Appendix A

Community detection

A.1 Spectral methods

A.1.1 Spectral methods

For a bipartition of a network into $g_i \in \{0, 1\}$, it can be shown [26, 35], that optimising the modularity (1.3) is equivalent to optimising the quantity $Q_s = \frac{1}{4m} s^\top Bs$, where

$$s_i = \begin{cases} +1 & \text{if } g_i = 0 \\ -1 & \text{if } g_i = 1 \end{cases}$$

and B is the modularity matrix $B = A - \gamma \hat{A}$. Let the eigenvalues λ_i of B and their corresponding eigenvectors u_i be ordered according to $\lambda_1 \geq \lambda_2 \geq \dots$. The leading eigenvector u_1 of B will give the maximum possible value of $v^\top Bv$ for all real-valued $v \in \mathbb{R}$. Choosing s such that

$$\begin{cases} s_i = +1 & \text{if } v_i > 0, \\ s_i = -1 & \text{if } v_i < 0 \end{cases}$$

gives the maximum value of Q_s for a bipartition of the network. The value of s_i for $u_i = 0$ is then chosen from $\{\pm 1\}$ to maximise Q_s. This is the simplest form of spectral partitioning.

For this dissertation, we optimise the modularity using the spectral bi- and tripartitioning methods developed by Richardson, Mucha, and Porter [35]:

Spectral bipartitioning of a network The classical spectral partitioning algorithm successively bipartitions the network into smaller and smaller subdivisions until no further improvement in modularity are achieved [35]. For a bipartition of a network into $g_i \in \{0, 1\}$, it can be shown [26, 35], that optimising the modularity (1.3) is equivalent to optimising the quantity $Q_s = \frac{1}{4m} s^\top B s$, where

$$s_i = \begin{cases} +1 & \text{if } g_i = 0 \\ -1 & \text{if } g_i = 1 \end{cases}$$

and B is the modularity matrix $B = A - \gamma\langle A\rangle$. Let the eigenvalues λ_i of B and their corresponding eigenvectors u_i be ordered according to $\lambda_1 \geq \lambda_2 \geq \ldots$. The leading eigenvector u_1 of B will give the maximum possible value of $v^\top B v$ for all real-valued $v \in \mathbb{R}$. Choosing s such that

$$\begin{cases} s_i = +1 & \text{if } v_i > 0, \\ s_i = -1 & \text{if } v_i < 0 \end{cases}$$

gives the maximum value of Q_s for a bipartition of the network. The value of s_i for $u_i = 0$ is then chosen from $\{\pm 1\}$ to maximise Q_s. This is the simplest form of spectral partitioning. In order to incorporate information from multiple eigenvectors, the p-eigenvector approach can be used [35]. A set of n *node vectors* are defined according to

$$[r_i]_j = \sqrt{\lambda_j - \lambda_n} U_{ij}$$

where λ_j is the eigenvalue corresponding to eigenvector u_j and $U = \begin{pmatrix} u_1 & u_2 & \ldots \end{pmatrix}$ Then the modularity may be approximated by

$$Q \approx \tilde{Q} = n\lambda_n + \sum_{i=1}^{k} |\mathbf{R}_{G_i}|^2,$$

where there are k communities and $\mathbf{R}_G = \sum_{g_i \in G} r_i$. If v_i is in G then r_i and $\mathbf{R}_G$ cannot be more than 90 degrees apart so $\mathbf{R}_G \cdot r_i > 0$. We also require that the communities are more than 90 degrees apart, so $R_G R_{G'} < 0$ for all $G \neq G'$. This means that in a p-dimensional space, there can be at most $(p+1)$ communities. To perform a bipartition of the node vector space that optimises the modularity one simply needs to find the *codimension-one* hyperplane (passing through the origin) which best bisects the vector space. Once again, this algorithm can be applied recursively to divide the network into an even number of successively smaller parit写itons.

Because spectral bipartitioning only considers bipartitions of a network, partitions with odd numbers of groups will not be considered. Thus the optimal partition may be missed. Richardson et al. propose a spectral *tri*-partitioning algorithm in [35] which at each iteration considers a subdivision of the network into three groups.

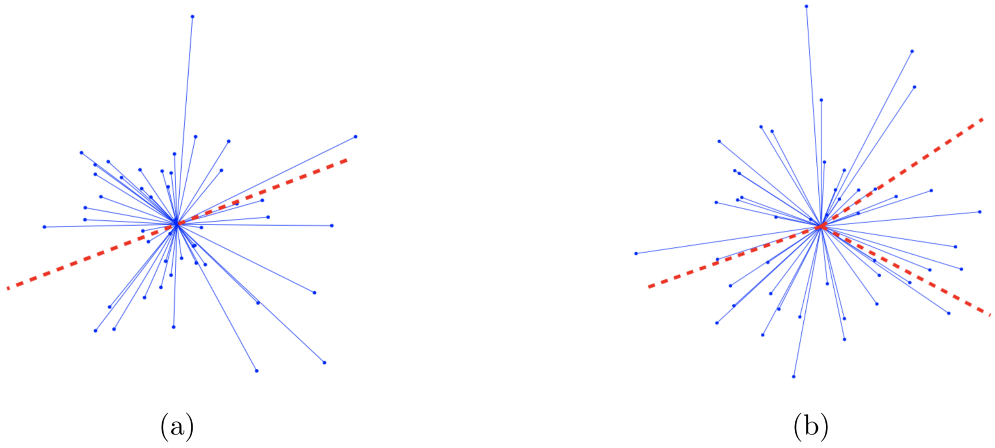


Figure A.1: Spectral bi-partitioning (a) and tri-partitioning (b). *Source:* [35].

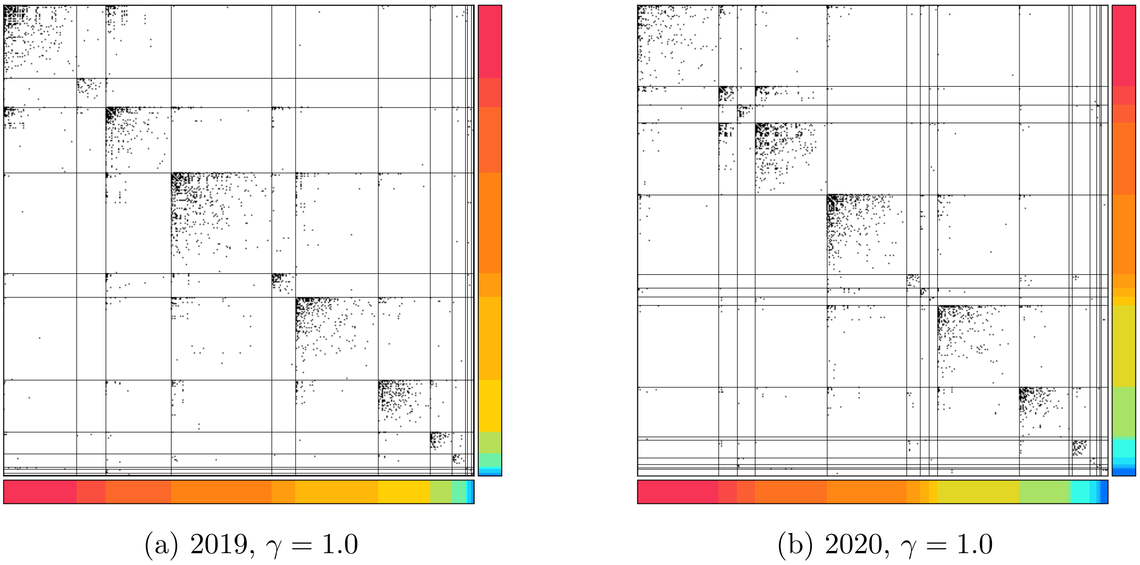
A.2 Classical community detection results

In Section 1.3.2 of the main text we explored the results of classical community detection applied to the 2019 maritime shipping network with resolution parameter $\gamma = 0.2$. In this section we include some results for this method applied to the 2019 and 2020 networks for $\gamma \in \{0.1, 0.2, 0.3, 0.5, 0.8\}$.

A.2.1 The maritime shipping network

In Figures A.2-A.6, we show the results obtained by optimising the Newman-Girvan modularity for a variety of resolution parameters. Setting the resolution parameter $\gamma = 1.0$ yielded 16 communities in both the 2019 and 2020 networks. In Figure A.2, we show the adjacency matrices for $\gamma = 1.0$, with nodes grouped by community and ordered by degree within communities. For γ less than 0.1, most ports were grouped into one large community.

Ports are plotted in their spatial locations and coloured according to their grouping in Figures A.3-A.6 with the three ports with the highest degree, Shanghai, Singapore, and Pusan (South Korea) indicated. As the resolution parameter decreases from 0.8 to 0.1 larger groups aggregate and smaller groups are absorbed into bigger groups. For most resolutions, we see groups dominated by either Asian, European, or North and South American groups. Ports in the Pacific Ocean and a cluster of around 30 South American ports are frequently separated from the major communities. For resolutions $\gamma = 0.5$ and $\gamma = 0.8$ in the 2020 network, London, U.K. is grouped with this group of South American ports, and San Francisco, U.S.A. is grouped with 17 Pacific Island ports. The partition of the larger groups remains relatively consistent across resolutions $0.3 \leq \gamma \leq 0.8$ while the different smaller groupings tend to be isolated at different resolutions. This suggests that several node groups are not as strongly connected to their communities as others. In the adjacency matrices in Figure A.2, there is an indication of core-periphery structures within groups. Nodes in the top-left of each square appear more densely connected than the rest of the group, resembling a core-like structure.

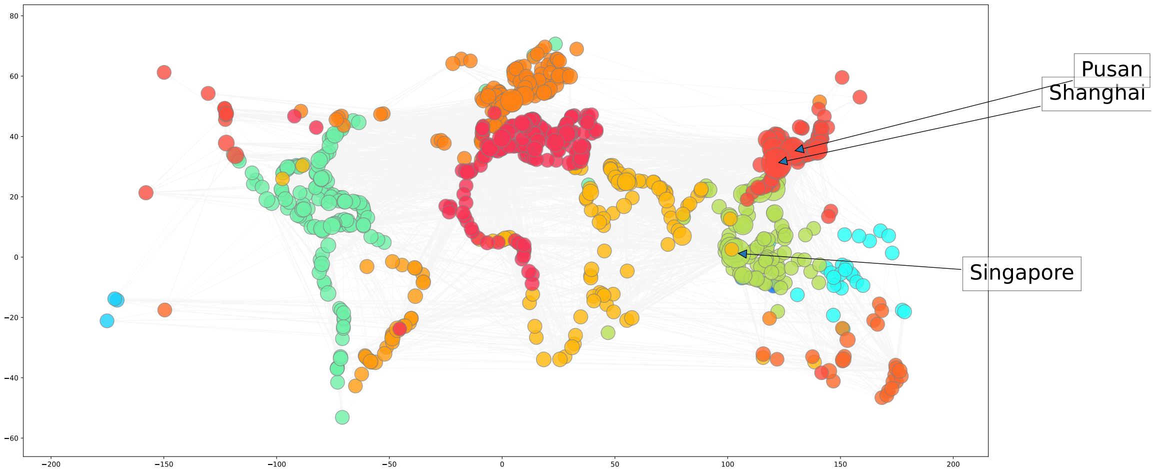


(a) 2019, $\gamma = 1.0$

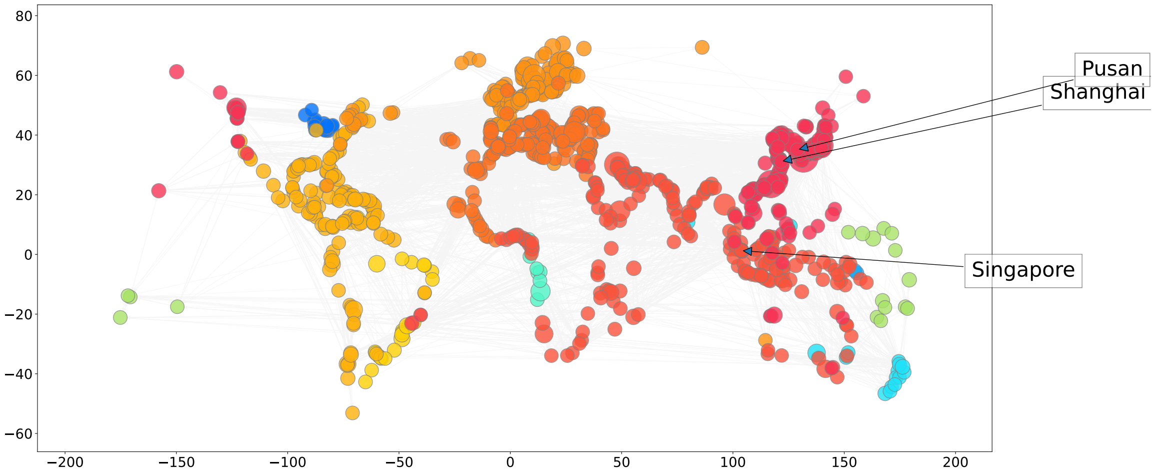
(b) 2020, $\gamma = 1.0$

Figure A.2: Adjacency matrices showing nodes ordered first by grouping, then by degree within a grouping. The colourbars below and to the right of each figure indicate community assignment while black dots indicate the presence of an edge and white space indicates the absence of an edge. *Note that group numbers are arbitrary so the ordering of groups in these adjacency matrices is meaningless.*

Resolution $\gamma = 0.8$ In 2019, 13 communities were found with 88% of ports divided into six major groups, while in 2020, 11 communities were found with 91% divided between five major groups. In both years, a group consisting primarily of Northern European and U.K. ports, a group of North and South American ports, a group consisting primarily of Middle Eastern, Asian and African ports, an Asian group dominated by Japanese and Chinese ports, and a group of mostly Mediterranean and Northwest African ports were isolated. In 2019, a group consisting of largely Southeast Asian ports was also identified. The smaller groups identified were less consistent. In 2019, a smaller group of 31 South American ports was identified, as well as 26 ports predominantly from Oceania. A group of six Pacific Island ports and four Australian ports is also isolated from the larger structure. The four Indonesian ports of Cirebon, Tanah Merah, Pelabuhan Ratu Coal Power Plant and Probolinggo were grouped together. Three ports from Samoa and Tonga are separated into one group, as well as a group of four Chinese ports and a group of only the two Indonesian ports of Waingapu and Ende. In 2020, there were two groups of size 18, the first consisting of San Francisco, U.S.A., and 17 ports in the Pacific Ocean. The second group consisted of mostly New Zealand ports with the Japanese ports of Susaki Ko and Niihama, the Filipino port of Masao and Whyalla, Newcastle, and Port Kembla from Australia. London, U.K. was grouped with the South American ports, which reduced to 29 ports. Buka and Kieta, from Papua New Guinea, are separated from the rest of the network. Eight ports from Gabon, Angola, and the Democratic Republic of the Congo have also been grouped away from the rest of the network.



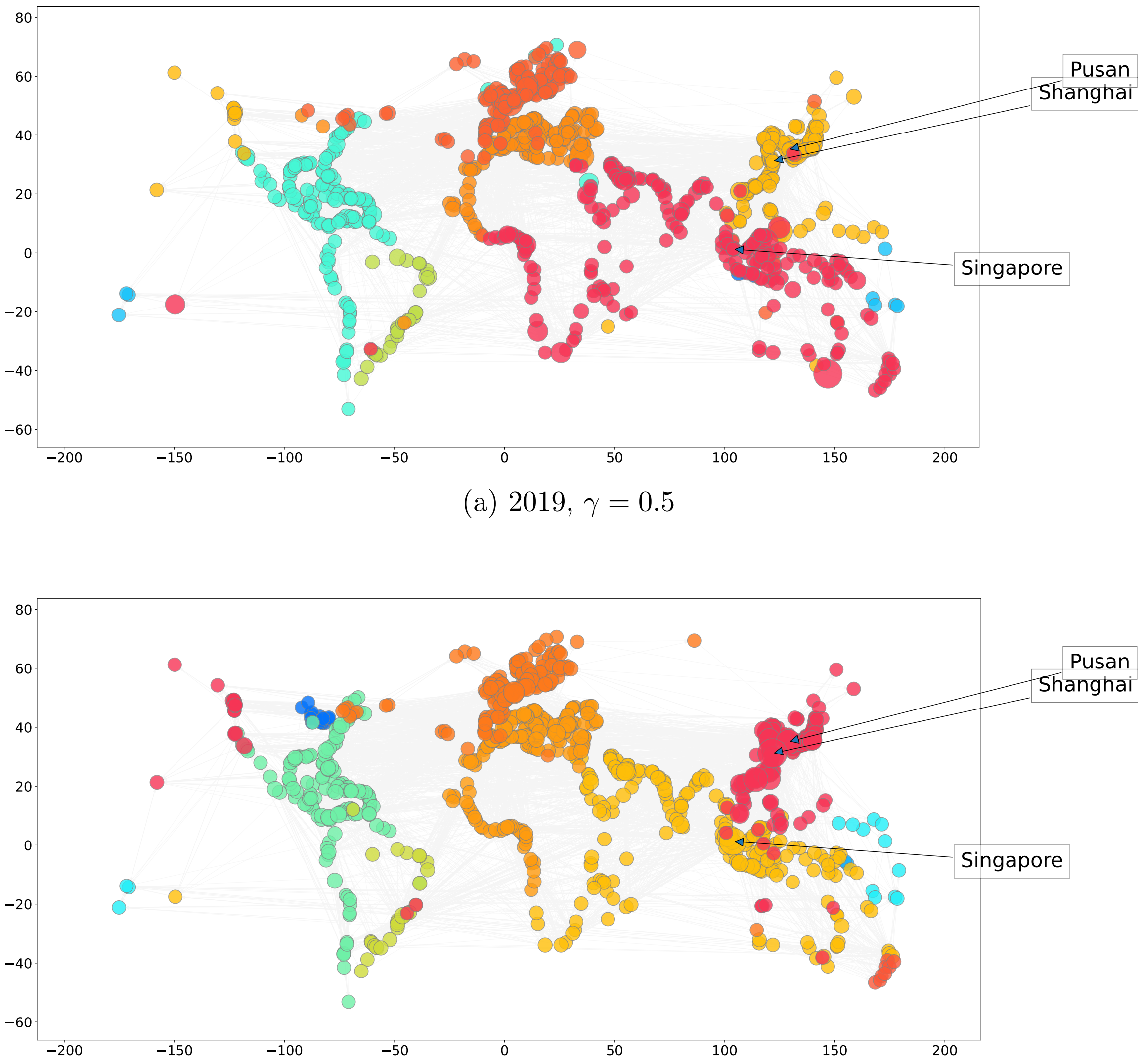
(a) 2019, $\gamma = 0.8$



(b) 2020, $\gamma = 0.8$

Figure A.3: **Visualisation of communities detected using Newman-Girvan modularity by spectral partitioning with resolution $\boldsymbol{\gamma = 0.8}$.**

Resolution $\gamma = 0.5$ In 2019, the algorithm identified eight communities with 95% of nodes placed divided between a similar five groups as for $\gamma = 0.8$. The group of 20 South American ports and the four Indonesian ports as well as a group of eight Pacific Island ports was separated from the larger groups. In 2020, ten communities were found with 93% of the ports split between the five major groups. London U.K. was again grouped with the South American group of 30 ports. There were two communities of size 15. The first consisted of four Canadian ports and 11 U.S. ports. San Francisco remained connected to a slightly smaller group of 14 Pacific Island ports. A group of eight New Zealand ports was also separated, and Buka and Kieta ports from Papua New Guinea were again placed in their own group.

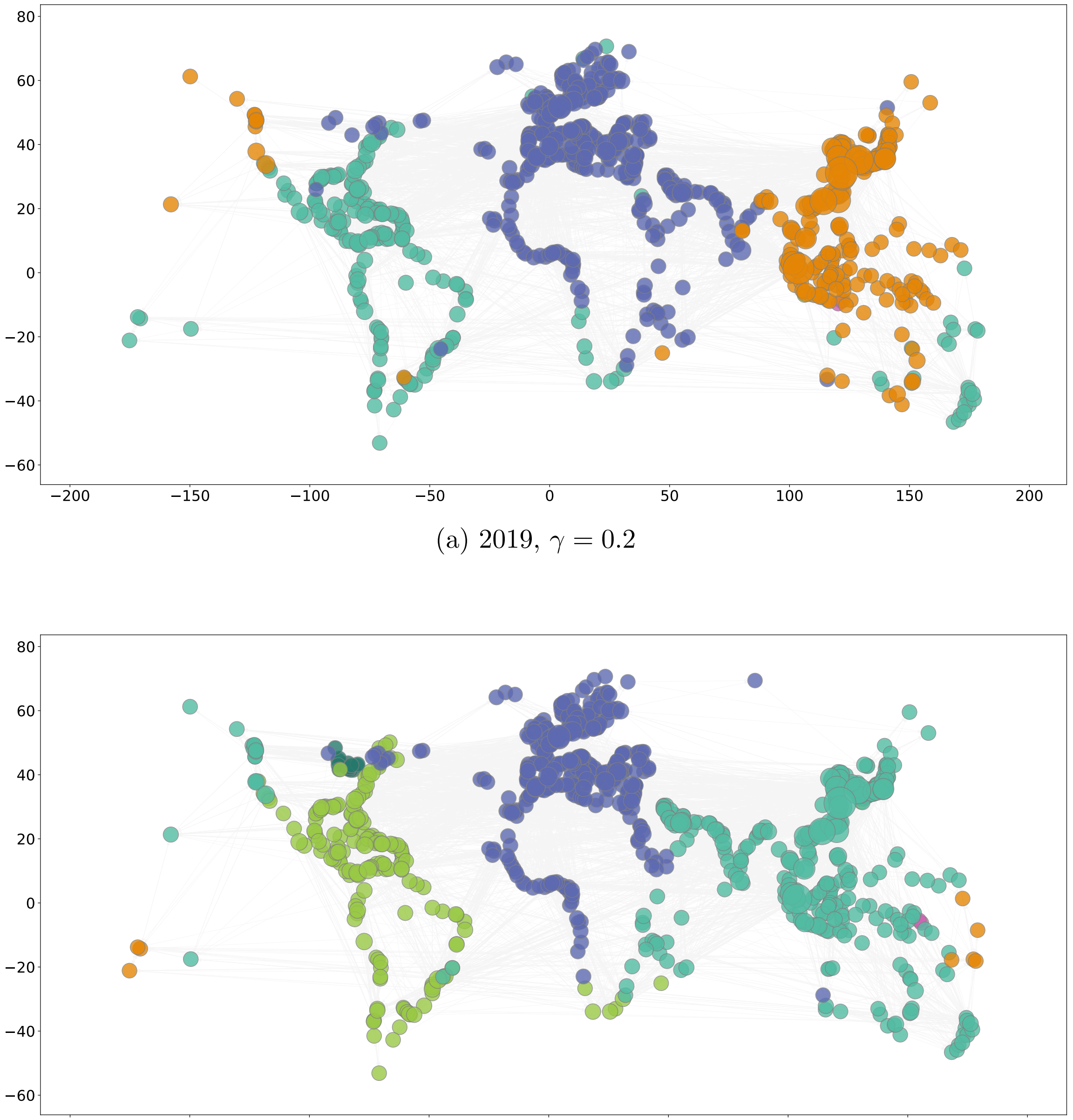


(a) 2019, $\gamma = 0.5$

(b) 2020, $\gamma = 0.5$

Figure A.4: **Visualisation of communities detected using Newman-Girvan modularity by spectral partitioning with resolution $\boldsymbol{\gamma = 0.5}$.**

Resolution $\gamma = 0.2$ For resolution $\gamma = 0.2$ 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. 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. In 2020 there were six communities with 97% of ports grouped into one of three major groups. Here, we note the majority of nodes from Southwest Africa and the Middle East were reassigned to the Asian/Middle Eastern group in 2020. Of the small communities, the first community formed consists of four Canadian and ten U.S. ports, all part of the Great Lakes Maritime System. The second small group consisted of eight Pacific Island ports. Lastly, the two ports of Buka and Kieta in Papua New Guinea were again grouped separately.

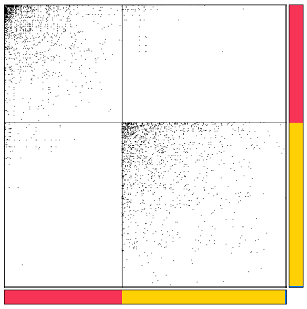


(a) 2019, $\gamma = 0.2$

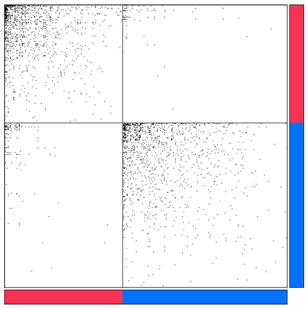
(b) 2020, $\gamma = 0.2$

Figure A.5: **Visualisation of communities detected using Newman-Girvan modularity by spectral partitioning with resolution $\gamma = 0.2$**.

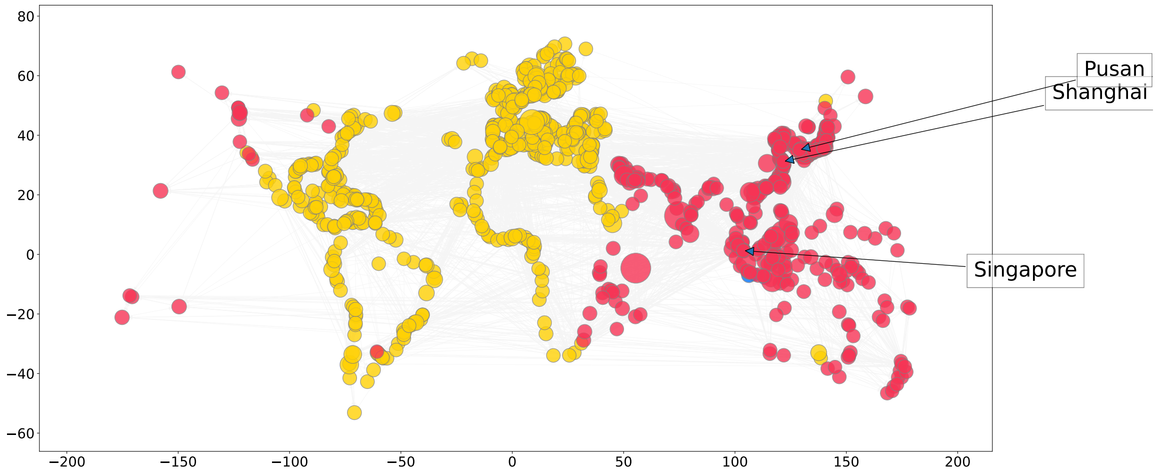
Resolution $\gamma = 0.1$ There were three communities found in the 2019 network with 99% of ports placed in two major groups. In both years, ports appear to be assigned to a group based on being East or West of the 100° West or 50° East lines or on a continent-basis. In 2019, the Indonesian ports of Cirebon, Tanah Merah, Pelabuhan Ratu Coal Power Plant, and Probolinggo were separated from the two major groups and in 2020 all the ports were divided into one of two main groups.



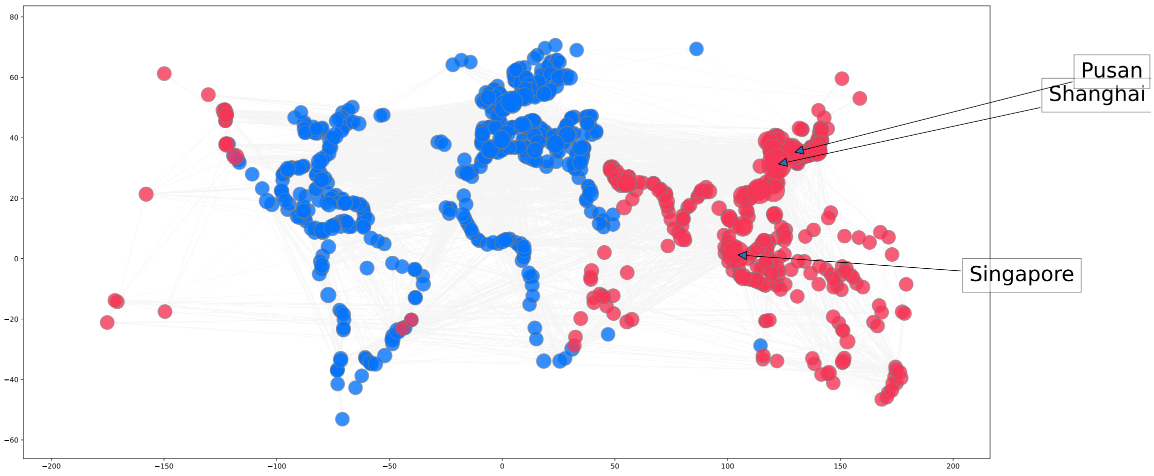
(a) 2019, $\gamma = 0.1$



(b) 2020, $\gamma = 0.1$



(c) 2019, $\gamma = 0.1$



(d) 2020, $\gamma = 0.1$

Figure A.6: **Visualisation of communities detected using Newman-Girvan modularity by spectral partitioning with resolution $\boldsymbol{\gamma = 0.1}$**.

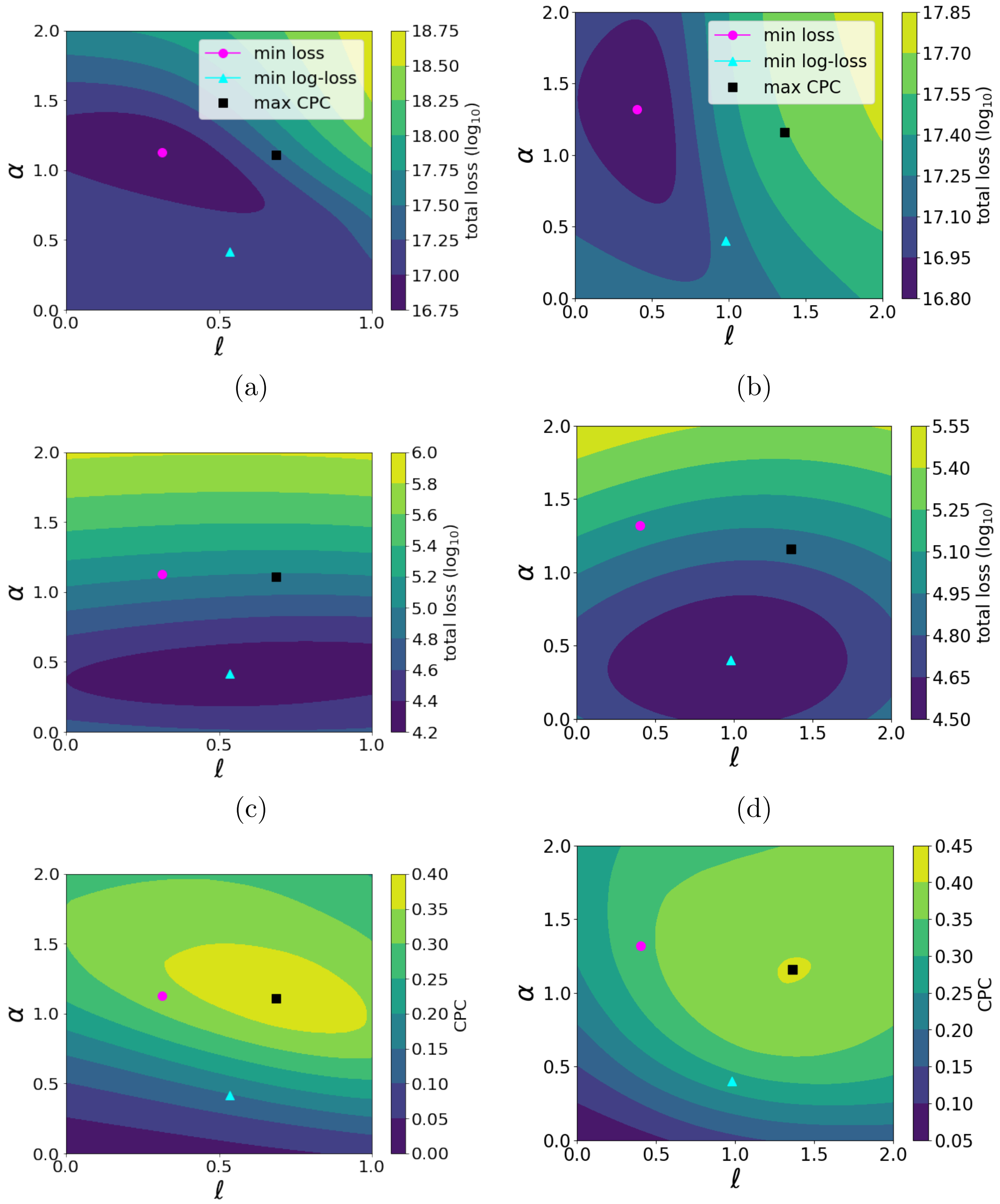
Appendix B

Spatially-corrected community detection

In Figure B.1, we visualise the full set of optimisation landscapes from Section 2.1.3 for the unconstrained (left panel) and attraction-constrained (right panel) models on the network of shipping data for three different metrics. The loss function (2.10), in the top row, the log of (2.10) in the middle row, and the common part of commuters (CPC) index [10] in the bottom row. The CPC index is widely used in Applied Mathematics and 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 and yellow points, and the maximum of the CPC is indicated by a black square in each plot.

B.1 Mobility models

B.1.1 Visualisations for the common neighbours model



(a)

(b)

(c)

(d)

(e)

(f)

Figure B.1: The optimisation landscapes for the unconstrained (left panel) and attraction-constrained (right panel) for three different metrics. The loss function (2.10), in the top row, the log of (2.10) in the middle row, and the CPC [10] in the bottom row. Minima of the loss functions are indicated by pink and yellow points, and the maximum of the CPC is indicated by a black square in each plot.

B.2 Synthetic spatial benchmarking networks

Throughout the dissertation, we use the normalised mutual information score to assess the similarity between two partitions. Here, we present its formal definition.

Definition B.2.1 (Normalised mutual information)**.** Normalised mutual information relies on the concepts of entropy and mutual information [13, 11]. Let X and Y be two random variables which can take values $x \in \Omega_X$ and $y \in \Omega_Y$, where Ω_X and Ω_Y are discrete sets. The entropy of X is given by

$$H(X) = -\sum_{x \in \Omega_X} p(x) \log_2(x)$$

The mutual information between X and Y is given by

$$MI(X, Y) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}.$$

This is normalised to give the normalised mutual information

$$\text{NMI}(X, Y) = \frac{2 \times MI(X, Y)}{H(X) + H(Y)}.$$

If the information in X is captured by the information in Y, then $MI(X, Y) = H(X) = H(Y)$, and $NMI(X, Y) = 1$. If X and Y contain completely different sets of information, then $MI(X, Y) = 0$ and $NMI(X, Y) = 0$.

B.3 Synthetic spatial benchmarking networks

B.3.1 The uniform model

The uniform model by Expert *et al.* 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.

The parameters used for assortativity and edge density are, respectively, λ and ρ. First, n nodes are randomly placed in some spatial location and assigned a binary attribute value of either $g_i = 0$ or $g_i = 1$. For any two nodes v_i and v_j, their community assignment is denoted by $g_i \in \{0, 1\}$. The expected number of flows between them is calculated according to

$$p_{ij}^{\text{Exp}} = \frac{1}{Z} \frac{\lambda_{g_i g_j}}{d^\ell} \tag{B.1}$$

where Z is a normalisation constant such that $\sum_{i \neq j} p_{ij} = 1$ and [10]. The Λ matrix takes the form

$$\Lambda = \begin{pmatrix} \lambda_{00} & \lambda_{01} \\ \lambda_{10} & \lambda 11 \end{pmatrix} = \begin{pmatrix} 1 & \lambda \\ \lambda & 1 \end{pmatrix}. \tag{B.2}$$

In this way, varying λ provides a simple way to control graph assortativity. For $\lambda < 1$ the graph is assortative, with disconnected communities for $\lambda = 0$. For all $\lambda > 1$ the graph has a primarily disassortative community structure. The parameter ρ controls the total number and weight of flows in the network, which will be given by $m = \rho n(n-1)$ in the directed case. A synthetic graph showing an assortative regime with $\lambda = 0.1$ and $\ell = 2.0$ is shown plotted with the nodes in their spatial locations in Figure B.2. Networks resulting from three different values of λ are shown in Figure 2.5. Figure 2.5a. shows a network where there are no connections between communities corresponding to $\lambda = 0$, Figure 2.5b. shows a predominantly assortative network corresponding to $\lambda = 0.1$ and Figure 2.5c. shows a network with a highly disassortative structure corresponding to $\lambda = 20$. A variation of this is proposed in [10] where $\lambda_{12} = \lambda + 0.1$ and $\lambda_{21} = \lambda - 0.1$ which induces a net flow between the communities.

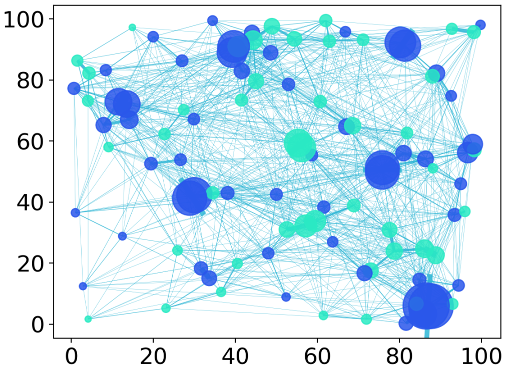


Figure B.2: Nodes plotted in their spatial locations for an assortative synthetic network generated according to the uniform model in eq. (B.1), [10, 16] with $\ell = 2$, $\lambda = 0.1$, $\rho = 1.0$ and $N = 100$.

B.3.2 The correlated group membership model

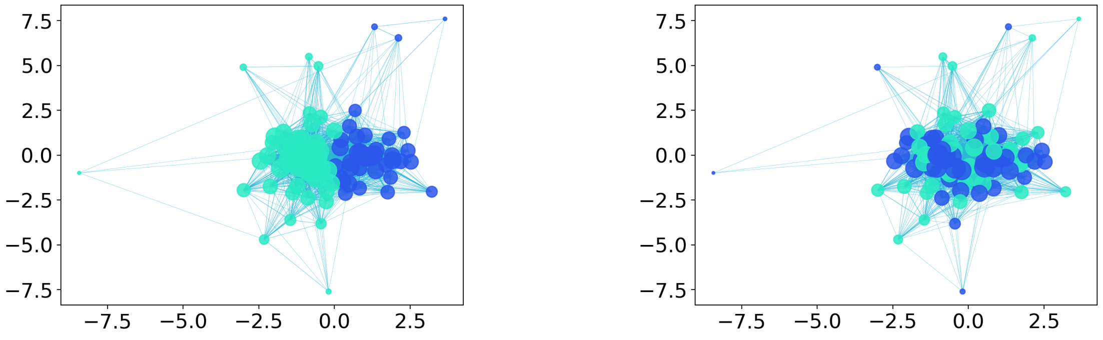
The correlated group membership model is constructed as follows: in an (x, y)-plane two centers are chosen, and as in [10], we place these centers at $(x \pm L, 0)$, where $L > 0$ is an value to be specified. Then, $n/2$ nodes are placed around each center and the distance d_{ci} of each node v_i from its center c is determined proportional to the exponential distribution with scale ζ, i.e. $p(d_{ci}) \propto e^{d_{ci}/\zeta}$.

To assign attributes in the fully correlated model (when a node's attributes are fully decided by its spatial location) we simply bisect the plane about $x = 0$. All nodes in $x < 0$ are assigned community $g_i = -1$ with probability $q = 1$ and all nodes in $x > 0$ are assigned to community $g_i = +1$ with probability $q = 1$. To introduce some randomness we can change q to $q = 1 - \epsilon$, with $\epsilon \in [0, 0.5]$. Thus, ϵ allows one to tune the degree of correlation between space and attribute assignment. We see that $\epsilon = 0.5$ results in the fully random case where space plays no role in community.
The links between nodes in the model are then assigned according to

$$p_{ij}^{\text{Cer}} = \frac{1}{Z}\exp\left[\beta\left(g_i g_j - \frac{d_{ij}}{\beta\zeta}\right)\right], \tag{B.3}$$

where Z is again the normalisation constant as in Eq. B.1. The number of edges is also determined by ρ as in Eq. B.1. The network produced by (B.3) is assortative [10] but may be made disassortative if we instead use

$$p_{ij}^{\text{Cer}} = \frac{1}{Z}\exp\left[\beta\left(-g_i g_j - \frac{d_{ij}}{\beta\zeta}\right)\right]. \tag{B.4}$$



(a)

(b)

Figure B.3: Two extremes of the correlated group membership model by Cerina *et al.* [9]: the first (a) has $\epsilon = 0.0$ which results in attribute assignment being fully dependent on space, and the second (b) uses $\epsilon = 0.5$ where space and attributes are entirely uncorrelated[1].

The value of the product $\beta\zeta$ controls the effect of space on link formation. For $\beta\zeta \ll 1$ space is the dominant factor and for $\beta\zeta \gg 1$ attribute assignment is more important [10]. For this dissertation, ζ is fixed at $\zeta \equiv 1$, so β controls the role of space in link formation.

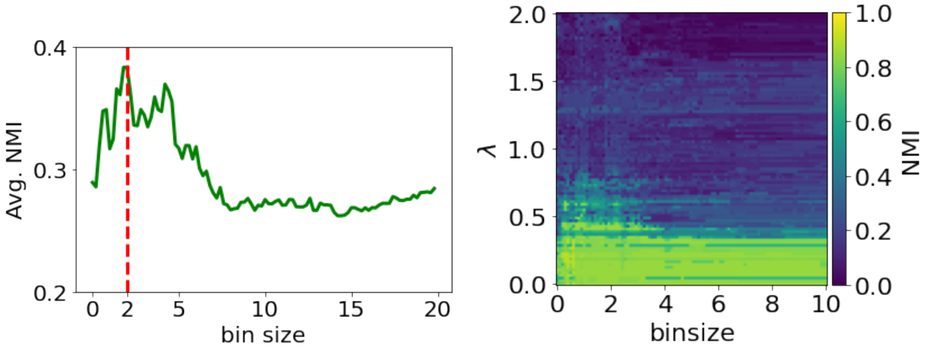
[1] code for this adapted from https://github.com/rodrigolece/spatial-nets

B.4 More results

B.4.1 Synthetic networks

Choosing a binning distance for the Expert model

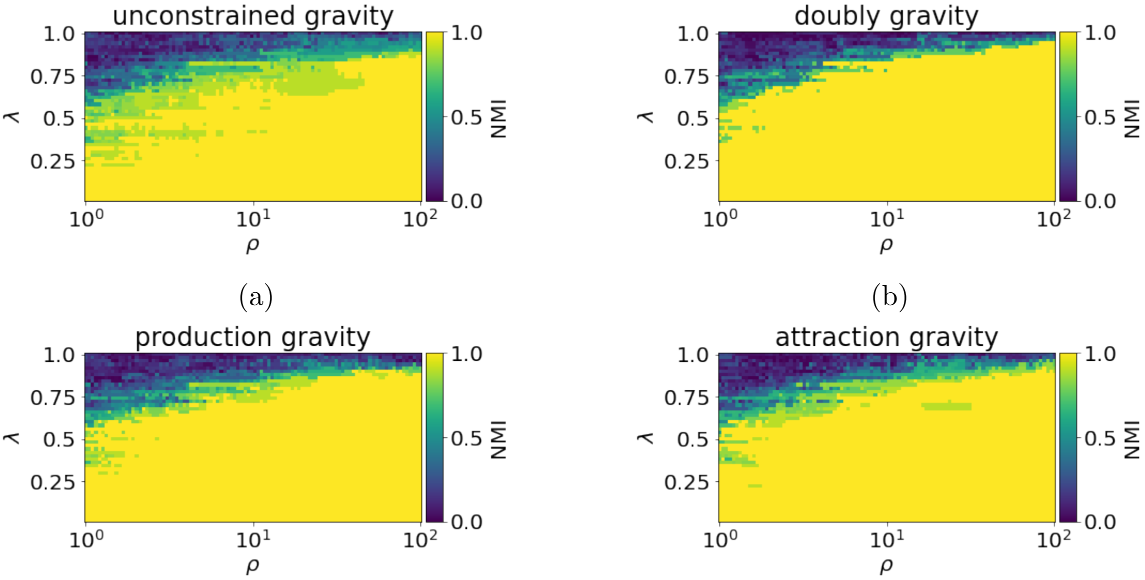
Since the Expert *et al.* distance decay function (2.2) depends on a binning procedure, we test to what degree bin size affects the results of the algorithm. In Figure B.4 we compare NMI scores between the true and predicted partitions for a range of different bin sizes and $\lambda \in [0, 2]$. For each parameter search, we construct synthetic networks using the uniform model as proposed by Expert *et al.* [16] with $n = 20$ nodes in a 10×10 square, and the random seed set to 0 for reproducibility. We see that for highly assortative graphs ($\lambda < 0.5$), there is no huge difference in performance between bin sizes. For slightly more disassortative graphs, the algorithm performs better for bin sizes less than or equal to two, so we chose this value of bin size to use in our comparative parameter searches in Section 2.3.1.



(a) (b)

Figure B.4: **NMI in (binsize, λ)-space with fixed $\rho = 1$.** The NMI scores are calculated for predictions of Expert *et al.*'s spatially-corrected modularity function on the uniform benchmarking network for a 100×100 gridsearch of $\lambda \in [0, 2]$ and binsize $\in (0, 10)$. It is clear from calculating the average NMI across each binsize (a) that the optimal binsize occurs near $\lambda = 2$.

Results of the gravity model in the one-step method Here we show some further results for the one-step method using the gravity model family in Chapter 2.



(a)

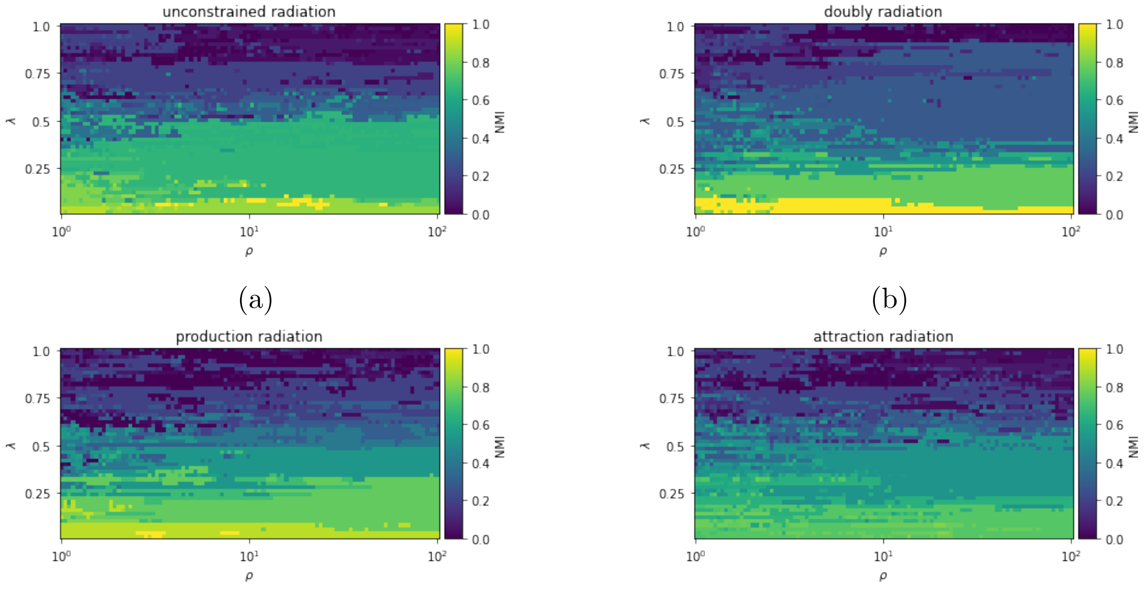
(b)

(c)

(d)

Figure B.5: **Normalised mutual information scores for the one-step method with the gravity model family applied to directed, uniform benchmarking networks.** Parameter searches were run with $\lambda \in [0, 1]$ and $\rho \in [1, 100]$. For each (ρ, λ)-pair a network of $n = 20$ nodes with a known binary partition was generated. Community detection was then performed using the one-step method with each member of the gravity model family (a)-(d) [49].

Results of the radiation model in the one-step method Here we show some further results for the one-step method using the radiation model family in Chapter 2.



(a)

(b)

(c)

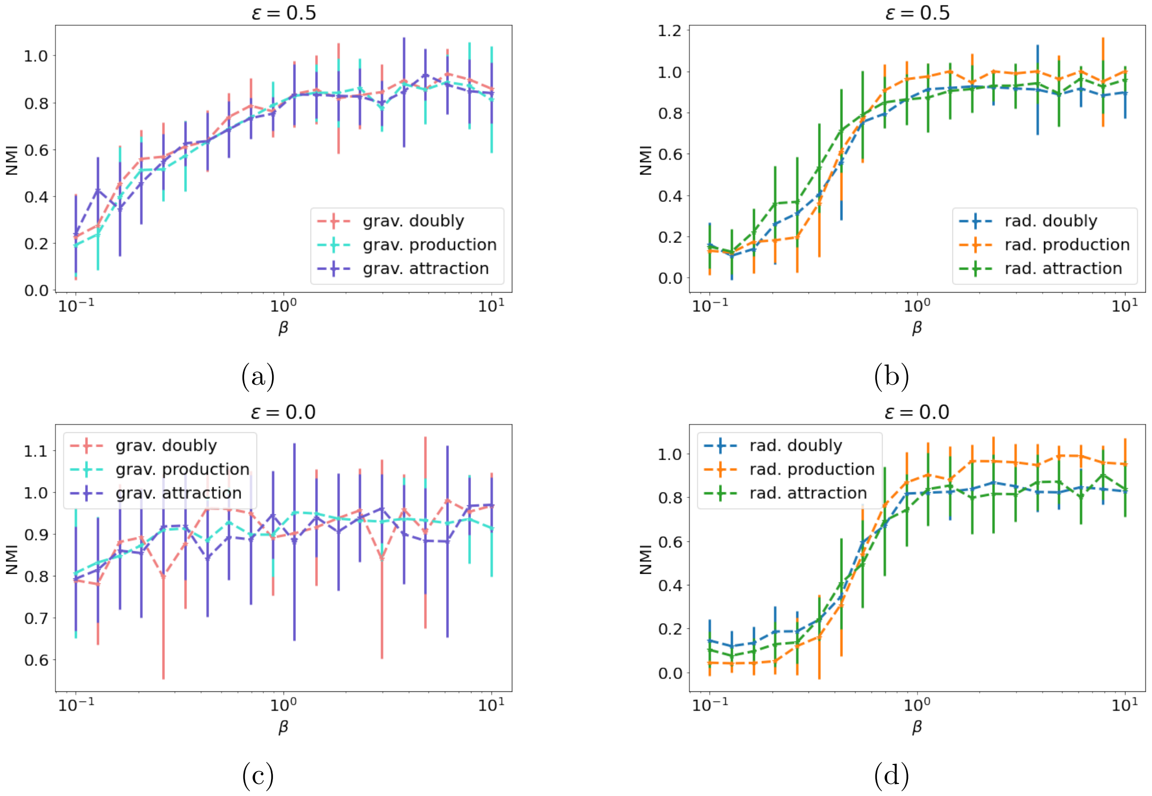
(d)

Figure B.6: **Normalised mutual information scores for the one-step method with the radiation model family, applied to directed, assortative synthetic graphs.** Parameter searches were run over a 100 × 100 grid with $\lambda \in [0, 1]$ and $\rho \in [0, 100]$. For each (ρ, λ)-pair a network of 20 nodes with a known binary partition was generated. Community detection was then performed on it using the one-step method with a member of the radiation model family [49].

Null Model	Avg. Time	Avg. Modularity	Avg. NMI
Unconstrained	0.0485	0.6009	0.4223
Production	0.0529	0.4561	0.4222
Attraction	0.0522	0.5735	0.3944
Doubly	0.0534	0.4491	0.4016

Table B.1: **Averaged results for one-step community detection using the radiation 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 for the two-step method Here we show visualisations of results for the two-step method in Chapter 2.



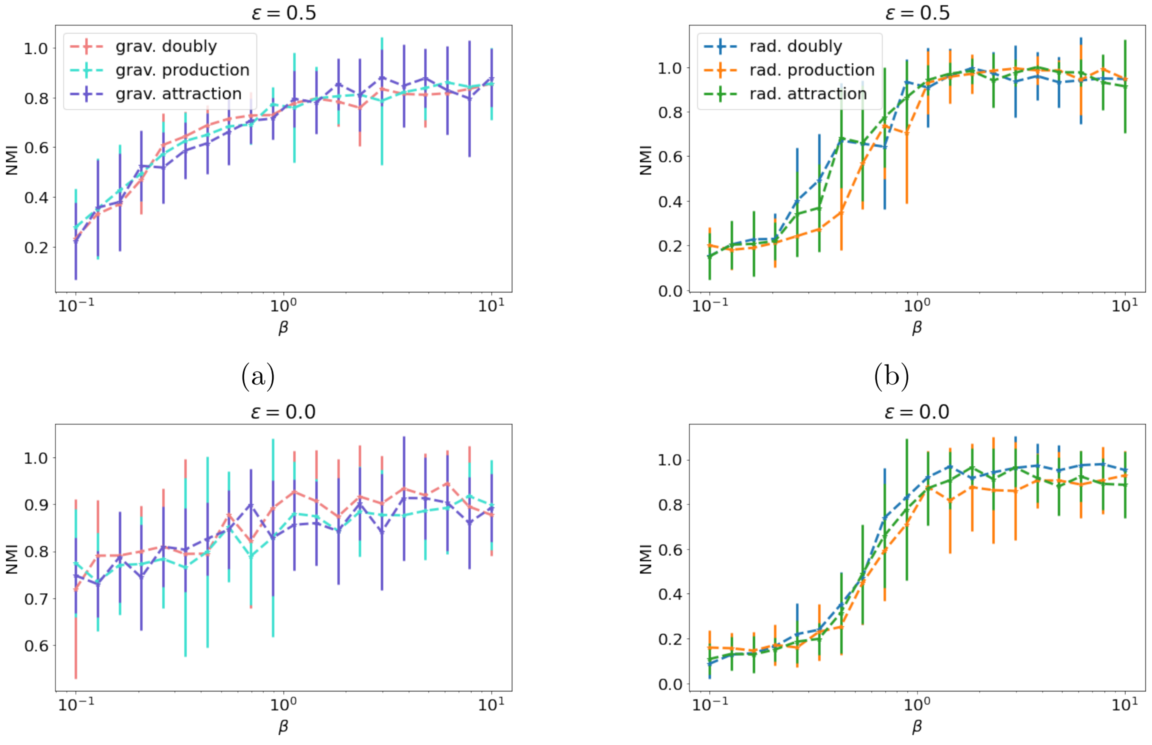
(a)

(b)

(c)

(d)

Figure B.7: **Results of the two-step method on correlated group membership model synthetic networks with Newman-Girvan modularity.** 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 $[10^{-1}, 10^{1}]$. Community detection was carried-out by first extracting the doubly-constrianed gravity (left) or radiation (right) spatial backbone, then performing community detection on the resulting signed network using the Newman-Girvan null model. 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.



(a)

(b)

(c)

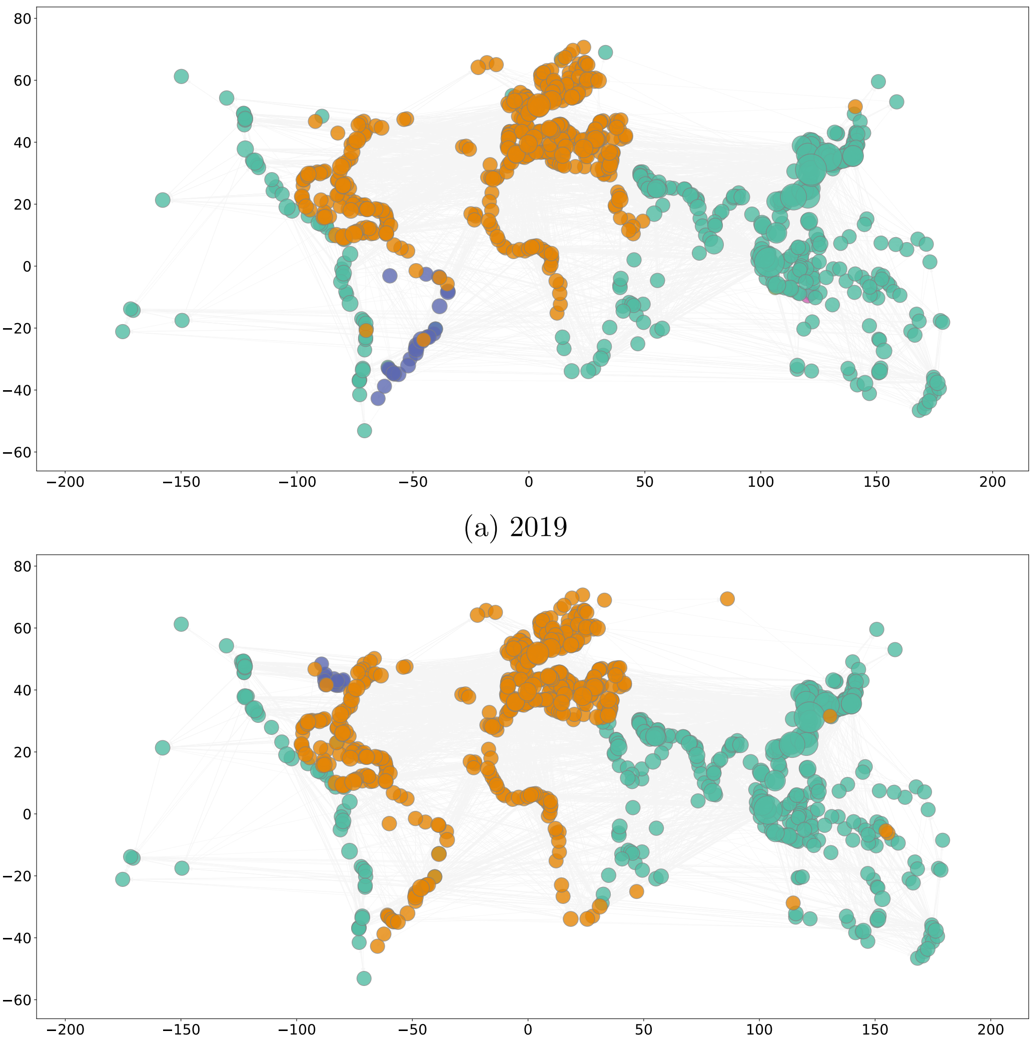
(d)

Figure B.8: **Results of the two-step method on correlated model synthetic networks with Erdős-Rényi modularity.** 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 $[10^{-1}, 10^{1}]$. Community detection was carried-out by first extracting the doubly-constrianed gravity (left) or radiation (right) spatial backbone, then performing community detection on the resulting signed network using the Erdős-Rényi null model. 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 $\beta \ll 1$ regime corrresponds to space having no impact on link formation while the $\beta \gg 1$ regime corresponds to space being the main factor. The error bars here represent one standard deviation.

B.4.2 The maritime shipping network

Results of the one-step method

Gravity model with resolution $\gamma = 0.2$ Figure B.9 shows ports plotted spatially for 2019 (a) and 2020 (b) networks, with community assignment denoted by colour. We include the 2019 network again to allow for easy comparison. Five communities are found at resolution $\gamma = 0.2$ in 2019 (b), 47% of ports are placed in a group that encompasses all trans-Pacific routes, and 49% of ports are placed in a group that contains all transatlantic routes. A group of 26 South American ports is shown in purple. Not visible on this figure, six Indonesian communities are also divided into two separate groups. In the 2020 network, there are just three communities at resolution $\gamma = 0.2$. The number of ports in both the trans-Pacific and transatlantic communities increased. A community consisting of 14 ports in the Great Lakes Maritime System is also shown in purple.



(a) 2019

(b) 2020

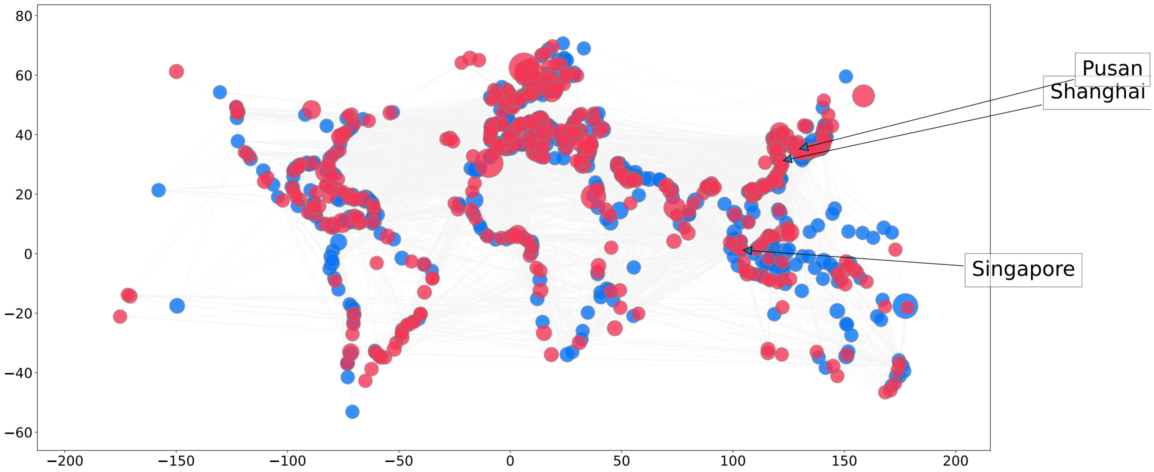
Figure B.9: **Visualisation of spatially-corrected communities detected using the one-step method with the doubly-constrained gravity model on the container ship networks.** Ports are shown in their spatial locations and groupings are denoted by colour. Five communities are found at resolution $\gamma = 0.2$ in 2019, 47% of ports are placed in a group which encompasses all trans-Pacific routes (teal), and 49% of ports are placed in a group which contains all transatlantic routes (orange).

Looking at the community structures in further detail, we make some more specific observations. Before spatial correction, six communities were detected at resolution $\gamma = 0.2$, of which 99% of ports belonged to one of three major groups. Using the gravity model, five communities are detected, of which 91% of ports belong to one of two major communities. The American continent becomes more divided, and the West Coast of America switches from a community that encompassed the entire continent of America in the Newman-Girvan communities, to a large community of

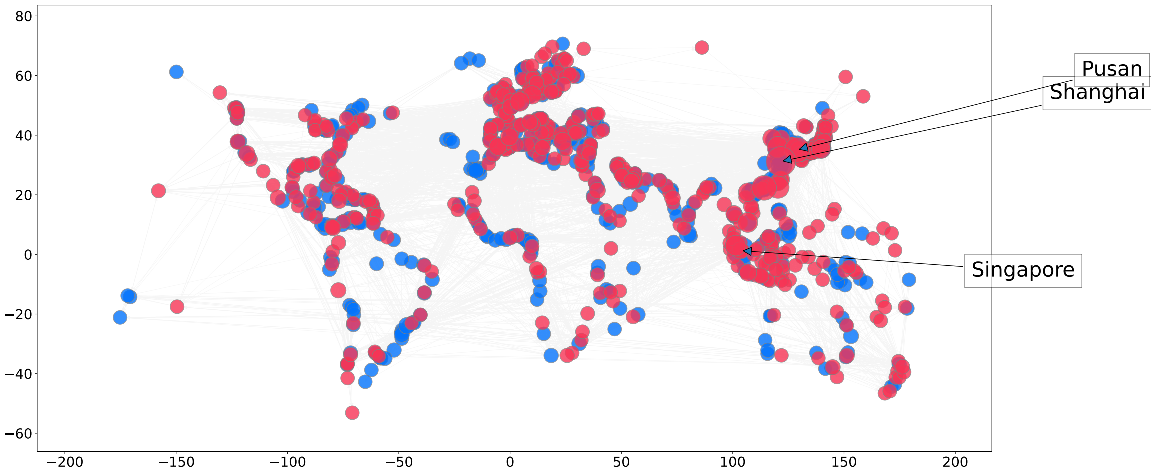
448 ports (∼47% of the network) containing most of Asia. Much of the Middle East and South and East Africa are removed from the European and North African groups and also placed in this community. Most of North America is now identified as part of the European and North African community, in a group of 470 ports (∼49%). A group of 26 South American ports, predominantly from Brazil, is isolated from the major communities in 2019 but rejoins the transatlantic community in 2020. While most ports experience a decline from January-August 2020, many Brazilian ports experienced an increase [48]. It appears this was sufficient to reintegrate Brazil into the network as this group of ports re-enters a major group in 2020.

Introducing the gravity model for the 2020 network Figure B.9(b) results in far fewer communities than detected by the classical algorithm. Instead of six, there are now just three communities, two large, and a North American community of just 14 ports. Again, we see transatlantic and trans-Pacific trade weighted more heavily, and the East and West coasts of the American continent are separated, with the Eastern coast reassigned to the group containing Europe and North and West Africa.

Radiation model with $\gamma = 0.8$ The results for $\gamma = 0.8$ are shown in Figure B.10b. Two communities are identified. The number of communities then increases rapidly with γ, seven communities are observed for $\gamma = 0.85$ and 29 for $\gamma = 1.0$. However, for both $\gamma = 0.85$, the major groups remain consistent with the 0.8 cases. Visualising the communities spatially (Figure B.10) is not as informative as for the gravity and Newman-Girvan modularities, so we include Tables B.2 and B.3 showing the ten ports the highest degree in each community for 2019 and 2020.



(a) Doubly-constrained radiation modularity, 2019, $\gamma = 0.8$



(b) Doubly-constrained radiation modularity, 2020, $\gamma = 0.8$

Figure B.10: **Visualisation of communities on the 2020 network detected using Newman-Girvan and doubly-constrained radiation modularity with resolution $\gamma = 0.8$**. The radiation model splits the network into two groups with a much less obvious interpretation.

2019	2020
Singapore, Singapore	Singapore, Singapore
Ningbo, China	Rotterdam, Netherlands
Rotterdam, Netherlands	Ningbo, China
Xiamen, China	Xiamen, China
Shekou, China	Shekou, China
Kobe, Japan	Tokyo Ko, Japan
Nagoya Ko, Japan	Mina Jabal Ali, United Arab Emirates
Colombo, Sri Lanka	Yiantian, China
Piraievs, Greece	Colombo, Sri Lanka
Osaka, Japan	Bremerhaven, Germany

Table B.2: Ports with the highest degree for 2019 and 2020, for group one as found by the doubly-constrained radiation modularity with resolution $\gamma = 0.8$.

2019	2020
Shanghai, China	Shanghai, China
Pusan, South Korea	Pusan, South Korea
Hong Kong, Hong Kong	Hong Kong, Hong Kong
Kao Hsiung, Taiwan	Kao Hsiung, Taiwan
Qingdao Gang, China	Qingdao Gang, China
Port Klang, Malaysia	Port Klang, Malaysia
Tokyo Ko, Japan	Thanh Ho Chi Minh, Vietnam
Thanh Ho Chi Minh, Vietnam	Antwerp, Belgium
Yokohama Ko, Japan	Yokohama Ko, Japan
Antwerp, Belgium	Tianjin Xin Gang, China

Table B.3: Ports with the highest degree for 2019 and 2020, for group two as found by the doubly-constrained radiation modularity with resolution $\gamma = 0.8$.

Communities detected by the common neighbours model with $\gamma = 0.2$ Here we include some results of the common neighbours model, where the production-constrained model has been used to detect communities in the 2020 network. While the code for using the production and attraction-constrained models for the one and two-step methods is complete, it has not been sufficiently tested on benchmarking, which is why these results have not been considered in any great detail.

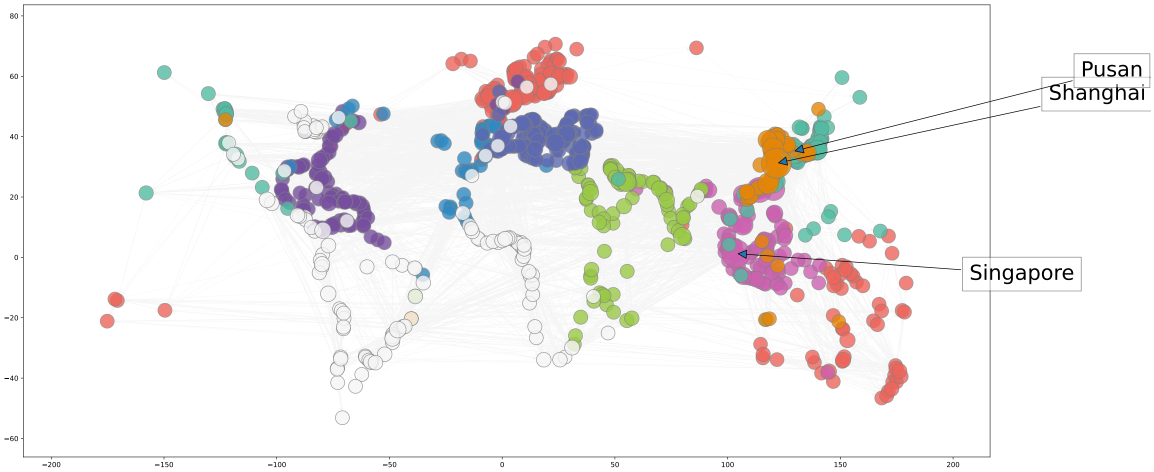
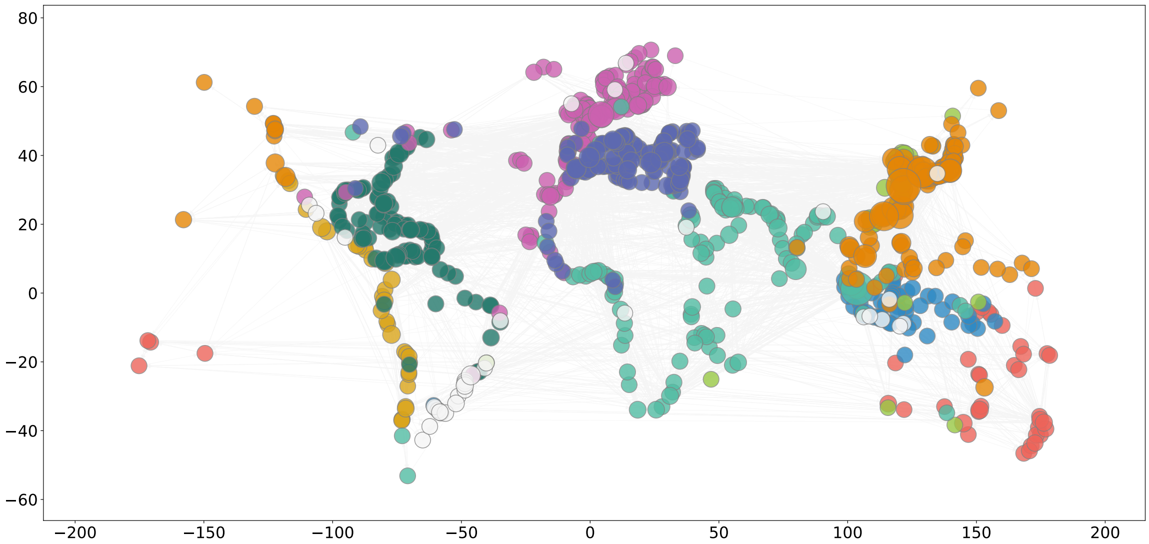


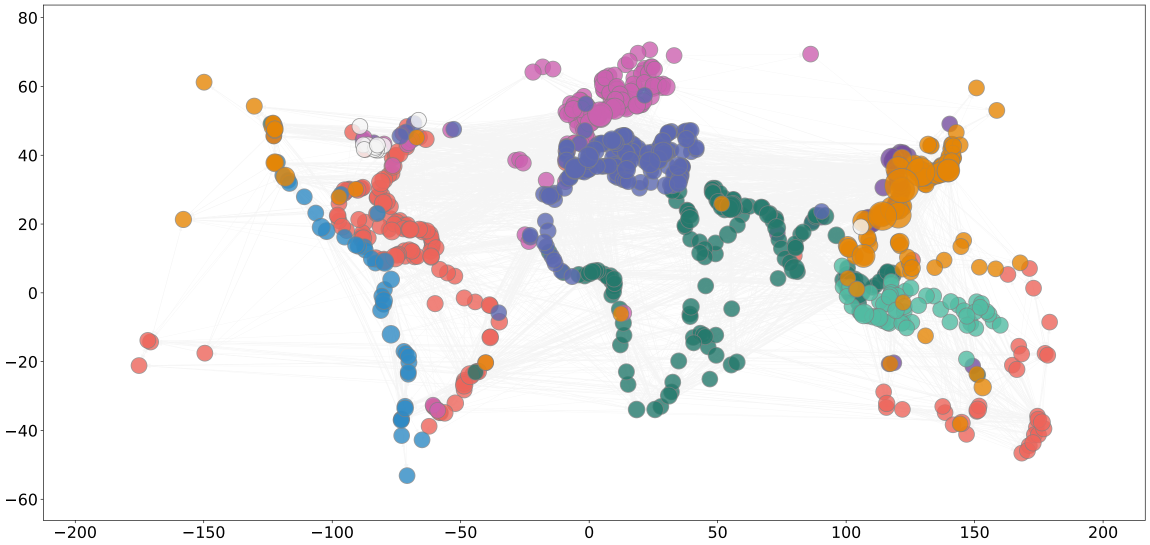
Figure B.11: **Visualisation of spatially-corrected communities detected using the one-step method with the production-constrained common neighbours+sea distance model and on the 2019 container ship networks.**

Results of the two-step method

We include the 2019 and 2020 results of the two-step method for community detection with the radiation model here to allow for easy comparison between years.



(a) Two-step method with radiation model, 2019, $\gamma = 0.2$



(b) Two-step method with radiation model, 2020, $\gamma = 0.2$

Figure B.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 container ship networks.** The methods detected 14 communities and 15 communities for the 2020 network at resolution $\gamma = 0.2$. Ports are shown in their spatial locations and groupings are denoted by colour. Ports in communites of size less than 30 are shown in white for visual clarity, which reduces the number of 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.

In general, the radiation model attributes less import to transatlantic and trans-Pacific crossings than the radiation model. While the gravity model groups Southern Africa with the transatlantic community, the radiation model groups it with Middle Eastern and Southern Asia.

Appendix C

Spatially-Corrected Core-Periphery Detection

C.0.1 Likelihood of the DCPM

Elliot *et al.* calculate the likelihood of this block structure without degree correction by following the procedure of Karrer and Newman [13, 19], where in this case there are four blocks $g_i \in \mathcal{P} = \{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$. In Karrer and Newman, this derivation is given for undirected networks so we briefly review the derivation of this likelihood in the directed case. For all (i, j), we assume the number of edges from node v_i to node v_j to be independently Poisson distributed with expected value ω_{rs} where r and s represent the group assignments $g_i = r$ and $g_j = s$. In other words, we consider A_{ij} as a random variable such that $A_{ij} \sim \text{Poi}(\omega_{rs})$. Thus, the probability mass function for each edge is

$$p(A_{ij}) = \frac{\omega_{rs}^{A_{ij}} \mathrm{e}^{-\omega_{rs}}}{A_{ij}!}$$

The probability of a graph G, given the parameters ω and the partition g is the probability of all its edges having weights A_{ij} and is given by the product of these probabilities

$$\mathrm{P}(G|\omega, g) = \prod_{i,j} \frac{\omega_{g_i g_j}^{A_{ij}} \mathrm{e}^{-\omega_{g_i, g_j}}}{A_{ij}!}.$$

This can be rewritten as

$$\mathrm{P}(G|\omega, g) = \frac{\prod_{r,s} \omega_{rs}^{m_{rs}} \mathrm{e}^{-n_r n_s \omega_{rs}}}{\prod_{i,j} A_{ij}!}.$$

where m_{rs} is the number of edges from group r to group s and n_r is the number of nodes in group r. Taking the logarithm of this expression and ignoring any terms that are independent of the parameters ω or the partition g, it is possible to obtain the expression

$$\log \mathrm{P}(G|\omega, g) = \sum_{r,s} \Big(m_{rs} \log \omega_{rs} - n_r n_s \Big). \tag{C.1}$$

Differentiating $\log \mathrm{P}(G|\omega, g)$ and setting this value to zero, we obtain a value for ω_{rs} which maximises the log-likelihood C.1,

$$\hat{\omega}_{rs} = \frac{m_{rs}}{n_r n_s}.$$

This estimate for $\hat{\omega}_{rs}$ can be substituted into equation C.1 to produce

$$\mathcal{L}(G|g) = \sum_{r,s} \Big(m_{rs} \log \frac{m_{rs}}{n_r n_s} \Big). \tag{C.2}$$

where $r, s \in \{P_{\text{out}}, C_{\text{in}}, C_{\text{out}}, P_{\text{in}}\}$.

Maximising the log-likelihood has been shown to be equivalent to minimising the microcanonical entropy [32, 34, 33], and this is the form that is used for the optimisation procedure in [13].

$$\mathcal{L}(G|g) = \sum_{(r,s)\in\text{`L'}} \Big(m_{rs} \log \frac{m_{rs}}{n_r n_s} + (n_r n_s - m_{rs}) \log \frac{(n_r n_s - m_{rs})}{n_r n_s} \Big) \tag{C.3}$$

$$+ \sum_{(r,s)\notin\text{`L'}} \Big(m_{rs} \log \frac{m_{rs}}{n_r n_s} + (n_r n_s - m_{rs}) \log \frac{(n_r n_s - m_{rs})}{n_r n_s} \Big). \tag{C.4}$$

C.0.2 Optimising the DCPM: Advanced HITS

The AdvHITS algorithm, instead of assigning two scores to each node as in HITS, *hubnesss* and *authority-ness*, assigns four scores, $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$ which are based on the reward-penalty matrix

$$\mathbf{D} = 2\mathbf{M} - 1 = \begin{bmatrix} -1 & 1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 \end{bmatrix} = \begin{bmatrix} d_1 & d_2 & d_3 & d_4 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}.$$

where $\mathbf{d}_j$ is the j^{th} column vector of $\mathbf{D}$, and $\mathbf{e}_i$ is the i^{th} row vector of $\mathbf{D}$. The iterative algorithm proceeds as follows:

(i) A $n{\times}4$ matrix $\mathbf{U}$ with independent, uniformly distributed entries $u_{ij} \sim \mathrm{Unif}(0, 1)$ is constructed and each column is considered as a raw score S_i^{raw} such that

$$\begin{bmatrix} S_1^{\text{raw}} & S_2^{\text{raw}} & S_3^{\text{raw}} & S_4^{\text{raw}} \end{bmatrix} = \begin{bmatrix} \mathcal{P}_{\text{out}}^{\text{raw}} & \mathcal{C}_{\text{in}}^{\text{raw}} & \mathcal{C}_{\text{out}}^{\text{raw}} & \mathcal{P}_{\text{in}}^{\text{raw}} \end{bmatrix}.$$

(ii) Each score is normalised such that for each node, the sum of its 4 scores is unity. This produces the set of normalised scores

$$\begin{bmatrix} S_1^{\text{norm}} & S_2^{\text{norm}} & S_3^{\text{norm}} & S_4^{\text{norm}} \end{bmatrix} = \begin{bmatrix} \mathcal{P}_{\text{out}}^{\text{norm}} & \mathcal{C}_{\text{in}}^{\text{norm}} & \mathcal{C}_{\text{out}}^{\text{norm}} & \mathcal{P}_{\text{in}}^{\text{norm}} \end{bmatrix}$$

according to the equation

$$S_i^{\text{norm}}(j) = \frac{S_i^{\text{raw}}(j) - S_{\text{min}}^{\text{raw}}(j)}{\sum_{k=1}^{4}(S_k^{\text{raw}} - S_{\text{min}}^{\text{raw}}(j))}$$

where $i \in \{1, \ldots, 4\}$ and $j \in \{1, \ldots, n\}$ and

$$S_{\text{min}}^{\text{raw}}(j) = \min\{S_1^{\text{raw}}(j), S_2^{\text{raw}}(j), S_3^{\text{raw}}(j), S_4^{\text{raw}}(j)\}.$$

Additionally, if the difference between the raw scores is less than some threshold (10^{-10} here) then each set is ascribed a score of 0.25, implying equal affinity to each set [13].

(iii) For the iterative step we now update the raw scores using the reward-penalty vectors from $\mathbf{D}$,

$$S_i^{\text{raw}} = \left(A - \hat{A}\right) S^{\text{norm}} e_i^{\top} + \left(A^{\top} - \hat{A}^{\top}\right) S^{\text{norm}} d_i.$$

This is the step where we incorporate the signed modification described in Section 3.0.2.

If the null model is symmetric, or a constant term as in this case, the equation can be rearranged to give

$$\begin{aligned} S_i^{\text{raw}} &= AS^{\text{norm}}\, e_i^{\top} + A^{\top} S^{\text{norm}}\, d_i - \hat{A} S^{\text{norm}}\, (e_i^{\top} + d_i) \\ &= AS^{\text{norm}}\, e_i^{\top} + A^{\top} S^{\text{norm}}\, d_i - \frac{L}{N^2} \cdot \mathbf{1} S^{\text{norm}}\, (e_i^{\top} + d_i) \\ &= AS^{\text{norm}}\, \mathbf{D}^{\top} + A^{\top} S^{\text{norm}}\, \mathbf{D} - \frac{L}{N^2}(\mathbf{D} + \mathbf{D}^{\top}) \sum_i S_i^{\text{norm}} \end{aligned}$$

which is used in practise.

(iv) Continue steps (ii)-(iii) and record the change in S_i^{norm} until the observed change for all four scores is less than a threshold of 10^{-8}.

(v) Use k-means++ [46] to divide the vertices into four clusters based on the latest S^{norm}.

(vi) Finally, simply iterate through the 24 possible permutations of the four clusters to assign each score to one of $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$, such that the log-likelihood of the DCPM, equation (3.1), is maximised, i.e., equation (C.3) is maximised.

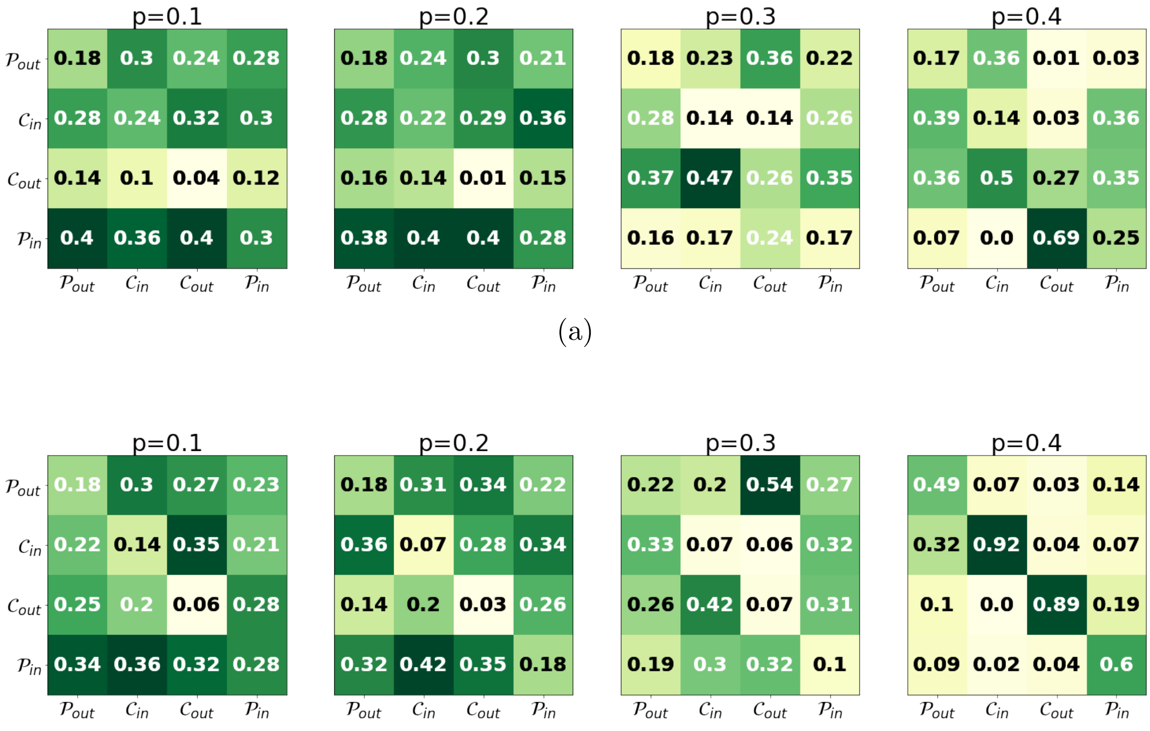
We slightly modify this code to now consider a new, signed version of the DCPM as in Equation (3.1) where the adjacency matrix $\tilde{A}$ is now $\tilde{A} = A^+ - A^-$ where $A^\pm$ are the positive(negative) spatial backbones, and the null model $\frac{m}{n^2}$ becomes

$$\frac{m^+ - m^-}{n^2}$$

and otherwise, the algorithm remains unchanged.

C.1 More results

C.1.1 Synthetic networks

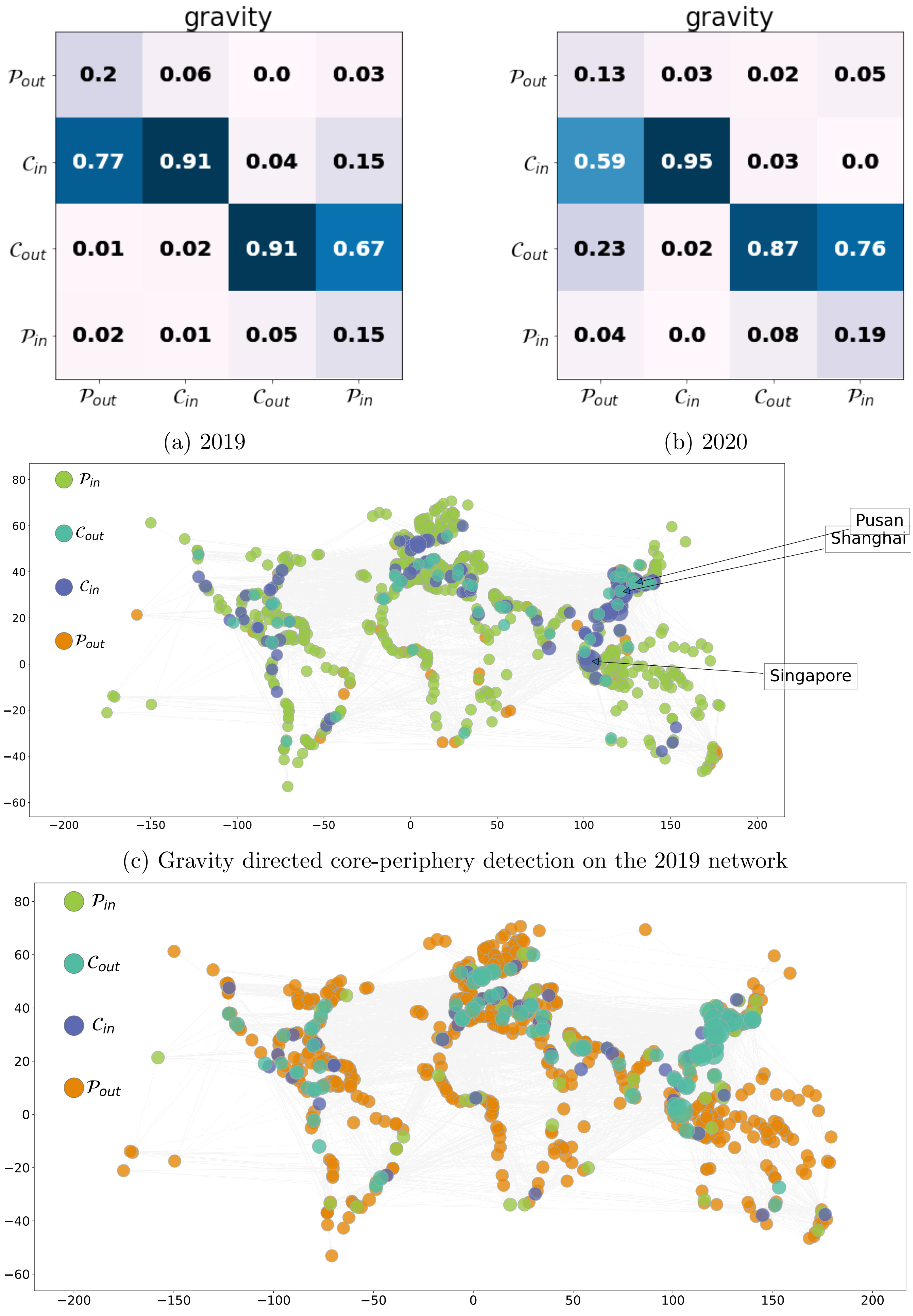


(a)

(b)

Figure C.1: **Confusion matrices for the directed core-periphery algorithm using the doubly-constrained radiation backbone of a directed graph with known core-periphery structure.** For each p-value, 20 random synthetic networks with known core-periphery structure were constructed and the directed core-periphery algorithm applied to them. The columns of the confusion matrices show the percentage of times core and periphery nodes were assigned to each of $\{\mathcal{P}_{\text{out}}, \mathcal{C}_{\text{in}}, \mathcal{C}_{\text{out}}, \mathcal{P}_{\text{in}}\}$. We see that removing the negative backbone in (b) improves results here, though the algorithm struggles more with peripheral sets for low p-values.

C.1.2 The maritime shipping network



(a) 2019

(b) 2020

(c) Gravity directed core-periphery detection on the 2019 network

(d) Gravity directed core-periphery detection on the 2020 network

Figure C.2

Specific cases We zoom in on three more specific cases to give a more detailed interpretation of the results.

(i) Many Australian ports are moved from $\mathcal{P}_{\text{out}}$ to $\mathcal{P}_{\text{in}}$ when spatial-correction using the gravity model is introduced. This means that the gravity model removed a significant number of outgoing links from Australian ports, i.e. links to nearby ports of high in-degree. This suggests that Australia's imports are largely intra-regional, which for container flows is likely the case. Australia has a large trade imbalance of manufacturing goods (more imports than exports). In terms of the local maritime system, many local lines serving the Pacific Islands go via Australia, New Zealand, and Fiji, which is all intra-regional traffic. When these links are corrected, Australia's status as a country that predominantly imports is revealed.

(ii) 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 shows that the correction using the radiation model can discover some cultural and economic affinities that were masked by space in the original network.

(iii) In Chile, when correction using the gravity model is introduced, eight ports are reassigned from $\mathcal{P}_{\text{out}}$ to $\mathcal{P}_{\text{in}}$, and the two biggest ports, Bahia De Valparaíso and San Antonio move from $\mathcal{C}_{\text{in}}$ to $\mathcal{C}_{\text{out}}$. Valparaíso is the largest container port in Chile and is a popular choice for exporters who wish to introduce their goods to the Pacific Side of Latin America. San Antonio moves the highest volume of goods in Chile. These ports very likely import most of their products from large core ports in Asia, after which they redistribute this to smaller ports in South America. Hence, they appear as in-core ports in the uncorrected network, but a clearer pattern of their local functions as out-cores emerges when spatial correction using the gravity model is introduced.