiation model

unveils some cultural or economic affinities that were not detected in the original network. Figures 3.6 and 3.9 show ports from the 2019 network colour-coded by their assignment to one of $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$ and plotted geographically.

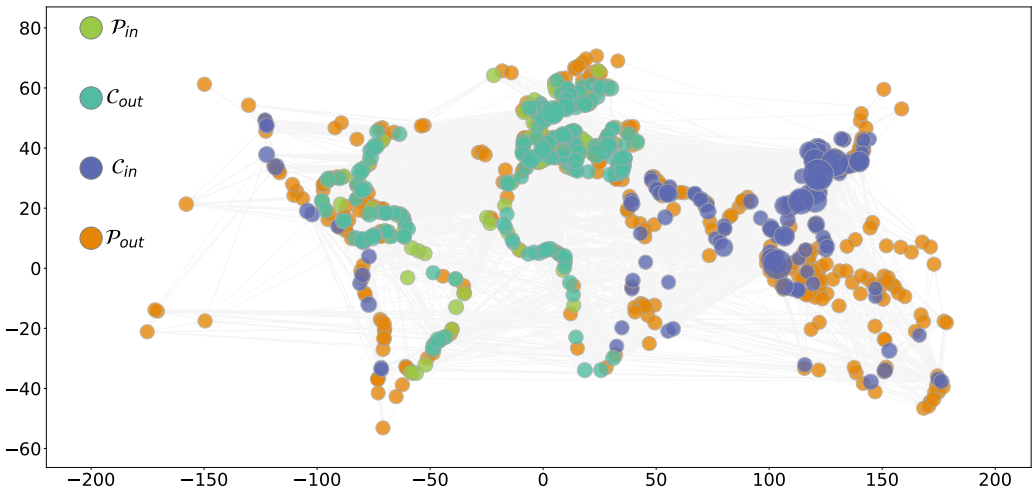


Figure 3.6: **Directed core-periphery (DCP) detection (aspatial) on the 2019 shipping network.** Container ship ports visualised in their spatial locations and colour-coded according to their assignment to one of $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$. The in-core group is shown in purple, the out-core group is shown in teal, the out-periphery group is shown in orange, and the in-periphery group is shown in green. Results were obtained using the methods of Elliot *et al.* in [13][1].

The left panel of Figure **??** visualises the block structure between groups detected by the aspatial method, and the right-hand panel of Figure **??**, and Figure 3.8 shows confusion matrices comparing results of the aspatial and spatial DCP methods to those of the classical, undirected core-periphery (CP) detection methods by Borgatti and Everett [7] and Rombach [36]. The left panel of Figure 3.10 visualises the NMI scores between the sets of labels detected by the DCP algorithms and the right panel of Figure 3.10 shows the distibution of ports across each group for each DCP method.

[1] https://github.com/alan-turing-institute/directedCorePeripheryPaper

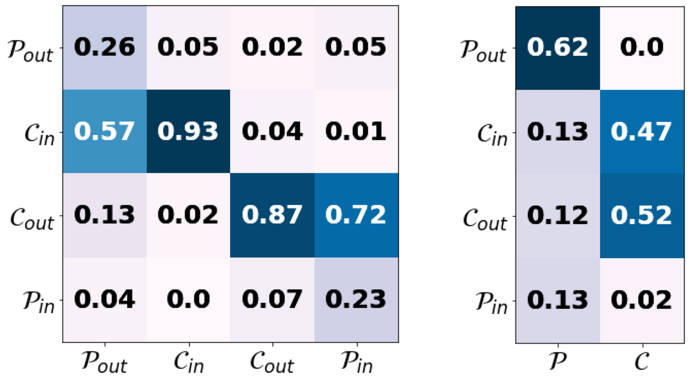


Figure 3.7: **Block structure (left) and confusion matrix with results of the Rombach method [36] (right) for aspatial core-periphery results on the 2019 shipping network.** Visualisation of the percentage of edges (column-normalised) between blocks detected by the DCP methods of Elliot *et al.* (left-panel) show a slightly different block structure to that in [13]. The confusion matrix on the right shows the distribution of nodes identified by the Rombach between the four sets $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$.

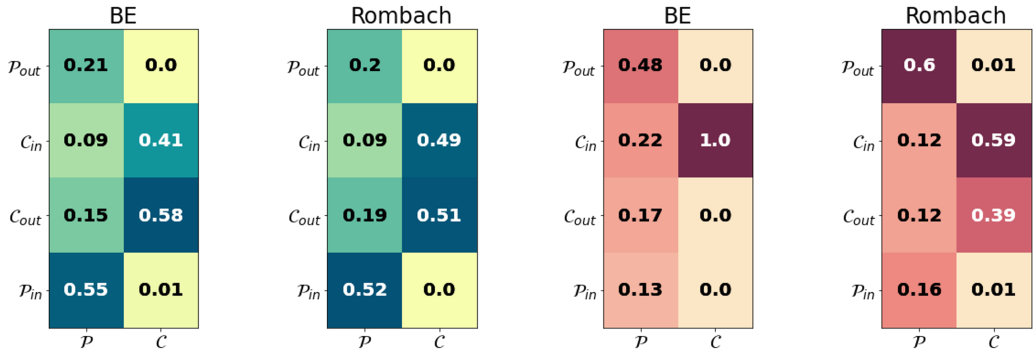
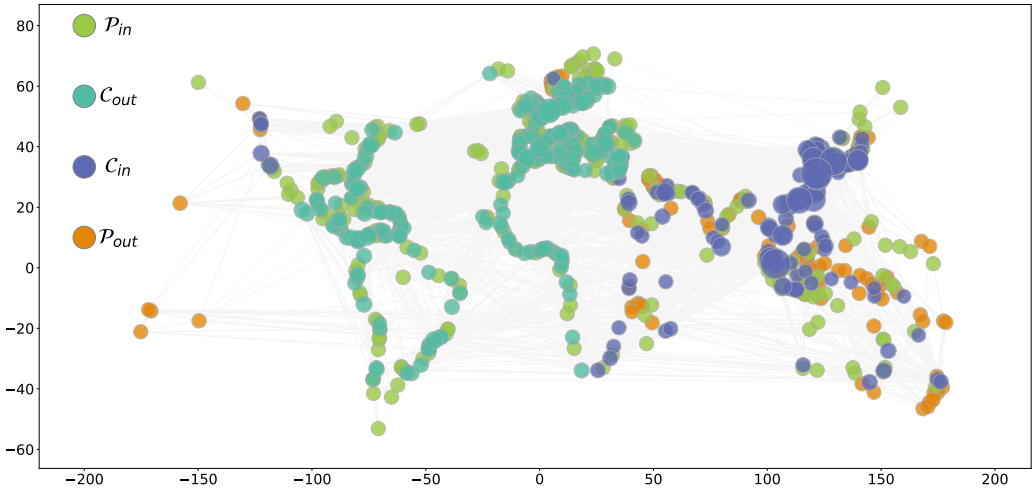
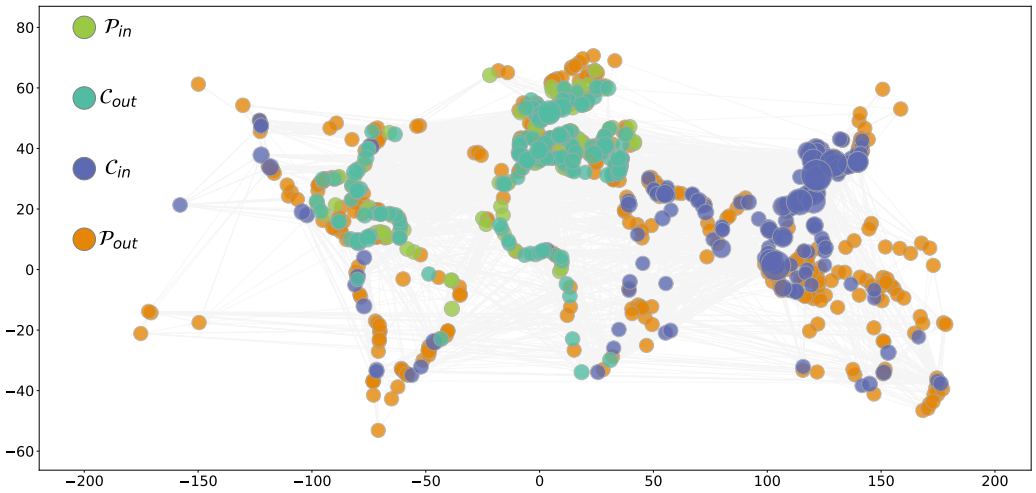


Figure 3.8: **Confusion matrices for the results of directed vs. undirected core-periphery detection on the 2019 shipping network with spatial correction.** Confusion matrices for the results of the DCP method vs. results of the Borgatti-Everett and Rombach algorithms [7, 36]. Methods were applied to the positive spatial backbone, which was extracted using the doubly-constrained the gravity model (left) and the doubly-constrained radiation model (right).



(a) Directed (gravity model-corrected) core-periphery detection on the 2019 network



(b) Directed (radiation model-corrected) core-periphery detection on the 2019 network

Figure 3.9: **Directed (spatially-corrected) core-periphery detection on the 2019 network.** Container ship ports visualised in their spatial locations and colour-coded according to their assignment to one of $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$. Results were obtained using the DCP methods of Elliot *et al.* in [13] and the spatial-backbone extraction methods of Leal Cervantes [10][1] to account for spatial effects using (a) the doubly-constrained gravity model, and (b) the doubly-constrained radiation model. Note that more ports are assigned to out-core and in-periphery when the gravity model is used, compared to the other two cases.

[1]https://github.com/rodrigolece/spatial-nets

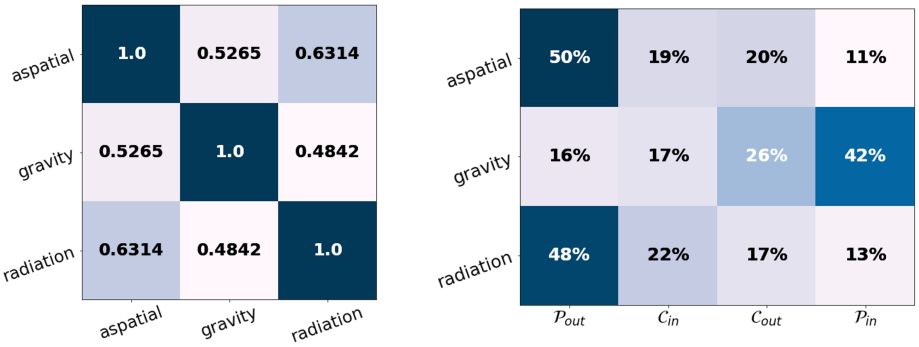


Figure 3.10: **Mutual NMI scores (left) and percentage of nodes assigned to each group (right) for all three DCP detection methods.** The NMI scores between the sets of labels produced by each of the aspatial DCP detection, and the spatially-corrected DCP detection using the doubly-constrained gravity and radiation backbones, are shown in the heatmap on the left. The percentage of nodes assigned to each of $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$ for each DCP detection method is shown in the heatmap on the right.

The overall DCP pattern is relatively consistent across all three methods. Asia and Middle Eastern ports, including the major ports of Singapore and Shanghai, are assigned to the in-core while the majority of European and North American ports are consistently assigned to the out-core. This is a surprising result, due to Asia's status as a manufacturing hub, we would expect it to be assigned to the out-core. Asia is, however, a major importer of raw materials so this may be down to some supply chain structures. From the symmetry in the block structure visualisations on the left panel of Figure 3.7, we can see the groups out- and in-connectivities only differ by a small number of edges, so there is not a huge difference between the functional roles played by core blocks. This block structure is actually more similar to a different case (A.7) mentioned in the Supplementary Information of Elliot *et al.* where there is an out-core, an in-core, and two *out*-peripheries. There is also a very low volume of flow between the out- and in-cores which suggests that the network might be better divided into two core-periphery pairs or global community structure with internal, local core-periphery structures.[4]

The spatially-corrected model with the radiation backbone is more similar to the aspatial case, but the gravity model reassigns 337 ports (∼40% of the network) in the 2019 network from $\mathcal{P}_{\text{out}}$ to $\mathcal{P}_{\text{in}}$ and reassigns ten ports (∼1%) from $\mathcal{C}_{\text{in}}$ to $\mathcal{C}_{\text{out}}$. We

[4]We note that the level of connectivity between the two cores may appear artificially low here. The edges in this network are unweighted so container ships of large or small capacities contribute the same values to the flow count. We expect major container ports to have many high-capacity container ships travelling between them, compared to smaller ports. Including this information may increase the observed connectivity between cores.

find that spatial correction using the gravity model uncovers a number of regional hubs in Latin America (e.g. Bahia De Valparaíso and San Antonio in Chile), Oceania, Scandinavia, and West and Central Africa, mainly serving local smaller ports. These core ports are identified as in-connectivity ports in the original network due to the large number of in-flows they receive from major ports such as Singapore, Shanghai, and Hong Kong. These links are removed when the spatial backbone is extracted using the gravity model, unveiling more regional, out-connectivity roles the ports play.

In the aspatial and gravity-corrected spatial results, Honolulu (Hawaii) is assigned to $\mathcal{P}_{\text{out}}$. The radiation model, however, reassigns Honolulu to $\mathcal{C}_{\text{in}}$ instead. Honolulu is not well-connected in the container ship network, however, it is part of the U.S.A. Culturally and economically it is much better integrated with North America than other Pacific Islands with a similar degree of geographic remoteness. This suggests that the correction using the radiation model can discover some cultural and economic affinities that were masked by space in the original network.

Chapter 4

Conclusions

Throughout this dissertation, we have considered a wide variety of methods to remove spatial bias from community and core-periphery detection methods. In Chapter 1 we surveyed the background theory for community detection and spatial networks and performed some preliminary, classical community detection on the network of container ships. In Chapter 2 we presented two methods for spatially-correcting community detection algorithms. The first method, which we referred to as the *one-step method*, involved directly modifying the null model of a modularity function to incorporate space. We took a new approach to this and extended the problem to directed networks by introducing the dimensionally constrained models of Wilson [49, 10] as null models, and using Leicht and Newman's approach to make the resulting asymmetric modularity matrices suitable for spectral partitioning [35]. In addition, we discussed a more general approach, which we named the *two-step method*. This method built on the methodology proposed by Leal Cervantes [10] to extract *spatial backbones* from a network, over which the modularity could then be optimised. We followed Traag's approach [44] to modify the modularity problem for signed networks. In addition to this, we formulated a novel mobility model inspired by the work of Kosowska-Stamirowska and Zusanna [23], the common neighbours+sea distance model. This model requires further testing but performed well on undirected benchmarking networks. In Chapter 3 we combined two recently proposed techniques for core-periphery detection and introduced both spatial-correction and directionality to the problem by making use of the spatial backbones and dimensionally-consistent spatial null models of Leal Cervantes [10], and the directed core-periphery methodology of Elliot *et al.*[13]. We developed a signed extension of the directed core-periphery modularity of Elliot *et al.* and optimised this over the extracted spatial backbones.

Overall, the gravity model-based methods outperformed the radiation model-based methods and produced more easily interpretable results on the shipping net-

work. Applied to the community detection problem, the gravity model prioritised mainlane East-West trading routes across major oceans which were not identified by the classical methods. In the core-periphery problem, introducing the gravity model uncovered the more regional roles played by some ports in the network. In the core-periphery detection problem, the radiation model also appeared to identify some ports, such as Honolulu (Hawaii), that were relatively well-connected despite their geographic isolation and as such, uncovered affinities that were not evident in the classical analysis.

Assumptions and limitations of this work A number of assumptions were made throughout this dissertation and we wish to acknowledge some of the limitations of the work. Firstly, the benchmarking networks used were highly simplistic. In particular, the spatial and directed core-periphery synthetic network in Chapter 3 could be made more realistic. A more realistic model would involve introducing an element of correlation between space and group assignment, similar to the benchmarking networks of Cerina *et al.*[9]. In transport networks, cores or hubs may develop in convenient, e.g., central, geographically central locations. Synthetic networks where geographically central nodes are more likely to gain core status would be more realistic. Additionally, as noted by the author of [10], the negative backbone procedure still needs improvement. At present, it does not consider zero edges as candidates for the negative backbone and it seems to be affected by a high degree of noise. For this reason, a number of the algorithms applied to the shipping network only used the positive backbone, as the results including the negative backbone were not easily interpretable. Once this methodology has been improved, it would be interesting to re-run these methods using both backbones. Additionally, we noted in Chapter 3 that the gravity model was not an ideal candidate to use in spatially-corrected core-periphery detection as it controls for degree heterogeneity. Since degree disparities between the core and the periphery are a key feature of core-periphery structures, controlling for them may remove too much information and obscure core-periphery structures. A more systematic exploration of the suitability of different mobility models for spatially-corrected core-periphery detection would be valuable in this respect.

Finally, in the interest of notational clarity, the shipping network used was unweighted and did not include information about the volume of cargo carried by ships. This is slightly misrepresentative of true trade intensity between ports, and in particular, underestimates trade between major ports that are frequented by large vessels with high capacities. A weighted version of this network is available [48] and the

algorithms do not need to be modified in order to be applied, so a further analysis using the weighted network may yield different insights.

Further work The results of this dissertation highlight many potential avenues for further work. The empirical network used is a large network of over 1000 nodes and interpreting the results requires both time and relevant expertise. It would be interesting to further investigate these results to see what more information they contain. In Chapter 3, we noted that the block structure observed when applying the directed core-periphery methods of Elliot *et al.* to the shipping network only loosely agreed with the ideal structure in their main paper. It bore a closer resemblance to a different block structure (A.7) which was given in the Supplementary Information. This raises the question of whether a different directed core-periphery structure might be more suited to this network. Modifying the algorithm of Elliot *et al.* to explore other possible structures is another direction for future work.

In addition, it has been observed that the global shipping network has a hierarchical structure [50]. A procedure that combines the methods of Chapter 2 and Chapter 3 to detect local core-periphery structures within communities, would be an interesting extension. In particular, it would be interesting to compare results of these methods to those of Kojaku [21, 22] which identify multiple core-periphery structures within a network. Finally, it would also be interesting to pursue the common neighbours+sea distance null model proposed in Chapter 2 further and to apply it to the shipping network. Cross-validation methods, such as those proposed by Leal Cervantes in [10], could be used to assess the suitability of the model to the shipping data.

References

[1] Iftikhar Ahmad et al. "Missing link prediction using common neighbor and centrality based parameterized algorithm". In: *Scientific reports* 10.1 (2020), pp. 1–9.

[2] James E Anderson. "The gravity model". In: *Annu. Rev. Econ.* 3.1 (2011), pp. 133–160.

[3] Armen S Asratian, Tristan MJ Denley, and Roland Häggkvist. *Bipartite graphs and their applications*. Vol. 131. Cambridge university press, 1998.

[4] Paolo Bartesaghi, Gian Paolo Clemente, and Rosanna Grassi. "Community structure in the World Trade Network based on communicability distances". In: *Journal of Economic Interaction and Coordination* (2020), pp. 1–37.

[5] Marc Barthélemy. "Spatial networks". In: *Physics Reports* 499.1-3 (2011), pp. 1–101.

[6] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[7] Stephen P Borgatti and Martin G Everett. "Models of core/periphery structures". In: *Social networks* 21.4 (2000), pp. 375–395.

[8] Gerald AP Carrothers. "An historical review of the gravity and potential concepts of human interaction". In: *Journal of the American Institute of Planners* 22.2 (1956), pp. 94–102.

[9] Federica Cerina et al. "Spatial correlations in attribute communities". In: *PLoS One* 7.5 (2012), e37507.

[10] Rodrigo Leal Cervantes. *Community Detection and Retail Networks, Confirmation Thesis for the degree of Doctor of Philosophy*. 2011.

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 1991.

[12] Edsger W Dijkstra et al. "A note on two problems in connexion with graphs". In: *Numerische mathematik* 1.1 (1959), pp. 269–271.

[13] Andrew Elliott et al. "Core–periphery structure in directed networks". In: *Proceedings of the Royal Society A* 476.2241 (2020), p. 20190783.

[14] Paul Erdös and Alfréd Rényi. "On the evolution of random graphs". In: *The structure and dynamics of networks*. Princeton University Press, 2011, pp. 38–82.

[15] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*. Vol. 3. Vsp, 1990.

[16] Paul Expert et al. "Uncovering space-independent communities in spatial networks". In: *Proceedings of the National Academy of Sciences* 108.19 (2011), pp. 7663–7668.

[17] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[18] Pablo Kaluza et al. "The complex network of global cargo ship movements". In: *Journal of the Royal Society Interface* 7.48 (2010), pp. 1093–1103.

[19] Brian Karrer and Mark EJ Newman. "Stochastic blockmodels and community structure in networks". In: *Physical review E* 83.1 (2011), p. 016107.

[20] Jon M Kleinberg et al. *Authoritative sources in a hyperlinked environment*. Princeton University Press, 2011.

[21] Sadamori Kojaku and Naoki Masuda. "Core-periphery structure requires something else in the network". In: *New Journal of physics* 20.4 (2018), p. 043012.

[22] Sadamori Kojaku et al. "Multiscale core-periphery structure in a global liner shipping network". In: *Scientific reports* 9.1 (2019), pp. 1–15.

[23] Zuzanna Kosowska-Stamirowska. "Network effects govern the evolution of maritime trade". In: *Proceedings of the National Academy of Sciences* 117.23 (2020), pp. 12719–12728.

[24] Paul Krugman and Anthony J Venables. "Globalization and the Inequality of Nations". In: *The quarterly journal of economics* 110.4 (1995), pp. 857–880.

[25] Sven Kurras. "Symmetric iterative proportional fitting". In: *Artificial Intelligence and Statistics*. PMLR. 2015, pp. 526–534.

[26] Renaud Lambiotte. *C5.4 Networks Lecture Notes*. 2021.

[27] Elizabeth A Leicht and Mark EJ Newman. "Community structure in directed networks". In: *Physical review letters* 100.11 (2008), p. 118703.

[28] Maxime Lenormand et al. "A universal model of commuting networks". In: (2012).

[29] Xin Liu, Tsuyoshi Murata, and Ken Wakita. "Detecting network communities beyond assortativity-related attributes". In: *Physical Review E* 90.1 (2014), p. 012806.

[30] Mark Newman. *Networks*. Oxford university press, 2018.

[31] Mark Newman. "The configuration model". In: *Networks*. Oxford university press, 2018.

[32] Tiago P Peixoto. "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models". In: *Physical Review E* 89.1 (2014), p. 012804.

[33] Tiago P Peixoto. "Entropy of stochastic blockmodel ensembles". In: *Physical Review E* 85.5 (2012), p. 056122.

[34] Tiago P Peixoto. "Parsimonious module inference in large networks". In: *Physical review letters* 110.14 (2013), p. 148701.

[35] Thomas Richardson, Peter J Mucha, and Mason A Porter. "Spectral tripartitioning of networks". In: *Physical Review E* 80.3 (2009), p. 036111.

[36] M Puck Rombach et al. "Core-periphery structure in networks". In: *SIAM Journal on Applied mathematics* 74.1 (2014), pp. 167–190.

[37] Marta Sarzynska et al. "Null models for community detection in spatially embedded, temporal networks". In: *Journal of Complex Networks* 4.3 (2016), pp. 363–406.

[38] Michael T Schaub et al. "Multiscale dynamical embeddings of complex networks". In: *Physical Review E* 99.6 (2019), p. 062308.

[39] Michael T Schaub et al. "The many facets of community detection in complex networks". In: *Applied network science* 2.1 (2017), p. 4.

[40] Sebastijan Sekulić, JA Long, and U Demšar. "The effect of geographical distance on community detection in flow networks". In: *Proceedings of the AGILE 2018 conference, Lund, Sweden*. 2018, pp. 12–5.

[41] M Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. "Patterns of dominant flows in the world trade web". In: *Journal of Economic Interaction and Coordination* 2.2 (2007), pp. 111–124.

[42] Paulo Shakarian et al. "Mining for geographically disperse communities in social networks by leveraging distance modularity". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, pp. 1402–1409.

[43] Filippo Simini et al. "A universal model for mobility and migration patterns". In: *Nature* 484.7392 (2012), pp. 96–100.

[44] Vincent Traag, Patrick Doreian, and Andrej Mrvar. "Partitioning signed networks". In: *Advances in network clustering and blockmodeling* (2019), pp. 225–249.

[45] United Nations Conference on Trade and Development (UNCTAD). *Review of Maritime Transport 2020*. 2020.

[46] Sergei Vassilvitskii and David Arthur. "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2006, pp. 1027–1035.

[47] Anthony J Venables. "Economic Geography and Trade". In: *Oxford Research Encyclopedia of Economics and Finance*. 2019.

[48] Jasper Verschuur, Elco E Koks, and Jim W Hall. "Global economic impacts of COVID-19 lockdown measures stand out in high-frequency shipping data". In: *PloS one* 16.4 (2021), e0248818.

[49] Alan Geoffrey Wilson. “A family of spatial interaction models, and associated developments”. In: *Environment and Planning A* 3.1 (1971), pp. 1–32.

[50] Mengqiao Xu et al. “Modular gateway-ness connectivity and structural core organization in maritime network science”. In: *Nature communications* 11.1 (2020), pp. 1–15.